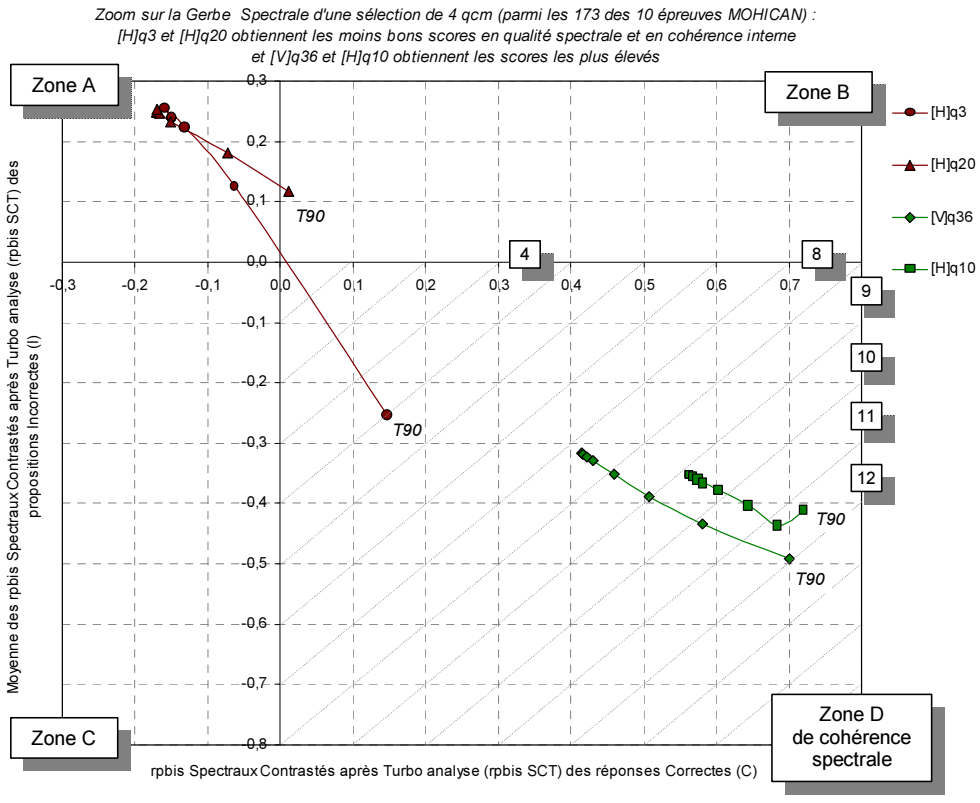
UNIVERSITE DE LIEGE
FACULTE DE PSYCHOLOGIE
ET DES SCIENCES DE L'EDUCATION

Qualité spectrale des tests standardisés universitaires

Mise au point d'indices édumétriques d'analyse de la qualité spectrale des évaluations des acquis des étudiants universitaires et application aux épreuves MOHICAN check up '99

Jean-Luc GILLES



THESE
PRESENTEE POUR L'OBTENTION DU GRADE DE
DOCTEUR EN SCIENCES DE L'EDUCATION
Promoteur : Professeur Dieudonné LECLERCQ
2002

L'illustration de la couverture montre l'aboutissement d'une série de raisonnements et de procédures utilisant des données récoltées à l'aide des pourcentages de certitude accompagnant les réponses à des QCM. Il s'agit de la visualisation des niveaux de cohérence spectrale (voir, p. 227) de quatre QCM provenant des épreuves MOHICAN (p. 93). Les quatre questions ont été sélectionnées du point de vue de leurs performances en cohérence interne, en cohérence spectrale et en réalisation des prédictions parmi les 173 QCM des 10 tests MOHICAN (pp. 343 et 350).

Quatre zones sont définies sur le graphique et identifiées par les lettres A, B, C et D. Les points qui apparaissent dans la zone A, sont représentatifs d'une situation problématique à la fois pour les distracteurs et pour la réponse correcte. La zone B permet quant à elle de mettre en évidence des QCM qui présenteraient des problèmes liés à une ou plusieurs réponses incorrectes, les réponses correctes ne présentant pas de problème particulier. A l'inverse de la précédente, la zone C pourrait contenir les points des QCM qui présentent des problèmes liés à la proposition correcte, mais non en ce qui concerne les distracteurs. Enfin, la zone D peut être qualifiée de « zone de cohérence spectrale » dans la mesure où les QCM qui y sont représentées récoltent des indices attendus pour les réponses correctes et pour les propositions incorrectes.

Les deux questions les moins « performantes » des 173 QCM ([H]q3 et [H]q20) apparaissent en haut à gauche du graphique.

Par contre, les deux autres questions les plus performantes ([V]q36 et [H]q10) figurent en bas à droite du graphique.

Les nombres placés en pourtour de la zone D dans les cadres ombrés désignent une série de niveaux de cohérence spectrale figurés par les espaces entre les diagonales que nous avons tracées dans la zone D. Nous remarquons qu'au palier de turbo analyse T90, les deux meilleures questions, [V]q36 et [H]q10, atteignent le niveau 12 tandis que [H]q3 ne dépasse pas le niveau 4.

REMERCIEMENTS

Cette thèse se situe dans le droit fil des travaux de recherche sur les degrés de certitude menés depuis plus d'un quart de siècle par le Professeur Dieudonné Leclercq. Je ne pouvais avoir de meilleur maître pour m'initier à cette problématique des certitudes et pour m'inculquer le goût de la recherche et de la rigueur scientifique. Je lui dois bien des choses dont la chance d'être chercheur au Service de Technologie de l'Education (STE) qu'il dirige ainsi que l'octroi de quatre mois sabbatiques qui m'ont permis de finaliser cette thèse. Merci à toi Dieudo, pour tes conseils, ta générosité et l'honneur que tu m'as fait de bien vouloir accepter d'être mon promoteur de thèse. Non seulement tu m'as donné des leviers, mais en plus tu m'as montré comment soulever des montagnes.

A plusieurs reprises pendant l'élaboration de cette thèse je me suis demandé ce qui avait bien pu m'amener à entreprendre un jour une telle aventure intellectuelle. Je crois pouvoir dire que l'étincelle initiale fut déclenchée par la rencontre avec mon professeur de pédagogie, M. Pierre Charlier, sur les bancs de l'école normale. Merci à toi Pierre, tu auras réussi à éveiller en moi une passion pour les Sciences de l'Education et c'est elle qui m'aura mené jusqu'ici.

Je remercie également les membres de mon Comité d'accompagnement de thèse. Les stimulations et suggestions du Prof. Daniel Defays m'ont beaucoup aidé à progresser dans ma thèse et plus particulièrement dans la partie statistique. Les nombreuses remarques et questions du Prof. Pol Dupont m'ont montré tout l'intérêt qu'il portait à ma recherche et ses conseils m'ont aidé à améliorer la qualité du travail. Enfin, si j'ai pu mener cette thèse à son terme, c'est aussi grâce au soutien et aux encouragements du Prof. Jean-François Leroy qui m'a fait l'amitié de bien vouloir faire partie du Comité d'accompagnement. Merci à vous trois d'avoir accepté cette tâche, vous dont les emplois du temps sont si chargés. Les deux premiers ont aussi accepté avec le Prof. Dany Laveault et le Prof. Filip Dochy de faire partie de mon jury, qu'il me soit permis de vous en remercier tous les quatre.

Merci aux nombreux professeurs et étudiants de l'université de Liège avec qui j'ai été amené à collaborer dans le cadre de la réalisation des examens standardisés. La docimologie et l'édumétrie appartiennent à ceux qui la pratiquent et vos pratiques m'ont beaucoup inspiré.

Ma gratitude s'adresse aussi à mes collègues du STE et plus particulièrement MM. Marc Pirson et Pascal Detroz, embarqués comme moi dans la dunette du Système Méthodologique d'Aide à la Réalisation de Tests. Merci à vous qui avez accepté, plus qu'à votre tour, d'assumer la relève du quart.

Je tiens aussi à remercier ma famille, ma belle famille et mes proches qui m'ont soutenu. Plus particulièrement le Prof. Jean Nuyts, qui depuis que nous nous connaissons est autant mon beau père que mon mentor et dont la relecture fut si précieuse. Merci à toi L'Ampère, merci à Laure, merci à vous mes parents et amis.

Enfin, *last but not least*, une personne remarquable sans laquelle cette thèse eut été impossible mérite d'être remerciée, il s'agit de Iolaine, mon épouse, qui a su me préserver des tâches domestiques et organiser notre vie familiale pour me permettre de me concentrer sur mes recherches. Plus que mes remerciements, c'est mon admiration que tu mérites, toi qui relève les défis du quotidien, assume ton rôle de mère et tes obligations de pédiatre avec brio et sans jamais perdre de ton dynamisme. Merci à toi Yoyo, notre rencontre est la meilleure chose qui me soit arrivée.

A Iolaine,

A Vicky, Max et Sacha,

SOMMAIRE

PARTIE II

Instrumentation de l'analyse de la qualité spectrale des examens standardisés universitaires

PARTIE III

Exploration spectrale et classique des "check up '99" MOHICAN

Introduction

Nous assistons ces dernières années à une massification des effectifs dans l'enseignement universitaire. Selon Debry, Leclercq & Boxus (1998, p. 57), « *... l'afflux d'étudiants dans le supérieur, et en particulier à l'université, pose notamment problème parce que les ressources d'encadrement n'ont pas suivi. Ainsi le rapport du Conseil des Recteurs Francophone (CReF) montre que depuis 1972, le budget des universités est resté identique (en francs constants) tandis que le nombre d'inscrits passait à 150%.* ».

Cette augmentation des effectifs sans suivi des moyens d'encadrement entraîne des répercutions sur la qualité des évaluations et, de façon connexe, des apprentissages. A ce propos, Romainville & Boxus (1998, p. 28) rappellent que « *…la qualité de l'apprentissage des étudiants dépend en partie du type d'évaluation organisée. Une évaluation par questions ouvertes, demandant à l'étudiant une production écrite personnelle d'une certaine ampleur favorise davantage une approche en profondeur des contenus du cours qu'un QCM simpliste à temps limité portant sur des faits.* ». Or, dans la situation actuelle, étant donné le sous financement de l'enseignement supérieur, la plupart des enseignants confrontés aux grands groupes ne disposent pas d'un nombre suffisant d'assistants pour permettre une correction objective et dans des délais raisonnables d'évaluations comportant de nombreuses questions ouvertes soumises à des centaines d'étudiants. Dès lors, beaucoup sont tentés de mettre en œuvre des évaluations ayant recours aux Questions à Choix Multiple (QCM), mais souvent en négligeant une série de règles méthodologiques et en n'étant pas au fait de procédures édumétriques qui permettent d'améliorer la qualité intrinsèque de ces évaluations standardisées (nous verrons plus loin que nous envisageons la construction de ces épreuves dans le cadre d'un cycle de gestion et de régulation de leur qualité).

Il se fait que depuis de nombreuses années nous participons (Leclercq & Gilles, 1993, 1994, 1995, 2001 ; Gilles, 1995, 1996a, 1996b, 1997, 1998a, 1998b ; Gilles & Leclercq, 1995 ; Gilles & Melon, 2000) à l'étude de la pertinence d'utilisation pour les QCM d'un concept novateur et important : les pourcentages de certitude et les indices qui y sont liés.

Cette thèse apporte un éclairage nouveau et, nous le pensons, *in fine*, très favorable à une utilisation des informations récoltées à l'aide des pourcentages de certitude pour fournir une série d'indications nouvelles sur la qualité des questions, et non plus seulement sur la qualité des performances des étudiants.

Mais précisons nos principes essentiels et le cheminement que nous proposons au lecteur.

Intuitions et principes essentiels

En octobre 1999, dans le cadre du projet de MOnitoring HIstorique de cohortes de CANdidatures universitaires (MOHICAN) dix épreuves de connaissances et de compréhension furent soumises à plusieurs milliers d'étudiants[1] entrant en 1ère candidature dans 8 universités et facultés universitaires de la Communauté française de Belgique (Leclercq & al., 2001), avec visée formative, d'où leur appellation « Check up ». Nous avons été chargé, par le groupe « réussite » du Conseil Interuniversitaire de la Communauté française de Belgique du traitement des données MOHICAN.

Ces épreuves standardisées étaient constituées de questions à choix multiple (QCM) pour lesquelles les étudiants furent invités à accompagner systématiquement le choix de chacune de leurs réponses d'un pourcentage de certitude. Voici l'échelle utilisée :

```
DC = 0        1        2        3        4        5
     |--------|--------|--------|--------|--------|
%C = 0%      20%      40%      60%      80%     100%
```

[1] entre 1.392 et 3.846 selon les tests.

Il s'agissait donc pour les étudiants d'indiquer pour chaque QCM, non seulement quelle était la proposition correcte, mais aussi quel était le pourcentage de chances qu'ils accordaient à chacune de leurs réponses d'être correcte.

Les tests (Check up) MOHICAN n'étaient pas cotés, l'anonymat était garanti. Le choix des pourcentages de certitude n'a donc pas été influencé par un barème de tarif de points ni même par l'octroi d'une cote finale qui aurait pu avoir une quelconque incidence sur le parcours académique ultérieur de l'étudiant.

Dans le cadre de cette thèse nous avons saisi l'opportunité de ces milliers de données, réponses et certitudes, recueillies à l'occasion des dix tests MOHICAN pour mettre au point de nouveaux indices d'analyse de la qualité des épreuves standardisées. Le caractère novateur de ces indices réside dans le fait qu'ils utilisent les informations liées à la façon dont les pourcentages de certitude ont été utilisés par les sujets pour nous renseigner sur la qualité des propositions et, plus globalement, sur la qualité des QCM ainsi qu'à un troisième niveau d'analyse, sur la qualité des tests.

Notre intuition de départ pour la construction de ces nouveaux indices est la suivante : logiquement les étudiants qui répondent correctement à une question devraient fournir des pourcentages de certitude plus élevés que les étudiants qui répondent incorrectement. Ainsi, pour une question à choix multiple qui fonctionne normalement du point de vue de l'utilisation des certitudes, nous devrions observer chez les sujets qui ont choisi la proposition correcte une tendance à fournir des pourcentages de certitudes en moyenne plus élevés que les pourcentages de certitude utilisés par les sujets qui se sont trompés. Parallèlement, pour chacune des propositions incorrectes, nous devrions aussi observer une tendance à choisir des pourcentages de certitude moins élevés que les pourcentages de certitude qui ont accompagné la réponse correcte. Nous dirons alors qu'il y a « cohérence spectrale ». Dès lors que cette situation ne se présente pas, par exemple lorsque les sujets ont tendance à fournir des certitudes plus élevées pour une des propositions incorrectes que pour la réponse correcte, nous nous trouvons face à un problème d'incohérence dans l'utilisation des pourcentages de certitude, nous parlerons alors « d'incohérence spectrale » (le terme « spectral » sera explicité plus loin).

A notre connaissance, jusqu'à présent, les informations fournies par les pourcentages de certitude n'avaient jamais été traitées en vue de déterminer les niveaux de qualité spectrale des épreuves standardisées, de leurs questions à choix multiple et des propositions au sein de celles-ci. C'est cette voie nouvelle de recherche docimologique que nous avons exploitée dans le cadre de cette thèse.

Nous avons créé deux nouveaux types d'indices au départ du principe de calcul du *rpbis classique*. Rappelons que dans le cas du *rpbis classique*, les choix ou les rejets (1 ou 0) de chaque proposition d'une QCM sont corrélés avec les nombres de réponses correctes obtenues à l'ensemble des questions du test. Dès lors, on s'attend à ce que la proposition correcte soit corrélée positivement avec les scores totaux et les distracteurs négativement avec les scores totaux (le *rpbis classique* est défini de manière précise p. 171).

Les deux nouveaux types d'indices de mesure de la cohérence spectrale sont : (1) le rpbis Spectral Contrasté (*rpbis SC*) et (2) le rpbis Spectral Contrasté calculé après Turbo analyse (*rpbis SCT*). Lors d'une recherche antérieure nous avions déjà utilisé les informations liées aux degrés de certitude pour calculer un nouveau type de coefficient de corrélation de point bisériale, le rpbis spectral ou *rpbis S* (Gilles, 1998a). Le *rpbis S* a été développé en vue d'analyser la cohérence d'utilisation des pourcentages de certitudes, c'est-à-dire la propension à utiliser des certitudes plus élevées dans le cas d'une réponse correcte que les certitudes utilisées dans le cas des réponses incorrectes. Dans le cadre de cette thèse nous proposons une première amélioration du *rpbis S* en mettant en œuvre un « traitement contrasté » pour les propositions incorrectes des QCM.

Nous utilisons l'appellation *rpbis SC* pour désigner les *rpbis S* qui bénéficient du « traitement contrasté ». Sans entrer ici dans les détails (voir p. 178), disons que le « traitement contrasté », consiste à faire intervenir dans le calcul du *rpbis SC* d'une proposition incorrecte les données des étudiants qui ont choisi cette proposition *en contraste* avec les seules données des étudiants qui ont choisi la proposition correcte. L'avantage réside dans l'élimination des données des étudiants ayant opté pour les autres

propositions incorrectes, ce qui évite d'introduire dans la mesure de la cohérence spectrale du distracteur envisagé, le « bruit » qu'engendreraient les données des autres propositions incorrectes.

En ce qui concerne le principe de la « turbo analyse » (voir détails p. 186), il consiste à opérer une sélection dans les données utilisées pour le calcul des *rpbis SC* sur la base du critère du niveau de réalisme atteint par les sujets. Disons pour faire bref dans le cadre de cette introduction, que nous sélectionnons les données des étudiants les plus réalistes de façon à ne calculer la cohérence spectrale que sur des données plus valides car contenant moins d'erreurs d'auto-estimations. Le niveau de performance des étudiants en matière d'auto-estimation des chances de leurs réponses d'être correctes est mesuré à l'aide de l'indice de réalisme des sujets (Leclercq & al., 2000). En sélectionnant les données des sujets qui atteignent un niveau de réalisme élevé[2], nous pouvons donc calculer des *rpbis SC* plus valides car fondés sur les informations fournies par des étudiants dont les données sont plus fiables. Nous utilisons l'appellation *rpbis SCT* pour désigner les *rpbis SC* calculés dans le cadre d'une Turbo analyse. Cette expression vient du concept de turbo machine, destiné à augmenter la pression d'un gaz. Ici par récupération des informations, de l'analyse précédente on ne garde que les données susceptibles d'augmenter la validité.

En plus des *rpbis SC* et *rpbis SCT*, qui sont au cœur de cette recherche, nous avons aussi adapté d'autres indices spectraux initialement prévus pour l'analyse des performances des étudiants de manière à ce que ces indices nous livrent des informations sur les performances des questions. Il s'agit essentiellement des indices de Réalisation des prédictions par question (*Rq*, p. 264) et de Centration par question (*Cq*, p. 254). Nous avons aussi appliqué le principe de la « turbo analyse » au calcul de ces indices spectraux.

Cheminement proposé au lecteur

Nous avons choisi de structurer notre propos en trois grandes parties. Cette structuration nous paraît refléter le mieux notre démarche de recherche tout en permettant au lecteur d'aborder progressivement les concepts spectraux (parfois un peu techniques) sans pour autant perdre de vue le contexte et les enjeux de l'analyse de la qualité spectrale des épreuves. Nous ne pouvions notamment pas imaginer de débuter le texte de cette thèse sans rappeler dans les ***Préliminaires intitulés « Le besoin : des épreuves universitaires de qualité*** » une série de faits et de problèmes qui nous paraissent fondamentaux et qui devraient interpeller tous ceux qui sont amenés à procéder à l'évaluation des compétences des apprenants dans l'enseignement supérieur.

Ensuite, dans la **première partie intitulée « *Pour une pragmatique de la qualité dans la réalisation des épreuves standardisées universitaires* »** nous avons souhaité planter le décor, décrire le contexte dans lequel nous envisageons l'utilisation des nouveaux indices spectraux. Nous y rappellerons par exemple que le monde universitaire vit une série de mutations depuis les années 60 (Dupont & Ossandon, 1994 ; Romainville & Boxus, 1998) et qu'actuellement, dans toutes les sphères de leurs activités, les universités doivent être en mesure de produire les preuves de la qualité des services qu'elles proposent. Les activités liées à la réalisation des examens n'échappent pas à cette tendance et dans notre esprit, les nouveaux indices spectraux dont il sera question dans cette thèse doivent permettre aux enseignants du supérieur de mieux réguler la qualité de leurs évaluations standardisées. Epinglons à ce propos le commentaire du GRIPU[3] (Blais & al., 1997, p. 7) : « *Dans le contexte social actuel les établissements d'enseignement supérieur sont de plus en plus souvent amenés à faire la démonstration de la qualité de la formation dispensée. En ce sens, l'évaluation des apprentissages, comme mécanisme de régulation et de contrôle exige une attention particulière. Par ailleurs, la régulation et le contrôle des apprentissages doivent prendre en considération un souci de plus en plus net d'assurer l'équité et la transparence des pratiques d'évaluation* ».

[2] L'indice de réalisme varie entre 0 et 100 et plus il est élevé, plus les étudiants sont réalistes, moins ils commettent d'erreurs dans leurs auto-estimations.

[3] Groupe de Recherche Interdisciplinaire en Pédagogie Universitaire, Faculté des Sciences de l'Education – Université de Montréal.

Dans le ***chapitre I*** nous exposerons le contexte du Système Méthodologique d'Aide à la Réalisation de Tests (SMART), une structure de l'Université de Liège qui offre aux enseignants une méthodologie de construction d'épreuves standardisées de qualité. Actuellement, dans le contexte du SMART, nous travaillons déjà au développement de programmes informatiques qui intègrent les retombées pratiques de l'analyse de la qualité spectrale des tests dans les procédures de traitement des évaluations standardisées des étudiants universitaires[4]. Dans le ***chapitre II*** nous évoquerons les techniques et les enjeux liés à l'utilisation des pourcentages de certitude. Pour ce qui est des enjeux, nous soulignerons qu'ils sont non seulement édumétriques, mais aussi sociaux, moraux, pédagogiques et épistémologiques. A ce propos, il nous paraissait utile de rappeler les fondements de l'utilisation des certitudes : l'incompétence est une situation normale de la vie, l'ignorance (connue) n'est pas dangereuse, la tendance à cacher l'ignorance est regrettable, le doute est le moteur même de la connaissance, être capable d'évaluer ses propres connaissances est une compétence qui se situe au niveau le plus élevé de la taxonomie des objectifs cognitifs… Le ***chapitre III*** nous permettra d'aborder les indices habituellement utilisés dans le cadre des analyses classiques de la qualité des épreuves. Bien que les *rpbis SC* et *rpbis SCT* mesurent la cohérence spectrale et non la cohérence interne (*rpbis classique*), nous rangeons les *rpbis Spectraux* dans la catégorie des indices corrélationnels de discrimination ; c'est la raison pour laquelle nous présenterons plus en détail dans ce chapitre les indices classiques du pouvoir discriminatif des questions.

Dans la **seconde partie intitulée « *Instrumentation de l'analyse de la qualité spectrale des examens standardisés universitaires* »** nous exposerons les instruments d'analyse de la qualité spectrale des épreuves. Nous les avons catégorisés en fonction de trois niveaux d'analyse. Nous distinguons en effet parmi les instruments spectraux, ceux qui permettent l'analyse de la qualité (1) des PROPOSITIONS au sein des QCM, (2) des QCM elles-mêmes et (3) des TESTS. Quel que soit le niveau d'analyse envisagé, les indices spectraux peuvent être calculés à différents niveaux de turbo analyse, ce qui améliore leur fiabilité. Nous appliquerons systématiquement le principe de la turbo analyse aux indices présentés dans cette partie. Les représentations graphiques que nous proposons en vue de faciliter l'interprétation des résultats spectraux seront également systématiquement décrites ici.

Nous commencerons par expliquer dans le ***chapitre IV*** comment nous réutilisons et améliorons le principe du *rpbis classique* pour évaluer la cohérence spectrale au niveau d'analyse des « PROPOSITIONS ». Nous y présenterons aussi en détails les principes du « *Traitement Contrasté* » appliqué au calcul des *rpbis SC* et de la « *Turbo analyse* » utilisée pour le *rpbis SCT*. Afin d'étudier ces *rpbis spectraux* et les autres instruments de mesure de la qualité spectrale des tests à partir des résultats des milliers de données des épreuves MOHICAN, nous avons été amené à programmer un logiciel intitulé SCANTEST 2.0[5]. Ce programme est décrit dans le ***chapitre V***. La mise au point des indices utilisés aux trois niveaux d'analyse a été réalisée à l'aide des données de l'épreuve MOHICAN de physique soumise à 2.497 étudiants. Le ***chapitre VI*** présente les matrices de résultats produites par SCANTEST 2.0 pour l'analyse du niveau « PROPOSITIONS ». Les indices et instruments d'analyse spectrale du niveau « QCM » sont quant à eux détaillés dans le ***chapitre VII***. Signalons que l'analyse spectrale du niveau QCM bénéficie entre autres d'une représentation graphique en « *Brins Spectraux par question (BSq)* » (voir figure en couverture) qui est expliquée dans ce chapitre. Cette seconde partie consacrée à l'instrumentation de la qualité spectrale se termine par le ***chapitre VIII*** où sont présentés les outils d'évaluation de la qualité spectrale du niveau « TEST ».

Dans la **troisième partie intitulée « *Exploration spectrale et classique des check up '99 MOHICAN* »** nous montrerons comment se comportent les instruments d'analyse spectrale et classiques présentés précédemment lorsqu'on les applique aux 10 épreuves MOHICAN (les *check up '99*). Quelles différences, quelles similitudes avec les résultats obtenus à l'aide des indices habituellement utilisés dans l'analyse classique des épreuves ? Quelles informations nouvelles les indices spectraux nous apportent-ils ? Nous commencerons par comparer les informations spectrales et classiques obtenues au niveau « TEST » pour ensuite explorer les niveaux « QCM » et « PROPOSITIONS ».

[4] Voir la partie « Perspectives » dans nos conclusions.
[5] le logiciel que nous avons conçu et programmé pour calculer les nouveaux indices spectraux.

Le ***chapitre IX*** est consacré à l'exploration du niveau « TEST ». Nous y comparerons les performances obtenues par les 10 épreuves aux indices spectraux présentés précédemment (chapitre VIII). Nous comparerons aussi les performances obtenues aux indices classiques de cohérence interne. Signalons à ce sujet que dans le cadre de l'exploration classique, nous avons utilisé deux types de matrices de résultats, l'une est binaire et l'autre spectrale. La première, habituellement utilisée, présente les données sous forme de « 1 » (la réponse est correcte) ou de « 0 » (réponse incorrecte). La seconde, plus nuancée, reprend les pourcentages de certitude en valeur positive lorsque la réponse est correcte et en valeur négative lorsqu'elle est incorrecte (Jans & Leclercq, 1999). Nous avons calculé les indices classiques de cohérence interne (dont l'alpha de Cronbach) à partir de chaque type de matrice et avons comparé les résultats.

Dans le cadre du ***chapitre X***, nous explorerons le niveau « QCM ». Nous commencerons par examiner les corrélations des valeurs obtenues par les 173 questions des 10 épreuves aux indices classiques (présentés au chapitre III). Nous envisagerons ensuite les corrélations entre les valeurs obtenues aux indices spectraux (présentés au chapitre VII) calculés à deux paliers de turbo analyse : T0 (à partir des données de tous les étudiants) et T80 (à partir des données des étudiants qui font preuve d'un réalisme élevé, $Rs \geqslant 80$). Nous terminerons cette étude des corrélations en comparant les valeurs des indices classiques et des indices spectraux. Nous exposerons les résultats des méthodes de comparaison visuelle des performances classiques et spectrales des 173 QCM. L'une d'elles utilise des ingénogrammes (graphiques à coordonnées polaires) pour faciliter l'analyse des performances des questions à trois indices spectraux (*NCSq T80*, *Rq T80* et *Cq T80*) et à deux indices classiques ($r_{qt\ mb}$ et $r_{qt\ ms}$). Des représentations en « *3D classico-spectral* » permettent aussi de visualiser les niveaux de qualité et de détecter les questions problématiques.

Le ***chapitre XI*** livrera les analyses des valeurs obtenues aux indices spectraux et classiques des propositions de 16 questions mises en évidence dans le chapitre précédent pour leurs performances globales faibles en cohérence interne et en qualité spectrale ou/et dont les valeurs récoltées par les propositions indiquent des incohérences. Nous comparerons les résultats des analyses classiques et spectrales aux diagnostics des experts consultés lors du débriefing des épreuves et examinerons dans quelle mesure les trois faisceaux d'informations (indices classiques – indices spectraux – avis des experts) concordent ou divergent.

En ce qui concerne nos ***conclusions***, nous avons préféré les scinder en deux parties. Dans la première partie intitulée « ***Conclusions détaillées*** », nous fournirons au lecteur un bilan précis des apports de cette thèse et envisagerons dans le contexte de la régulation de la qualité des épreuves standardisées universitaires, une série de perspectives offertes par l'utilisation de l'analyse spectrale. Les nouveaux indices spectraux que nous proposons doivent en effet aider les professeurs à prendre des décisions de rectifications de leurs épreuves et nous verrons dans la section « ***Perspectives*** » comment nous envisageons de leur fournir l'information liée au contrôle de la qualité spectrale de leurs tests, QCM et propositions au sein de celles-ci. Enfin, nous terminerons par les « ***Conclusions générales*** » où nous prendrons de façon plus synthétique la mesure du chemin parcouru dans le cadre de cette thèse.

Signalons pour terminer cette introduction qu'étant donné le caractère parfois assez technique et peu familier des appellations et indices utilisés dans certaines sections de cette thèse, nous avons placé en fin d'ouvrage deux glossaires et un index des formules. Le premier glossaire concerne les principaux indices et instruments d'analyse spectrale utilisés dans cette recherche (p. 468). Le second glossaire reprend les principaux symboles et abréviations (p. 473). L'index des formules (p. 476) donne les intitulés des formules ainsi que les pages où elles sont expliquées en détails.

Préliminaires

Le besoin :
des examens universitaires de qualité

Sommaire

A. Problèmes liés aux examens oraux ou écrits ayant recours aux Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL)

Le sort académique (réussite ou échec) et par suite le sort professionnel de millions d'étudiants par le monde dépend d'évaluations sommatives certificatives. Pour des raisons d'équité, relayées par des raisons de prudence juridique (afin d'éviter les recours en justice), ces évaluations sont de plus en plus standardisées. On sait en effet depuis des décennies, notamment depuis les travaux de recherche du courant de docimologie critique mené par Pieron (1963), à quel point la correction des examens où les étudiants sont interrogés à l'aide de Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL), que ce soit par écrit ou oralement, souffre de nombreux biais. Ces biais expliquent en grande partie le succès rencontré par les épreuves standardisées dont le contenu et les conditions d'administration sont identiques pour tous les examinés et dont la correction peut être automatisée. Pour mieux comprendre les enjeux de la standardisation des épreuves universitaires, nous dressons ici un inventaire (non exhaustif) des principales faiblesses des examens ayant recours aux QROM/QROL.

1. Le manque de concordance intra et inter-correcteurs dans la correction des réponses ouvertes

La correction des Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL) ne peut être automatisée et est donc confiée à un ou plusieurs correcteurs. Nous allons voir que cette caractéristique entraîne des problèmes liés au manque d'accord intra-correcteur ou/et inter-correcteurs.

Travaillant entre autres sur le matériel fourni par le baccalauréat français, Pieron (1963) et ses collaborateurs ont montré que les notes obtenues par un examiné dépendaient de l'« attitude typologique » de l'examinateur lors de la correction des examens. Leclercq (1986, p.22) résume cette équation personnelle du correcteur en quatre caractéristiques principales qui pourraient être mesurées soit d'une épreuve à l'autre, d'une année à l'autre ou même au sein d'une seule épreuve : « *(1) le centrage ou la moyenne de ses notes (certains examinateurs se révèlent trop sévères et d'autres excessivement généreux) ; (2) l'ampleur de la distribution de leurs notes (dispersion exprimée par l'écart type) ; (3) la forme de la distribution de leurs notes (normale, symétrique ou non) ; (4) la constance dans le temps de ces trois paramètres* ».

Les ouvrages traitant de docimologie fourmillent d'études où sont mises en évidence les attitudes typologiques divergentes des correcteurs. Nous mentionnons ici à titre d'exemple le cas mis en évidence par Agazzi (1967, relaté par De Landsheere 1979, p. 33) pour un ensemble de branches où six correcteurs ont, chaque fois, noté les examens du baccalauréat[6] :

	Refusés par les six correcteurs	Admis par les six correcteurs	Admis par les uns et refusés par les autres
Version latine	40%	10%	50%
Composition française	21%	9%	70%
Anglais	37%	16%	47%
Mathématiques	44%	20%	36%
Philosophie	9%	10%	81%
Physique	37%	13%	50%

On voit que pour l'examen de philosophie, dans 81% des cas l'admission ou le refus d'un candidat dépendait du correcteur, la concordance entre correcteurs n'apparaît donc que dans 19% des cas. C'est en mathématique que le pourcentage de concordance est le plus élevé mais avec cependant seulement 66% de décisions de refus ou d'admissions sur lesquels les correcteurs s'accordent.

[6] Note de refus/échec : moins de 10 sur 20.

Des problèmes de discordance peuvent également s'observer chez une seule personne chargée de la correction, on parlera alors d'infidélité chez un même correcteur. L'étude de Hartog et Rhodes (1936), relatée par De Landsheere (1979, p. 39), est exemplaire : quatorze historiens furent invités à corriger une seconde fois quinze compositions, douze à 19 mois après la première correction (après que toute trace de la première correction fut effacée). De Landsheere rapporte : « *Dans 92 cas sur 210, le verdict a été différent d'une fois à l'autre. Il faut toutefois insister sur le fait que des résultats aussi pauvres sont dus au manque de directives rigoureuses précisant les aspects à considérer par les notateurs.* »

Leclercq (1986, p. 22) distingue trois catégories de biais liés à la correction avec deux effets dans chaque catégorie. Nous les résumons dans le tableau ci-dessous.

Biais dus au seul correcteur		Biais dus aux interactions prof.-élève		Biais dus aux séries de copies	
Effet de sévérité	Effet de tendance centrale	Effet de halo	Effet de stéréotypie	Effet de séquence	Effet de relativisation
Sévérité systématiquement plus élevée ou au contraire moins élevée chez certains correcteurs	*Evitement des notes extrêmes et concentration des scores au milieu de l'échelle*	*Des caractéristiques de l'étudiant influencent sa note (aspect physique, présentation, …)*	*Tendance à attribuer à un examiné les notes que celui-ci a acquise antérieurement*	*La copie qui suit une copie brillante risque d'être désavantagée, et inversement*	*Parmi toutes les copies jugées moyennes quelques mois auparavant le correcteur distinguera des faibles et des bonnes*

Ce bref aperçu des biais liés à la subjectivité des examinateurs montre que les problèmes d'infidélité lors de la correction sont énormes : les examens traditionnels offrent peu de garanties qu'un travail corrigé et classé dans la catégorie « excellent » bénéficierait de la même mention s'il était corrigé dans d'autres conditions (autres correcteurs ou quelques semaines plus tard par un même correcteur).

Signalons ici que les constats négatifs de la docimologie critique (tels que ceux liés à l'inconstance intra et inter-correcteurs que nous venons d'évoquer) ont cependant permis l'avènement d'un courant de docimologie constructive où les docimologistes ont tenté d'apporter des solutions (malheureusement trop peu mises en pratique). Citons à titre d'exemple les travaux sur les « Echelles descriptives en évaluation » de De Bal, De Landsheere et Beckers (1977). A l'aide de la technique des échelles descriptives ces chercheurs ont mis en évidence la possibilité d'améliorer la cohésion inter-correcteurs en découpant le « trait » à évaluer en plusieurs facettes et en veillant à définir clairement chaque échelon des échelles mises au point pour mesurer les aspects à évaluer.

2. Le manque de validité

Lors d'un examen ayant recours aux QROM/QROL, lorsque le nombre d'étudiants est élevé (parfois plus de 600 dans une institution comme l'Université de Liège) l'examinateur est contraint d'utiliser un nombre réduit de questions ouvertes étant donné le temps considérable que nécessite la correction des réponses fournies. Le faible nombre de questions posées entraîne dès lors le problème de l'absence d'une couverture large de tous les points importants du cours. Il est en effet difficile avec seulement quelques QROL de balayer l'ensemble du cours.

Une autre critique liée au problème de la validité et qui peut aussi être adressée à l'encontre des examens ayant recours aux QROM/QROL concerne la tendance à exiger la simple restitution de faits abordés dans le cours malgré la possibilité d'évaluer des processus mentaux plus complexes. En effet, parmi les processus mentaux qui sont évalués lors des examens, force est de constater que le plus souvent c'est la connaissance de mémoire qui est sollicitée et plus rarement la capacité d'analyser, de synthétiser, d'élaborer des jugements critiques, etc. De Landsheere (1979, p. 52) signale que dès 1911 un rapport de la Commission Consultative sur les Examens dans l'Enseignement Secondaire de Grande Bretagne déplorait que les élèves consacraient trop d'énergie à reproduire les idées des autres au lieu de développer leur propre créativité.

Si un examinateur souhaite évaluer le niveau « connaissance » chez les étudiants, il est préférable qu'il le fasse à l'aide d'épreuves standardisées ayant recours aux Questions à Choix Multiple (QCM). Le nombre de questions qu'il pourra poser sera bien plus élevé et lui permettra de couvrir une large partie de la matière enseignée. Nous verrons par la suite qu'il existe des formes de QCM plus sophistiquées que les QCM classiques et qui permettent d'évaluer plus systématiquement des processus mentaux plus élevés (dans la taxonomie de Bloom, 1969) que la simple connaissance de mémoire (en fait de la 'reconnaissance' dans le cas des QCM où l'examiné doit choisir, donc reconnaître, parmi *x* solutions qui lui sont proposées, celle qui est correcte).

3. Le manque de sensibilité des mesures qui ignorent les états de connaissances partielles

Tout le monde s'accorde sur le fait que les mesures des acquis des étudiants devraient refléter des phénomènes subtils mais bien peu d'examinateurs se soucient de sonder méthodiquement la conviction avec laquelle les examinés maîtrisent le sujet sur lequel ils sont interrogés.

Rarement on demande aux étudiants d'exprimer systématiquement leurs certitudes à propos des réponses qu'ils fournissent lors d'un examen. C'est ignorer qu'en termes de connaissances partielles, un fossé peut séparer la performance de deux étudiants qui fournissent pourtant la réponse correcte à une même question. En effet, la performance de l'examiné qui répond correctement et de façon très assurée (en accompagnant sa réponse d'une probabilité élevée d'être correcte) est bien meilleure que celle d'un autre examiné qui lui aussi fournit la réponse correcte mais en lui attribuant une probabilité si peu élevée de l'être que cette (pseudo) connaissance en devient inutilisable parce que le sujet lui-même ne peut se fonder sur elle pour prendre des décisions, pour agir. De même, les étudiants qui avouent leur ignorance devraient être moins sanctionnés (et même encouragés dans cette démarche) par rapport à ceux qui prétendent avec assurance fournir des réponses correctes qui s'avèrent erronées (ce qu'Ebel a appelé « *unwarranted pretense of confidence* »). Soulignons que dans ce dernier cas les « prétentions de connaissances erronées » constituent des comportements particulièrement dangereux dans les domaines où la vie des gens est en jeu (par exemple en médecine, en pilotage d'avion, de véhicules,…).

Malheureusement, les examinateurs se donnent trop peu souvent les moyens d'évaluer les connaissances partielles des étudiants, et, lorsqu'ils le font dans le cadre d'examens traditionnels c'est en général de façon fort peu systématique. Signalons qu'on commence à en comprendre l'intérêt dans l'industrie (Shufford, 1993) et dans l'éducation du patient (D'Ivernois et Gagnayre, 1995).

Or, la recherche et la pratique ont montré qu'en respectant une série de règles méthodologiques (De Finetti, 1965; Shufford & al., 1966; Van Naerssen, 1962; Leclercq, 1975, 1982, 1993; Hunt, 1977; Bruno, 1986, 1987) il est possible de mesurer de manière subtile, systématique et objective les états de connaissances partielles des étudiants à l'aide des pourcentages de certitude. Lorsqu'ils sont utilisés, ceux-ci permettent à l'examiné d'exprimer son doute en accompagnant chacune de ses réponses de la probabilité qu'il lui accorde d'être correcte. Nous passons alors d'une conception binaire et frustre de la mesure des compétences à une conception spectrale et subtile où il est enfin possible de distinguer selon la terminologie proposée par Jans & Leclercq (1999, p. 307) entre (de la situation la plus catastrophique à l'idéal) : (1) méconnaissance erronée (réponse incorrecte et certitude élevée), (2) confusion (réponse incorrecte et certitude moyenne), (3) méconnaissance reconnue (réponse incorrecte et certitude faible), (4) ignorance (réponse correcte et certitude zéro), (5) connaissance douteuse (réponse correcte et certitude faible), (6) connaissance partielle (réponse correcte et certitude moyenne) et (7) connaissance parfaite (réponse correcte et certitude élevée).

Nous reviendrons plus loin sur les enjeux (p. 101) et les techniques de recueil (p. 106) des pourcentages de certitude qui sont à la base des indices de qualité spectrale des épreuves exposés dans les chapitres suivants (p. 169 et p. 211).

Bien qu'il soit possible de demander aux étudiants d'accompagner leurs réponses ouvertes moyennes ou longues par des pourcentages de certitude, cette procédure n'est que très rarement utilisée lors

d'exercices et quasi jamais lors d'examens où le professeur sollicite des réponses ouvertes[7]. C'est dans le cadre d'épreuves standardisées ayant recours aux QCM avec pourcentages de certitude que l'évaluation systématique des états de connaissances partielles est la plus aisément mise en œuvre.

4. Le manque de diagnosticité des épreuves sommatives classiques qui ont recours aux QROM ou aux QROL

Les examinateurs se préoccupent en général très peu de renvoyer après l'épreuve un feedback détaillé et individualisé vers les examinés de manière à permettre à ces derniers d'effectuer un bilan précis de leurs compétences. Certains prétexteront que les examens certificatifs sont là pour vérifier si l'étudiant est capable de réaliser les tâches qu'on attend de lui en fin d'enseignement et non pour diagnostiquer où se situent les éventuels problèmes qu'il rencontre, qu'il ne faut pas confondre évaluation formative à visée diagnostique et évaluation sommative à visée certificative.

Cependant, dans notre contexte universitaire les étudiants qui subissent un échec dans un cours en 1ère session disposent en général d'une deuxième chance sous la forme d'une seconde épreuve en 2ème session (souvent du même type que la première) et de quelques mois pour améliorer leurs connaissances. Dès lors, pourquoi ne pas renvoyer un maximum d'informations vers ces étudiants en situation d'échec, et, le plus tôt possible après l'épreuve, de manière à leur permettre d'ajuster leur étude en fonction des faiblesses décelées lors du premier examen ? Ne pas le faire est à notre avis condamnable et relève de la « *non assistance à étudiant en danger d'échec* ».

Il existe des explications moins avouables à ce manque de rétroaction vers les examinés (explications d'autant plus difficiles à comprendre qu'un des chevaux de bataille de la plupart des institutions universitaires est la lutte contre l'échec). Ainsi, on avancera le fait que les épreuves sommatives traditionnelles ayant recours aux QROM/QROL ne sont en général pas conçues pour permettre des feedbacks détaillés. Il est vrai que quelques questions ouvertes ne couvrant qu'une partie très limitée de la matière (voir le manque de validité des épreuves traditionnelles) ne permettent guère d'informer l'étudiant sur ce qui est maîtrisé ou non dans son étude du cours.

La vérité est que les examinateurs évitent de poser de nombreuses questions à réponses ouvertes notamment parce que la correction de celles-ci prend un temps considérable. Un des avantages des épreuves standardisées ayant recours à des questions fermées est de permettre de poser de nombreuses questions couvrant une large partie de la matière et dont la correction peut être automatisée (donc effectuée en peu de temps). Nous verrons plus loin que cette automatisation de la correction des épreuves standardisées permet aussi d'envisager l'envoi de feedbacks détaillés (notamment via Internet) vers les étudiants, et ce, dans des délais très courts.

5. Le manque d'équité des épreuves traditionnelles, en particulier les oraux

En période d'examen on entend souvent, dans les couloirs jouxtant les bureaux des examinateurs, des commentaires de la part des étudiants sur la chance ou la malchance qui fut la leur en ce qui concerne le tirage au sort des questions. Il faut reconnaître que le facteur chance peut en effet jouer un rôle important dans la réussite lorsque seulement deux ou trois questions sont posées lors d'une épreuve.

Les étudiants savent aussi que lors des examens oraux il vaut mieux ne pas suivre un condisciple particulièrement brillant car l'effet de contraste risque d'être défavorable. Inversement, il vaut mieux suivre un condisciple qui a échoué pour autant que la contre-performance de ce dernier n'ait pas mis le professeur de trop mauvaise humeur… La chance peut donc aussi jouer au niveau de l'ordre de passage chez l'examinateur.

[7] D. Leclercq l'a fait dans une épreuve de vocabulaire français en novembre 2000 avec des étudiants de 1ère candidature à la FAPSE-ULg et Jans (1997) dans le domaine du vocabulaire anglais.

Ce type de phénomène a été observé par Bonniol (1972) dans le cadre de la correction de copies d'épreuves écrites. Ce dernier introduit dans la correction de travaux de valeur moyenne des ancres (copies de valeur soit excellente, soit médiocre) et constate un effet de contraste sur la note attribuée aux travaux suivants : une copie de qualité moyenne sera surévaluée après une copie médiocre ou sous-évaluée après une copie excellente.

Ces quelques exemples et les problèmes liés au manque de concordance intra-correcteur et/ou inter-correcteur exposés plus haut (p. 23) montrent que les examinés peuvent se retrouver dans des situations fort injustes parce qu'elles manquent de standardisation (tous ne sont pas traités de la même façon) et parce qu'une trop grande place est laissée au facteur chance.

Il existe un sentiment d'injustice chez les étudiants lié à ce manque d'équité des examens traditionnels. Comme le soulignent les auteurs de l'enquête sur les pratiques d'examen à l'université de Montréal (Blais & al., 1997, pp.126-127) : « *…lors des entrevues de groupe avec les étudiants, la plainte la plus fréquente est celle qui concerne le manque de standardisation de l'évaluation et de la notation (…). Les étudiants en ont contre le fait que les professeurs les évaluent comme bon leur semble. Ils dénoncent la subjectivité qui intervient dans l'élaboration et la correction des travaux ou examens. Pourquoi dans certains domaines les étudiants trouvent-ils facile de « passer » et d'obtenir des notes élevées (on pourrait parler d'inflation de notes), alors que dans d'autres domaines, les étudiants sont continuellement sous le stress d'une évaluation exigeante leur demandant, selon leur perception, de fournir des efforts plus grands que leurs collègues ? Que dire des programmes où il existe des cours « éliminatoires », plus difficiles que les autres et faisant office de mécanisme de sélection ? De leur point de vue, il s'agit d'injustices flagrantes, du point de vue de l'enseignant, il s'agit de choix « personnels » en relation avec l'absence de balises contraignantes (voir institutionnelles) quant aux exigences de réussite des cours* ».

Après cette revue rapide des biais potentiels liés aux examens traditionnels, on comprendra probablement mieux l'ironie et l'humour provocateurs qui caractérisent les propos des auteurs du livre intitulé « *Le bouton du mandarin* », Didier & al. (1966), lorsqu'ils définissent le sens primitif et le sens dérivé du terme « examen » :

Sens primitif	Sens dérivé
« *Epreuve, ayant pour objet de vérifier l'état physique, intellectuel ou moral, de la chose ou de la personne examinée* »	« *Loterie instituée pour la distribution de lots appelés 'diplômes'* »

La situation est cependant pire, le hasard dans les examens, surtout dans les oraux, où les QROM/QROL sont utilisées, intervient moins que ne le laisse penser le terme « *loterie* » employé par les auteurs.

En effet, les épreuves traditionnelles ne sont pas socialement neutres. On sait que les examinés selon qu'ils proviennent de classes sociales aisées ou défavorisées n'emballeront pas leurs compétences de la même façon. Ce phénomène a été bien mis en évidence par Passeron (1970). Ce dernier (p. 12) propose la thèse de l'examen instrument d'immobilisme social : « *Les procédures de notation et les types d'épreuves utilisées prennent en compte au moins autant que les aptitudes techniques certains aspects gratuits de la performance, qui n'ont aucune importance technique, mais qui sont en revanche très fortement liés aux habitudes culturelles de telle classe sociale plutôt que telle autre* ». Tourneur (1988, p. 4) commente le texte de Passeron en signalant : « *La critique vaut surtout pour l'examen oral si prisé dans l'enseignement supérieur et qui accorde la part belle à la présentation, à la correction du langage et à l'élégance de la diction* ».

Avec l'accentuation de la diversification des origines sociales des étudiants universitaires (voir plus loin : *Les facteurs de mutation des universités*, p. 60), les examinés sont de plus en plus conscients des problèmes liés à l'effet de halo lors des oraux. Voici l'extrait d'une interview d'un étudiant en pleine session d'examen parue dans le quotidien Le Soir[8] du 29 mai 1999 : « Daniel, en deuxième licence philo et

[8] Article intitulé « *Tremblez maintenant !* » de Michel Verlinden.

lettres, s'inquiète pour des raisons économiques. *Je ne peux pas rater: c'est le CPAS qui m'a pris en charge. C'est très paniquant, surtout maintenant que je suis si proche du but. J'ai très peur des "oraux", parce que je n'arrive pas à m'exprimer facilement. J'ai grandi dans la rue avec beaucoup d'immigrés autour de moi. On a toujours parlé une langue qui n'a rien à voir avec celle qu'on pratique ici. Devant un prof, j'ai peur de ce que je dis, surtout que certains ne se gênent pas pour vous faire remarquer vos points faibles... ».*

Même si nous pensons que le déterminisme social n'est pas inéluctable, il faut cependant bien admettre que l'effet de halo ne joue pas en faveur des étudiants provenant de couches sociales moins favorisées. Si dans un pays comme les Etats-Unis les tests standardisés connaissent un tel succès, c'est notamment parce qu'ils offrent des garanties de non discrimination raciale (*culture-free test*) et de non discrimination de classe sociale (*class-free test*).

***

En synthèse :

A la lecture des problèmes posés par les épreuves traditionnelles ayant recours aux Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL) on aura compris que l'enjeu principal des épreuves standardisées est d'éviter que la réussite ou l'échec des étudiants soit tributaire de la mauvaise qualité des procédures et des instruments d'évaluation qui font la part belle à la subjectivité des examinateurs. Faut-il s'étonner que les USA nous précèdent depuis des décennies dans cette démarche de standardisation, eux qui sont obnubilés par la problématique de la non discrimination et de l'objectivité des processus de sélection ?

Malgré les biais bien connus des QROM/QROL et les difficultés liées à leur correction lorsque les professeurs doivent faire face à de grands groupes avec peu de moyens, nous ne préconiserons pas l'abandon de cette modalité de questionnement mais son amélioration et son usage en complémentarité avec les QCM standardisés. La procédure des « échelles descriptives » (De Bal, De Landsheere et Beckers, 1977) permet d'améliorer la fidélité de la correction des QROM/QROL qui demeurent indispensables pour mesurer certains types de performances complexes.

B. Les examens standardisés permettent-ils de faire mieux ?

L'exposé des points faibles des examens traditionnels ayant recours aux Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL) pourrait amener certains à conclure qu'il faut définitivement abandonner cette forme de questionnement et s'orienter exclusivement vers des épreuves standardisées ayant recours aux questions fermées. Nous ne le pensons pas. Nous commencerons par montrer dans cette section que les examens standardisés présentent eux aussi des faiblesses liées à l'utilisation des Questions à Choix Multiples (QCM) classiques[9]. Il existe des parades à ces problèmes, elles seront brièvement évoquées ici et décrites plus en détail dans le chapitre suivant. Plutôt que d'opposer les examens traditionnels à réponses ouvertes aux examens standardisés ayant recours aux QCM nous préconisons une complémentarité entre ces deux types d'approches lorsque les objectifs de l'évaluateur l'exigent. Nous terminerons en insistant sur le fait qu'il existe des épreuves standardisées dont les niveaux de qualité peuvent être très différents. Nous verrons à ce propos qu'on peut opposer une approche « amateuriste » à une autre plus « professionnelle » pour la construction des épreuves standardisées universitaires.

1. Les examens standardisés classiques sont sensibles à une série d'inconvénients inhérents aux questions à choix multiple classiques

Bien que l'automatisation de la correction permette d'échapper aux incohérences intra-correcteur ou/et inter-correcteurs, les examens standardisés ont aussi leurs inconvénients. Ces problèmes sont inhérents aux questions à choix multiple. Nous évoquerons ici cinq problèmes qui nous paraissent devoir être signalés mais qui ont tous leur parade.

a) Le piège de la parcellisation des connaissances

Les concepteurs de QCM sont souvent tentés de poser des questions de détail car plus la réponse attendue est précise et limitée à un contexte très particulier et moins elle sera susceptible d'être contestée. Les QCM doivent être « décidables », elles ne peuvent pas être sujettes à interprétation. Noizet et Caverni (1978) cités par Leclercq (1986, p. 31) signalent à ce propos : « *Cette condition de décidabilité est capitale puisque la réponse sollicitée est de nature 'vrai' ou 'faux'. Ainsi, par exemple, la question de savoir si tel écrivain a écrit ou non telle œuvre constitue un item décidable dans la mesure où la paternité d'une œuvre est susceptible d'être attribuée sans erreur. Mais la question de savoir si en dernière analyse tel écrivain peut être considéré comme un romantique constituerait un item indécidable* ».

Cette caractéristique entraîne une tendance à poser des questions de détails dans les examens standardisés. Les effets de cette tendance se manifestent alors aussi dans la façon dont les étudiants abordent l'étude du cours : ils se focalisent sur les détails en ne prenant par exemple plus la peine d'effectuer les liens entre les différentes parties du cours. Des solutions existent pour contrer cette *tendance à la parcellisation des connaissances*.

D'une part, Leclercq (1993) a montré qu'on peut améliorer les QCM à l'aide des Solutions Générales Implicites (SGI), (voir *infra* p. 67), ce qui permet à l'évaluateur d'éviter de ne poser que des questions de connaissance. Mais cette sophistication des consignes entraîne alors aussi d'autres difficultés (surmontables) liées à la nécessité d'un entraînement préalable des examinés. Nous verrons plus loin quels dispositifs nous mettons en place pour permettre aux étudiants de s'entraîner avant l'examen final (p. 79).

[9] Nous ajoutons l'appellation « classique » à QCM de façon à les différencier des Questions à Choix Multiple avec Solutions Générales Implicites (QCM-SGI) (Leclercq, 1993) qui permettent de palier à une série d'inconvénients que présentent les QCM *classiques*.

D'autre part, un accompagnement méthodologique des examinateurs par des personnes expérimentées dans le domaine de la création des QCM, diminue le risque de tomber dans le piège de la rédaction des questions de détails. Des procédures de contrôle de la qualité des questions existent et permettent d'améliorer la qualité des épreuves (qualité *a priori*). Il existe parmi les services offerts par le SMART (voir p. 59) des possibilités d'aide à l'analyse de la qualité *a priori* des QCM, notamment par la relecture formelle des questions en fonction d'une série de règles méthodologiques décrites par Leclercq (1986, pp. 79-144).

b) Le danger de la mémorisation des réponses incorrectes aux questions fermées

On peut craindre que les étudiants mémorisent les solutions incorrectes des QCM lors des examens standardisés. Cette crainte est ancienne et avait déjà été exprimée il y a une quarantaine d'années par Skinner (1961). Elle fut ensuite confirmée par Preston (1965) mais lors de son expérience, ce dernier n'avait pas fourni les solutions correctes aux sujets testés après l'épreuve.

C'est Karraker (1967) qui a démontré que ce danger n'est plus à craindre lorsqu'on communique les réponses correctes et que, au contraire, le testing via des QCM améliore les performances à une épreuve ultérieure ayant recours à des questions ouvertes construites sur la même matière. Les expériences de Preston et de Karraker liées au problème de la mémorisation des solutions fausses ont été décrites en détail par Leclercq (1986, pp. 35-40).

La parade de ce danger de mémorisation des réponses incorrectes se situe donc dans la rétroaction qui doit suivre rapidement la passation de l'examen. Nous verrons plus loin que le principal reproche que nous adressent les étudiants lorsqu'on leur demande leur avis sur les examens, se situe justement au niveau du manque de feedback (voir, p. 87). Nous présenterons au chapitre suivant une solution technologique que nous avons mis en place dans le cadre du SMART pour permettre une diffusion rapide et détaillée vers les étudiants et via l'Internet des résultats d'un examen (p. 85).

c) Le manque d'équité lorsque les possibilités de fraudes ne sont pas suffisamment prises au sérieux

La nature des réponses fournies aux examens standardisés ayant recours aux QCM : une lettre ou un chiffre désignant la réponse choisie, permet de poser beaucoup de questions et de simplifier la correction en l'automatisant. Malheureusement, le revers de la médaille réside dans le fait que ce type de réponses peut être assez facilement communiqué. Il suffit par exemple que les tricheurs conviennent avant l'épreuve d'un code gestuel différent pour chaque proposition d'une QCM. La communication des réponses devient alors un jeu d'enfant lors de la passation de l'épreuve. L'épreuve n'est dès lors plus équitable dans la mesure où les examinés qui ne cèderont pas à la tricherie seront pénalisés.

Des parades existent. Elles consistent à fournir des versions parallèles d'un même questionnaire. Les étudiants restent soumis aux mêmes questions, mais celles-ci sont mélangées au sein des différentes versions. Le danger d'épreuves standardisées biaisées par des fraudes est aussi lié à des conditions d'administration inadéquates. Les locaux où ont lieu ces épreuves doivent être adaptés à la taille des groupes d'étudiants évalués, assez grands que pour permettre d'espacer suffisamment les examinés. Le nombre de surveillants et le sérieux avec lequel ces derniers accomplissent leur tâche constituent aussi des facteurs à prendre en compte pour éviter la fraude.

En fait, l'organisation d'un examen, particulièrement lorsqu'on est confronté à un grand groupe d'examinés, demande un effort de préparation intense où de multiples facteurs doivent être passés en revue dont les mesures anti-fraude. En guise d'exemple à ne pas suivre, on se rappellera des évènements qui eurent lieu lors d'un grand concours de recrutement de la Communauté européenne en septembre 1998. Suite à ce concours, le quotidien Le Soir du 21 septembre 1998 titrait : « *Tricheries et contestations au*

concours de recrutement - La Commission européenne a raté son examen ». Voici un extrait de l'article[10] « *…Tous les témoignages concordent. Au Heysel, beaucoup ont pu tricher sans encourir la moindre sanction (…) on s'échangeait des réponses, on téléphonait carrément à l'extérieur, à l'aide d'appareils portables, pour obtenir des renseignements. (…) Dans un des centres d'examen à Rome, la situation était tellement confuse … qu'un candidat a appelé la police pour venir constater les irrégularités. Ce qui fut fait…* ». L'absence de procédures anti-fraude sérieuses et le manque de contrôle de la situation par les organisateurs de ce concours qui réunissait plus de 30.000 candidats à travers les quatre coins de l'Europe provoqua l'annulation de l'épreuve et la Commission européenne fut amenée à produire des excuses publiques.

Si l'on calcule en heures de travail perdues dans la préparation par les organisateurs et surtout dans la passation par les candidats, la perte est gigantesque sur le plan matériel. Sans parler des dommages moraux (pour les individus) et de l'image de l'Union Européenne.

d) Le problème des réponses aux hasard

Les solutions les plus anciennes au problème des choix au hasard dans les QCM ont consisté à augmenter le nombre de propositions ou/et à pénaliser les erreurs en appliquant la méthode de la *correction for guessing* qui sera exposée en détail plus loin (p. 104). Cependant ces deux solutions ont leurs inconvénients. En ce qui concerne l'augmentation du nombre de propositions, il est bien connu que certaines QCM n'ont « naturellement » que peu de solutions[11] et en ajouter revient à créer une proposition dont l'attractivité sera très faible. Pour ce qui est des pénalités en cas de réponses au hasard, les travaux de Leclercq, (1987) ont montré que la *correction for guessing* est d'autant plus un instrument à bannir, qu'il peut être remplacé par l'utilisation des probabilités subjectives qui, elles, (1) sont basées sur un modèle théorique plus pertinent, (2) ont un principe de notation plus équitable, (3) sont plus formatives pour l'apprenant et (4) plus informatives pour l'examiné et l'enseignant.

e) Les QCM ne permettent pas de mesurer tous les types de performances

Cette critique est souvent formulée à l'encontre des QCM et il est vrai que ce type de questionnement ne permet pas actuellement de mesurer la capacité à rédiger, à inventer, à s'exprimer oralement... A ce propos, Leclercq (1986, p. 34) rappelle : « *Les QCM sont un outil parmi d'autres : il importe de recourir au mode d'évaluation le plus adéquat à chaque situation. Les QCM conviennent moins bien pour les performances complexes (réponse longue où la structure et l'expression jouent un grand rôle) que pour les performances isolables* ».

Si l'évaluation de certains types de performances ne peut en effet s'effectuer qu'à l'aide de réponses ouvertes, il faut associer cette forme de questionnement aux examens standardisés quand les objectifs de ces derniers l'exigent. C'est la solution que nous proposons aux enseignants qui font appel aux services du Système Méthodologique d'Aide à la Réalisation de Tests (SMART) à l'université de Liège.

On pourrait également craindre qu'avec la généralisation des examens standardisés et l'impossibilité d'exprimer une réponse longue à l'aide des QCM, les étudiants soient de moins en moins sollicités à s'exprimer par écrit ou oralement et donc « perdent » cette compétence. Ceci constitue à notre avis un argument supplémentaire en faveur de formules d'évaluation faisant appel à des combinaisons QCM-QROM/QROL.

[10] Article signé par André Riche.

[11] Par exemple, le genre d'un nom commun ne peut être que féminin, masculin ou neutre, dès lors, inutile d'inventer une quatrième solution étant donné que les propositions sont « imposées » par la nature du contenu de la question.

2. La nécessité d'une complémentarité entre QCM (de qualité) et QROM/QROL (améliorées) lorsque les objectifs de l'évaluation l'exigent

L'exposé des principaux biais liés aux Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL) pourrait amener le lecteur à croire que ce type de questionnement est à proscrire. Mais comme nous venons de le voir, les QROM/QROL demeurent indispensables pour évaluer certains types de performances complexes : capacité à synthétiser, développement de raisonnements, créativité, aptitude à s'exprimer par écrit, etc. qui ne peuvent pas être mesurées à l'aide de questions fermés.

Malheureusement, la façon dont les QROM/QROL sont habituellement corrigées fait la part belle à la subjectivité des examinateurs. Cependant, la fidélité des correcteurs peut être améliorée, De Bal, De Landsheere et Beckers (1977) ont en effet montré à l'aide des « *Echelles descriptives en évaluation* » qu'il était possible d'améliorer la cohésion inter-correcteurs en découpant le « trait » à évaluer en plusieurs facettes et en veillant à définir clairement chaque échelon des échelles mises au point pour mesurer les aspects à évaluer. Cependant, c'est un autre inconvénient qui apparaît alors : la tâche demandée à l'évaluateur est encore plus complexe vu l'effort d'analyse nécessaire pour séparer les différentes catégories de réponses. Cette technique des échelles descriptives augmente la fidélité de la correction mais ne diminue pas le temps de correction. Ceci ne favorisera pas l'augmentation du nombre de QROM/QROL dans les épreuves lorsque de grands groupes d'étudiants sont évalués. C'est dommage car plus de QROM/QROL lors des examens permettrait de couvrir un plus large éventail de matières enseignées, donc d'améliorer la validité de contenu liée à ce type de questionnement.

D'autres techniques permettant d'augmenter la fidélité de la correction des QROM/QROL existent. Par exemple celle qui implique une double correction et qui consiste à faire corriger les copies indépendamment par deux correcteurs (A et B) qui se sont préalablement concertés sur les critères de correction. Après avoir corrigé les copies, les correcteurs confrontent question par question chaque paire de note (celle du correcteur A avec celle du correcteur B), lorsque les résultats coïncident la note est confirmée. Lorsque deux résultats diffèrent, les correcteurs se concertent en vue de comprendre ce qui a fait les différences d'appréciation. Malheureusement, cette méthode est coûteuse en temps et en personnes.

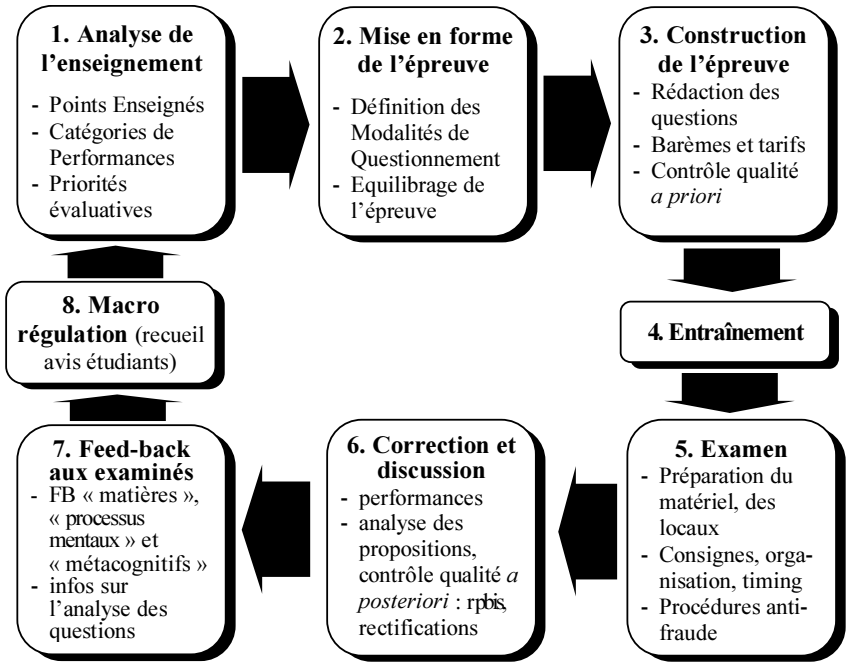
En fait, les questions à réponses ouvertes moyennes ou longues écrites améliorées à l'aide de techniques telles que les échelles descriptives, devraient être utilisées en complémentarité avec les épreuves standardisées ayant recours aux QCM. C'est ce que nous préconisons dans le cadre de l'aide méthodologique que nous proposons aux enseignants de l'Université de Liège via le Système Méthodologique d'Aide à la Réalisation de Tests (SMART) (voir p. 57). Dans ce contexte, nous sommes souvent amené à conseiller aux examinateurs lorsque leurs objectifs évaluatifs l'imposent, d'utiliser une ou deux QROM/QROL en combinaison avec de nombreuses QCM (au minimum une trentaine). Ainsi, des QROM/QROL corrigées avec la technique des échelles descriptives permettent de mesurer des performances complexes du type de celles que nous avons évoquées plus haut tandis que le nombre élevé de QCM permet une couverture large de la matière. Une partie du gain de temps considérable lié à la correction automatisée des QCM peut ainsi être réinvestie dans la correction plus poussée d'une ou deux QROM/QROL avec échelles descriptives.

Nous verrons plus loin qu'à l'aide de QCM avec Solutions Générales Implicites (QCM-SGI) (voir p. 67) nous pouvons évaluer d'autres niveaux de processus mentaux qui ne peuvent l'être, ou alors moins systématiquement, avec les QCM classiques.

3. Approche « amateuriste » et approche « professionnelle » dans la réalisation des examens standardisés

Nous pensons qu'il est nécessaire d'examiner de plus près huit étapes dans la réalisation des examens standardisés universitaires (Gilles et Leclercq, 1995). Ces huit étapes font partie d'un processus cyclique en « spirale de qualité » qui sera décrit en détail au chapitre suivant.

Les huit étapes sont résumées dans les encadrés du schéma ci-contre (détails p. 74).



Selon nous, la construction d'un examen standardisé de qualité doit passer par ces huit étapes. Tout au long de la réalisation, le(s) concepteur(s) de l'évaluation devront veiller à rencontrer dans les procédures qu'ils mettent en œuvre, une série d'objectifs « qualité ».

D'une part, un examen standardisé ne pourra offrir des garanties sérieuses en ce qui concerne l'atteinte des objectifs « qualité » que nous listons ci-dessous (ils seront aussi détaillés plus loin dans le premier chapitre, p. 63) qu'à la condition que la construction de l'épreuve soit envisagée dans le cadre d'un processus global comportant une série d'étapes dont les procédures et les dispositifs d'ingénierie docimologiques sont étudiés pour rencontrer ces objectifs « qualité ». Il est important de prévoir, au sein des procédures utilisées dans les phases de construction de l'examen, des micro régulations qui permettent d'améliorer la qualité du produit/service fourni.

D'autre part, nous pensons que la qualité des examens peut être améliorée d'épreuve en épreuve notamment à l'aide des avis des étudiants. Cette amélioration de la qualité des tests passe aussi par le perfectionnement du processus qui a permis de les construire, c'est l'idée du cycle en « spirale de qualité » de réalisation des examens standardisés. A ce point de vue, la dernière étape « 8. Macro régulation » est cruciale car elle permet de récolter les remarques des examinés en vue d'améliorer la réalisation des examens suivants.

Voici nos **objectifs « qualité »**. Lors de la construction d'un examen standardisé universitaire de qualité il s'agit d'offrir des garanties en ce qui concerne

- la ***validité*** : les scores des étudiants doivent refléter ce que l'enseignant veut mesurer ;
- la ***fidélité*** : un travail corrigé et classé dans une catégorie donnée doit bénéficier de la même mention s'il est corrigé dans d'autres conditions, par exemple par d'autres correcteurs ou quelques semaines plus tard ;
- la ***sensibilité*** : la mesure doit être précise ;
- la ***diagnosticité*** : le diagnostic précis des difficultés d'apprentissage, des processus maîtrisés et de ceux qui ne le sont pas doit être possible ;
- la ***praticabilité*** : la faisabilité en termes de temps, de ressources en personnel et en matériel doit est assurée ;

- l'***équité*** : tous les étudiants doivent être traités de façon juste, en principe de la même manière (standardisation) ;
- la ***communicabilité*** : les informations non confidentielles relatives au déroulement du processus doivent être communiquées et comprises par les partenaires, enseignants, étudiants, équipe de soutien docimologique (par exemple le SMART), …, engagés dans la réalisation des épreuves.

On peut dès lors opposer deux types d'approches. La première, « professionnelle », a l'ambition d'envisager le problème de la réalisation des examens universitaires de façon globale et avec la volonté de garantir la qualité et le sérieux auxquels les étudiants ont droit. C'est en effet leur sort académique qui est en jeu dans ces épreuves. La seconde, que nous qualifierons (sans doute avec excès) d'« amateuriste » consiste à négliger ou à ne pas prendre en compte une série d'objectifs « qualité » et à faire l'impasse sur des étapes pourtant cruciales dans la construction d'un examen. Cette dernière approche est souvent liée à un contexte de surcharge de travail chez certains enseignants universitaires. Dans ce cas, beaucoup préfèrent alors mettre la priorité sur les tâches de recherche plutôt que sur celles liées à l'enseignement. Il faut dire que le corps académique était, jusqu'il y a peu, essentiellement évalué sur le nombre et la qualité de ses publications dans les revues scientifiques et très peu sur la qualité des enseignements dispensés.

Depuis quelques années, en Communauté Française de Belgique (CFB), le Conseil des Recteurs des institutions universitaires Francophones (CReF) a confié à un groupe de travail la mission de mettre en place une évaluation de la qualité de ses neuf institutions universitaires (Boucher & al. 1997). Dans les rapports d'auto-évaluation et plus tard d'évaluation externe, les institutions doivent et devront aussi répondre à des questions qui envisagent la qualité des enseignements. La qualité des examens universitaires dispensés au sein d'une institution pèsera probablement encore plus qu'aujourd'hui dans ces évaluations. Ce mouvement d'étude de la qualité des institutions universitaires est mondial et nous aurons l'occasion d'y revenir plus loin (p. 60)

On voit bien tout le bénéfice que peuvent retirer les professeurs (et par-delà l'institution universitaire de plus en plus amenée à devoir fournir la preuve de la qualité de ses enseignements) d'une structure d'appui méthodologique et logistique d'aide à la réalisation d'examens standardisés comme celle qui fonctionne actuellement à l'Université de Liège. Nous décrirons cette structure, le Système Méthodologique d'Aide à la Réalisation de Tests (SMART), en détail dans le chapitre suivant. Signalons ici que le SMART fournit une série de services qui s'insèrent dans la perspective du cycle en « spirale de qualité » et qui tiennent compte des objectifs « qualité » décrits plus haut. Le SMART propose aux professeurs de les aider dans :

- le choix du(des) type(s) de questionnements, de la(des) méthode(s) de test la(les) plus appropriée(s) ;
- la gestion des banques de questions ;
- l'analyse de la qualité *a priori* des épreuves (tests formatifs ou examens) ;
- l'entraînement des étudiants aux procédures d'évaluation ;
- la préparation et la logistique des épreuves ;
- la correction des tests à l'aide de procédures informatisées ;
- l'analyse de la qualité *a posteriori* des questions ;
- la mise en œuvre de solutions en vue de rectifier les épreuves lorsque des problèmes sont détectés ;
- la réalisation et le renvoi vers les étudiants des feedbacks individualisés relatifs à leurs performances et à la qualité de l'épreuve ;
- le recueil et l'analyse des avis des étudiants sur la qualité des épreuves.

***

En synthèse :

Les examens standardisés présentent une série de points faibles. Ceux-ci sont liés aux problèmes (bien connus) que peuvent présenter les QCM classiques. Nous verrons plus loin que les QCM-SGI (p. 67) et les pourcentages de certitude (p. 68) permettent d'améliorer notablement la qualité des épreuves ayant recours aux QCM.

Les examens standardisés offrent potentiellement de nombreux avantages : étudiants tous traités de la même façon (équité), correction automatisée (fidélité), large éventail de la matière évalué (validité de contenu), systématisation des pourcentages de certitude (sensibilité), rétroactions rapides à l'aide de feedbacks détaillés automatisés (diagnosticité et communicabilité), …

Malgré tous ces avantages, il n'en reste pas moins vrai que certaines performances complexes ne peuvent être évaluées à l'aide des QCM (par exemple la capacité à s'exprimer par écrit). Dès lors, lorsque les objectifs de l'évaluateur l'exigent, nous préconisons une utilisation combinée de l'approche standardisée et de l'approche plus traditionnelle d'évaluation à l'aide de questions ouvertes (améliorées notamment à l'aide des « Echelles descriptives en évaluation » de De Bal, De Landsheere et Beckers, 1977).

Ceci étant, les avantages potentiels liés aux épreuves standardisées et à une éventuelle approche combinée avec les questions ouvertes ne constituent pas en soi des garanties automatiques de qualité. Nous pensons que cette qualité ne peut être obtenue que dans le contexte d'une approche « professionnelle » de la réalisation des examens universitaires. Cette approche est selon nous sous-tendue par deux principes de base : (1) la qualité d'un examen standardisé universitaire ne peut être établie qu'à la condition que sept objectifs « qualité » soient atteints : validité, fidélité, sensibilité, diagnosticité, praticabilité, équité, communicabilité et (2) pour atteindre ces objectifs « qualité » il faut concevoir la construction de l'épreuve à l'aide d'un processus où les régulations permettent non seulement l'amélioration de l'examen en cours de construction, mais aussi l'amélioration du processus de construction lui-même (« spirale de qualité », voir p. 74).

La mise en pratique de cette approche d'amélioration de la qualité dans la construction des examens universitaires est difficilement réalisable sans qu'une aide méthodologique et logistique soit fournie aux professeurs. Le système Méthodologique d'Aide à la Réalisation de Tests (SMART) de l'Université de Liège (ULg) propose une série de services qui vont dans ce sens.

Dès lors, la réponse à la question que nous nous étions posée dans le titre de cette section « Les examens standardisés permettent-ils de faire mieux ? » peut se résumer, à ce stade, de la façon suivante.

Les épreuves standardisées permettent d'évaluer un large spectre de performances, mais pas toutes. Dans certaines situations une approche combinée incluant des questions à « réponse ouverte » dont la fidélité de la correction sera améliorée doit être recommandée. Dans tous les cas, pour faire mieux, les objectifs « qualité » entrevus ci-dessus (voir détails p. 63) devront être atteints, ce qui implique de concevoir la construction des examens à l'aide d'un processus méthodique dont les régulations permettent une amélioration continue des épreuves. L'analyse de la qualité a posteriori des questions est effectuée lors de la sixième étape « correction et discussion » de ce processus. C'est dans cette étape que s'insèrent les nouveaux indices d'analyse de la qualité spectrale qui sont au cœur de cette thèse.

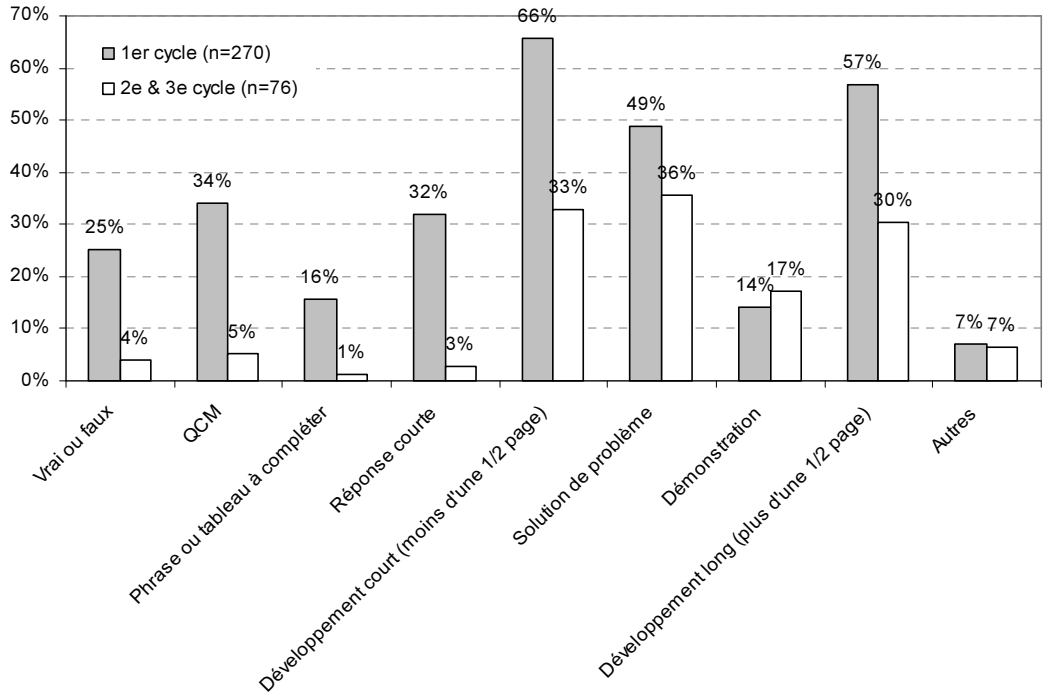
C. Fréquences des modalités de questionnement utilisées dans les examens universitaires

Dans cette première partie nous exposerons les données d'une enquête réalisée à l'Université de Montréal (Blais & al., 1997) relatives à l'utilisation des modalités de questionnement dans les examens universitaires. Nous verrons que dans le 1er cycle d'études de cette institution les questions à réponses fermées ou semi-fermées sont utilisées mais moins fréquemment que les questions à réponses ouvertes moyennes ou longues. L'enquête montre également que dans le second cycle les questions fermées ou semi fermées ne sont pratiquement pas utilisées. Nous observerons ensuite à l'aide des données d'une autre étude réalisée dans le 1er cycle de la Faculté de Psychologie et des Sciences de l'Education (FAPSE) de l'Université de Liège (Debry & al., 1999), la répartition des modalités d'examen en 1ère et 2ème candidature à la FAPSE ainsi que les taux de réussite qui sont associés à ces différentes modalités.

1. Modalités d'examen final à l'Université de Montréal et ses écoles affiliées

Une enquête effectuée à l'Université de Montréal et dans ses écoles affiliées (Blais & al., 1997) par le Groupe de Recherche Interdisciplinaire en Pédagogie Universitaire[12] (GRIPU) montre qu'il existe des différences dans la répartition des types de tâches d'examen selon que ceux-ci se déroulent au 1er ou au 2ème cycle dans cette institution. Voici la répartition des réponses[13] (270 professeurs interrogés au 1er cycle et 76 aux 2ème et 3ème cycles).

Tâches d'examen au 1er et aux 2e & 3e cycles à l'Université de Montréal
(d'après BLAIS & al., 1997, pp. 48-49)



Les quatre premières modalités de questionnement : « *Vrai ou faux* », « *QCM* », « *Phrase ou tableau à compléter* » et « *Réponses courtes* » peuvent être assimilées à des types de questionnements à réponses fermées ou semi-fermées. Les quatre suivantes : « *Développement court (moins d'une ½ page)* »,

[12] Faculté des Sciences de l'Education, Université de Montréal, Québec - Canada.

[13] Les professeurs intérrogés pouvaient exprimer le fait qu'il utilisent plusieurs types de tâches d'examen, ceci explique que lorsqu'on effectue la somme des pourcentages pour le 1er cycle ou les 2ème & 3ème cycle, on dépasse 100%.

« *Solution de problème* », « *Démonstration* » et « *Développement long (plus d'une ½ page)* » sont quant à elles du type Questions à Réponses Ouvertes Moyenne (QROM) ou Longues (QROL).

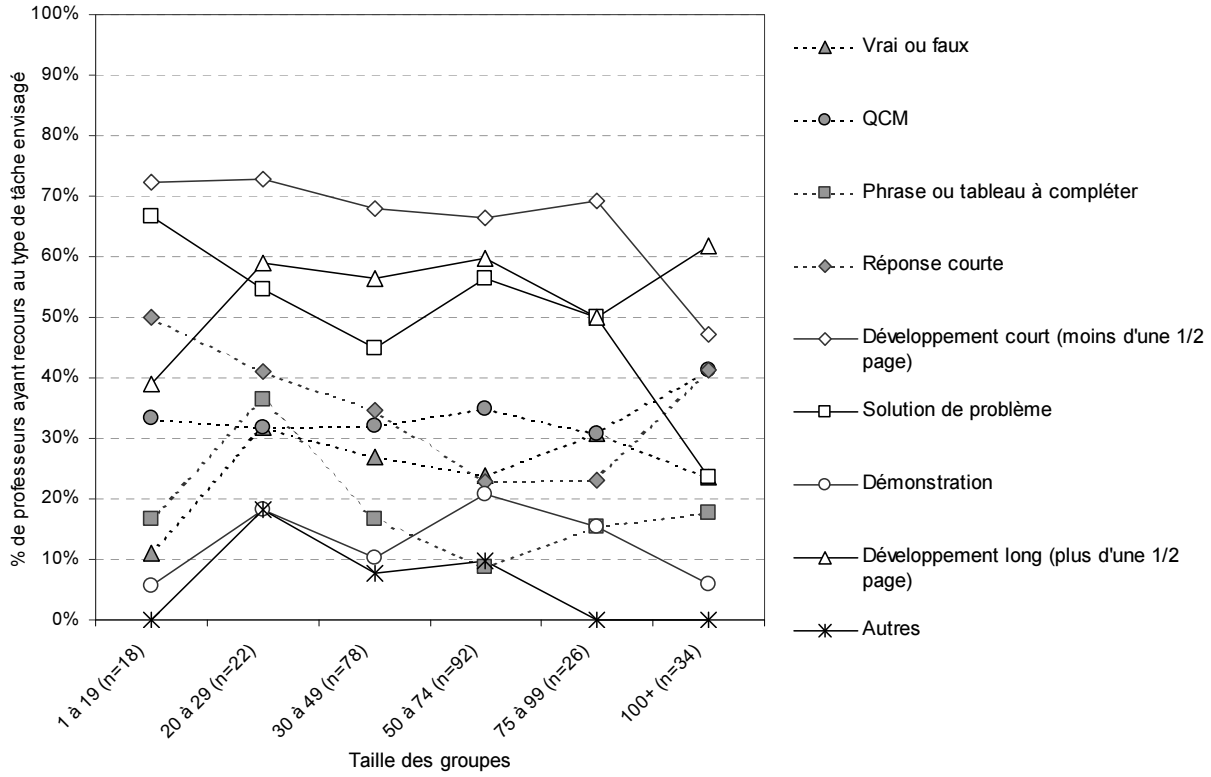
On voit que dans le 1er cycle toutes les modalités de questionnement sont utilisées. La plus fréquente est celle qui consiste à demander un développement court de moins d'une ½ page (66% des interrogés du 1er cycle l'utilisent). On remarque des pourcentages d'utilisation moins élevés pour les formes de questionnement à réponses fermées ou courtes tels que les « *Vrai ou faux* » (25%), les « *QCM* » (34%), les « *Tableaux ou phrases à compléter* » (16%) et les questions à « *Réponse courte* » (32%) que pour la plupart des formes de questionnement à réponses ouvertes moyennes ou longues : « *développement court…* » (66%), « *Solution de problème* » (49%), « *Démonstration* » (14%) et « *Développement long…* » (57%).

La situation est différente aux 2ème et 3ème cycles. Les modalités de questionnement à réponses fermées ou courtes sont très peu utilisées : entre 1% et 5% des interrogés y ont recours. Les questions à réponses ouvertes moyennes ou longues sont les plus prisées : « *Développement court…* » (33%), « *Solution de problème* » (36%), « *Démonstration* » (17%) et « *Développement long…* » (30%).

Donc, on rencontre fréquemment les modalités de questionnement à réponses fermées ou courtes dans le 1er cycle mais pratiquement pas dans les 2ème et 3ème cycles. Le nombre d'étudiants évalués peut-il expliquer la présence des questions à réponses fermées ou semi-fermées dont la correction est moins longue ?

Le graphique ci-dessous montre la répartition des tâches d'examen en fonction de la taille des groupes dans le 1er cycle (le nombre de professeurs interrogés diffère selon la taille des groupe).

Tâches d'examen en fonction de la taille des groupes, 1er cycle Université de Montréal (d'après BLAIS & al., 1997, p. 49)

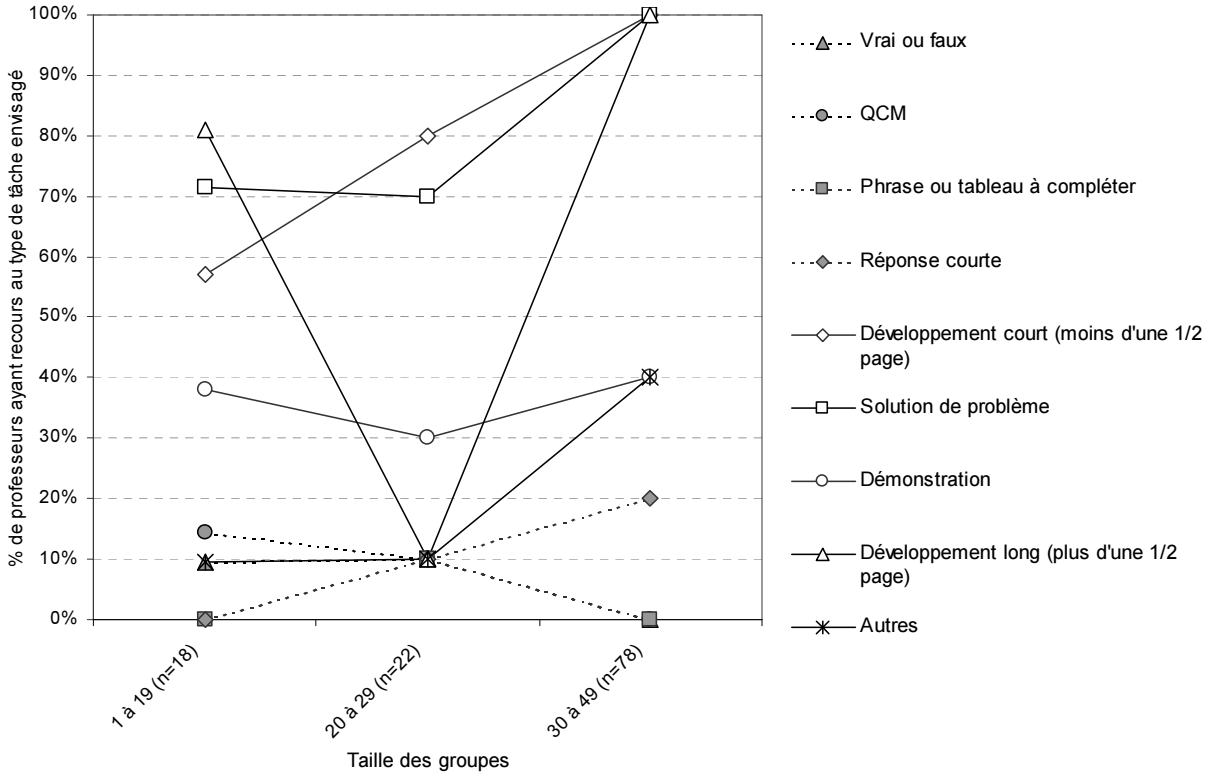


On remarque que même lorsque la taille des groupes est supérieure ou égale à 100 étudiants [100+ (*n*=34)], les questions à réponses ouvertes longues (« *Développement long (plus d'une ½ page)* ») et moyennes (« *Développement court (moins d'une ½ page)* ») sont les plus utilisées par les professeurs.

Le graphique suivant montre la répartition des tâches d'examen en fonction de la taille des groupes dans les 2ème et 3ème cycles. Nous remarquons l'absence de groupes dont la taille est supérieure à une cinquantaine d'étudiants, les auteurs signalent à ce propos (*op. cit.*, p. 48) : « *Au 2ème cycle, nous avons étudié uniquement les groupes dont la taille était inférieure à 50 (il y a trop peu de groupes nombreux pour que la description ait un sens)* ».

Là où la taille des groupes est la plus élevée (entre 30 et 49 étudiants), les modalités de questionnement à réponses ouvertes moyennes ou longues sont utilisées par tous les professeurs interrogés alors que les questions à réponses fermées telles que les « *QCM* » et les « *Vrai ou faux* » ne sont pas utilisées. Parmi les questions à réponses fermées ou semi-fermées, seules les questions à réponses courtes sont un peu plus utilisées lorsque la taille du groupe est plus importante.

Tâches d'examen en fonction de la taille des groupes, 2e & 3e cycle Université de Montréal (d'après BLAIS & al., 1997, p. 49)



L'équipe du GRIPU commente les chiffres : « *Dans l'ensemble, on peut dire que la taille du groupe a peu d'impact sur la composition de l'examen au 1er cycle* (…). *Une assez grande variété de tâches, des plus simples et des plus complexes, entrent dans la composition de celui-ci. D'autre part, au deuxième cycle, on favorise presque exclusivement, peu importe la taille du groupe, des tâches complexes comme le développement long et la solution de problème* » (*op. cit.* p. 48).

Dans leurs conclusions, les auteurs ajoutent à propos des modalités d'évaluation :

« *Les analyses des données permettent de dire qu'il existe un lien entre la taille des groupes et le type d'habileté qu'on évalue dans les examens. Egalement, on semble favoriser par le biais de l'évaluation le développement d'habiletés différentes selon les cycles. Dans les grands groupes, au premier cycle, la production et la création sont les habiletés les moins visées par les modalités d'évaluation, sauf dans les cours de formation pratique et les stages. Les échecs sont les plus rares au niveau des cycles supérieurs* ». (*op. cit.* p. 120)

Nous retiendrons que la situation est donc assez différente dans le 1er cycle par rapport aux 2ème et 3ème cycles. Contrairement à ce qu'on pourrait croire, au 1er cycle, malgré la taille des groupes plus

importante, les questions à réponses ouvertes moyennes et longues sont plus utilisées que les questions à réponses fermées et semi-fermées. Dans le second cycle, même quand la taille des groupes est relativement élevée, les questions fermées et semi-fermées ne sont pas ou très peu utilisées. Il faut cependant ajouter que pour les 2ème et 3ème cycles l'étude réalisée à l'Université de Montréal ne mentionne pas les données relatives aux groupes d'étudiants dont la taille est supérieure à 50 étudiants.

Enfin, signalons qu'au 1er cycle, les sections dont la taille est supérieure ou égale à 100 étudiants sont regroupées dans une seule catégorie « 100+ (*n*=38) ». On ne nous dit donc pas si les conclusions quant aux modalités d'examens sont encore valables lorsque de très grands groupes sont concernés (par exemple entre 300 et 400 étudiants). Dans la partie suivante nous verrons qu'en 1ère candidature à la Faculté de Psychologie et des Sciences de l'Education (FAPSE) de l'Université de Liège (ULg), la taille du groupe est de 379 étudiants[14] pour l'année académique 1997-1998.

[14] Debry & al. (1999, p. 8).

2. Constats à propos des examens organisés en 1997-1998 dans le premier cycle FAPSE-ULg

Voyons maintenant quelle est la répartition des modalités d'examens dans le premier cycle de la FAPSE-ULg. Nous nous baserons sur les chiffres de l'année académique 1997-1998 fournis par l'enquête intitulée « Observations sur les Etudes Universitaires à la Faculté (ŒUF) » (Debry & al., 1999). Nous allons envisager les taux d'utilisation ainsi que les taux de réussite, d'une part pour les épreuves écrites et orales ayant recours aux Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL) et, d'autre part, pour les tests standardisés ayant recours aux vrai-faux (VF) ou aux Questions à Choix Multiple (QCM).

a) Répartition des modalités d'examens

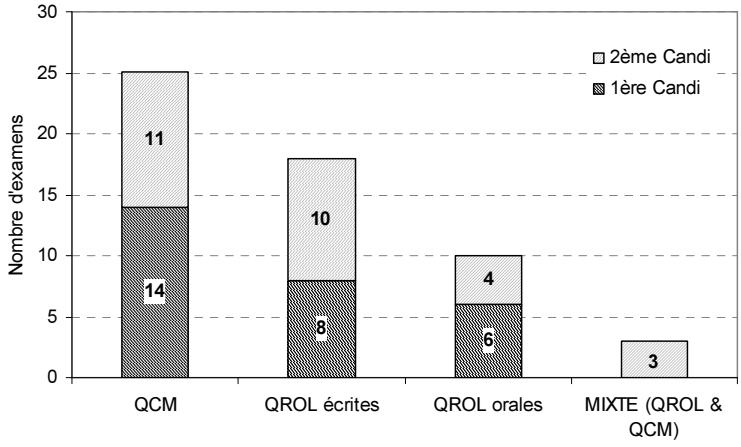
Dans le cadre d'une étude réalisée sur les données des épreuves de l'année académique 1998-1999 (Debry & al., 1999), une équipe de professeurs, chercheurs et étudiants a analysé les taux de réussite et les modalités d'examen dans le 1er cycle d'étude de la FAPSE-ULg.

Ce travail a abouti à un rapport intitulé « *Observations sur les Etudes Universitaires à la Faculté (ŒUF)* ». Il montre notamment (*op. cit.*, pp. 14-16) que trois modalités d'examens sont essentiellement utilisées en candidatures :

- les examens oraux, donc avec Questions à Réponses Ouvertes Moyennes ou Longues (QROM ou QROL) orales, désignés par « ORAL » dans les tableaux ci-dessous ;
- les examens écrits, donc avec QROM ou QROL écrites, désignés par « ECRIT » dans les tableaux qui suivront ;
- les examens ayant recours aux vrai-faux ou aux Questions à Choix Multiple (désignés par « QCM »).

Le graphique ci-contre montre la répartition de ces différentes modalités de questionnement dans les 56 épreuves sommatives certificatives que comptent les deux candidatures du 1er cycle FAPSE (1ère et 2ème sessions confondues)

Répartition des modalités de questionnement dans le 1er cycle FAPSE (année académique 1997-1998)



En 1ère candidature (1ère et 2ème sessions confondues), sur les 28 épreuves, en 1997-1998 la moitié étaient des épreuves avec QROM/QROL (8 QROM/QROL écrites et 6 QROM/QROL orales), l'autre moitié étaient des examens standardisés ayant recours à des VF/QCM. Rappelons que la taille de la section en 1ère candidature est de 379 étudiants.

Nous constatons qu'en 2ème candidature 11 épreuves étaient standardisées avec VF/QCM, 10 étaient de type QROM/QROL écrites, 4 furent réalisées à l'aide de QROM/QROL orales et 3 en ayant recours aux QROM/QROL et aux QCM.

En résumé, les QROM/QROL écrites ou orales sont donc utilisées dans le cadre de 28 examens sur 56 (un examen sur deux). Un peu moins de l'autre moitié des examens étant constituée d'épreuves standardisées ayant recours aux VF/QCM.

Comparée à la situation décrite par le GRIPU au niveau des premiers cycles de l'Université de Montréal, le premier cycle de la FAPSE fait donc un peu plus appel aux épreuves standardisées ayant recours aux VF/QCM.

b) Taux de réussite liés aux examens avec QROL et aux examens standardisés avec QCM

L'équipe de la FAPSE qui a réalisé le rapport « ŒUF » s'est aussi longuement penchée sur les taux de réussite en 1ère et 2ème session pour les différents cours qui font partie du programme du 1er cycle.

(1) En 1ère candidature

Voici les données portant sur l'étude de la réussite par cours en 1ère et en 2ème session pour la 1ère candidature de l'année académique 1997-1998 (*op. cit.*, pp. 14-15). Nous avons mis en évidence à l'aide de caractères gras les examens qui ont recours aux QCM. Remarquons qu'en 1ère session 8 examens sur 14 utilisent les QCM et en 2ème session 6 sur 14.

	1ère SESSION (*n* = 379)						2ème SESSION (*n* = 202)					
Code Cours	% absents ou cotes de présence	% réussite des présen-tants	Moy.	Ecart type	Modalité d'examen	Indice de réussite[15]	% absents ou cotes de présence	% réussite des présen-tants	Moy.	Ecart type	Modalité d'examen	Indice de réussite
1C1	30%	54%	10,8	3,7	ORAL	37,8	45%	68%	9 ,7	3,3	ORAL	38
1C2	10%	48%	9	5,2	ECRIT	43,2	35%	62%	9	4,3	ECRIT	40,3
1C3	1%	57%	10,6	2,5	**QCM**	56,4	48%	72%	9,7	2,4	**QCM**	37,4
1C4	20%	65%	11,5	4	ORAL	52	46%	80%	10,9	3,9	ORAL	43,2
1C5	4%	55%	11,1	2	ECRIT	52,8	36%	66%	10	1,9	ECRIT	42,2
1C6	5%	51%	9,4	3,4	**QCM**	48,4	32%	75%	10,1	3,7	**QCM**	51
1C7	14%	65%	12	3	**QCM**	55,9	44%	76%	10,8	2,8	**QCM**	42,5
1C8	0,5%	62%	11,9	3	**QCM**	61,7	56%	84%	9,2	3,2	ORAL	36,9
1C9	6%	66%	10,7	3	ECRIT	62	38%	61%	9,3	2,2	ECRIT	37,8
1C10	11%	64%	11,3	3,9	**QCM**	56,9	48%	89%	11,4	5	**QCM**	46,3
1C11	7%	65%	12,4	3,1	**QCM**	60,4	49%	84%	10,8	2,5	ORAL	42,8
1C12	10%	76%	11,7	3,5	ECRIT	68,4	39%	79%	9,4	4	ECRIT	48,2
1C13	12%	85%	11,5	3,1	**QCM**	82,3	57%	94%	9,4	3	**QCM**	40,4
1C14	3%	77%	12,5	2,7	**QCM**	74,7	41%	87%	9,7	3	**QCM**	51,3
Moyenne	8,8%	63%	11,6	3,5		58	44%	77%	10	3,2		42,7

Le tableau ci-dessous montre des taux de réussites différents en fonction du type d'épreuves.

	QROL ORALES		QROL ECRITES		QCM STANDARDISES	
	% réussite des présentants	Indice de réussite	% réussite des présentants	Indice de réussite	% réussite des présentants	Indice de réussite
1ère SESSION	60%	44,9	61%	56,6	66%	62,1
2ème SESSION	79%	40,2	67%	42,1	82%	44,8

Nous remarquons qu'aussi bien en 1ère et qu'en 2ème session les pourcentages de réussite et l'indice de réussite (le pourcentage d'étudiants présentant l'examen multiplié par le pourcentage de réussite) sont en moyenne meilleurs pour les QCM (épreuves standardisées). Par contre, en 1ère session les oraux obtiennent

[15] Indice de réussite = pourcentage de présentants x pourcentage de réussites.

en moyenne les moins bons pourcentages et indices de réussite, tandis qu'en deuxième session ce sont les écrits qui sont en moyenne moins bons en termes de réussite.

(2) En 2ème candidature

Voici maintenant les données portant sur l'étude de la réussite par cours en 1ère et en 2ème session pour la 2ème candidature (op. cit., p. 16). Des appellations « MIXTE » apparaissent dans ce tableau, elles ne concernent que 3 épreuves où différentes modalités de questionnements[16] sont utilisées lors des examens de 1ère session.

	1ère SESSION (*n* = 205)						2ème SESSION (*n* = 131)					
Code Cours	% absents ou cotes de présence	% réussite des présen-tants	Moy.	Ecart type	Modalité d'examen	Indice de réussite	% absents ou cotes de présence	% réussite des présen-tants	Moy.	Ecart type	Modalité d'examen	Indice de réussite
2C1	20%	41%	9,8	3,9	ECRIT	32,8	23%	31%	9	3,1	ECRIT	23,8
2C2	17%	47%	9,3	4,6	ECRIT	39	21%	50%	9,7	3,5	ECRIT	39,5
2C3	20%	76%	12,6	4	MIXTE	60,8	20%	62%	9,4	4	ORAL	49,6
2C4	3%	63%	10,7	3,4	ECRIT	61	17%	69%	11	3,2	ORAL	57,2
2C5	4%	69%	12,3	2,2	**QCM**	66,2	15%	69%	9,7	3,2	ECRIT	58,6
2C6	17%	72%	12,1	4,2	ECRIT	59,7	21%	83%	12,3	4,2	ECRIT	65,6
2C7	1%	61,5%	10,4	2,3	ECRIT	60,8	8%	79%	11,3	2,3	ECRIT	72,7
2C8	0%	62%	12,5	3,6	**QCM**	62	6%	80%	11,4	2,8	**QCM**	75,2
2C9	5%	66%	11,8	1,8	**QCM**	62,7	13%	87%	12,3	1,5	**QCM**	75,7
2C10	3%	70%	13	3	**QCM**	67,9	11%	87%	13,1	3,5	**QCM**	77,4
2C11	0%	59%	11,4	3,1	**QCM**	59	3%	89%	11,3	2,8	**QCM**	82
2C12	0%	70%	12,3	2,4	**QCM**	70	6%	81%	11,9	2,8	**QCM**	76,1
2C13	7%	91%	13,6	2,4	MIXTE	84,6	13%	80%	11,9	1,7	ORAL	69,6
2C14	0%	90%	14	2	MIXTE	90	11%	99%	13,3	1,4	ORAL	88
Moyenne	7%	67%	11,8	3		63,3	13,5%	74,5%	10,5	2,9		58,5

Les questionnements utilisés étaient de type QROM/QROL écrites dans 5 cas, de type VF/QCM dans 6 cas et dans les 3 cas restant plusieurs types de questionnements furent utilisés. En 2ème session les oraux (QROM/QROL orales) réapparaissent dans 4 épreuves, 5 examens sont des écrits ayant recours aux QROM/QROL et 5 autres sont des épreuves standardisées avec VF/QCM.

Les chiffres du tableau ci-dessous[17] montrent que les examens standardisés obtiennent les meilleurs taux et indices de réussite quelle que soit la session (comme en 1ère candidature). En 2ème session les examens avec QROM/QROL orales obtiennent un pourcentage et un indice de réussite en moyenne plus élevés que les QROM/QROL écrites.

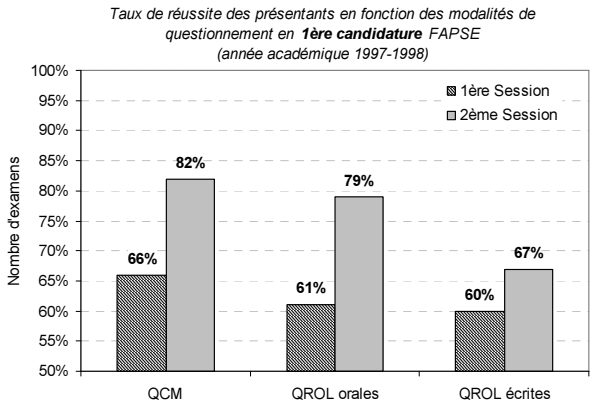
	QROM/QROL ORALES		QROM/QROL ECRITES		VF/QCM STANDARDISES	
	% réussite des présentants	Indice de réussite	% réussite des présentants	Indice de réussite	% réussite des présentants	Indice de réussite
1ère SESSION	-	-	57%	50,7	66%	64,6
2ème SESSION	78%	66,1	62%	52	85%	77,3

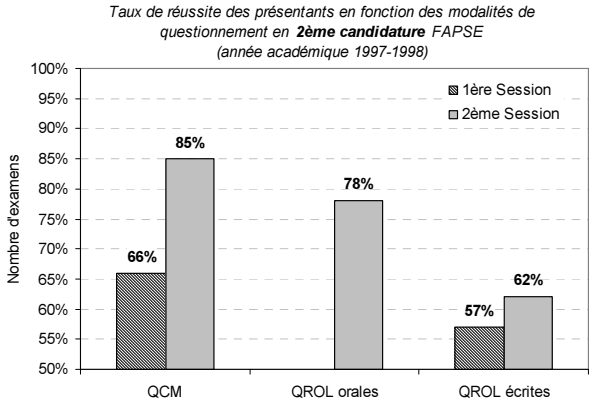
[16] Le cours 2C13 étant évalué à l'aide d'un examen dispensatoire écrit et d'un examen oral et le 2C14 par un QCM et une épreuve orale.

[17] Les données ne tiennent pas compte des 3 examens de type « MIXTE ».

(3) Des taux de réussite meilleurs lorsque les modalités d'examen font appel à des tests standardisés ayant recours aux VF/QCM

Voici les données exprimées sous forme de graphiques[18] (à gauche les taux de réussite en 1ère candidature et à droite en 2ème candidature, nous ne tenons pas compte des trois épreuves « MIXTES »).





En résumé, dans le contexte du 1er cycle d'étude de la FAPSE-ULg en 1997-1998, la comparaison des taux de réussite par type de questionnement utilisé lors des épreuves montre que les taux de réussite sont meilleurs lorsque le questionnement utilisé est de type QCM, ensuite viennent les QROL orales et puis les QROL écrites.

[18] Nous ne tenons pas compte des trois épreuves « MIXTES » en 2ème candidature car l'étude de Debry & al. (1999) ne nous fournit pas les taux de réussite des sous-épreuves (ECRIT, ORAL ou QCM) qui composent ces trois examens « MIXTES ».

3. Quelles contingences orientent les différentes pratiques d'examen à l'université ?

Le Groupe de Recherche Interdisciplinaire en Pédagogie Universitaire (GRIPU) de l'Université de Montréal nous donne dans les conclusions de son enquête sur les pratiques d'évaluation, quelques pistes qui peuvent expliquer les différentes modalités d'examens utilisées par les professeurs (Blais & al., 1997, pp. 123-128).

a) La culture d'évaluation

Nous savons qu'il existe des cultures de recherches, d'enseignement et aussi d'évaluation parmi les groupes de chercheurs, dans les différentes facultés et même à l'échelle des institutions[19]. « *Ainsi, on peut se demander ce qui arrive lorsqu'un professeur ou un chargé de cours veut trop s'éloigner de cette culture et introduire des pratiques différentes ? Sera-t-il encouragé ou ostracisé ? Tout dépend probablement des nouveautés introduites, de la distance entre ce qui est nouveau et ce qui est 'traditionnel'* » (*op. cit.*, p. 125).

Les auteurs signalent aussi un phénomène d'isomorphisme qui peut contribuer à expliquer la perpétuation d'une culture d'examen propre à un groupe donné dans une institution : « *...les années de formation des étudiants les mettent en présence d'une culture de l'évaluation propre à leur domaine. Pour celui qui fait des études de premier cycle dans un domaine et qui revient y enseigner après des études graduées, il sera naturel dans un premier temps de reproduire ce que les autres font autour de lui et ce que lui-même a vécu lors de ses études, une forme d'auto-contrainte inconsciente* » (*op. cit.*, p. 125).

b) La taille du groupe

Il est clair que la taille du groupe exerce une pression sur les pratiques d'évaluation. Il faut rappeler qu'à l'université l'enseignement ne constitue qu'une partie des tâches des professeurs. Dès lors, il est logique que l'enseignant confronté à de grands groupes d'étudiants comme c'est souvent le cas dans les premiers cycles, aura tendance à s'orienter vers des formes de questionnement dont la correction ne lui prendra pas trop de temps (des questions fermées de type QCM ou vrai-faux).

Cette tendance inquiète les auteurs de l'enquête : « *En corollaire, si la taille des groupes limite le recours à une plus grande variété de modalités, ne limite-t-elle pas aussi la mise en œuvre d'une plus grande variété d'habiletés cognitives par les étudiants ? Cette contrainte ne touche pas seulement la façon d'évaluer mais également ce qui est évalué en termes d'habiletés sollicitées et développées. Les répondants à l'enquête ont peut-être raison de s'inquiéter de l'influence exercée par les modalités d'évaluation sur l'orientation des stratégies d'apprentissage par les étudiants* ».

D'une part ces inquiétudes nous paraissent dans une certaine mesure justifiées, notamment parce qu'il est en effet impossible d'évaluer à l'aide de questions fermées certaines compétences complexes (expression écrite, créativité, développement de raisonnements, …). Nous avons développé précédemment les limites des épreuves standardisées ayant recours aux questions fermées (p. 29).

Mais d'autre part, il faut aussi signaler que :

- Les autres formes de questionnement telles que les Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL), qu'elles soient orales ou écrites, ont aussi leurs limites. Nous les avons rappelées précédemment (p. 23) ;

[19] Par exemple, il existe un monde entre la culture d'enseignement des universités traditionnelles et celle qui régit une institution telle que l'Université du Limbourg à Maastricht aux Pays-Bas où l'on pratique une approche pédagogique basée sur les problèmes (« *Problem Based Learning* »).

- Les épreuves standardisées ayant recours aux questions à réponses fermées offrent une parade à une série de biais dont souffrent les QROM/QROL (notamment le manque de concordance intra-correcteur et/ou inter-correcteurs) ainsi que des avantages (par exemple une couverture large de la matière) ;
- Des formes plus sophistiquées de questions à choix multiple existent. Il s'agit des QCM avec Solutions Générales Implicites (SGI) et pourcentages de certitude. Ce type de question permet d'évaluer systématiquement des niveaux de processus mentaux plus élevés que la simple connaissance. Nous décrirons en détail plus loin ces modalités de questionnement (pp. 67, 68 et 101) ;
- une complémentarité entre QROM/QROL et QCM-SGI avec pourcentages de certitude devrait être plus systématiquement envisagée lorsque ces formes de questionnement sont en adéquation avec les objectifs de l'examinateur.

Dès lors, si les stratégies d'apprentissage des étudiants peuvent en effet souffrir d'une utilisation accrue de questionnements à choix de réponses classiques (notamment parce que les QCM classiques ne favorisent pas les jugements critiques), c'est parce que les examinateurs ignorent ou refusent des modalités plus sophistiquées et plus performantes de questionnement standardisés tels que les QCM-SGI ainsi que les avantages qu'offrent une utilisation combinée avec les QROM/QROL.

c) Le règlement pédagogique

Des règlements pédagogiques existent dans chaque institution universitaire, ces règlements contiennent des contraintes en matière d'organisation des examens auxquelles les enseignants doivent se plier.

d) La contrainte des disciplines

Signalons ce que les membres du GRIPU rapportent au sujet de l'utilisation des QCM[20] : « … *autant dans certains domaines (et peu importe le nombre d'étudiants) on honnit les questions à choix de réponse (on l'a observé lors des entrevues avec les professeurs), autant dans d'autres domaines ces stratégies sont utilisées sans qu'on les perçoive comme intrinsèquement 'dangereuses'. Elles remplissent bien la fonction pour laquelle elles sont utilisées. Elles permettent de bien couvrir la matière et se corrigent rapidement avec un minimum de subjectivité. Elles font plus appel à la mémoire qu'à la capacité de raisonnement* » (*op. cit.*, p. 125).

e) La lutte contre une dissonance pédagogico-docimologique

On peut imaginer que des modalités d'examen soient plus en harmonie avec certaines formes de pédagogie universitaire. Voici ce qu'en disent les auteurs : « *…dans une approche participative, où les étudiants travaillent très souvent en groupe sur des projets de développement, il y a fort à parier que l'évaluation sera teintée par ce choix d'approche pédagogique. Le contraire pourrait amener les étudiants à contester la validité des modalités d'évaluation en regard des habiletés développées dans le cadre du cours* » (*op. cit*., p. 126). Ce problème nous semble aussi relever de la « culture d'évaluation ». Par exemple, si dans le cadre d'une approche participative les étudiants sont amenés à la fin de chaque séance à répondre à une série de QCM lors d'un quiz (bilan des acquis) immédiatement suivi d'un débriefing (afin d'éviter l'imprégnation de réponses incorrectes), les étudiants seraient probablement moins enclins à contester la validité de ce type de questionnement.

***

[20] Il s'agit de QCM de type classiques. On verra plus loin que d'autres modalités de questionnement à choix multiple sont envisageables (voir p. 6).

En synthèse :

A l'Université de Montréal, c'est dans le 1er cycle d'études (là où les étudiants sont les plus nombreux) que les questions à réponses fermées et semi-fermées sont utilisées alors qu'elles ne le sont pratiquement pas dans les 2ème et 3ème cycles. On remarque aussi une utilisation plus fréquente des questions à réponses ouvertes dans le 1er cycle.

A la FAPSE-ULG (données de 1997-1998), dans le 1er cycle d'études, un examen sur deux environ est un oral ou un écrit où les acquis des étudiants sont exclusivement évalués à l'aide de Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL). Un peu moins de la moitié des examens utilise le testing standardisé avec Questions Vrai-Faux (VF) ou à Choix Multiple (QCM).

Cinq types de contraintes peuvent expliquer les modalités d'examen différentes d'un enseignement à l'autre : le règlement pédagogique, les contraintes liées à la discipline, la culture d'évaluation, la taille du groupe et la volonté d'éviter des problèmes de dissonance pédagogico docimologique.

Lorsque nous comparons les taux de réussite des épreuves standardisées ayant recours aux VF/QCM avec ceux des épreuves ayant recours aux QROM/QROL à la FAPSE-ULg, nous observons des pourcentages de réussites moins élevés pour ce dernier type d'examens en 1ère et en 2ème candidature, et ce, que ce soit en 1ère ou en 2ème session.

Les facteurs qui pourraient expliquer les meilleurs taux de réussite des examens standardisés à la FAPSE-ULg sont divers : questions plus faciles étant donné les processus mentaux moins complexes qui seraient sollicités (nous verrons plus loin qu'il existe des QCM-SGI qui permettent d'évaluer systématiquement des processus mentaux plus élevés que ceux habituellement évalués par les QCM classiques), sévérité moins forte des examinateurs, présence de biais importants dans les épreuves orales et écrites ayant recours aux QROL/QROM (voir plus loin)…

Un autre facteur qui pourrait expliquer les meilleurs taux de réussite des épreuves standardisées à la FAPSE pourrait être lié au fait que la qualité des QCM peut être régulée à l'aide des procédures proposées par le Système Méthodologique d'Aide à la Réalisation de Tests (SMART) qui offre aux professeurs de la FAPSE (et de l'ULg) qui font appel à ce service une aide à la réalisation et à la correction des examens. Dans la suite, nous observerons au travers d'un cas, l'impact positif sur les taux de réussite d'une régulation de la qualité des QCM lors de la correction d'une épreuve standardisée.

Mais avant, voyons quels biais peuvent apparaître dans les examens oraux ou écrits ayant recours aux Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL) et ensuite dans les épreuves standardisées ayant recours aux Questions à Choix Multiple (QCM) et aux Vrai-Faux (VF).

D. Un exemple de régulation de la qualité des questions

Le Système Méthodologique d'Aide à la Réalisation de Tests (SMART, description p. 57) propose un processus en « spirale de qualité » (p. 74) pour la réalisation des épreuves standardisées. Ce processus comprend une procédure de régulation de la qualité des questions, cette procédure est utilisée à l'étape de la correction de l'examen. C'est dans cette phase de correction que s'insèrent les nouveaux indices de qualité spectrale dont la mise au point est au cœur de cette thèse. Dans cette section, nous décrirons, à l'aide d'un cas, la mise en œuvre de la procédure de régulation de la qualité des QCM avec les indices habituellement utilisés par le SMART. Nous observerons également les effets de cette procédure sur les taux de réussite des étudiants.

1. Réguler la qualité des questions d'un test standardisé : description d'un cas

Les faits qui vont être exposés ont réellement eu lieu et n'ont rien d'exceptionnel dans la mesure où ils font partie d'une routine de fonctionnement dans le cadre du cours d'*Approche technologique de l'éducation et de la formation* (ATEF) donné en 1ère candidature à la Faculté de Psychologie et des Sciences de l'Education (FAPSE) de l'Université de Liège (ULg). La procédure de correction avec régulation de la qualité des questions qui sera présentée est aussi utilisée dans d'autres enseignements de l'ULg (y compris dans le second cycle).

En ce qui concerne le cours ATEF, le professeur[21] organise régulièrement en collaboration avec le Système Méthodologique d'Aide à la Réalisation de Tests (SMART, voir p. 57) des tests formatifs en fin de cours (quiz). L'exemple qui suit concerne le quiz qui eut lieu le 29 novembre 2000. Ce jour là, à la fin de son cours, l'enseignant présente 10 questions à choix multiple à l'aide de transparents, les étudiants sont invités à répondre en utilisant des formulaires spéciaux prévus pour la lecture optique de marques (*formuloms*).

Lorsque le quiz est terminé, les *formuloms* sont ramassés et on passe au débriefing de l'épreuve : une par une les réponses aux 10 questions sont commentées et discutées avec les étudiants.

Pour une des questions, la septième, le professeur se rend compte lors de ce débriefing de la pertinence des remarques des étudiants. La question est en effet mal rédigée. Sensible aux arguments invoqués, il décide de ne pas tenir compte de cette question pour la correction et l'annonce séance tenante à l'auditoire.

Les étudiants formulent aussi une série de remarques à propos des questions 4 et 6 pour lesquelles certains donnent quelques arguments en faveur de propositions incorrectes qui, selon eux, pourraient aussi être considérées comme correctes (en plus de celle que le professeur a prévue comme correcte). En l'occurrence la proposition « AUCUNE (6) » pour la quatrième question et la proposition « TOUTES (7) » pour la sixième question. Le professeur retient les arguments et signale qu'il rectifiera les erreurs lors de la correction en fonction des résultats de l'analyse statistique des propositions des QCM.

[21] Il s'agit du Professeur Dieudonné Leclercq que nous tenons à remercier pour nous avoir permis d'utiliser les données et informations relatives à son quiz du 29 novembre 2000.

Après traitement des formuloms par le SMART, le professeur reçoit le tableau ci-dessous qui contient une série d'informations à propos de la qualité des questions du quiz.

	0	1	2	3	4	5	6	7	8	9
Q1	0.0	2.5	5.7	15.1			**56.0**	15.1	4.4	0.6
rpbis	0.00	-0.02	-0.15	-0.18			**0.48**	-0.16	-0.33	-0.11
Cmoy	0.00	52.5	62.5	64.27			**59.13**	52.81	40.36	77.5
Q2	0.0	6.3	0.6	0.6	4.4		**87.4**	0.0	0.0	0.0
rpbis	0.00	-0.19	-0.2	-0.06	-0.03		**0.22**	0.00	0.00	0.00
Cmoy	0.00	60.0	12.5	60.0	78.21		**76.15**	0.00	0.00	0.00
Q3	1.9	3.1	1.3	17.0			28.9	0.6	0.6	**46.5**
rpbis	-0.05	-0.15	-0.12	-0.29			-0.16	-0.11	-0.06	**0.48**
Cmoy	12.5	35.0	48.75	43.33			46.3	12.5	77.5	**71.15**
Q4	1.3	2.5	3.1	0.0	0.6		30.8	2.5	8.8	**50.3**
rpbis	-0.08	-0.12	-0.22	0.00	-0.06		-0.18	-0.07	0.08	**0.25**
Cmoy	12.5	30.63	22.0	.00	60.0		56.68	31.88	47.5	**47.59**
Q5	2.5	0.0	0.0	0.0	0.0		11.9	**71.7**	5.0	8.8
rpbis	-0.12	0.00	0.00	0.00	0.00		-0.3	**0.34**	-0.15	-0.06
Cmoy	12.5	0.00	0.00	0.00	0.00		51.71	**67.83**	57.81	48.57
Q6	1.3	0.0	0.0	8.2	3.1		6.3	26.4	**12.6**	42.1
rpbis	-0.16	0.00	0.00	-0.15	-0.06		-0.07	0.17	**0.24**	-0.14
Cmoy	12.5	0.00	0.00	44.42	41.0		46.5	64.88	**57.13**	62.13
Q7	3.1	2.5	18.9	13.8			18.9	31.4	1.9	**9.4**
rpbis	-0.11	-0.07	0.03	-0.15			0.06	-0.07	-0.01	**0.27**
Cmoy	12.5	48.75	46.83	44.32			37.0	55.35	45.0	**76.0**
Q8	1.3	1.3	1.3	15.7			32.1	0.0	1.3	**47.2**
rpbis	-0.05	-0.09	-0.16	-0.02			-0.21	0.00	-0.26	**0.34**
Cmoy	12.5	57.5	68.75	47.1			50.59	0.00	36.25	**58.47**
Q9	0.0	**50.9**	18.9	11.9			2.5	0.0	12.6	2.5
rpbis	0.00	**0.37**	-0.13	-0.21			0.00	0.00	-0.17	-0.02
Cmoy	0.00	**57.56**	55.25	60.92			53.13	0.00	59.63	54.38
Q10	3.1	2.5	25.8	7.5			**40.9**	10.7	3.8	5.0
rpbis	-0.20	-0.02	-0.33	-0.12			**0.5**	0.03	-0.22	-0.02
Cmoy	12.5	56.25	58.23	62.08			**54.46**	50.74	54.17	36.88

Expliquons brièvement comment lire les données. Chaque encadré reprend les résultats d'une question, la lettre « Q » en gras suivie d'un chiffre indique le numéro de la QCM. Les numéros des colonnes, de 0 à 9, indiquent les numéros des propositions au sein de chacune des questions. La colonne « 0 » concerne les omissions, la colonne « 1 » la première proposition, etc.

Lorsque des blancs apparaissent dans les colonnes en regard d'une question cela signifie que celle-ci ne contenait pas de proposition correspondant au numéro de la colonne (par exemple, la première question ne contient pas de propositions « 4 » et « 5 »). Le fait que chaque QCM contiennent systématiquement les propositions « 6 », « 7 », « 8 » et « 9 » est lié à l'emploi de Solutions Générales Implicites (SGI), ces solutions seront expliquées en détail au chapitre suivant (p. 67). Pour chaque question, les données de la réponse correcte sont signalées en caractères gras (par exemple, en ce qui concerne la première question, c'est la 6ème proposition qui est correcte).

L'encadré correspondant à chaque question contient trois lignes de données. La première ligne nous donne le pourcentage d'étudiants ayant choisi la proposition concernée (par exemple, pour la première question aucun étudiant n'a omis, 2,5% ont choisi la 1ère proposition, 5,7% la 2ème proposition, … etc.). La seconde ligne intitulée « *rpbis* » nous donne la valeur du *rpbis classique*, le niveau de discrimination de chaque proposition (le *rpbis classique* a été arbitrairement mis à zéro lorsque aucun étudiant n'a choisi la proposition). La problématique du *rpbis classique* fait l'objet d'un exposé détaillé (voir p. 171) et est à la base des nouveaux indices de cohérence spectrale que nous proposons plus loin (pp. 178 et 184) et qui sont calculés à l'aide des informations fournies par les pourcentages de certitude. Pour faire bref dans le cadre de cette introduction, disons que le *rpbis classique* de la réponse correcte nous permet d'évaluer dans quelle mesure les sujets qui ont choisi cette solution obtiennent, en moyenne, un nombre de réponses correctes au total de l'épreuve plus élevé que les sujets qui se sont trompés. Lorsque le *rpbis classique* de la réponse correcte (en gras) est positif et dépasse un seuil en fonction du nombre de questions, alors cette

proposition discrimine correctement (ici, le seuil vaut $\sqrt{1/10} = 0{,}32$, la problématique du seuil calculé pour contrer le problème du recouvrement de la question dans le score total sera expliquée plus loin, p. 176). En ce qui concerne chaque proposition incorrecte, lorsque le *rpbis* est négatif, ce coefficient indique les sujets qui se sont trompés obtiennent, en moyenne, un nombre de réponses correctes au total de l'épreuve moins élevé que les sujets qui ont répondu correctement.

Enfin, la présence d'une troisième ligne intitulée « Cmoy » s'explique par le fait que le professeur a utilisé les degrés de certitude (les étudiants ont été invités à accompagner chaque réponse d'un pourcentage de certitude) et cette 3ème ligne nous donne le pourcentage moyen de certitude des étudiants pour chaque proposition (lorsque aucun étudiant n'a choisi la proposition, ce « Cmoy » a aussi été mis à zéro).

Terminons cette présentation des données en insistant sur le fait qu'elles fournissent des informations précieuses en vue de déterminer le niveau de qualité des questions, mais qu'en aucun cas elles ne constituent un diagnostic à elles seules. Les éventuels problèmes qui peuvent être mis en évidence par ce type de tableau doivent systématiquement être recoupés par une relecture fine de la question de la part de l'examinateur. Dans le cadre du SMART, nous préconisons aussi de compléter ces résultats par d'autres informations moins chiffrables et plus qualitatives qui proviennent des commentaires effectués par les étudiants ou par les collègues maîtrisant le contenu de l'épreuve et à qui le professeur peut demander un avis.

Nous avons reproduit sur le tableau les cercles et les soulignements qui mettent en évidence des problèmes éventuels[22] pour certaines questions (rappelons que le seuil du *rpbis classique* vaut ici 0,32).

Nous constatons que quatre questions pourraient poser des problèmes (Q2, Q4, Q6 et Q7). Pour les 6 autres (Q1, Q3, Q5, Q8, Q9 et Q10) les indices *rpbis classiques* des réponses correctes sont à chaque fois positifs et supérieurs au seuil 0,32, en ce qui concerne les propositions incorrectes ils sont négatifs ou proches de zéro. Ces 6 questions discriminent donc bien les étudiants dont le nombre de réponses correctes (au total de l'épreuve) est faible par rapport à ceux dont le nombre de réponses correctes (au total de l'épreuve) est élevé.

Nous remarquons que la 7ème question supprimée pendant le débriefing, suite aux arguments invoqués par les étudiants lors du débat, est aussi la plus mal réussie du test avec seulement 9,4% de sujets qui choisissent la proposition correcte. De plus, la valeur du *rpbis classique* de la réponse correcte est inférieur au seuil 0,32.

En ce qui concerne la 2ème question, nous observons aussi pour la réponse correcte une valeur du *rpbis classique* sous le seuil de 0,32. Mais ici, après relecture, le professeur ne trouve rien d'anormal dans la question qui pourrait expliquer cette valeur inférieure au seuil. De plus, les étudiants n'avaient pas fait de remarques au sujet de cette question à laquelle un pourcentage élevé de sujet (87,4%) répond correctement avec un pourcentage moyen de certitude élevé (Cmoy = 76.15%). Il constate aussi que les *rpbis classiques* des autres propositions sont tous négatifs ou proches de zéro. Après cette analyse, il décide de ne rien modifier en ce qui concerne les résultats de cette question.

Pour ce qui est de la 4ème question, il décide de tenir compte des commentaires des étudiants à propos de la proposition « 6 », et, après réflexion et relecture et malgré le *rpbis classique* négatif, il valorise la proposition « 6 » comme étant aussi correcte.

Enfin, pour la 6ème question, il constate que le *rpbis classique* du distracteur « 7 » est positif et que celui de la réponse correcte (« 8 ») est inférieur au seuil 0,32. De plus, les peu nombreux étudiants (12,7%) qui ont opté pour la réponse correcte l'ont fait avec une conviction moins forte (Cmoy = 57,13%) que ceux qui ont choisi la proposition « 7 » (Cmoy = 64,88%). Le professeur décide alors de valoriser cette proposition « 7 » comme étant aussi correcte.

[22] Ces problèmes sont signalés par la personne de l'équipe du SMART qui a effectué le traitement et peuvent être discutés avec le professeur si ce dernier le souhaite.

Voici la seconde version du tableau des statistiques des questions après les modifications demandées par l'enseignant au SMART. On remarque que : (1) le pourcentage de réponses correctes de la 4ème question est passé de 50,3% à 81,1% après valorisation de la proposition 6, (2) le pourcentage de réponses correctes de la 6ème question passe de 12,6% à 39% après valorisation de la proposition 7 et (3) les statistiques de la 7ème question ne figurent plus dans le tableau.

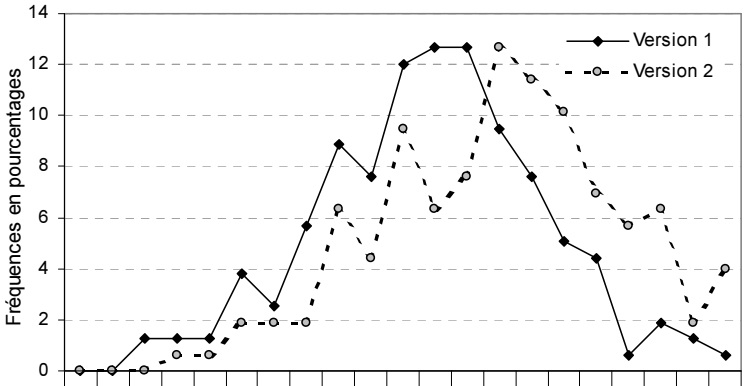
	0	1	2	3	4	5	6	7	8	9
Q1	0.0	2.5	5.7	15.1			**56.0**	15.1	4.4	0.6
rpbis	0.00	-0.02	-0.15	-0.18			**0.48**	-0.16	-0.33	-0.11
Cmoy	0.00	52.5	62.5	64.27			**59.13**	52.81	40.36	77.5
Q2	0.0	6.3	0.6	0.6	4.4		**87.4**	0.0	0.0	0.0
rpbis	0.00	-0.19	-0.2	-0.06	-0.03		**0.22**	0.00	0.00	0.00
Cmoy	0.00	60.0	12.5	60.0	78.21		**76.15**	0.00	0.00	0.00
Q3	1.9	3.1	1.3	17.0			28.9	0.6	0.6	**46.5**
rpbis	-0.05	-0.15	-0.12	-0.29			-0.16	-0.11	-0.06	**0.48**
Cmoy	12.5	35.0	48.75	43.33			46.3	12.5	77.5	**71.15**
Q4	1.3	2.5	3.1	0.0	0.6		0.0	2.5	8.8	**81.1**
rpbis	-0.08	-0.12	-0.22	0.00	-0.06		0.00	-0.07	0.08	**0.21**
Cmoy	12.5	30.63	22.0	.00	60.0		0.00	31.88	47.5	**51.05**
Q5	2.5	0.0	0.0	0.0	0.0		11.9	**71.7**	5.0	8.8
rpbis	-0.12	0.00	0.00	0.00	0.00		-0.3	**0.34**	-0.15	-0.06
Cmoy	12.5	0.00	0.00	0.00	0.00		51.71	**67.83**	57.81	48.57
Q6	1.3	0.0	0.0	8.2	3.1		6.3	0.0	**39.0**	42.1
rpbis	-0.16	0.00	0.00	-0.15	-0.06		-0.07	0.00	**0.48**	-0.14
Cmoy	12.5	0.00	0.00	44.42	41.0		46.5	0.00	**62.38**	62.13
Q8	1.3	1.3	1.3	15.7			32.1	0.0	1.3	**47.2**
rpbis	-0.05	-0.09	-0.16	-0.02			-0.21	0.00	-0.26	**0.34**
Cmoy	12.5	57.5	68.75	47.1			50.59	0.00	36.25	**58.47**
Q9	0.0	**50.9**	18.9	11.9			2.5	0.0	12.6	2.5
rpbis	0.00	**0.37**	-0.13	-0.21			0.00	0.00	-0.17	-0.02
Cmoy	0.00	**57.56**	55.25	60.92			53.13	0.00	59.63	54.38
Q10	3.1	2.5	25.8	7.5			**40.9**	10.7	3.8	5.0
rpbis	-0.20	-0.02	-0.33	-0.12			**0.5**	0.03	-0.22	-0.02
Cmoy	12.5	56.25	58.23	62.08			**54.46**	50.74	54.17	36.88

2. Impact sur les taux de réussite de cette procédure de régulation de la qualité des questions

Voici les courbes de fréquences en pourcentages des scores avant (version 1 en trait continu) et après (version 2 en pointillés) ces modifications.

Après la valorisation de certaines propositions pouvant aussi être considérées comme correctes pour les 4ème et 6ème questions ainsi qu'après la suppression de la 7ème question mal formulée, la courbe des cotes de la version 2 des résultats (en pointillés) est plus décalée vers la droite et montre une amélioration des scores après régulation de la qualité des questions.

Courbes de fréquences en pourcentages des cotes avant (version 1) et après (version 2) rectification de l'épreuve



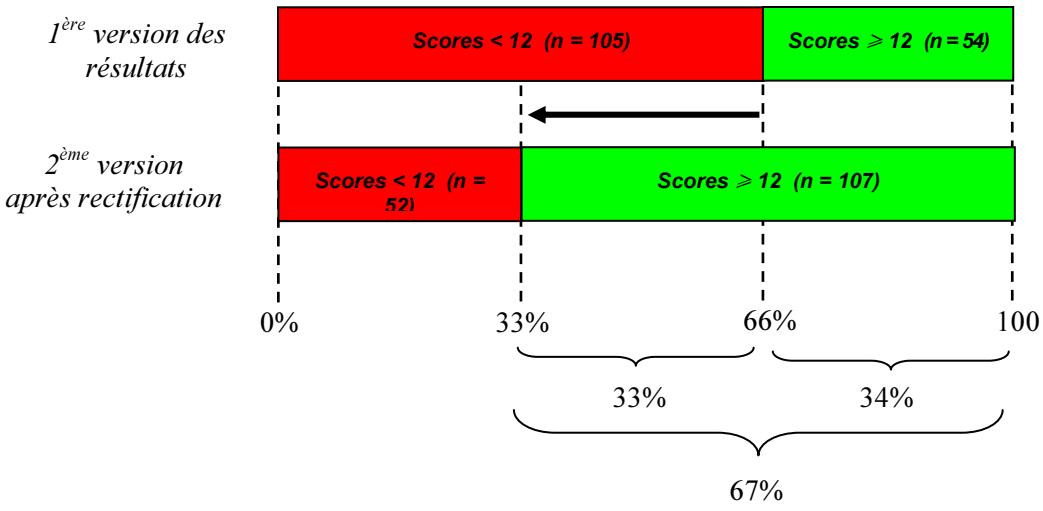
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Version 1	0	0	1	1	1	4	3	6	9	8	12	13	13	9	8	5	4	1	2	1	1
Version 2	0	0	0	1	1	2	2	2	6	4	9	6	8	13	11	10	7	6	6	2	4

Cotes

Lorsque nous comparons le pourcentage de sujets obtenant une cote supérieure ou égale à 12 (la cote

de passage habituellement définie pour les épreuves universitaires) nous constatons qu'on passe de 34% à 67% d'étudiants en situation de réussite, soit un gain de 33% !

Le quiz du 29 novembre 2000 dont nous venons de relater la correction a touché 159 étudiants. Dans la première version des résultats, 105 étudiants (66%) sur ces 159 examinés étaient en situation d'échec. Après régulation de la qualité des questions, le nombre d'étudiants qui n'ont pas atteint la note de réussite (12/20) descend à 52 (33%). Le gain de 23% correspond donc à 53 apprenants (105 - 52) qui passent du statut « d'étudiants recalés » à celui « d'étudiants ayant réussi » rien que parce que les questions de l'épreuve ont été régulées du point de vue de leur qualité !



Ces observations doivent être nuancées. Disons dès à présent que l'amélioration des taux de réussite (on « sauve » la moitié des étudiants qui ont échoué en passant de 66% d'échecs à 33%) serait probablement plus faible dans le cas d'un examen pour au moins trois raisons :

- L'examen comporterait 4 à 5 fois plus de questions[23] ;
- Les professeurs apportent en général plus de soin à la rédaction des questions d'examen ;
- Les étudiants sont mieux préparés lors des examens ;
- Tous les examens n'ont pas forcément recours au questionnement à choix multiple.

En synthèse :

On comprend mieux par cet exemple de l'impact positif que peut avoir un examen standardisé avec un processus de régulation de la qualité des questions sur les taux de réussite des étudiants (nous pourrions estimer l'impact d'une régulation en situation d'examen à 10% d'amélioration des taux de réussite).

Insistons sur le fait que ce processus de régulation de la qualité de l'épreuve fait appel à la fois aux indices statistiques des questions et au jugement de l'enseignant/examinateur. Répétons l'importance d'éclairer les informations livrées par les indices à la lumière des autres informations plus qualitatives détenues par l'enseignant/examinateur (il sait mieux que personne quelles performances sont sollicitées, il est l'expert du contenu, il a vécu le cours et l'épreuve, il a entendu ce que les examinés lui ont dit lors de l'examen et lors du débriefing, …). Dès lors, il est primordial que ce soit lui qui, in fine, prenne les décisions de rectifications et que celles-ci ne soient pas le résultat d'une interprétation aveugle et sans éclairage qualitatif des seuls indices statistiques.

Une structure de soutien telle que le Système Méthodologique d'Aide à la Réalisation de Tests (SMART) ne peut prendre à elle seule les décisions finales quant à la rectification des épreuves. Le rôle du SMART lors de la correction consiste à mettre en évidence des problèmes éventuels et à en alerter le professeur.

[23] Par exemple, à l'examen D. Leclercq propose 40 QCL et 30 QCM, soit 7 fois plus de questions.

PARTIE I

Pour une pragmatique de la qualité dans la réalisation des épreuves standardisées universitaires

Chapitre I :

Contexte institutionnel

Sommaire

A. Le Système Méthodologique d'Aide à la Réalisation de Tests (SMART)

B. Pénétration des concepts « qualité » dans les activités du SMART

C. Objectifs « qualité » des évaluations standardisées universitaires

D. « Spirale de qualité » et dispositifs d'ingénierie docimologique pour la réalisation des examens standardisés universitaires

E. Le contexte du projet MOHICAN

A. Le Système Méthodologique d'Aide à la Réalisation de Tests (SMART)

Notre recherche sur la mise au point des principes et indices d'analyse de la qualité spectrale des tests standardisés universitaires se situe au carrefour de deux projets : d'une part celui du MOnitoring HIstorique de cohortes de CANdidatures universitaires (MOHICAN) qui sera présenté plus loin et d'autre part celui du Système Méthodologique d'Aide à la Réalisation de Tests (SMART) que nous allons exposer ici. C'est à l'aide du SMART que les réponses des étudiants soumis aux 10 épreuves MOHICAN ont été informatisées, corrigées et traitées de manière à permettre l'envoi d'un feedback individualisé à chaque répondant (entre 1.392 et 3.846 sujets, selon les épreuves). Dans le cadre du SMART nous avons défini un cycle en « spirale de qualité » pour la réalisation d'une épreuve standardisée universitaire (nous l'exposerons plus loin) et c'est lors de la phase de correction de ce cycle que nous situons l'utilisation des indices spectraux, nouveaux instruments d'analyse de la qualité spectrale des tests. Mais d'abord, situons le contexte du SMART.

1. Contexte

Les processus de formation se complexifient et font de plus en plus appel aux Technologies de l'Information et de la Communication (TIC). Souvent très complexes, impliquant une multitude d'opérations techniques, ces processus quand ils s'insèrent dans un contexte pédagogique sont dans la plupart des cas pris en charge par des équipes spécialisées qui proposent leur soutien aux enseignants. Beaucoup d'institutions universitaires en ont pris conscience, en particulier lorsqu'il s'agit de fournir un appui pédagogique aux professeurs, en particulier à ceux qui doivent faire face à de grands groupes d'étudiants (parfois plus de 600) car ceux là ne peuvent plus gérer à eux seuls tous les problèmes de la sphère de leurs activités liée aux prestations de cours et aux évaluations.

Dans et en dehors de l'institution, les TIC sont de plus en plus présentes et des spécialistes interviennent de plus en plus fréquemment lorsqu'elles sont utilisées dans les processus d'enseignement/apprentissage (réalisation des supports pédagogiques, de documents audiovisuels, de sites Internet, de tests informatisés, ...). Parallèlement, l'utilisation des TIC dans un contexte pédagogique se banalise et de plus en plus d'étudiants équipés de micro-ordinateurs, de connections au web, en exploitent les ressources.

Pratiquement toutes les grandes universités se dotent aujourd'hui de cellules spécialisées dont la mission est d'aider les enseignants. Deux exemples dans le paysage des universités francophones de Belgique : le Centre de Méthodologie des Technologies de l'Information pour l'Education (MéTIE) à l'Université Libre de Bruxelles, l'Institut de Pédagogie Universitaire et des Multimédias (IPM) à l'Université Catholique de Louvain.

Dans le domaine de l'évaluation standardisée des grands groupes d'étudiants, le Système Méthodologique d'Aide à la Réalisation de Tests (SMART) constitue une réponse de l'Université de Liège (ULg) au besoin d'évaluer rapidement et de façon fiable de grands auditoires. Grâce au SMART, les enseignants peuvent recourir à des moyens techniques modernes pour procéder à l'évaluation des acquis des grands groupes (souvent dans le 1er cycle d'étude, là où les étudiants sont les plus nombreux). Le SMART propose aussi des services aux étudiants en ayant recours aux TIC, par exemple la consultation via Internet des feedbacks individualisés de leurs performances cognitives et métacognitives après une évaluation, la réalisation d'entraînements aux procédures d'examen en amphithéâtre à l'aide de boîtiers de vote électronique ou sur Internet (nous détaillerons plus loin ces services).

2. Historique

La reconnaissance officielle du SMART par les autorités académiques de l'ULg remonte à l'année académique 1998-1999, année lors de laquelle un budget fut accordé par M. le Recteur Legros pour mener à bien deux missions : (1) aider tous les enseignants ULg (quelle que soit leur faculté) dans la réalisation de leurs évaluations standardisées et (2) mettre en place le soutien logistique nécessaire à l'évaluation des enseignements de l'ULg organisée par la Commission EVALENS[24] mise en place par le Conseil Général des Etudes (CGE).

Historiquement, on peut dire que le SMART est né du succès rencontré par les méthodes d'évaluation des grands groupes proposées par le Centre d'Auto-Formation et d'Evaluation Interactives Multimédias (CAFEIM) basé à la FAculté de Psychologie et des Sciences de l'Education (FAPSE) de l'ULg.

Le CAFEIM[25] fut créé en 1990 par le Professeur Leclercq du Service de Technologie de l'Education (STE) avec le soutien de la FAPSE et en collaboration avec les professeurs Crahay (Service de Pédagogie Expérimentale), Henry (Service de Développement et Evaluation de Programmes Pédagogiques) et Reginster-Haneuse (Ecole de Santé Publique). Lors de l'année académique 1994-1995, le CAFEIM devint une structure officielle de la FAPSE-ULg. Sous l'impulsion du Doyen de la FAPSE, à l'époque le Professeur De Keyser, le centre fut chargé d'aider les enseignants à réaliser leurs examens en mettant à leur disposition un dispositif ayant recours aux techniques d'évaluation informatisées, notamment par Questions à Choix Multiples avec Solutions Générales Implicites[26] (QCM-SGI) et degrés de certitude (D. Leclercq, 1975, 1983, 1986, 1993). Ces interventions avaient lieu en priorité au premier cycle, là où les étudiants sont les plus nombreux : plus de 400 en première année (parfois plus de 600 lorsque des sections sont regroupées) et plus de 200 en deuxième année.

La mise en œuvre de ce service d'aide à la réalisation d'évaluations standardisées fut rendue techniquement possible grâce à des crédits de recherche octroyés par le FNRS[27] au STE en vue d'équiper ce service d'un dispositif complet de Lecture Optique de Marques (LOM, voir p. 81). Cette chaîne LOM fut mise à la disposition du CAFEIM qui fut ainsi en mesure de répondre efficacement aux nombreuses demandes. A l'époque, l'augmentation des effectifs dans le premier cycle d'études amena des changements dans les pratiques d'évaluation chez des enseignants confrontés à des groupes d'étudiants de plus en plus nombreux. Vu la taille des auditoires, le recours aux épreuves standardisées avec questions à choix multiple devint incontournable pour certains. Sur 61 épreuves organisées en première et deuxième candidatures en 1997-1998 à la FAPSE, 24 (39,3%) étaient de type QCM (Debry & al., 1999) et corrigées à l'aide du dispositif LOM, à l'exception d'une seule épreuve[28] qui avait recours à un système de testing informatisé *WINCHECK* [29] où les questions étaient posées et les réponses traitées via les ordinateurs de la salle multimédias du CAFEIM (les étudiants y disposent chacun d'un ordinateur et sont évalués par cohortes de

[24] Président de la Commission EVALENS : Prof. Claude Houssier.

[25] Le CAFEIM (Président : Prof. Dieudonné Leclercq, directeur de 1995 à 2000 : Jean-Luc Gilles) s'est fixé pour objectifs :
- de promouvoir l'apprentissage autonome et coopératif recourant aux multimédias ;
- de faciliter les évaluations de grands groupes d'étudiants en recourant aux technologies appropriées (QCM, lecture optique, etc.) ;
- de permettre les animations et/ou les évaluations interactives individuelles ou collectives (théâtre électronique) ;
- de développer une activité de recherche scientifique en matière d'application des technologies de l'information à l'apprentissage, l'enseignement et l'évaluation ;
- de favoriser des activités de conception et de réalisation de produits de formation.

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

[27] Fonds National de la Recherche Scientifique

[28] L'examen du cours de Psychologie développementale, 1ère candidature, Prof. Michel Born.

[29] *WINCHECK* est un logiciel d'évaluation interactive multimédias développé par le STE.

24). Dès le démarrage de ces activités, des professeurs donnant des cours dans le second cycle ou provenant d'autres facultés de l'ULg, marquèrent aussi leur intérêt pour cette aide méthodologique et logistique dans la réalisation des examens standardisés. En deux années académiques, 1994-1995 et 1995-1996, c'est plus d'une septantaine d'évaluations standardisées qui furent ainsi organisées avec l'aide du CAFEIM (Gilles, 1996a; 1996b).

L'accroissement des demandes et le caractère de plus en plus « interfacultaire » de celles-ci (elles émanent des 8 facultés de l'ULg) provoqua une prise de conscience en ce qui concerne la nécessité de subventionner cette réponse aux besoins des professeurs de toute l'ULg par un financement provenant des autorités académiques. Il faut dire que jusqu'en 1998 l'aide fournie était financée par des crédits de la FAPSE (qui lui octroyait aussi un assistant facultaire) et rendue possible par des apports en équipement et en personnel technique provenant du STE. Devant le succès rencontré et l'accroissement des demandes, cette situation mettait en péril l'équilibre budgétaire du CAFEIM (structure interne à la FAPSE) et le fonctionnement interne du STE. Le problème fut résolu par la création d'une nouvelle cellule à vocation interfacultaire, le Système Méthodologique d'Aide à la Réalisation de Tests (SMART), qui fut financée en 1998 par les autorités académiques grâce à un budget octroyé par M. le Recteur Legros.

On le voit, la naissance du SMART[30] fut possible grâce aux efforts conjugués du Service de Technologie de l'Education (STE), de la Faculté de Psychologie et des Sciences de l'Education (FAPSE), du Centre d'Auto-Formation et d'Evaluation Interactives Multimédias (CAFEIM) ainsi que des autorités académiques de l'ULg (Rectorat). Mais avant tout, ce qui explique l'émergence et le succès de cette cellule interfacultaire, c'est le fait qu'elle réponde à des besoins rencontrés par de plus en plus de professeurs confrontés aux mêmes problèmes méthodologiques et logistiques que posent les évaluations des grands groupes d'étudiants.

Les services proposés par le SMART ne se sont jamais cantonnés à l'unique phase de correction des examens. En fait, dès le départ, l'aide aux enseignants consista en une série de services docimologiques qui forment un tout cohérent et s'inscrivent dans un cycle de gestion des épreuves standardisées que nous détaillerons plus loin. Voici en bref les différents types de services proposés. Il s'agit d'aider les professeurs à :

- choisir le(s) type(s) de questionnements, la(les) méthode(s) de testing la(les) plus appropriée(s) ;
- gérer la(les) banque(s) de questions ;
- analyser la qualité *a priori* des épreuves (tests formatifs ou examens) ;
- entraîner les étudiants aux procédures d'évaluation ;
- préparer la logistique des épreuves ;
- corriger les tests à l'aide de procédures informatisées ;
- analyser la qualité *a posteriori* des questions ;
- mettre en place des solutions pour rectifier les épreuves lorsque des problèmes sont détectés ;
- fournir des feedbacks individualisés aux étudiants relatifs à leurs performances et aux performances de l'épreuve ;
- recueillir les avis des étudiants sur la qualité des épreuves.

En plus du dispositif de la lecture optique de marques, le SMART a été doté d'autres dispositifs technologiques que nous présenterons plus loin : les Boîtiers de Vote Electroniques (BVE) et les Systèmes de Testing Informatisés Multimédias où les étudiants répondent aux questions via un ordinateur (STIM). Plus récemment le SMART s'est équipé d'un serveur Internet de bases de données pour lequel une série d'applications ont été créées sur mesure en vue de fournir une aide à distance aux enseignants et aux étudiants. Ces services à distances sont regroupés sous l'appellation Traitements Via Internet (TVI).

[30] Responsable académique : Prof. Dieudonné Leclercq, directeur : Jean-Luc Gilles.

B. Pénétration des concepts « qualité » dans les activités du SMART

Dans cette partie nous envisagerons une série de facteurs qui ont progressivement fait émerger les préoccupations liées à l'évaluation de la qualité dans la sphère des activités universitaires. Nous clarifierons ensuite le concept de « spirale de qualité » que nous avons intégré dans notre modèle de construction des épreuves standardisées universitaires que nous présenterons plus loin dans ce chapitre.

1. Facteurs de mutation de l'université depuis les années 60

Le SMART se soucie de la qualité des épreuves qui sont proposées via son aide méthodologique. Cette préoccupation relative à la qualité des services offerts aux étudiants et aux enseignants s'insère dans un mouvement qui depuis le début des années '90, touche les institutions universitaires belges engagées dans une démarche d'évaluation de la qualité (Kaufman & al., 1995 ; Segers & Dochy, 1996). Ce mouvement participe lui-même à une tendance mondiale d'introduction du management de la qualité dans la sphère des activités universitaires (Ramsden, 1991 ; Nightingale & O'Neil, 1994 ; Zink & Schmidt, 1995).

Les pratiques évaluatives de l'université n'échappent évidemment pas à cette tendance récente d'évaluation de la qualité, c'est ce que soulignent Blais & al. (1997) : « *Dans le contexte social actuel les établissements d'enseignement supérieur sont de plus en plus souvent amenés à faire la démonstration de la qualité de la formation dispensée. En ce sens, l'évaluation des apprentissages, comme mécanisme de régulation et de contrôle exige une attention particulière. Par ailleurs, la régulation et le contrôle des apprentissages doivent prendre en considération un souci de plus en plus net d'assurer l'équité et la transparence des pratiques d'évaluation* ». Conscient de cette ligne de force actuelle, le SMART a entamé une réflexion sur la mise en place de méthodes de gestion de la qualité des épreuves standardisées en vue de satisfaire au mieux les enseignants et les étudiants. Notre réponse se situe dans un contexte universitaire qui s'est fortement modifié en une génération.

En effet, l'institution universitaire a considérablement évolué depuis les années 60. Divers facteurs expliquent les mutations de l'université. Dupont & Ossandon (1994, p.13) citent notamment la diversification de l'origine sociale des étudiants : « … *alors que les universitaires d'antan provenaient de filières secondaires sélectionnées, on observe que l'origine scolaire secondaire appartient à un éventail extrêmement varié de formations* ». Ces auteurs font l'hypothèse d'une diminution de la motivation en lien avec cette diversification sociale : « *la motivation sans doute plus affirmée de l'élite d'antan s'accompagne maintenant de groupes d'étudiants plus nombreux qui entreprennent des études universitaires sans véritable projet de carrière* ». A cette diversification sociale on peut aussi ajouter une tendance nouvelle de diversification des âges de ceux qui viennent se former à l'université. En effet, l'évolution rapide des sciences et des techniques entraîne la nécessité d'une formation continue des universitaires : aujourd'hui l'offre des formations en troisième cycle est en pleine explosion (surtout dans les domaines en pleine mutation). On remarque également la création de passerelles de plus en plus nombreuses entre les formations, favorisant ainsi encore plus de diversité dans les publics. Et à ce constat de la diversification des publics, s'ajoute celui de la massification. Récemment à l'Université de Liège, des amphithéâtres pouvant contenir plus de 600 étudiants ont été construits. Dès lors, la pédagogie des grands groupes s'impose par elle-même et comme le soulignent Dupont & Ossandon (1994, p.15) : « *par la force des choses, la pédagogie universitaire, prise sous l'angle des méthodologies utilisées, est de plus en plus inadéquate ; la plupart des universités européennes sont confrontées à une pédagogie des grands groupes à laquelle la majorité des enseignants n'ont pas été sensibilisés ; on est loin d'un taux d'encadrement de 1 enseignant pour 15 à 20 élèves rencontré dans les prestigieuses universités américaines* ».

La pédagogie des grands groupes est souvent connotée négativement, il en va de même pour les évaluations standardisées des grands auditoires. Une cellule comme le SMART se doit dès lors de mettre en place des stratégies qui permettent d'une part de garantir un niveau de qualité *a priori* des épreuves et d'autre part de faire la preuve *a posteriori* de la qualité des examens standardisés. Elle doit également veiller à faire connaître les niveaux de qualité atteints et mettre en œuvre des procédures docimologique de

rectification des épreuves lorsque des problèmes se posent (nous verrons plus loin quelles procédures ont été mise en place dans le cadre du SMART). La mise en œuvre de canaux de communication entre les enseignants, les étudiants et le SMART est nécessaire pour maintenir la confiance dans les procédures proposées. La preuve de la qualité et la communication de cette preuve (qui permet sa discussion) constituent à notre avis deux socles indispensables à la gestion de la qualité des épreuves standardisées universitaires.

Revenons aux facteurs de mutation de l'université, Romainville & Boxus (1998, p.23) invoquent d'autres facteurs se situant au niveau de la conception même de l'université. Ils font le constat du passage d'une université instrument de justice sociale dans les années 60-70 « … *fièrement indépendante des milieux économiques …* » à une université soumise aux exigences du monde industriel dans les années 80, période où l'état des finances publiques impose une sacro-sainte rigueur budgétaire tout en confiant à l'université une nouvelle mission, celle de favoriser la reconversion économique : « … *on exige que l'université participe activement au développement de la compétitivité du pays* ». D'après les auteurs, l'introduction de préoccupations nouvelles comme l'évaluation de la qualité au sein de l'université serait à mettre en lien avec cette volte-face de l'institution se rapprochant des milieux économiques. Un autre facteur évoqué est lié à l'idée que la mondialisation n'est pas qu'économique : « … *il existe des solutions internationales aux problèmes que rencontrent les universités d'un pays donné. Ainsi, les Pays-Bas ont largement imité les procédures en cours aux Etats-Unis.* ». On peut aussi ajouter qu'en plus de la concurrence intra-nationale, une concurrence internationale s'installe entre les universités. Les universités « captent » des publics provenant d'horizons de plus en plus éloignés dépassant les frontières nationales. C'est aussi une nouvelle tendance que soulignent Dupont & Ossandon (1994, p.14), « … *de plus en plus, l'enseignement passera du monoculturel et du monolingue au pluriculturel et au plurilingue en raison d'une part des programmes Erasmus et, d'autre part, de la libre circulation des personnes* ». Le mélange dans les amphithéâtres d'étudiants « du cru » avec des étudiants provenant d'autres institutions universitaires posera tôt ou tard le problème de l'uniformisation et de la qualité des programmes de cours dispensés dans les différentes universités.

Ces facteurs de mutation liés à la diversification et la massification des publics, au rapprochement avec les milieux industriels et à l'internationalisation de l'offre de formation ont amené de façon progressive la pénétration au sein de l'université de concepts nouveaux liés à la problématique de « l'évaluation de la qualité » historiquement née outre-atlantique dans les milieux industriels au début du siècle. Au milieu des années 90, Barnabe (1995), puisant chez une série d'auteurs spécialisés dans le domaine de la qualité, faisait le pronostic suivant : « *Tôt ou tard, notre système d'éducation devra s'intéresser sérieusement à la qualité totale. Rhodes (1992) ainsi que Glasser (1992) et Deming (1992) prétendent que l'approche qualité totale est la seule réponse aux problèmes scolaires. Comme toute autre organisation, selon Kaufman (1992), un système d'éducation utilise des ressources, développe des moyens d'action et fournit des diplômés à l'attention des clients externes. Un fait demeure, un système d'éducation entretient la promesse qu'il produira quelque chose qui sera utile à la société qui paie dans l'espoir d'obtenir des diplômés de valeur. En dernier ressort, les commentaires reçus et les pressions exercées de la part des clients externes à l'égard de ces diplômés l'obligeront à rechercher la qualité totale* ».

En parallèle avec l'apparition des préoccupations relatives à l'évaluation de la qualité, on assiste aussi aujourd'hui à l'émergence d'un véritable « marketing universitaire ». Les institutions soignent leur image de marque et cherchent à se positionner sur le marché de l'offre des formations en mettant en avant leurs atouts respectifs. Dans ce contexte, la mise en évidence de la qualité des enseignements et services offerts aux étudiants constitue un argument susceptible d'alimenter les politiques de communications vers ceux qui sont les futurs « clients » de l'université. La qualité des examens auxquels les étudiants sont soumis pourrait bien dans un avenir proche constituer un nouvel argument des institutions d'enseignement supérieur vis à vis de leurs publics cibles.

On le voit, le contexte des années 90 était mûr pour qu'une cellule spécialisée dans les problèmes d'évaluation des grands groupes voie le jour. Tout comme le recours à des méthodes et des dispositifs de testing utilisant les Technologies de l'Information et de la Communication (TIC) est incontournable pour répondre efficacement aux problèmes d'évaluation standardisée des grands auditoires ; la mise en œuvre d'une gestion qualité des épreuves est essentielle pour garantir un service optimal adapté aux exigences des

enseignants et étudiants. Dans ce contexte, c'est aussi assez « naturellement » qu'une réflexion sur la problématique de la qualité des épreuves standardisées universitaires s'est construite au sein du SMART.

2. Définir le concept « qualité »

Produire une définition de la qualité qui soit acceptable par une large majorité semble constituer un exercice difficile voire impossible. Pour le Conseil de l'Education et de la Formation (CEF)[31] : « *...la qualité est un concept complexe, qui diffère selon le produit, selon les attentes des utilisateurs, selon les processus et selon les résultats* ». Si on se réfère à la définition officielle donnée par la norme ISO 8402[32], la qualité : « *est l'ensemble des propriétés et caractéristiques d'un produit ou service qui lui confèrent l'aptitude à satisfaire des besoins exprimés et implicites* ».

Il est heureux que le document ISO officiel complète cette définition d'une série de notes pour en préciser le sens. Ces notes mettent l'accent sur des nuances importantes :

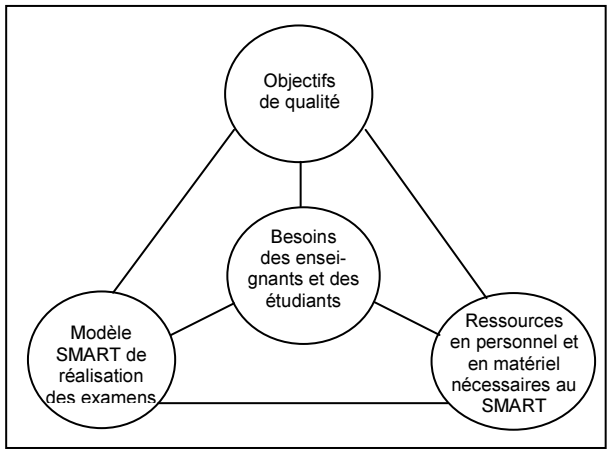
- la nécessité d'identifier et de définir les besoins implicites ;
- la traduction des besoins en propriétés et caractéristiques (ce qui permet d'envisager des comparaisons besoins-produits). Ces besoins pouvant, entre autres, comporter des aspects de sûreté, de fiabilité, de « maintenabilité » ;
- l'adaptation des spécifications en fonction de l'évolution des besoins ;
- la nécessité d'accompagner le terme « qualité » d'un qualificatif lorsqu'il s'agit de l'utiliser pour exprimer un degré d'excellence dans un sens comparatif : « *qualité relative lorsque les produits ou services sont classés par degré d'excellence* » ou « *niveau de qualité et mesure de qualité lorsque des évaluations techniques précises sont effectuées quantitativement* » ;
- l'influence des phases du processus de réalisation : « *la qualité d'un produit ou service est influencée par de nombreuses phases d'activités interdépendantes, telles que la conception, la production ou le service après-vente et la maintenance.* » ;
- l'influence des apports à la qualité des différentes phases du processus réalisation est aussi mis en rapport avec le concept de « *boucle de la qualité* » ou « *spirale de la qualité* » lui-même défini un peu plus loin dans la norme ISO 8402 (3.3 p.6) comme étant le « *Modèle conceptuel des activités interdépendantes qui exercent leur influence sur la qualité d'un produit ou service tout au long des phases qui vont de l'identification des besoins jusqu'à l'évaluation de leur satisfaction* ». Dès lors, il est possible d'identifier l'apport en qualité dû à une phase donnée, par exemple la « *qualité due à la conception* » ou la « *qualité due à la mise en œuvre* ».

Ainsi nuancée, la définition ISO 8402 de la qualité est beaucoup moins éloignée qu'il n'y paraît de prime abord de la définition de la Fédération des Etudiants Francophones (FEF)[33] : « *La qualité n'existe pas une fois pour toutes. Elle relève de la gestion de projet. Un projet se définit avec l'ensemble des acteurs concernés, est ensuite mis en application... avant d'être évalué. Après évaluation, le projet est reprécisé (redéfini) et ainsi de suite... La qualité s'inscrit dans un processus progressif infini qui fait que sa limite (l'excellence) n'est jamais atteinte.* »

La définition ISO 8402 et celle de la FEF se rejoignent dans l'idée que la qualité peut être constamment améliorée, qu'elle doit être envisagée comme faisant partie d'un processus cyclique où après chaque cycle de réalisation mais au sein de chaque phase de sa construction, le produit, le service, l'examen (dans le cadre des épreuves standardisées universitaires) s'améliorent pour tendre vers une qualité maximale (l'excellence) qui n'est jamais atteinte. C'est le concept de « *spirale de la qualité* ». Ce concept est intégré dans notre modèle de gestion des évaluations universitaires standardisées que nous présenterons plus loin. Mais avant, clarifions quels sont les objectifs du SMART en ce qui concerne la « qualité des épreuves standardisées ».

[31] Conseil de l'Education et de la Formation, CEF/avis, n°52/9, janvier 1998.
[32] Institut Belge de Normalisation (IBN), 1990.
[33] Fédération des Etudiants Francophones, *Les modules capitalisables*, Bruxelles, 17 novembre 1997.

C. Objectifs « qualité » des évaluations standardisées universitaires



Les recommandations de la norme internationale ISO 9004-2[34] (Gestion de la qualité et éléments de système qualité - Lignes directrices pour les services) mettent en évidence les facteurs clés d'un système qualité. Dans le contexte du SMART ces facteurs sont illustrés par le schéma ci-contre. Au centre : les « Besoins des enseignants et des étudiants », destinataires des services proposés par le SMART et points de convergence des autres facteurs clés. A la base : d'une part le modèle de construction en spirale de qualité des examens que nous utilisons dans le cadre du « Modèle SMART de réalisation des examens » (voir p. 74) et, d'autre part, les « Ressources en personnel et en matériel nécessaires au SMART » pour la mise en œuvre du processus de réalisation des épreuves. En haut du schéma nous avons placé les « Objectifs de qualité », c'est d'eux dont il sera question dans cette partie.

Succinctement, en ce qui concerne le SMART, notre politique de qualité est orientée vers la mise en place d'évaluations qui devraient offrir aux enseignants et étudiants des garanties de :

- ***Validité***, le contenu des questions du test doit correspondre à ce que l'enseignant veut mesurer ;
- ***Fidélité***, un travail corrigé et classé dans une catégorie donnée bénéficierait de la même mention s'il était corrigé dans d'autres conditions, par exemple par d'autres correcteurs ou quelques mois plus tard ;
- ***Sensibilité***, la mesure doit être précise, refléter des phénomènes subtils ;
- ***Diagnosticité***, permettre le diagnostic précis des difficultés d'apprentissage, des processus maîtrisés et de ceux qui ne le sont pas ;
- ***Praticabilité***, la faisabilité en termes de temps, de ressources en personnel et en matériel est assurée ;
- ***Equité***, tous les étudiants doivent être traités de façon juste, en principe de la même manière (standardisation) ;
- ***Communicabilité***, les informations non confidentielles relatives au déroulement du processus doivent être communiquées et comprises par les partenaires (enseignants, étudiants, équipe SMART) engagés dans la réalisation des épreuves.

Envisageons maintenant de manière plus détaillée ces différents objectifs « qualité » à atteindre dans la réalisation des examens standardisés universitaires.

1. La validité

Selon De Landsheere (1980, p. 178), valider une épreuve consiste à « *apporter la preuve (...) que l'examen fournit une évaluation correcte de ce qu'il prétend mesurer ou prédire* ». En d'autres termes, il s'agit de se poser la question : les questions posées aux étudiants correspondent-elles bien à ce que l'examinateur veut mesurer ? Répondre à cette question implique que la clarté ait été préalablement faite sur ce qui doit être mesuré, que les questions de l'examen soient clairement mises en lien avec les objectifs

[34] *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.

du cours. On parlera alors de validité de contenu. Nous allons voir qu'il existe différentes formes de validité et qu'on s'accorde aujourd'hui à en définir trois grands types.

a) Evolution des conceptions en matière de validité

Le concept de validité a considérablement évolué depuis un demi siècle. Laveault & Gregoire (1997, p. 189) signalent (nous soulignons) : « *Au début des années 50 (Messick, 1988, pp. 18-19), la validité était envisagée de manière relativement morcelée. Ainsi, les 'Technical Recommendations' de l'American Psychology Association (1954) se limitaient à codifier des types de validité (de contenu, prédictive, concomitante et conceptuelle). (…) Il faut attendre les années 70 pour qu'un effort important soit réalisé dans le sens d'une intégration des différents types de validité. L'aboutissement de cet effort est manifeste dans les 'Standards for Educational and Psychological Testing' publiés conjointement par l'American Psychological Association et l'American Educational Research Association en 1985. Les 'Standards' constituent aujourd'hui une référence incontournable pour les spécialistes de la mesure en psychologie et en éducation. (…)* la validité est présentée comme un concept unitaire se rapportant non au test lui même mais aux inférences faites à partir des résultats de celui-ci*. Dans cette perspective… Seules sont valides les inférences en faveur desquelles suffisamment d'arguments et de données empiriques ont pu être rassemblés. (…)* Suivant cette conception de la validité, la validation est un processus toujours continu d'accumulation de preuves*. Les types de validité, définis dans les ouvrages des années 50 et 60 sont aujourd'hui envisagés comme* moyens *de validation servant à rassembler des arguments en faveur de telle ou telle inférence.* »

Ainsi, les 'standards' de 1985 distinguent trois grands types de validation : (1) la validation relative au contenu, (2) la validation en référence à un critère externe et (3) la validation en référence à un concept ou un modèle théorique.

(1) La validation relative au contenu

Il s'agit dans ce cas de demander l'avis des experts du contenu, la question qui leur est posée étant « Les items de l'épreuve sont-ils représentatifs du concept ou du domaine évalué ? ».

Une épreuve sera donc jugée valide du point de vue du contenu si les arguments (subjectifs) des experts s'accordent sur le fait que le test de connaissances qui ambitionne de faire le bilan des acquis liés à un enseignement couvre réellement les aspects importants de cet enseignement.

De Landsheere (1979, p. 291) introduit l'idée d'une sélection plus rationnelle des aspects d'un enseignement qui devront être évalués dans une épreuve (nous soulignons) : « *L'appréciation de l'importance repose soit sur un jugement de valeur, soit* sur un raisonnement *(…) dans quelle mesure tel apprentissage est-il nécessaire pour accéder à tel autre, jugé important ? (C'est pourquoi on parle parfois dans ce dernier cas de* validité rationnelle ou logique*)* ».

De Landsheere signale également qu'on peut considérer qu'un test manque de validité de contenu s'il n'explore pas un éventail suffisant de comportements, par exemple s'il ne fait appel qu'à la mémoire (*op. cit.* p. 291). C'est tout le problème de la rupture entre enseignement et examen qui est posé. De Landsheere (1980, p.29) évoque ainsi le cas du professeur de chimie dont 4 élèves sur 22 obtenaient la moitié des points à l'examen : « *Le professeur ne faisait que deux ou trois interrogations par an. Sur les trois questions d'examen, deux impliquaient interprétation et transfert, démarches qui, au niveau de difficulté où le maître se plaçait, nécessitaient une compréhension profonde des phénomènes et un entraînement, long et finement contrôlé, à la solution de problèmes originaux. Une discussion amicale avec le jeune professeur en question révéla qu'il se croyait obligé de traiter en détail de tous les points du programme, qu'il n'avait pas conscience de la nécessité de l'entraînement systématique et au moins semi-individualisé à la solution de problèmes originaux et que, de toute façon, il ne possédait pas les notions psychopédagogiques de base nécessaire à la conduite d'un tel entraînement ».*

Nous décrivons plus loin, dans le cadre du cycle en « spirale de qualité » de réalisation des épreuves universitaires sommatives standardisées (voir p. 74), une procédure de validation rationnelle relative au contenu à couvrir par un examen. Cette procédure consiste à identifier dans un premier temps, d'une part les Points Enseignés (PE) et d'autre part les Catégories de Performances (CP). On crée ainsi une liste de binômes unissant à chaque fois un point enseigné avec une catégorie de performances : [PE x CP] et on en détermine quel sont les binômes essentiels. Ensuite lors de la mise en forme de l'épreuve les Modalités de Questionnement (MQ) sont définies. Ces MQ sont ensuite associées aux Catégories de Performances (CP) mises en évidence à l'étape précédente. On obtient alors une autre série de binômes [CP x MQ]. Enfin, pour chaque Point Enseigné mis en évidence précédemment, on peut alors concevoir un trinôme [PE x CP x MQ]. Chaque trinôme est donc formé d'un Point Enseigné, d'une Catégorie de Performance et d'une Modalité de Questionnement. Cela donne par exemple le trinôme suivant : « *Identification sur l'image d'une échographie* [CP] *d'un kyste ovarien* [PE] *évaluée à l'aide d'une QCM-SGI à livre ouvert* [MQ] ».

Ceci dit, et malgré le caractère structuré de cette démarche, il faut bien reconnaître que déterminer quels sont les binômes essentiels qui devront être évalués relève finalement du jugement de valeur, de croyances en l'importance des PE, et, en dernier ressort, c'est le bon sens qui fini par emporter la décision. Laveault & Gregoire (*op. cit.*) font remarquer à propos de la validation relative au contenu : « *Par définition cette modalité de validation des tests est subjective. Toutefois, si elle respecte une méthodologie rigoureuse, elle permet d'arriver à des conclusions solides qui pourront trouver confirmation dans des recherches empiriques ultérieures* ».

Si il est fait appel à des juges non entraînés, pas forcément experts, pour évaluer si les questions d'une épreuve couvrent bien un domaine donné et si ces juges n'utilisent pas de méthodologie rigoureuse on parlera alors plutôt de validation apparente (*face validity*). Cette procédure est donc moins scientifique mais peut dans certains cas être utile pour orienter la façon dont les questions devraient être posées et faciliter ainsi l'acceptation d'épreuves par les utilisateurs. De Landsheere (1979, p. 290) relate l'exemple d'un test soumis à un groupe de candidats aviateurs : « *…Ainsi, R. Thorndike et E. Hagen observent qu'un groupe de candidats aviateurs sera plus disposé à accepter un test d'arithmétique dont les problèmes portent sur la vitesse du vent ou sur la consommation de carburant qu'un test faisant porter les mêmes types de problèmes sur l'agriculture* ».

(2) La validation en référence à un critère externe

Laveault & Gregoire (1997, p. 190) distinguent deux types de validation en référence à un critère externe, d'une part la validation concomitante et d'autre part la validation prédictive.

Dans le cas de la validation concomitante, il s'agit de vérifier si les corrélations entre les scores à l'examen et les mesures prises comme critère de référence sont suffisantes. On examinera par exemple la corrélation entre les scores à un test et ceux obtenus à une autre épreuve ou une série d'autres épreuves considéré(s) comme une référence solide (voir plus loin la validité concomitante des QCM-SGI, p. 67).

Dans le cas de la validation prédictive l'idée est de vérifier si le pronostic formulé par le test se vérifie ou non. Selon Laveault & Gregoire (*op. cit.*) « *La validation prédictive consiste, quant à elle, à évaluer la qualité des prédictions faites sur base des scores au test. dans ce cas, le critère est la mesure de ce qui a été prédit. Par exemple, la validation d'un test d'admission consistera en la comparaison des scores au test et des résultats obtenus à la fin d'un programme d'étude* ».

(3) La validation en référence à un concept ou un modèle théorique

Les scores des étudiants doivent permettre des inférences solides (validité de *construct* ou théorique ou conceptuelle). Les épreuves reposent sur des modèles implicites ou explicites de la réalité qu'elles sont sensées évaluer. Ce sont ces modèles qui permettent de donner du sens aux résultats des tests.

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

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

b) Trois techniques qui renforcent la validité de contenu des épreuves standardisées universitaires

Précédemment nous avons évoqué la nécessité de distinguer les QCM classiques d'un autre type de QCM dont les modalités améliorent considérablement la qualité de cette forme de questionnement fermé.

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

Les Questions à Choix Multiple (QCM) habituelles permettent de mesurer les niveaux (1) connaissance, (2) compréhension et (3) application. Mais, dans le cadre d'épreuves standardisées destinées à de grands groupes, lorsqu'il s'agit d'évaluer (4) l'analyse c'est à un autre type QCM, les QCM-SGI (que nous allons aborder ci-dessous) qu'il faut faire appel.

Nous présenterons aussi plus loin la technique des degrés de certitude (Leclercq, 1983, 1993) qui constitue une réponse proposée par le SMART en ce qui concerne l'évaluation du niveau taxonomique « (6) l'évaluation ou le jugement sur base de critères personnels » à propos de la matière enseignée.

Enfin, nous terminerons en abordant le questionnement à livre ouvert. Cette technique est en effet plus adéquate lorsque l'examinateur souhaite mesurer le niveau (2) de compréhension et non le niveau (1) de connaissance.

(1) Les QCM-SGI

Leclercq (1993, p. 210) a montré qu'il était possible d'évaluer (4) l'analyse de façon systématique en ayant recours aux Questions à Choix Multiple avec Solutions Générales Implicites (QCM-SGI) qui autorisent, en plus des solutions habituellement proposées, les quatre possibilités suivantes : Rejet (aucune solution proposée n'est correcte), Toutes (toutes sont correctes), Manque (il manque des données dans *l'énoncé pour que l'on puisse choisir UNE solution comme correcte)*, Absurdité (il y a une contrevérité dans l'énoncé *à dénoncer en priorité !)*. L'auteur souligne (*op. cit.*, p. 211) que la performance demandée à l'étudiant dans le cadre des QCM-SGI ne relève plus de la simple (re)connaissance, mais de processus mentaux plus élevés dans la taxonomie de Bloom : « *These type of question force the student to consider the relevance of the data or of the question themselves, i.e. to understand the problem and analyse the way it is stated instead of limiting him/herself at plain knowledge and application levels (in the terms of Bloom's taxonomy of cognitive objectives). We have adopted this principle, making it systematic, i.e. applying it in all our questions* ».

Gilles et Melon (2000) ont procédé à une comparaison des résultats au test de Compréhension d'un Texte Scientifique[35] (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.

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**
	20 Q. ORTH					**,42**	**,54**	**,39**	**,23**
	20 Q. VOCA						**,69**	**,48**	**,38**
	20 Q. SYNT							**,57**	**,33**
	20 Q. COMP								**,30**

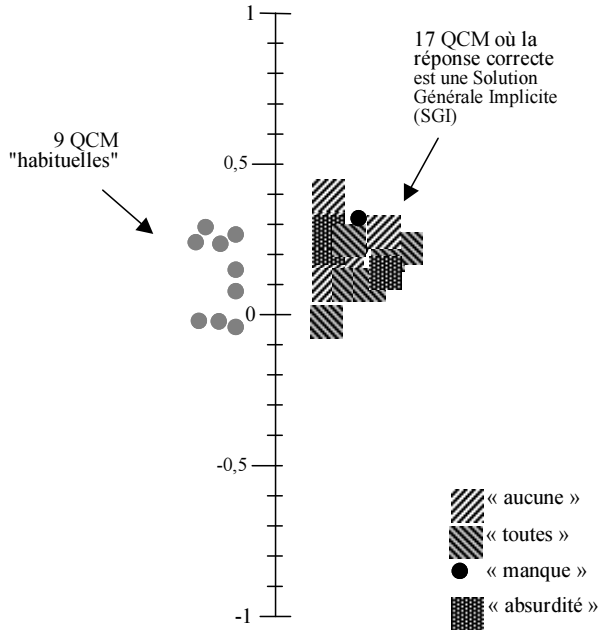
Tableau 2 : corrélations des scores aux partiels et des scores avec degrés de certitude aux testsMF&CTS

Le test CTS comportait quand à lui 27 questions réparties en 17 QCM dont les réponses correctes était des Solutions Générales Implicites (SGI) et 9 autres QCM dont les réponses correctes étaient habituelles (HABI).

Les résultats à ces tests ont aussi été comparés avec les performances académiques des étudiants aux examens partiels de janvier (dernière colonne du tableau).

Les corrélations présentées dans le tableau ci-dessus montrent que ce sont les performances au test CTS qui sont les plus corrélées (r = .55) avec les résultats obtenus aux examens partiels de janvier même lorsqu'on le compare au test de Maîtrise du Français (MF) qui comporte cependant plus de questions.

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) mais ces questions étaient aussi plus nombreuses dans le



[35] Des mesures gouvernementales visant à fixer le nombre de médecins praticiens 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[35] avec le SMART, 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.

test. Dès lors, qu'en est-il lorsqu'on compare question par question la liaison avec les performances aux examens partiels ?

Nous avons corrélé les scores des étudiants (n = 198) obtenus à chacune des questions du test CTS avec la moyenne de leurs résultats aux partiels.

Le schéma ci-contre permet de visualiser la position des 26 questions (chaque rond) en fonction de la corrélation obtenue. La ligne graduée verticale présente le continuum des valeurs possibles de 1 à –1. A gauche nous avons positionné les 9 QCM habituelles, celles où la réponse correcte figure parmi les propositions qui ne sont pas SGI. A droite se trouvent les 17 QCM où la réponse correcte est une SGI.

Nous remarquons que les résultats obtenus aux QCM avec réponse correcte SGI ont tendance à être un peu mieux corrélés avec la moyenne des scores aux examens partiels, mais cette différence n'est pas très marquée sur le schéma.

Voici les tableaux récapitulatifs des corrélations récoltées par chaque question avec les partiels. La moyenne de toutes les corrélations des QCM « SGI » est peu élevée (0,22 avec un écart type égal à 0,1) mais presque du double de celle de toutes les QCM « habituelles » (0,13 avec un écart type égal à 0,13). La validité concomitante (voir p. 65) est donc faible entre les performances des étudiants aux partiels et leurs scores aux QCM, et, encore plus faible quand la réponse correcte n'est pas une SGI.

Corrélations des scores aux 17 QCM « SGI » avec la moyenne des scores aux examens partiels (n = 198, corrélations significatives à p<0,05 marquées en gras)

aucune						toutes						manque		absurdité		
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0,30	**0,42**	**0,36**	0,12	**0,17**	**0,19**	**0,25**	0,12	**0,22**	0,12	**0,27**	0,00	**0,31**	**0,21**	**0,25**	**0,16**	**0,30**

Corrélations des scores aux 9 QCM « habituelles » avec la moyenne des scores aux examens partiels (n = 198, corrélations significatives à p<0,05 marquées en gras)

1	2	3	4	5	6	7	8	9
-0,04	0,08	**0,27**	**0,15**	**0,24**	-0,02	**0,29**	-0,02	**0,24**

(2) Les pourcentages de certitude

A l'aide des pourcentages de certitude l'examinateur peut évaluer de façon objective et systématique la capacité des étudiants à auto-évaluer leurs compétences (De Finetti, 1956; Shufford & al., 1966; Van Naerssen, 1962; Leclercq, 1975, 1982, 1993; Hunt, 1977; Bruno, 1986). En demandant aux étudiants de porter systématiquement un jugement sur leurs propres compétences (en ce qui concerne la matière enseignée) à chaque question d'une épreuve, nous nous situons au niveau (6) de la taxonomie de Bloom habituellement très peu évalué. Nous reviendrons plus loin sur les enjeux (voir p. 101) et les techniques (voir p. 106) du recueil des pourcentages de certitude. Nous rappellerons ici les conditions méthodologiques qui doivent être rencontrées pour garantir un recueil sans biais des données liées à l'auto-estimation de ses compétences et que Shufford & al. (1966) appellent *"Admissible probability measurement procedures"*. Ces procédures sont mises en œuvre dans le cadre du SMART.

(a) Lorsque des tarifs sont utilisés ils doivent être conformes à la théorie des décisions

Il s'agit de gratifier une réponse correcte accompagnée d'un degré de certitude élevé d'un meilleur score que si elle était accompagnée d'une certitude faible et inversement pour les réponses incorrectes. Les tarifs du barème des points doivent être calculés de manière à favoriser une seule stratégie : celle qui consiste à dire la vérité. Le barème des points figurant dans le tableau ci-après garantit que l'expression de son intime conviction rapporte plus de points que tout autre stratégie.

(b) La consigne doit être "probabiliste".

Demander à l'étudiant d'indiquer sa certitude par des termes vagues du type "peu sûr", "moyennement sûr", "très sûr", etc. est à proscrire car ces expressions recouvrent des réalités différentes en fonction des sujets. De plus, avec des termes aussi flous la variabilité est telle chez un même étudiant qu'on ne peut même pas recourir à des traitements ordinaux intra-sujets.

Voici la consigne mise au point par Leclercq (1983, 1993, 1998) et qui est proposée aux enseignants qui souhaitent avoir recours aux pourcentages de certitude dans le cadre des évaluations standardisées réalisées avec l'aide du SMART.

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 (RC)	réponse incorrecte (RI)
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

(c) Des zones correspondant à la précision humaine

Les coupures sur l'axe ne sont pas équidistantes ce qui permet une expression du degré de certitude plus nuancée à l'extrémité supérieure de l'échelle. Ainsi, l'étudiant peut faire la distinction entre 90 % (valeur centrale de la certitude 4) et 97,5 % (valeur centrale de la certitude 5) bien que la différence soit de 7,5 % seulement. Dans le premier cas (90 %) il n'a qu'1 chance sur 10 (1/10) de se tromper tandis que dans le second (97,5 %) il n'a qu'1 chance sur 40 (1/40), soit 4 fois moins. Etablir la même différence au milieu de l'échelle, par exemple entre 40 % (1/1,7) et 47,5 % (1/1,9) , n'est pas pertinent car nous ne sommes pas capable de distinguer ces deux derniers « rapports »...

(d) Le calcul d'indices métacognitifs doit être possible

La consigne utilisée autorise le calcul d'un indice de réalisme basé sur les différences entre les taux d'exactitude et les valeurs centrales des intervalles de probabilité ainsi que le calcul d'un indice de centration basé sur la différence entre la certitude moyenne et le taux d'exactitude moyen et dont le signe détermine la surestimation (+) ou la sous-estimation (-). Nous reviendrons plus loin en détail sur les méthodes et formules de calcul du réalisme (pp. 184, 272 et 274) et de la centration (p. 277).

(3) Le questionnement à livres ouverts

L'examinateur qui souhaite mesurer le niveau (2) compréhension et non le niveau (1) de connaissance (ou la reconnaissance parmi une série de propositions dans le cas des QCM) devrait permettre aux étudiants de répondre aux questions de compréhension en ayant les ressources documentaires à disposition. Leclercq (1993 , p. 18) souligne la proximité de cette situation d'évaluation par rapport à la vie courante : « *Open-Book work is the most common situation in adult work, as well as in adult learning. Often, we discover new things to understand and memorise from written text. We read them, we try them (on the computer, the engine or the objects dealing with the content) and, only if we do not understand, do we ask others* ».

2. La fidélité

Dans quelle mesure un correcteur peut-il prétendre qu'un travail corrigé et classé dans la catégorie « excellent » bénéficierait de la même mention s'il était corrigé dans d'autres conditions (autres correcteurs ou quelques mois plus tard) ? Les problématiques liées à la subjectivité de la correction ont été l'objet d'un courant de docimologie dite « négative » ou « critique » mené par Pieron (1963). Ce dernier et d'autres

chercheurs à sa suite ont relevé une série de biais d'évaluation tels que l'inconstance d'un même évaluateur[36] et la discordance entre évaluateurs[37].

La note (d'une épreuve corrigée) classée dans une catégorie donnée doit l'être de la même façon si elle est traitée par d'autres correcteurs (concordance interjuges) ou/et à un autre moment (consistance intrajuge).

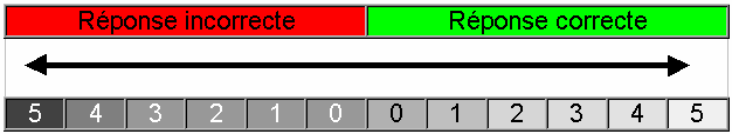
Des procédures d'évaluation automatisées visent à garantir la fidélité des mesures. Les QCM permettent d'échapper à la subjectivité des correcteurs et contribuent ainsi à augmenter la fidélité des évaluations. De plus, la simplicité de correction autorise un traitement informatisé rapide grâce à la lecture optique de marques[38] (voir p. 81) ou au questionnement interactif (les questions sont alors posées via l'écran d'un ordinateur ce qui permet notamment un feedback après chaque réponse fournie ou/et immédiatement en fin de test). Une autre source de fluctuations est liée à l'impossibilité habituelle d'exprimer son doute ou la sûreté de ses connaissances et nous avons vu qu'il existe des solutions pour permettre un recours systématique aux pourcentages de certitude dans le cadre d'épreuves standardisées à condition de respecter une série de règles méthodologiques (voir p. 68).

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

3. La sensibilité

La mesure doit être précise, refléter des phénomènes subtils.

Habituellement, on ne permet pas aux étudiants d'exprimer leur doute. Dans le contexte des épreuves standardisées ayant recours aux questions fermées (QCM classiques), on se situe alors dans un système de codage binaire des réponses. Soit une réponse est correcte, soit elle est incorrecte (1ère ligne de la figure).



Lorsque les étudiants peuvent exprimer leur doute à l'aide des pourcentages de certitude, la situation devient plus nuancée. En effet, nous avons vu plus haut (p. 68) que la consigne utilisée dans le cadre des épreuves organisées à l'aide du SMART autorisait l'emploi de 6 degrés de certitude (de 0 à 5). Dès lors, une réponse incorrecte accompagnée d'un degré de certitude maximum (« 5 ») présente la pire des situations, celle où l'étudiant donne une réponse incorrecte en estimant qu'elle a un maximum de chances d'être correcte (à gauche sur le spectre de la figure), on parlera alors de méconnaissance ignorée. A l'opposé (à droite sur le spectre grisé

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

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

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

de la figure), on trouve la situation de l'étudiant qui répond correctement avec une certitude maximale, dans ce cas il fait preuve d'une connaissance parfaite. Entre ces deux extrêmes s'ouvre tout l'espace de l'analyse spectrale.

Jans et Leclercq (1999, pp. 308-309) ont étudié en détail l'analyse spectrale appliquée aux performances des étudiants et présentent le caractère édumétrique de l'analyse spectrale : « *Nos propositions sont, dans notre esprit, édumétriques par sept aspects :*

- *Elles s'attachent à rendre compte de la subtilité des phénomènes mesurés. Ainsi la performance n'est plus affaire de tout (1) ou rien (0) : entre ces deux extrêmes, la connaissance partielle et ses divers degrés est désormais pleinement reconnue. Les spectres de performances sont l'illustration de cette approche.*
- *Elles visent à créer, pour les étudiants des indices individuels indépendants du groupe des étudiants …« group free scores » (…).*
- *Elles visent à apprécier des modifications de situation, d'où le besoin de termes tels que « analyse bi-statique, analyse dynamique … » (…).*
- *Elles lient l'approche standardisée à l'approche clinique dans le but de comprendre au moins autant que de mesurer (…).*
- *Elles sont conçues pour être communiquées aux intéressés, pour faire comprendre, un peu comme une psychanalyse « didactique » prépare le patient à devenir un jour expert ; ici à être un expert métacognitif transversal (Jans et Leclercq, 1997) et pas seulement un meilleur maîtriseur du contenu spécifique. (…).*
- *Elles assument la complexité et la multidimensionnalité des compétences et performances par le dégagement de profils de performances et/ou d'apprentissage, par le regroupement des ces profils dans des typologies.*
- *Elles ont une perspective interventionniste, à savoir le changement (l'amélioration) du profil des performances.* ».

En améliorant la sensibilité des instruments de mesure des acquis des étudiants, on augmente les possibilités de différenciation des étudiants grâce à l'affinement de l'échelle des niveaux de maîtrise. En effet, en augmentant la sensibilité de l'instrument par l'emploi des pourcentages de certitude, on peut par exemple distinguer entre (de la situation la plus catastrophique à l'idéal) : (1) méconnaissance erronée (réponse incorrecte et certitude élevée), (2) confusion (réponse incorrecte et certitude moyenne), (3) méconnaissance reconnue (réponse incorrecte et certitude faible), (4) ignorance (réponse correcte et certitude zéro), (5) connaissance douteuse (réponse correcte et certitude faible), (6) connaissance partielle (réponse correcte et certitude moyenne) et (7) connaissance parfaite (réponse correcte et certitude élevée) (Jans & Leclercq 1999, in Depover & Noël, Eds., p. 307). On améliore aussi potentiellement la mesure des progressions individuelles de chaque individu.

4. La « diagnosticité »

L'idée est ici de permettre le diagnostic précis des difficultés d'apprentissage, des processus maîtrisés et de ceux qui ne le sont pas.

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

VANDEVELDE (1971) signale à ce sujet : « *Sans doute 'ce qui se conçoit bien s'énonce-t-il clairement', encore ne peut-on pas confondre comprendre et savoir. L'école se livre à des extrapolations abusives au travers de la connaissance. La connaissance étant la condition sine qua non de la réussite, elle*

constitue un écran à l'égard de ce que nous voulons mesurer. L'individu examiné peut ne pas posséder la matière sur laquelle porte l'épreuve, n'avoir pas dépensé l'énergie nécessaire à sa mémorisation. Ceci peut bien entendu lui être reproché, mais le grief ne porte dès lors plus sur son inaptitude à comprendre. Nous nous trouvons dans l'impossibilité de conclure à cet égard. (…) C'est pour avoir trop souvent réuni en une même technique d'investigation une gamme importante de facteurs très différents (connaissance, compréhension, application, analyse, etc.) que l'image que nous nous faisons de l'élève est si imprécise. Nos examens, nos épreuves de contrôle scolaire gagneraient beaucoup à se présenter de manière analytique et davantage sur le plan des activités mentales que sur le plan des matières. La question à choix multiple est susceptible de fournir une information précieuse dans cette perspective. »

Le « recoupement » de plusieurs réponses à plusieurs questions permet lui, un diagnostic, comme cela se passe en médecine : la température à elle seule ne permet pas d'identifier la maladie, mais COMBINEE avec d'autres observations, elle le permet. C'est le principe qui a présidé à la conception de « double check » (Leclercq, 1993), une procédure d'évaluation interactive, qui consiste à poser une question en deux volets *prim* et *bis* et que nous avons présentée plus haut dans le cadre de la validation en référence à un modèle théorique (p. 66).

Ceci dit, même dans les épreuves non interactives il est possible d'améliorer le diagnostic des compétences des étudiants si pour chaque question les catégories de processus mental et de rubrique matière qui lui sont associées ont été identifiées. On est alors en mesure de renvoyer vers l'étudiant des feedbacks automatisés plus analytiques du type : « *Votre taux de réussite en ce qui concerne le chapitre I est de 45%. Vous avez répondu correctement à 9 questions sur 20 dont 8 questions sur 10 de connaissance et 1 question sur 10 d'application…* ». Cet affinement du diagnostic peut aussi aboutir à des prescriptions du type : « *Etant donné ces taux de réussite revoyez les exercices proposés en fin du chapitre 1…* ».

L'utilisation des pourcentages de certitude permet aussi de proposer un diagnostic métacognitif. Nous verrons plus loin qu'on peut calculer le réalisme des étudiants (indice *Rs*, voir p. 184 et pp. 272-274), leur centration (tendance à se sur ou sous-estimer, indice *Cs*, voir p. 277) au départ des informations qu'ils nous fournissent en accompagnant chacune de leurs réponses d'un pourcentage de certitude.

5. La « praticabilité »

Garantir l'objectivité des corrections, l'évaluation systématique des processus mentaux en jeu ainsi que l'augmentation de la sensibilité et de la « diagnosticité » impliquent dans le cadre d'examens classiques[39] que l'on y consacre de l'énergie et du temps. Il est possible de poser une question ouverte à réponse écrite longue notée selon *x* points de vue différents à l'aide de *y* critères préalablement opérationnalisés pour chacun, et il en résulte une correction complexe. Or, comme dans la plupart des universités européennes (Gibbs, Jenkins & al., 1992), on assiste aujourd'hui à une explosion du nombre d'inscriptions d'étudiants à la FAPSE-ULg[40], et, si on multiplie par le nombre d'étudiants le temps passé à corriger un examen dans ces conditions on en arrive rapidement à un constat d'impraticabilité.

Le SMART propose des techniques d'évaluation informatisées, notamment par QCM-SGI avec degrés de certitude et lecture optique de marques ainsi qu'un soutien méthodologique (voir encadré, p. 59).

De plus en plus d'enseignants de l'Université de Liège et de Hautes Ecoles font appel au SMART. Dans plusieurs examens, deux types de questions sont proposés en symbiose. D'une part, un grand nombre de QCM-SGI (en général une trentaine) avec degrés de certitudes, offrent une série d'avantages décrits par Leclercq (1986) : représentativité de l'échantillon des questions, simplicité de correction, objectivité de la correction, possibilité d'évaluer systématiquement les niveaux de compréhension, d'analyse et d'évaluation/jugement. D'autre part, un petit nombre de questions ouvertes (en général une ou deux)

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

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

permettent d'évaluer l'esprit de synthèse, l'originalité, la créativité, la capacité à organiser une réponse, ce qui n'est pas possible en ayant recours uniquement aux QCM fussent-elles SGI.

6. L'équité

Tous les étudiants doivent être traités de façon juste, en principe de la même manière. Veiller à ce qu'une épreuve soit équitable c'est notamment poser le problème de la standardisation du test.

Une standardisation parfaite implique :

- que tous les étudiants d'une section soient soumis à la même épreuve (mêmes questions pour tous) ;
- que les conditions d'administration de l'examen soient identiques pour tous les examinés ;
- que les modalités de correction soient uniformisées, que toutes les copies soient corrigées de la même manière.

Dans les faits, une standardisation idéale : même heure, même dispositions matérielles (sièges, éclairage, ...), même degré de familiarité des examinés avec les techniques de questionnement, même consignes, même entraînement, même durée, etc. est difficilement réalisable, voire impossible, mais il importe d'essayer de s'en approcher le plus possible.

7. La communicabilité

Les informations non confidentielles relatives au déroulement du processus SMART doivent être communiquées et comprises par les partenaires engagés dans la réalisation des évaluations.

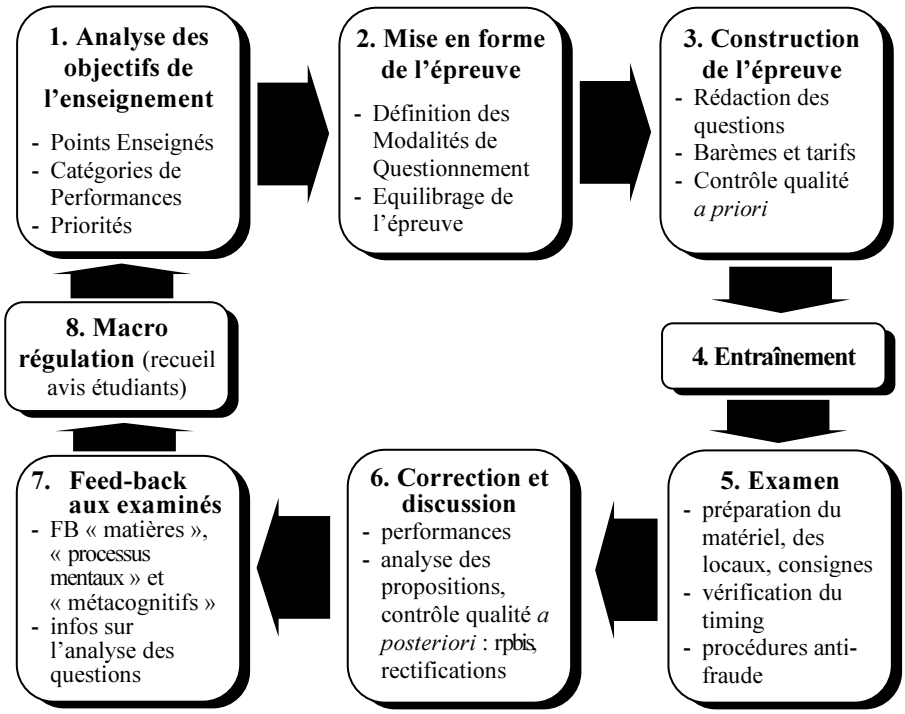
Cet objectif de « communicabilité » pose notamment le problème des feedbacks vers les étudiants après les épreuves. Nous verrons plus loin que le SMART y a apporté une solution technologique en permettant la consultation des feuilles de résultats individualisées après les épreuves via Internet.

D. « Spirale de qualité » et dispositifs d'ingénierie docimologique pour la réalisation des examens standardisés universitaires

Le modèle que nous présentons est à situer dans le contexte de la construction des examens universitaires standardisés visant à évaluer les acquis des étudiants et en particulier lorsqu'il s'agit d'évaluer de grands groupes. Il est proposé dans le cadre de la construction des examens standardisés de l'Université de Liège lorsque les professeurs souhaitent recourir aux services du Système Méthodologique d'Aide à la Réalisation de Tests (SMART). Il existe bien entendu d'autres modèles, par exemple celui du processus de construction d'un instrument de mesure proposé par Laveault & Gregoire (1997, pp. 79-122). Dans leur livre, ces auteurs proposent 5 étapes pour la réalisation d'un instrument de mesure. Etape 1 : détermination des utilisations prévues du test. Etape 2 : définition de ce que l'on souhaite mesurer. Etape 3 : création des items. Etape 4 : évaluation des items. Etape 5 : détermination des propriétés métriques du test définitif. Ils distinguent aussi des procédures de construction différentes selon que l'on élabore des tests d'acquis scolaires à visées sommatives ou à visées formatives.

Notre modèle en 8 étapes est plus ciblé sur les épreuves universitaires sommatives standardisées destinées à évaluer les acquis de grands groupes d'étudiants et englobe d'autres phases non couvertes par Laveault & Gregoire. Il met notamment l'accent sur l'entraînement des étudiants et sur la fourniture de feedbacks informatifs et rapides à ces derniers. L'idée étant de fournir autant que possible des informations aux examinés en situation d'échec en vue de les aider à atteindre le niveau de compétence exigé lors d'une session d'examen ultérieure. Le modèle qui sera exposé s'insère aussi dans une perspective de travail en « spirale de qualité » qui permet d'améliorer tout le processus de construction des épreuves.

Les huit étapes de la gestion des examens dans l'enseignement supérieur de Gilles et Leclercq (1995) présenté ci-dessous, permettent de décrire un processus cyclique de réalisation des examens standardisés. Elles sont inspirées du schéma des grandes phases de construction d'un examen décrit par De Landsheere (1980, p. 65).



Nous proposons ce processus cyclique en « spirale de qualité » dans la réalisation des épreuves universitaires sommatives standardisées. Le respect de ces huit phases de construction et des sept objectifs

« qualité » précédemment décrits nous paraît crucial si nous souhaitons réaliser des examens standardisés universitaires de qualité.

1. Analyse des objectifs de l'enseignement

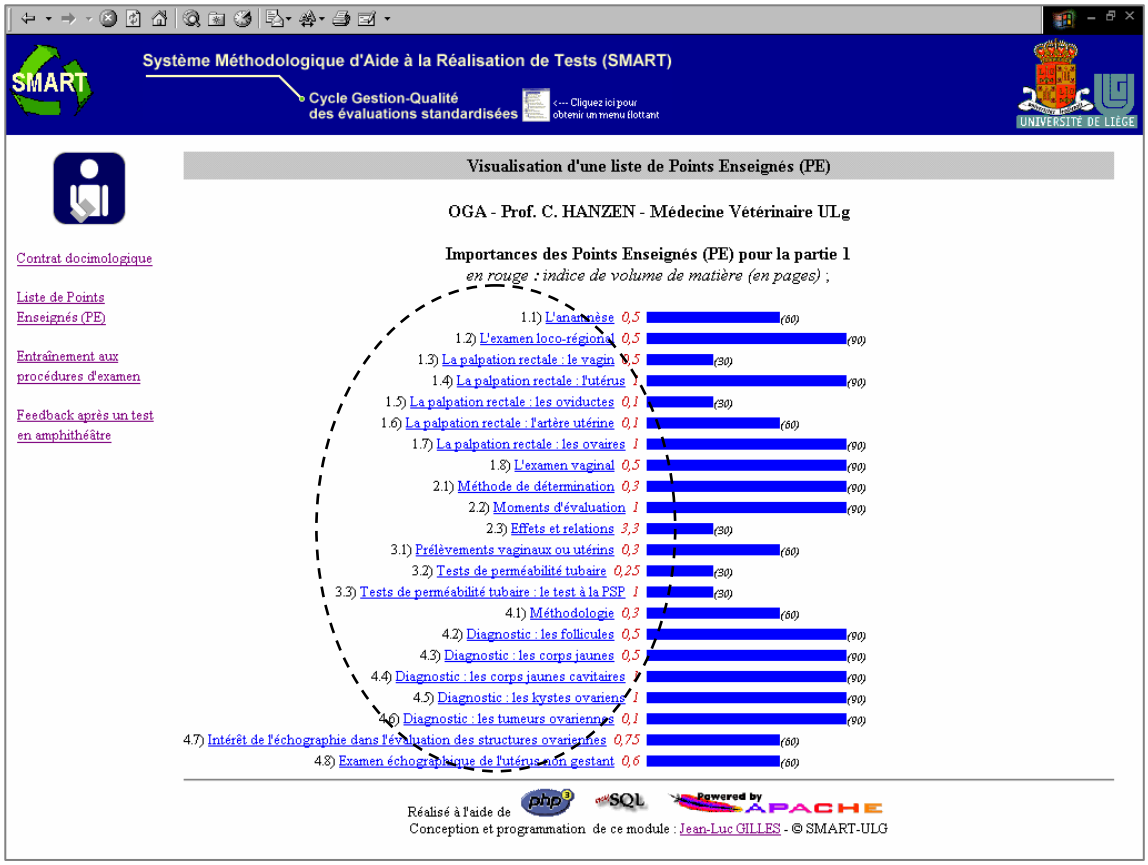
Cette étape a pour but de clarifier et de mettre en évidence les objectifs susceptibles d'être évalués dans le contexte d'une épreuve standardisée universitaire.

D'un point de vue pragmatique, l'idée est de proposer une méthode qui permet un inventaire rapide des objectifs de l'enseignement pour lequel un examen standardisé est envisagé. La mise en œuvre de cette méthode sera rendue encore plus aisée si le professeur prend soin de baliser ses notes de cours par des codes lui permettant de repérer les points qu'il considère comme formant des unités de matière.

La procédure d'analyse des objectifs de l'enseignement que nous proposons dans le contexte du SMART se déroule en quatre phases.

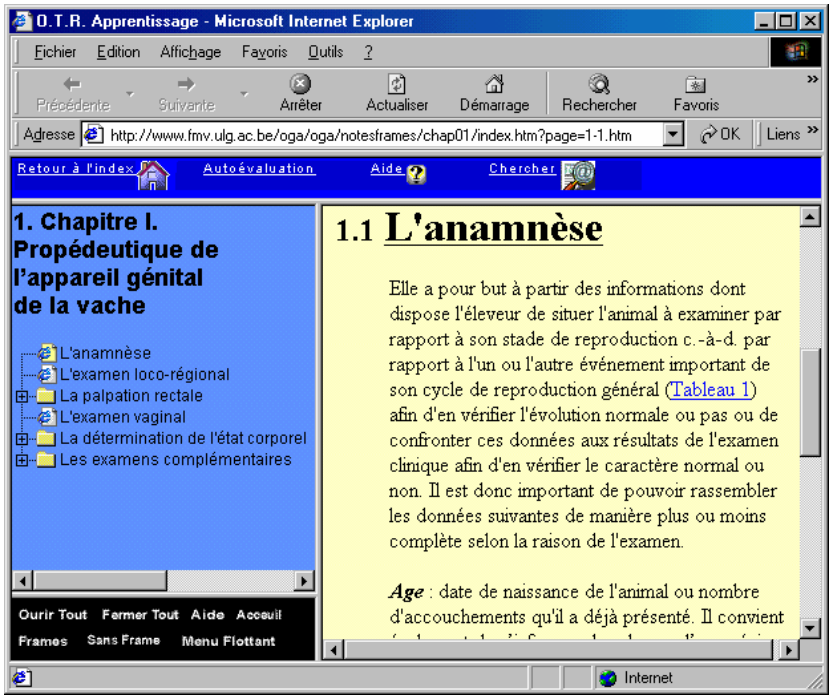
Phase 1 : dans un premier temps l'enseignant liste les Points Enseignés (PE) (ou les rubriques de matière qui ont été abordées au cours) et fixe leur importance.

Voici un exemple de liste de Points Enseignés liés au cours d'Obstétrique des Grands Animaux (OGA) pour lequel nous avons utilisé cette méthode lors de la construction d'épreuves standardisées (Castaigne & Gilles, 2000 ; Castaigne, Gilles & Hansen, 2001). Les traits bleus à droite permettent de visualiser le degré d'importance attribué par le professeur à chacun des points enseignés (0 = importance minimum et 100 importance maximum).



Signalons que dans le cadre des épreuves OGA cette liste des Points Enseignés (PE) était accessible aux étudiants vétérinaires via le serveur Internet du SMART. Nous avons entouré en pointillés sur l'écran ci-dessous une série de PE sélectionnés dans le cadre du cours OGA.

Grâce au support de l'enseignement à distance offert par le service du Prof. Hansen aux étudiants du cours OGA, des liens entre l'interface du SMART et le site Internet professeur ont pu être créés. En cliquant sur un point enseigné (écran précédent) l'étudiant pouvait obtenir la matière du cours concernée[41] (exemple ci-contre).



Phase 2 : dans un second temps l'enseignant qui procède à la validation du contenu de son épreuve identifie les Catégories de Performances (CP), c'est-à-dire les processus mentaux que les étudiants doivent exercer sur les points enseignés.

Codes PE	Intitulés PE	Connaissance	Compréhension	Application	Analyse	Synthèse	Evaluation
4.1	Méthodologie	☑	☐	☐	☐	☐	☐
4.2	Diagnostic : les follicules	☑	☑	☐	☑	☐	☐
4.3	Diagnostic : les corps jaunes	☑	☐	☐	☐	☐	☐
4.4	Diagnostic : les corps jaunes	☑	☑	☐	☐	☐	☐
4.5	Diagnostic : les kystes ovariens	☑	☐	☐	☐	☐	☐
4.6	Diagnostic : les tumeurs ovariennes	☑	☐	☐	☐	☐	☐
4.7	Intérêt de l'échographie dans	☑	☐	☐	☐	☐	☐
4.8	Examen échographique de	☑	☑	☐	☐	☐	☐

[41] Le contenu du cours du professeur C. Hansen a été porté sur Internet par le Dr J.-L. Castaigne. Nous tenons à les remercier tous deux pour leur précieuse collaboration.

C'est la taxonomie des objectifs cognitifs de Bloom (1969) qui fut choisie dans le cadre de la construction des épreuves du cours OGA. L'identification des CP permet d'établir une liste des binômes [PE x CP] où pour chaque point enseigné l'examinateur est amené à associer une ou plusieurs catégories de performances. L'interface ci-dessous développée par le SMART est disponible via l'Internet et permet à l'enseignant d'associer à chaque Point Enseigné (PE) une ou plusieurs catégories de Performances (CP)

Dans le cadre de l'exemple ci-dessus, le point enseigné *« 4.2 Diagnostic : les follicules »* est associé à trois catégories de performances issues de la taxonomie de Bloom : « *Connaissance* » (réactualisation de contenus mémorisés), « *Compréhension* » (interprétation, traduction avec accès aux ressources documentaires) et « *Analyse* » (décomposition, classification, organisation). On obtient ainsi trois binômes [PE x CP].

Ces binômes peuvent ensuite être classés en fonction de l'importance accordée par l'enseignant (établissement des priorités évaluatives).

2. Mise en forme de l'épreuve

Phase 3 : Lors de cette étape l'enseignant qui utilise la procédure de validation du contenu définit les Modalités de Questionnement (MQ) qu'il compte utiliser. Les MQ sont composées d'un type de questionnement et d'une série de caractéristiques associées. Exemples de MQ : le type de questionnement « QCM » accompagné de « Solutions Générales Implicites (SGI) » à « livres ouverts[42] » ou le type de questionnement « QROC[43] » accompagné de la technique des « degrés de certitude », …

Ces Modalités de Questionnement (MQ) sont ensuite associées aux Catégories de Performances (CP) mises en évidence à l'étape précédente d'analyse de l'enseignement. On obtient alors une deuxième série de binômes intitulés « [CP x MQ] ». Dans le cadre de la méthode de validation du contenu, l'interface proposée par le SMART permet à l'enseignant de créer ses binômes [CP x MQ] en sélectionnant les éléments constitutifs dans des listes (les informations sont ensuite mémorisées dans une base de données).

Les nombreuses possibilités offertes par la combinaison des modalités de questionnement et des catégories de performances ont été décrites par ailleurs dans Leclercq & Gilles (1995).

Phase 4 : Enfin, pour chaque Point Enseigné mis en évidence à l'étape de l'analyse des objectifs de l'enseignement (*Phase 1*), le professeur peut concevoir un ou plusieurs trinômes [PE x CP x MQ]. Chaque trinôme est donc formé d'un Point Enseigné, d'une Catégorie de Performance et d'une Modalité de Questionnement. Cela donne par exemple le trinôme : « *Connaissance (réactualisation de contenus mémorisés)* [CP] » x « *liée au diagnostic des follicules* [PE] » x « *à l'aide d'une QCM-SGI avec degrés de certitude* [MQ] ».

Une liste des trinômes peut alors être générée et proposée à l'examinateur qui sélectionne en fonction de l'importance qu'il leur accorde les trinômes qui feront l'objet d'une question lors de l'épreuve. On aboutit ainsi à un plan d'épreuve qui guidera la création des questions à l'étape suivante « 3. Construction de l'épreuve ».

---

[42] Les étudiants disposent alors lors de l'examen de ressources documentaires spécifiées par l'examinateur, dès lors les questions peuvent réellement mesurer les performances des étudiants liées au niveau taxonomique « compréhension » (selon la taxonomie des objectifs cognitifs de Bloom, 1969).

[43] QROC = Question à Réponse Ouverte Courte.

3. Construction de l'épreuve

On procédera ici à la réalisation des questions en tenant compte de règles méthodologiques de construction lorsqu'elles existent (par exemple les règles de rédaction des QCM de Leclercq, 1986).

Voici un plan d'épreuve conçu par le professeur Lion dans le cadre des interrogations hebdomadaires (année académique 1999-2000) proposées aux étudiants ingénieurs qui ont suivi le cours de Physique (Lion, 2000). Dans le contexte de ces épreuves, il n'y avait pas de définition de Modalités de Questionnement (MQ) dans la mesure où toutes les questions étaient de type « QCM » avec « Solutions Générales Implicites » (SGI, voir p. 67) et « pourcentages de certitude » (voir p. 68).

Interro 14B **Nombre de questions : 10**

Pour revenir à l'accueil ou imprimer, cliquez sur "continuer".
Légende : RC = Réponse Correcte ; NSP = Nombre de solutions proposées ; RM = Rubrique matière ; PM = Processus Mental.

Code	Titre	Type	RC	NSP	RM	PM
5 094 451	Couple sur boucle dans B	Problème simple	**8**	4	5	4
5 095 466	Interaction entre 2 boucles	Compréhension	**2**	5	5	2
5 101 441	Para- et diamagnétisme	Connaissance	**4**	4	5	1
5 103 460	Energie d'orientation d'un dipôle	Compréhension	**4**	4	5	2
5 112 446	Faraday - Boucle tournant dans B	Problème simple	**4**	4	5	4
5 112 452	Unité de B	Connaissance	**4**	5	5	1
5 112 453	Bobine dans B variable	Problème simple	**1**	4	5	4
5 113 461	Barre sur rails dans B	Problème simple	**8**	4	5	4
5 114 443	Loi de Faraday	Compréhension	**3**	4	5	2
5 114 450	Unité de la circulation de E	Connaissance	**4**	4	5	1

Un tel plan permet de guider et de rationaliser la réalisation des questions, voici la 1ère question de cette épreuve dont le PE est « *Interaction entre deux boucles* » et la CP « *Compréhension* » :

Q2. Deux fils conducteurs circulaires se trouvent dans deux plans parallèles; la droite passant par les centres des cercles ainsi constitués est perpendiculaire aux deux plans; si les deux boucles sont parcourues par des courants de même sens, elles vont :

1. se repousser;
2. s'attirer;
3. n'exercer aucune force l'une sur l'autre;
4. s'attirer si les intensités des courants sont égales;
5. se repousser si les intensités des courants sont égales.

Lors de cette étape, un contrôle qualité *a priori* peut avoir lieu en soumettant le questionnaire à la critique de collègues experts du contenu. Des docimologistes peuvent aussi effectuer à la demande des professeurs une relecture « formelle » de leurs QCM suivie d'un entretien où ils discutent des pistes à suivre en vue d'améliorer la qualité des questions.

L'enseignant définira également les barèmes et tarifs liés aux questions. Un « poids » peut aussi être associé à chaque question en fonction de l'importance que lui accorde l'examinateur (voir plus haut).

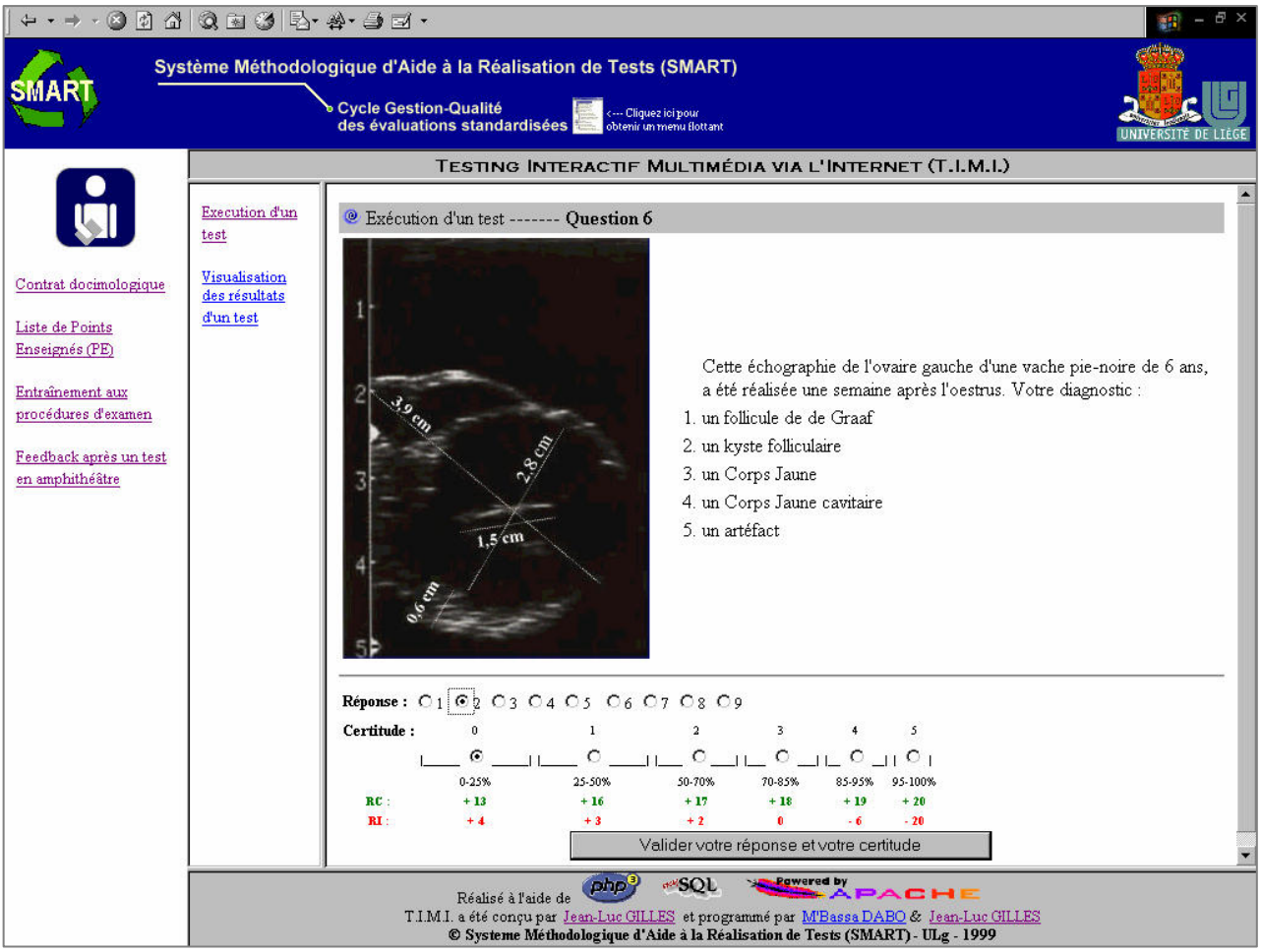
4. Entraînement des étudiants

Cette étape est particulièrement importante lorsqu'on utilise des méthodes de questionnement sophistiquées auxquelles les étudiants sont peu habitués à leur entrée à l'université. Par exemple, il est recommandé d'entraîner les étudiants à l'utilisation des Solutions Générales Implicites (SGI, voir p. 67) et des degrés de certitude (p. 68) avant la première épreuve certificative.

Nous préconisons au moins un test (quiz) d'entraînement en amphithéâtre pour les étudiants qui entrent à l'université et qui seront amenés à utiliser les dispositifs du SMART de Lecture Optique de Marques (LOM). Un quiz en fin de cours permet ainsi de se familiariser avec les formulaires spéciaux prévus pour la lecture optique (*formuloms,* voir p. 81) et les méthodes de questionnement associées.

Pour certains cours, les questions des examens qui ont eu lieu lors des années académiques précédentes (et qui sont donc « brûlées ») sont disponibles via l'Internet. Nous avons intitulé cette forme d'entraînement « *WebQuiz* » (Gilles, 1998b). A l'aide de ces derniers, l'étudiant peut répondre de chez lui ou d'une des salles publiques informatisées de l'ULg. Ces tests peuvent se faire entre deux cours et ne sont pas sanctionnés. L'ordre de réponse aux questions est déterminé par l'étudiant qui peut lire les questions avant de répondre et commencer par celle de son choix.

Récemment, tirant parti d'expériences précédentes (Leclercq & Gilles, 1993 ; Leclercq, Rommes & al., 1998) et plus particulièrement d'une utilisation du multimédia en intranet (Gilles & al., 1999), des dispositifs de Testing Interactifs Multimédias via Internet (*TIMI*) ont été mis au point dans le cadre du SMART. Ces *TIMI* permettent l'utilisation combinée dans les tests de séquences multimédias enregistrées sur CD-ROM (fournis aux étudiants) en complémentarité avec les QCM-SGI stockées dans des bases de données accessibles via l'Internet.



Les *TIMI* permettent aussi aux étudiants de recevoir un feedback immédiat à propos de la réponse correcte attendue après chaque question ainsi qu'un bilan de leurs performances cognitives (taux de réponses correctes, moyenne) et métacognitives (réalisme, tendance à se sur ou sous-estimer).

Enfin, il existe aussi un entraînement à l'utilisation des pourcentages de certitude qui est accessible via l'Internet[44] sous la forme d'une activité intitulée GUESS (Leclercq & Gilles, 1994). Lors de cet entraînement les étudiants sont amenés à accompagner de nombreuses réponses de leurs pourcentages de certitude. Un feedback est fourni en direct tout au long de l'exercice à chaque étudiant qui peut voir l'évolution de son réalisme, de sa tendance à se sur ou sous-estimer.

L'entraînement des étudiants est aussi pris au sérieux ailleurs dans le monde. Par exemple le Chicago Board of Education (1999) édite pour les enseignants des « Chicago Public Schools » (USA) un manuel d'instruction en vue de préparer leurs étudiants à la passation des tests standardisés. L'ouvrage s'intitule « *Preparing Your High School Students to Take Standardized Tests* » et est accessible via l'Internet[45], les auteurs y exposent une « *test-taking skills and abilities checklist* » de neuf conseils à répercuter auprès des étudiants ou à effectuer avec eux : « *(1) Follow the directions closely, (2) Budget time appropriately, (3) Check work carefully, (4) Read the entire item and all answers, (5) Answer easier questions first and persist to the end, (6) Make educated guesses, (7) Use test item formats for practice, (8) Review the practice items and answers with students, (9) Practice using answers sheets* ».

Enfin, signalons que depuis l'année académique 1998-1999, le SMART met aussi à la disposition des enseignants un système de boîtiers de vote électronique qui permet aussi d'entraîner les étudiants en amphithéâtre (Leclercq & al., 1999 ; Gilles & al., 2000). L'équipement comprend une centaine de boîtiers, une station de réception, un ordinateur et un vidéo projecteur portables.

Les boîtiers de vote ressemblent à des télécommandes sans fil à 10 chiffres (0 à 9) permettant à 100 étudiants (ou à 100 groupes d'étudiants) réunis dans un même amphithéâtre de fournir une réponse codée par un chiffre (donc à des QCM). Ces réponses sont captées par un "concentrateur" et traitées par un ordinateur portable contenant les fichiers de questions (rédigées au préalable par le professeur). Un logiciel (*INTERACT*[46]) qui permet d'afficher immédiatement sur l'écran du professeur les résultats d'ensemble. Ces résultats sont mémorisés dans un fichier.

Le professeur peut, s'il le veut, projeter à l'aide du vidéo projecteur son propre écran pour que les étudiants eux aussi aient, en direct, l'image de l'ensemble des réponses.

[44] Adresse Internet : http://www.ulg.ac.be/cafeim/guess/guessdc6.htm. La version Internet de GUESS a été conçue par D. Leclercq & J.-L. Gilles et programmée dans le langage JAVA par M. Hurard.
[45] Adresse Internet : http://intranet.cps.k12.il.us/Assessments/Preparation/preparation.html
[46] *INTERACT* est un logiciel produit par la firme Soft Concept.

5. Mise en œuvre de l'examen

L'examen proprement dit nécessite une série de préparatifs. La standardisation d'une épreuve implique une définition précise des modalités d'utilisation du test, des consignes. Le SMART apporte aussi son expertise dans ce domaine.

L'examinateur veillera lors de l'épreuve à mettre en place des procédures anti-fraude qu'il aurait tort de négliger (voir les problèmes liés à la fraude, p. 30).

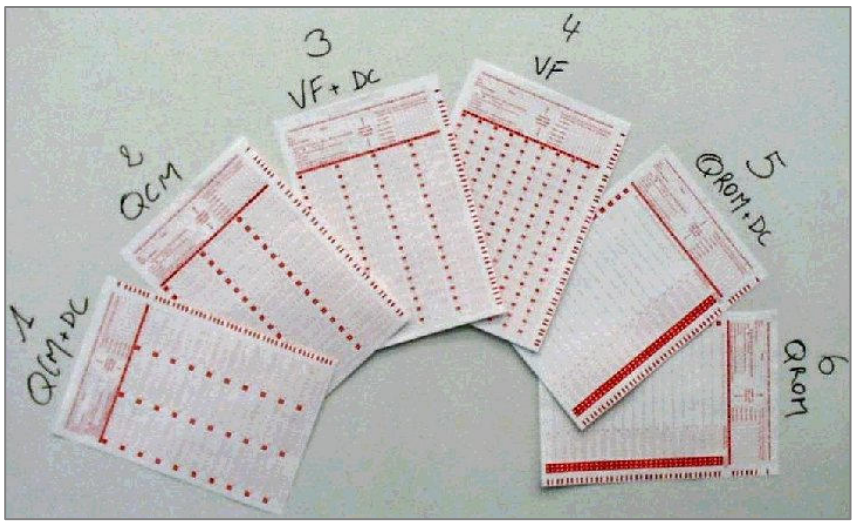
a) Les examens ayant recours à la Lecture Optique de Marques (LOM)

Plusieurs modalités de questionnement sont disponibles lorsque la lecture optique de marque est utilisée dans le cadre de l'aide proposée par le SMART :

- *Les QROC ou Questions à Réponses Ouvertes Courtes* : l'étudiant répond par un mot ou une locution lu(e) par un correcteur humain qui marquera sa note dans la case *ad hoc* (correct / incorrect) sur le *formulom* de l'étudiant, avant traitement de la feuille.
- *Les QCL ou Questions à Choix Large* : l'étudiant a le choix entre des centaines de réponses possibles (999 maximum), ce qui constitue une procédure intermédiaire entre les QCM et les QRO (Réponses Ouvertes) car l'étudiant doit d'abord penser (se rappeler, évoquer) la solution correcte AVANT d'aller voir dans la liste (index sous forme de mini dictionnaire) comment la coder.
- *Les QCM SGI ou à Solutions Générales Implicites* : où en plus des (3 à 5) solutions dactylographiées, l'étudiant doit considérer comme possible les 4 SGI que sont « 6. Aucune », « 7. Toutes », « 8. Manque de données » et « 9 Absurdité dans l'énoncé ». Cette consigne convient particulièrement aux interrogations à livres ouverts, car elle permet de tester spécialement la compréhension et la vigilance cognitive.
- *Les QVF ou Questions Vrai Faux*
- *Les degrés de certitude* peuvent s'ajouter à toutes les formes de questionnement proposées ci-dessus et permettent de calculer outre les résultats cognitifs classiques (taux d'exactitude), un indice métacognitif de réalisme (sur ou sous-estimation) pour chaque étudiant.

Voici une vue des différents *formuloms*[47] développés par le SMART pour permettre un traitement automatisé des modalités de questionnements qui viennent d'être présentés.

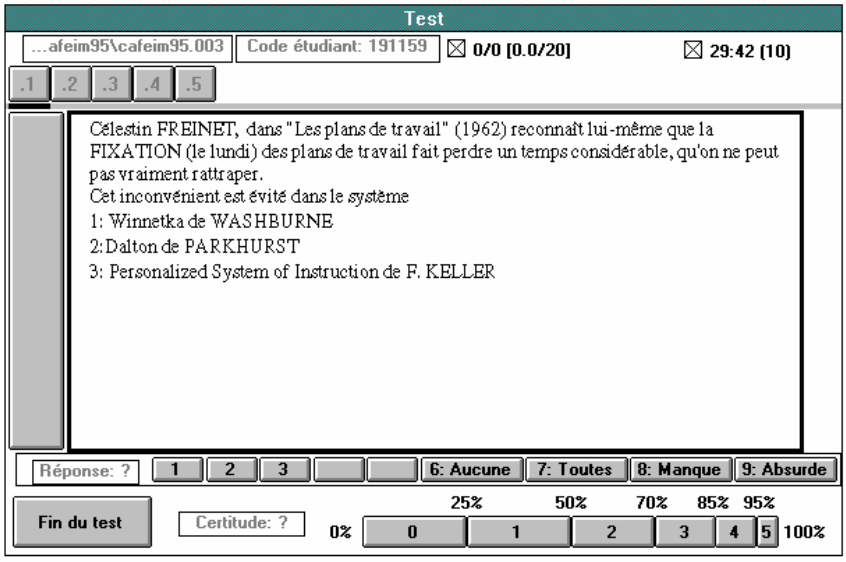
Lors de certaines épreuves, les évaluateurs autorisent de brèves justifications des réponses sur des feuilles ad hoc. Il est en outre convenu que le(s) correcteur(s) ne lira(ont) que les commentaires concernant les réponses incorrectes. La justification ne peut donc QUE bénéficier à l'étudiant.



[47] *Formulom* est le terme technique utilisé par le SMART pour désigner le formulaire de réponses spécial destiné à la lecture optique de marques (ils sont imprimés en encre aveugle rouge sur du papier spécial équerré).

b) Les examens ayant recours au testing interactif en intranet (*WINCHECK*)

WINCHECK (Leclercq & al., 1994) est un logiciel d'évaluation interactive développé par le Service de Technologie de l'Education de l'Université de Liège (ULg). Le programme fonctionne dans l'environnement Windows et permet l'utilisation du « double check » (voir plus haut, pp. 66 et 72). *WINCHECK* est disponible via les ordinateurs et le réseau Intranet du Centre d'Auto-Formation et d'Evaluation Interactives Multimédias (CAFEIM) de la FAPSE-ULg. L'évaluateur peut se constituer des banques de questions (QCM-SGI) et créer des tests en sélectionnant les questions en fonction de critères liés aux contenus, aux processus mentaux et aux objectifs.



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

A la FAPSE-ULg les étudiants prennent rendez-vous pour l'examen en réservant un ordinateur au CAFEIM dans une plage horaire qui leur convient.

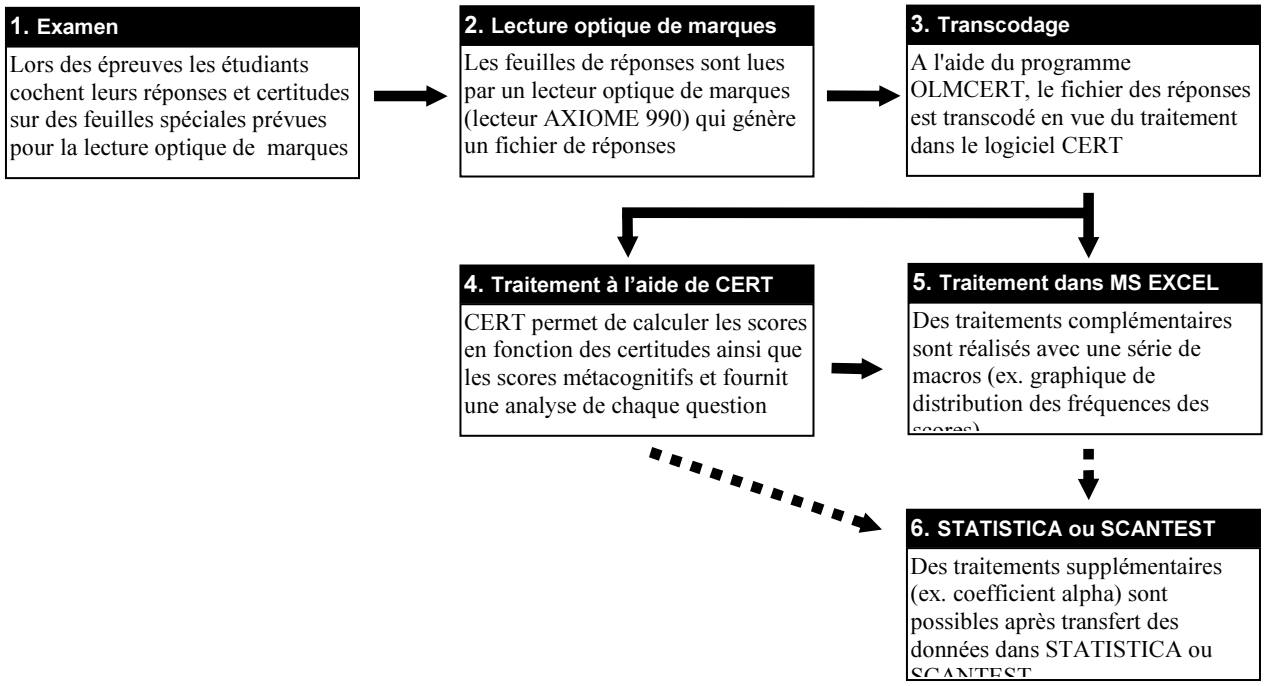
Enfin, signalons aussi l'importance de recueillir des informations sur la qualité des questions lors de l'examen proprement dit. Les étudiants n'hésitent pas à signaler les problèmes qu'ils perçoivent dans les questions lors des épreuves. Collecter leurs avis est utile dans la mesure où ils apportent un éclairage qualitatif qui permet de mieux interpréter les indices de qualité des questions calculés lors de l'étape suivante de correction et discussion.

6. Correction et discussion

Durant l'étape de correction une série de contrôles de qualité ont lieu lors de la lecture informatisée des réponses fournies par les étudiants. En ce qui concerne la qualité des questions, des contrôles de qualité *a posteriori* sont effectués à l'aide d'indices de discrimination tels que la corrélation bisériale de point classique (*rpbis classique*, voir p. 171) calculée pour chaque proposition au sein des QCM. Nous verrons plus loin que c'est dans le cadre de cette étape de correction et discussion que s'insère l'utilisation des nouveaux indices d'analyse de la qualité spectrale des questions (voir pp. 178 et 184).

Lors de cette étape, le SMART communique les résultats de l'analyse de la qualité des questions à l'examinateur. Lorsque la situation l'exige, cet envoi d'information peut être suivi d'une discussion entre ce dernier et un chercheur du SMART à propos des mesures correctives (suppression de questions, valorisation de propositions au départ considérées comme incorrectes, …) à mettre en place pour rectifier l'épreuve (voir le cas de régulation des questions d'un test exposé plus haut, p. 47).

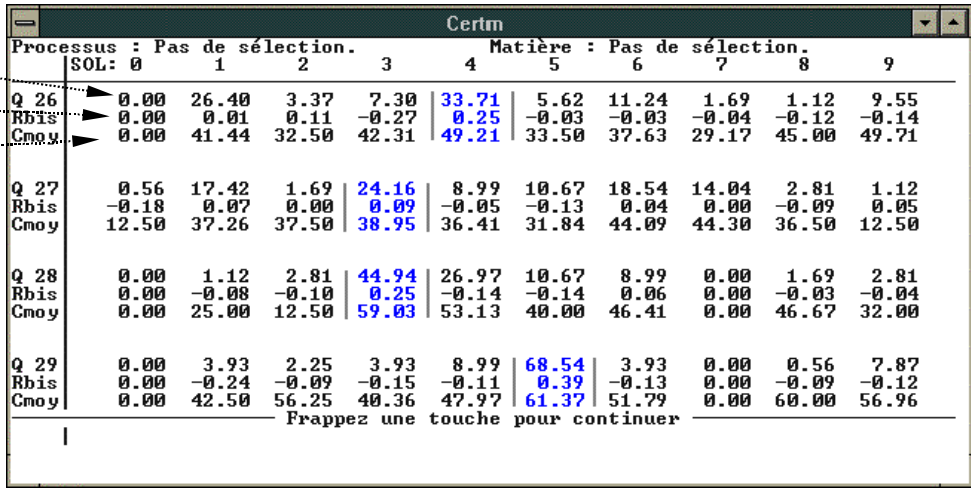
Le SMART est équipé d'une chaîne complète de lecture optique de marques performante et fiable pour produire des questionnaires, les lire et les traiter (duplicatrice digitale, lecteur optique et logiciels de traitement *ad hoc*). Voici un schéma qui montre les différentes étapes de la correction d'un examen.



Après l'examen, la lecture optique des réponses des étudiants et le transcodage du fichier au format du logiciel de traitement CERT[48] (Boxus & al., 1991), l'analyse de la qualité des questions peut réellement débuter.

a) Le traitement à l'aide du programme CERT

Les formuloms sont lus par un lecteur optique de marques qui génère un fichier informatisé contenant les réponses des étudiants. Ce fichier est ensuite converti au format du logiciel CERT utilisé pour la correction. CERT fournit l'analyse de chaque QCM de l'épreuve en trois lignes par question.



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

Frappez une touche pour continuer

Pour chaque solution proposée (de 1 à 9 pour ces QCM-SGI) sont en effet fournis (0 = omission) dans ces trois lignes respectivement :

(1) le pourcentage d'étudiants qui ont choisi la proposition ;

(2) le coefficient *rpbis classique* (voir l'exposé de la problématique du *rpbis classique*, p. 171) ;

(3) la certitude moyenne. L'écran présenté ci-dessus montre comment se présente l'écran ou le listing (les statistiques concernant la réponse correcte sont encadrées de deux barres verticales).

[48] CERT a été conçu en 1991 par le Service de Technologie de l'Education (STE) sous les auspices de la Commission des Communautés européennes dans le cadre du programme EUROTECNET (Boxus & al., 1991). Le logiciel a été programmé dans le langage Clipper par MM. Philippe de Brogniez et Henri Wuidar.

Sans entrer dans les détails ici, disons que le *rpbis classique* (voir p. 171) est un indice de discrimination qui varie dans l'intervalle [-1, 1] qui permet d'évaluer dans quelle mesure une proposition discrimine les étudiants « forts » des étudiants « faibles » (du point de vue du score total obtenu au test). L'analyse d'une épreuve à l'aide de ce dispositif peut amener l'évaluateur à supprimer une question, par exemple lorsque la réponse correcte récolte un *rpbis classique* négatif ou proche de zéro et certains distracteurs des *rpbis classiques* positifs marqués (voir le cas présenté dans l'introduction, p. 47). Dans l'exemple ci-dessus, la question 27 devrait être analysée car le rpbis classique de la réponse correcte (SOL 3) est faible (0,09) et en particulier celui du premier distracteur (SOL 1) presque aussi élevé (0,07).

Il arrive que l'enseignant décide d'accepter le choix d'un distracteur comme réponse correcte au vu des *rpbis classiques* et des éventuels commentaires de justification des étudiants.

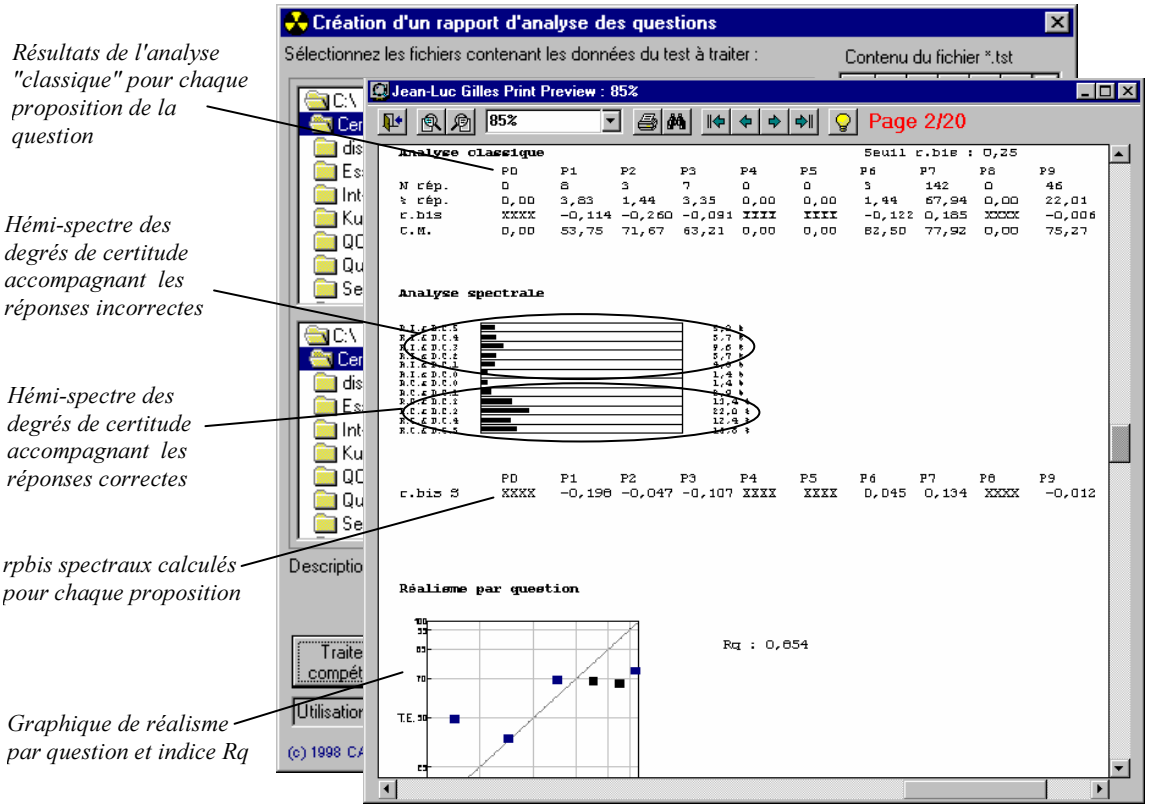
Il est également possible d'exporter les résultats des évaluations dans les logiciels EXCEL et STATISTICA en vue de traitements plus poussés.

Nous verrons plus loin qu'il est possible d'utiliser les informations fournies par les pourcentages de certitude qui ont accompagné les réponses des étudiants pour mesurer la cohérence spectrale des questions. C'est un autre logiciel de traitement intitulé « *SCANTEST 1.0* » qui est alors utilisé.

b) Le logiciel *SCANTEST* (version 1.0)

L'utilisation des informations liées à l'emploi des pourcentages de certitude en vue d'évaluer la qualité des QCM et QROC nous a amener à programmer un nouveau type de logiciel de traitement qui permet de calculer des rpbis Spectraux Contrastés (*rpbis SC*, p. 178) et des rpbis Spectraux Contrastés après Turbo analyse (*rpbis SCT*, p. 184) et donc d'effectuer une analyse spectrale des examens.

Voici un aperçu du premier logiciel que nous avons créé pour accélérer les traitements, il s'agit de la première version (1.0) de *SCANTEST*[49] (Gilles, 1998a).



Exemple de rapport de traitement pour une question - logiciel SCANTEST 1.0

[49] Nous avons conçu et programmé le logiciel SCANTEST à l'aide du langage *Microsoft Visual Basic 5.0.*

Dans sa version 1.0, le programme *SCANTEST* utilise les fichiers des réponses des étudiants et de paramétrage des 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 du SMART. Les sorties peuvent être visualisées à l'écran ou imprimées. Le rapport final comprend pour chaque question :

- une analyse "classique" des propositions (ces résultats étaient déjà disponibles dans CERT): nombre d'étudiants (N rép.) qui ont choisi la proposition, pourcentage (% rép.), corrélation bisériale de point (*rpbis classique*) 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 certitudes 0, 1, 2, 3, 4, 5) ;
- les *rpbis spectraux* (*rpbis 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 Réalisation des prédictions par question (*Rq*) reprenant pour chaque degré de certitude (en abscisse) les taux d'exactitude correspondant (en ordonnée). Le graphique de chaque question est accompagné de la valeur de son indice *Rq* (voir p. 242).

Notons aussi ici de manière un peu trop préliminaire que des synthèses sont également proposées :

- un récapitulatif des certitudes moyennes pour les réponses correctes et incorrectes accompagné d'un graphique de cohérence métacognitive ;
- un récapitulatif des *rpbis classiques* pour les réponses correctes et incorrectes avec un graphique de qualité des *rpbis classiques* reprenant ces données pour l'ensemble des questions ;
- un récapitulatif des *rpbis spectraux* pour les réponses correctes et incorrectes accompagné d'un graphique de qualité des *rpbis S*.

Cependant dans cette première version du logiciel *SCANTEST* les *rpbis spectraux* n'étaient pas contrastés (voir p. 178) et le principe de la turbo analyse n'était pas encore implémenté (voir p. 186).

Nous verrons plus loin que ces fonctionnalités de traitement contrasté et de turbo analyse ont été introduites dans une nouvelle version (2.0) du logiciel *SCANTEST* (voir présentation, p. 199) programmée en vue d'étudier et de mettre au point ces nouveaux instruments d'analyse spectrale de la qualité des épreuves. La nouvelle version de SCANTEST a été programmée pour permettre l'analyse spectrale des résultats des milliers d'étudiants (entre 1.410 et 3.846 sujets selon les tests) ayant participé aux 10 épreuves du projet de MOnitoring HIstorique de cohortes de CANdidatures universitaires (MOHICAN, voir p. 93).

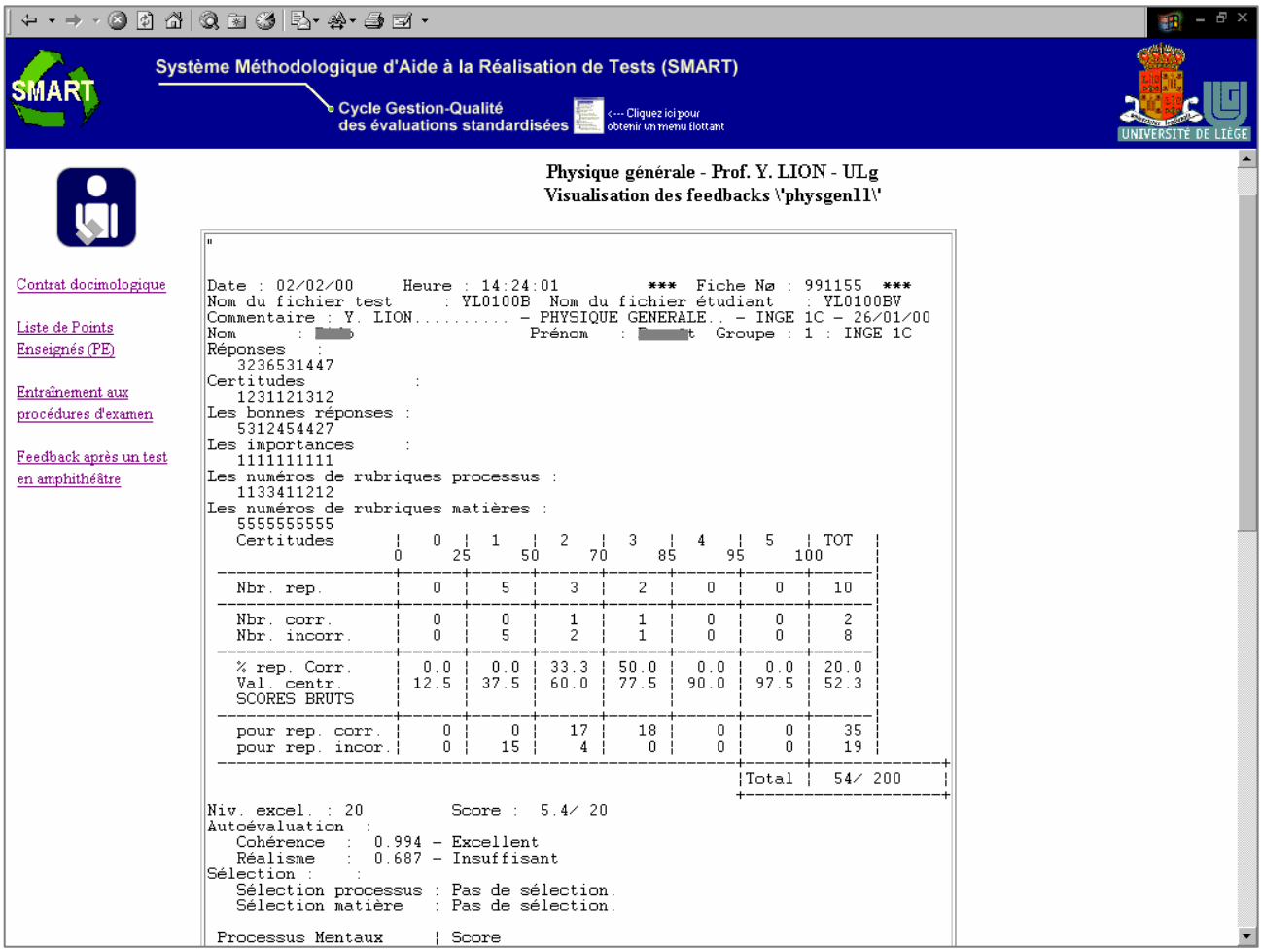
7. La phase des feedbacks aux examinés

Cette étape consiste en la communication aux étudiants, d'une part, des résultats des performances individuelles et, d'autre part, des résultats des contrôles de qualité relatifs aux questions (les performances de l'épreuve, c'est-à-dire des listings provenant du programme CERT (voir plus haut) que le professeur annote en expliquant les éventuelles mesures correctives qui ont été prises en vue d'améliorer la qualité de l'épreuve. Ces listings annotés sont ensuite affichés aux valves du service.

Nous verrons plus loin que lorsqu'on demande aux étudiants du premier cycle de la FAPSE-ULg leur avis sur la qualité des examens qu'ils ont subi, leurs reproches les plus marqués se situent au niveau du manque de rétroactions en ce qui concerne leurs performances aux examens (voir p. 90). Ce problème a aussi été mis en avant par les étudiants de l'Université de Montréal dans le cadre d'une enquête sur les pratiques d'évaluation des apprentissages menée par le Groupe de Recherche Interdisciplinaire en Pédagogie Universitaire (GRIPU) (Blais & al., 1997). Cette équipe de chercheurs signale à propos des avis émis par les étudiants interrogés (n = 54) : « *Leurs demandes même si elles ne peuvent pas être généralisées, pointent des dimensions où une fragilité des pratiques se manifeste : le manque de rétroaction (absence de commentaires ou commentaires réduits ne permettant pas de comprendre les erreurs et d'apprendre de celles-ci)…* » (Blais & al., p. 100).

Dans le but d'essayer d'améliorer la communication des résultats après les examens, nous avons mis en place depuis l'année académique 1998-1999 dans le cadre des épreuves standardisées réalisées à l'aide du dispositif de lecture optique de marques du SMART, une procédure d'accès aux feedbacks individualisés via l'Internet. Les étudiants peuvent obtenir leurs résultats après un examen standardisé après s'être connectés au site web du SMART (et après avoir fourni leur code d'accès).

Voici un aperçu de l'interface proposée par le SMART et utilisée notamment avec les étudiants ingénieurs ayant participé aux interrogations hebdomadaires du cours de physique lors de l'année académique 1999-2000 (Lion, 2000).



Le succès rencontré par cette possibilité de consulter son feedback après une épreuve nous fait penser que ce service offert par le SMART aux étudiants répond à un réel besoin chez ces derniers. Les statistiques de fréquentation du site web du SMART lors de la période de mise en œuvre de la première expérience de diffusion des feedbacks via l'Internet dans le cadre des interrogations hebdomadaires du cours de Physique destiné aux étudiants de 1ère candidature ingénieurs (de novembre 1999 à février 2000) ont montré une augmentation sensible du nombre de connexions au serveur en parallèle avec une augmentation de la consultation des pages du site qui permettaient l'accès aux feedbacks individualisés.

8. La macro-régulation

La dernière étape « Macro-régulation » permet d'organiser l'amélioration de tout le processus pour un cycle ultérieur de construction d'examen. C'est dans le cadre de cette 8ème étape que sont recueillis les avis des étudiants dans le contexte du SMART.

Depuis 1997, une Commission interfacultaire d'EVALuation des ENSeignements (EVALENS) mise en place par le Conseil Général des Etudes (CGE) de l'Université de Liège organise à l'échelle de l'institution une évaluation systématique des cours, travaux pratiques et examens. De 1997 à 1999, au sein de chaque faculté, des sous-commissions ont été chargées de rédiger les items qui devaient figurer sur les questionnaires d'avis (les questionnaires ont donc été créés sur mesure pour chaque faculté durant cette période). Depuis le début des opérations, le SMART a été sollicité pour fournir une aide méthodologique et logistique à la Commission EVALENS. Des formulaires spéciaux permettant le dépouillement par lecture optique de marques des avis des étudiants ont ainsi été conçus en étroite collaboration avec les facultés.

A la Faculté de Psychologie et des Sciences de l'Education (FAPSE-ULg), trois types de questionnaires furent proposés par la commission facultaire d'évaluation des enseignements : un premier pour les cours, un second pour les travaux pratiques et un troisième pour les examens. Dans cette partie nous allons d'abord exposer la façon dont les avis des étudiants ont été récoltés et traités en vue d'améliorer la qualité des examens à la FAPSE-Ulg ainsi que le mode de calcul des moyennes des avis. Ensuite nous présenterons les faits saillants liés aux avis récoltés depuis 1997.

a) Procédure de recueil des avis des étudiants à propos des examens à la FAPSE-ULg

Nous présentons en annexe le questionnaire (*formulom*) de recueil des avis des étudiants à propos des examens utilisés lors des années académiques 1996-1997 et 1997-1998 (voir p. 481). Ce *formulom* proposé par la sous-commission EVALENS de la FAPSE-ULg[50] a été soumis au Conseil de Faculté qui a accepté de l'utiliser dans le cadre des enseignements du 1er cycle.

A la fin du dernier examen de la 1ère session en juin 1998 chaque étudiant inscrit en 1ère ou en 2ème candidature reçut une série de 15 *formuloms*. Pour des raisons budgétaires, en juin 1999 l'opération eu lieu uniquement en 1ère candidature. Les questionnaires furent complétés sur place (un *formulom* par examen évalué). Chaque *formulom* contenait 13 items (p. 481) qui étaient répartis dans 3 catégories : *« 1. Mode d'évaluation* », « *2. Attitude de l'examinateur* » et « *3. Feedback après examen* ». Le choix des 13 items présentés dans les *formuloms* fut le résultat d'un compromis au niveau :

- du nombre maximum d'items auxquels il était possible de répondre en un temps raisonnable à la fin du dernier examen (13 questions * 15 cours cela représente 195 items à cocher) ;
- du nombre minimum de questions à poser pour donner une vue d'ensemble suffisamment détaillée de la façon dont les étudiants percevaient les examens (tous les objectifs « qualité » que nous avons présentés ne sont malheureusement pas couverts par ces 13 items, ni d'ailleurs toutes les 8 étapes du cycle de réalisation d'une épreuve) ;
- de l'importance à accorder aux différentes questions possibles dans le contexte facultaire (dans le contexte FAPSE, le choix des 13 questions sélectionnées fut discuté entre les membres de la Commission facultaire EVALENS, puis soumis au Conseil des études et ensuite voté en Conseil de faculté).

[50] Membres de la Commission facultaire d'évaluation des enseignements : Annick Combain, Marianne Debry, Marc Demeuse, Brigitte Denis, Jean-Jacques Detraux, Pascal Detroz, Rebekka Dobbels, Anne-Marie Etienne, Caroline Geuzaine, Jean-Luc Gilles (Président de la commission), Véronique Jans, Olivier Jurdan, Dieudonné Leclercq, Bernadette Mouvet et Ezio Tirelli.

Voici les 13 items utilisés en 1997-1998 et 1998-1999 :

1. Mode d'évaluation

[1.1] Le mode d'évaluation (QCM, oral, écrit, travaux personnels) était adéquat

[1.2] L'entraînement à la procédure d'évaluation avant l'examen était suffisant

[1.3] Les questions d'examen étaient clairement formulées

[1.4] Les questions d'examen étaient bien adaptées à la matière

[1.5] Le mode d'évaluation choisi permet au professeur d'avoir une bonne représentation des compétences acquises par l'étudiant

[1.6] L'évaluation est équitable et impartiale

2. Attitude de l'examinateur

[2.1] Les exigences de l'enseignant sont clairement présentées aux étudiants

[2.2] Les exigences de l'enseignant sont présentées en temps utiles

[2.3] L'enseignant met l'étudiant à l'aise à l'examen oral (si écrit, indiquez SO)

[2.4] L'horaire fixé pour l'examen est respecté

3. Feedback après examen

[3.1] Les réponses correctes sont communiquées aux étudiants après l'examen

[3.2] L'analyse statistique de la qualité des questions (r.bis) est communiquée

[3.3] Après l'examen, l'étudiant peut obtenir des explications sur la qualité de ses réponses auprès de l'enseignant.

Chaque item se présente donc sous la forme d'une affirmation relative à l'examen évalué. L'étudiant exprime son avis en se référant à l'échelle ci-dessous. Cette échelle est reprise sur le *formulom* (voir annexe, p. 481) où il coche son choix.

SO = sans objet	1 = pas du tout d'accord	2 = pas d'accord	3 = plutôt pas d'accord	4 = plutôt d'accord	5 = d'accord	6 = tout à fait d'accord	SA = sans avis

Après lecture optique et traitement des avis des étudiants deux types de feedbacks sont diffusés. Le premier qui est public et anonyme est destiné à être affiché aux valves de la faculté. Voici un exemplaire de ces données publiques relatives à l'évaluation des examens par les étudiants de 1ère candidature en 1997-1998.

Faculté de Psychologie et Sciences de l'Education - 1 ère candidature (1997-1998)

Tableau des moyennes (anonymes) pour l'ensemble des examens de la section évaluée

Les chiffres qui figurent dans les cases indiquent le nombre d'examens de cette section qui récoltent la moyenne correspondant à la case.

		Moyennes des avis											
		1		1,5		2		2,5		3		3,5	
1.	**Mode d'évaluation**												
1.1	Le mode d'évaluation (QCM, oral, écrit, travaux personnels) était adéquat									2	2	9	2
1.2	L'entraînement à la procédure d'évaluation avant l'examen était suffisant					1	3	5	1	5			
1.3	Les questions d'examen étaient clairement formulées							2	5	6	2		
1.4	Les questions d'examen étaient bien adaptées à la matière							2	3	8	2		
1.5	Le mode d'évaluation choisi permet au professeur d'avoir une bonne représentation des compétences acquises par l'étudiant							3	7	4	1		
1.6	L'évaluation est équitable et impartiale						1	3	4	6	1		
2.	**Attitude de l'examinateur**												
2.1	Les exigences de l'enseignant sont clairement présentées aux étudiants						1	3	4	6	1		
2.2	Les exigences de l'enseignant sont présentées en temps utiles						1	3	3	6	2		
2.3	L'enseignant met l'étudiant à l'aise à l'examen oral (si écrit, indiquez SO)							4	7	4			
2.4	L'horaire fixé pour l'examen est respecté								2	9	4		
3.	**Feed-back après examen**												
3.1	Les réponses correctes sont communiquées aux étudiants après l'examen	1	8	2	3	1							
3.2	L'analyse statistique de la qualité des questions (r.bis) est communiquée	1	10	2	1	1							
3.2	Après l'examen, l'étudiant peut obtenir des explications sur la qualité de ses réponses auprès de l'enseignant.		4	5	4	2							

A droite des items on remarque une série de 10 cases. Dans le cas du premier item « *1.1 Le mode d'évaluation (QCM, oral, écrit, travaux personnels) était adéquat* » la cinquième case comprise entre 3 et 3,5 sur l'échelle proposée contient un « 2 ». Cela signifie que 2 enseignements de 1ère candidature récoltent une moyenne située entre 3 et 3,5 à cet item. Etant donné qu'il s'agit ici de données publiques (anonymes) on ne peut savoir de quels enseignements il s'agit.

Dans la procédure EVALENS, cette synthèse des avis relatifs aux examens (ainsi que celles des cours et travaux pratiques) est non seulement affichée aux valves de la faculté, mais aussi présentée lors d'un Conseil de faculté où tous les membres du corps académique ainsi que les représentants du corps scientifique et les délégués des étudiants sont présents et peuvent intervenir pour obtenir des clarifications ou émettre des suggestions.

Le second type de feedback est personnel et destiné à chaque enseignant. C'est le même que celui qui figure ci-dessus à la différence que la position de l'enseignant est signalée par une case grisée. Chaque enseignant reçoit son feedback personnel sous pli fermé et une copie est envoyée au Doyen de la faculté. Ce dernier organise une discussion à propos des avis des étudiants avec chaque professeur.

En plus de sa synthèse personnelle, le professeur reçoit les formulaires sur lesquels les étudiants ont noté leurs commentaires dans la case « *Suggestion(s) à propos de l'évaluation du cours* ».

b) Mode de calcul des moyennes pour l'ensemble des examens d'une section

La moyenne des avis récoltée par chaque examen à chaque item du formulom est calculée de la façon suivante. Voici un exemple (fictif) de répartition des avis dans le cadre d'un item pour lequel 235 étudiants ont donné leur avis :

SO = sans objet	1 = pas du tout d'accord	2 = pas d'accord	3 = plutôt pas d'accord	4 = plutôt d'accord	5 = d'accord	6 = tout à fait d'accord	SA = sans avis
	1	2	3	4	5	6	
n = 5	*n* = 10	*n* = 30	*n* = 35	*n* = 70	*n* = 55	*n* = 15	*n* = 15
	Seuls les avis de « 1 » à « 6 » sont pris en compte dans le calcul de la moyenne (*n* = 215)						

Le tableau montre que 5 étudiants ont coché « SO = sans objet », 10 ont coché « 1 = pas du tout d'accord », 30 ont coché « 2 = pas d'accord », etc. Au total, pour cet enseignement, 215 avis ont été retenus après décompte des SO et SA. Les avis « 1 = pas du tout d'accord » à « 6 = tout à fait d'accord » sont considérés comme des points accordés à l'enseignement par les étudiants, ce qui donne :

$$(10 * 1) + (30 * 2) + (35 * 3) + (70 * 4) + (55 * 5) + (6 * 15) = 820$$

Dès lors, la moyenne récoltée par l'examen à cet item vaut 820/215 = 3,8 (le minimum pourrait être 1 et le maximum 6).

Remarquons qu'étant donné l'échelle d'avis proposée, la note charnière se situe à 3,5 c'est-à-dire entre le négatif « *3 = plutôt pas d'accord* » et le positif « *4 = plutôt d'accord* ».

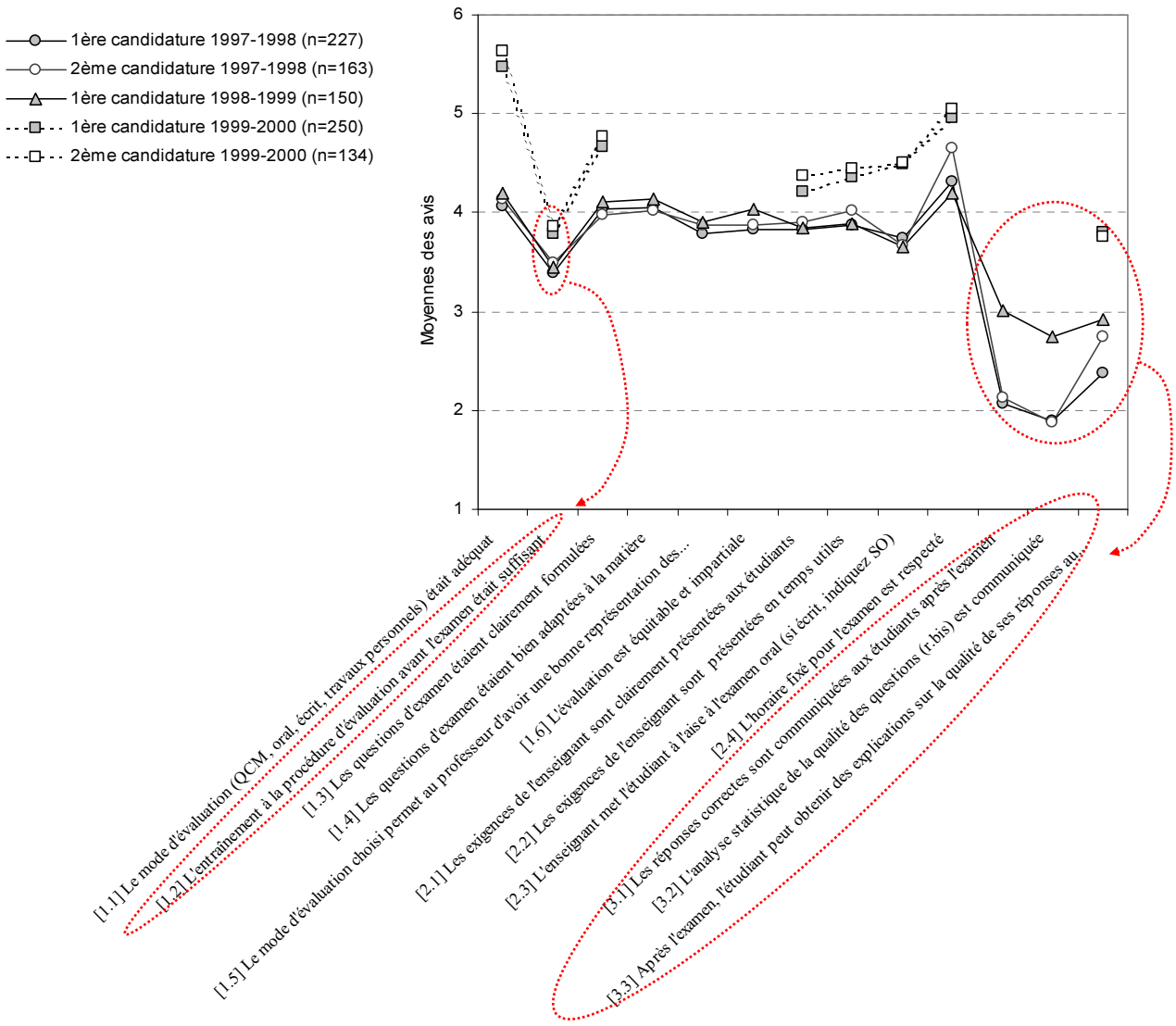
c) Faits saillants liés aux avis récoltés depuis 1997 à propos des examens à la FAPSE-ULg

Le graphique ci-dessous reprend les moyennes des avis des étudiants calculées pour l'ensemble des examens organisés en 1ère candidature ou/et en 2ème candidature à la FAPSE-ULg de 1997 à 2000 (Gilles & al., 1998). Chacun des 13 items rappelés sous le graphique reçoit une moyenne qui peut aller de 1 à 6. Les points représentant les moyennes obtenues par une candidature ont été reliés afin d'aider le lecteur à visualiser les résultats section par section.

En 1999-2000, la procédure de récolte des avis fut modifiée ainsi que le questionnaire qui ne comportait plus tous les items proposés les deux années académiques précédentes (5 items furent supprimés et 11 nouveaux items furent ajoutés). Ceci explique l'absence de carrés (et de traits pointillés) pour les deux candidatures relatives à 1999-2000 en ce qui concerne les items 1.4 à 1.6, 3.1 et 3.2. En 1998-1999, la 2ème candidature ne fut pas évaluée (pour des raisons budgétaires).

Les « *n* » qui figurent dans la légende représentent les nombres moyens d'étudiants qui ont donné leurs avis sur les examens dans les différentes sections envisagées.

Comparaison des moyennes des avis des étudiants à propos des examens dans le 1er cycle de la FAPSE-ULg (de 1997 à 2000)



Nous remarquons que pour la 1ère et la 2ème candidature de l'année 1997-1998 (ronds grisés et blancs reliés par des traits continus) les courbes des moyennes des avis sont très proches. La courbe des avis récoltés en 1ère candidature en 1998-1999 (triangles grisés et trait continu) est aussi fort proche de ces deux dernières.

Globalement, nous remarquons que les items qui récoltent les moyennes d'avis les moins favorables sont liés d'une part à l'entraînement à la procédure d'évaluation avant l'examen et d'autre part aux feedback après l'examen. Ces résultats montrent que les étudiants ne se sentent pas suffisamment préparés aux modalités d'évaluation et de façon encore plus marquée que les feedbacks qu'ils peuvent obtenir après l'examen sont insuffisants. On constate cependant une nette amélioration en 1998-1999 par

rapport à l'année précédente en ce qui concerne les 3 derniers items relatifs à la catégorie « *3. Feedback après l'examen* ».

Rappelons que ces résultats sont présentés et discutés en Conseil de faculté et que des entrevues individuelles sont organisées entre chaque professeur et le Doyen de la faculté lors desquelles des actions pour améliorer la situation sont envisagées. Les enseignants peuvent alors s'expliquer à propos des moyennes que récoltent leurs examen et c'est aussi pour eux l'occasion d'exprimer des besoins qui permettraient d'augmenter la qualité de leurs épreuves.

On remarque en 1999-2000 une amélioration aux huit items repris des questionnaires d'avis présentés les années précédentes. Lorsqu'on compare les moyennes obtenues aux huit items, on remarque que les deux items relatifs à l'entraînement et au feedback après examen restent ceux qui obtiennent les avis les moins favorables.

d) Faits saillants pour l'année académique 1999-2000

Précédemment nous avons signalé que le questionnaire utilisé en 1999-2000 était différent. Voici les 19 items relatifs aux examens qui furent proposés aux étudiants.

1. Préalables à l'examen
- **[1.1]** L'entraînement à la procédure d'évaluation avant l'examen était suffisant
- **[1.2]** Les exigences de l'enseignant sont clairement présentées aux étudiants
- **[1.3]** Les exigences de l'enseignant sont présentées aux étudiants en temps utile

2. Mode d'évaluation :
- **[2.1]** QCM : ce mode d'évaluation était adéquat
- **[2.2]** Certitudes : ce mode d'évaluation était adéquat
- **[2.3]** Examen oral : ce mode d'évaluation était adéquat
- **[2.4]** Examen écrit rédigé : ce mode d'évaluation était adéquat
- **[2.5]** Présentation écrite d'un travail (rapport, fiche de lecture, …) : ce mode d'évaluation était adéquat

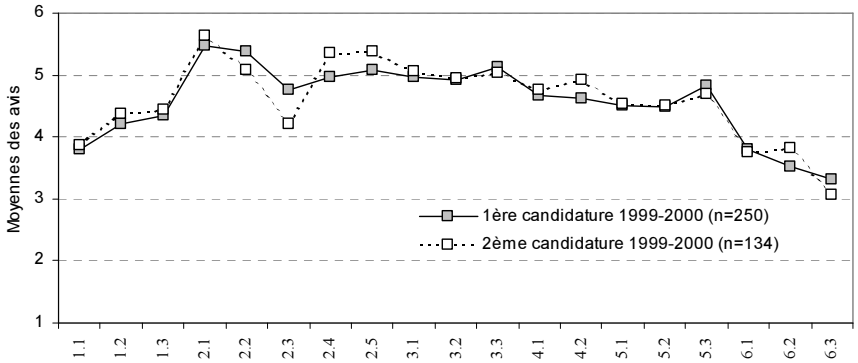
3. Organisation de l'examen
- **[3.1]** L'horaire fixé pour l'examen est respecté
- **[3.2]** La durée de l'examen est suffisante
- **[3.3]** La surveillance anti-fraude est suffisamment dissuasive

4. Qualité du questionnement
- **[4.1]** Les questions de l'examen sont clairement formulées
- **[4.2]** La procédure d'examen recouvre bien toute la matière

5. L'examinateur(trice)
- **[5.1]** …Adopte une attitude correcte à l'égard de l'étudiant, lors de l'examen oral
- **[5.2]** …Met l'étudiant à l'aise lors de l'examen
- **[5.3]** …Est impartial

6. Après l'examen …
- **[6.1]** …L'étudiant peut obtenir auprès de l'enseignant des explications sur la qualité de ses réponses
- **[6.2]** …Les résultats (points ou grades) sont affichés
- **[6.3]** …Pour les examens écrits, les réponses correctes sont communiquées dans les délais

Comparaison des moyennes des avis des étudiants à propos des examens dans le 1er cycle de la FAPSE-ULg lors de l'année académique 1999-2000



Les points des courbes ci-contre montrent les moyennes des avis des étudiants récoltées par les 19 items en 1999-2000.

Nous observons deux courbes très proches en 1ère et 2ème candidature (comme en 1997-1998).

On remarque à nouveau que l'item lié à l'entraînement (1.1) et ceux qui sont liés aux rétroactions après l'examen (6.1, 6.2 et 6.3) obtiennent les moyennes les plus faibles.

Nous constatons aussi que l'item qui obtient la moyenne la plus élevée en 1ère et en 2ème candidature est celui qui concerne l'adéquation du mode de questionnement par QCM. Les certitudes sont aussi considérées comme un mode d'évaluation adéquat. En 1ère candidature les examens écrits rédigés et les travaux écrits rédigés sont considérés moins adéquats qu'en 2ème candidature. La moyenne obtenue par l'examen oral est quant à elle la moins élevée en ce qui concerne l'adéquation des modes d'évaluation, avec des différences entre la 1ère et la 2ème candidature dans la mesure où les étudiants de 2ème trouvent l'examen oral moins adéquat que les étudiants de 1ère.

E. Le contexte des « check up » du projet de MOnitoring HIstorique de cohortes de CANdidatures universitaires (MOHICAN)

1. Historique et objectifs

Au début des années 90, le Conseil Inter Universitaire de la Communauté française de Belgique (CIUF) confia au groupe de travail « Réussite »[51] la mission d'étudier le phénomène de réussite et d'échec dans le 1er cycle d'études des institutions universitaires francophones. Ce travail aboutit à un rapport intitulé « *Franchir le cap des candis* » (Leclercq & al., 1997). Dans ses conclusions le groupe de travail propose quatre catégories d'hypothèses pour expliquer l'affaissement des taux de réussite depuis 1991-1992. Il s'agit d'hypothèses liées (1) à l'université, (2) aux étudiants, (3) à l'enseignement secondaire et (4) à la société globale. Dans le but de vérifier une série d'hypothèses émises à l'époque, les membres du groupe « Réussite » proposèrent le projet MOHICAN : « *Le système universitaire ne sera jamais en mesure de répondre à ces diverses questions, de mettre à l'épreuve ces hypothèses s'il ne se dote pas d'un mécanisme anticipatif de récolte de données. On ne peut se contenter de les récolter a posteriori. Il s'agit de se doter d'un système permanent d'observation (en anglais monitoring), avec une visée rétrospective (historique) des cohortes successives d'étudiants de candidature. Le projet MOHICAN (Monitoring Historique de Cohortes de Candidatures universitaires) répond à cet objectif. Ce projet comporte le recueil d'avis, d'opinions (données subjectives), de données objectives (épreuves de rendement), de données contextuelles.* » (op. cit. , p. 89).

Comme on ne dispose pas d'épreuve de type « baccalauréat » telles qu'il en existe en France, on ne peut en Communauté française de Belgique se prononcer sur l'évolution de l'état des compétences des apprenants au sortir du secondaire. Dès lors, comme le souligne Leclercq (2000), : « *Ceci laisse la porte ouverte aux déclarations les plus extrêmes du type 'ils ne savent plus rien' ou 'l'enseignement secondaire fait mal son travail', déclarations difficilement réfutables en l'absence de données empiriques représentatives de la population* ».

Voici les objectifs que s'est donné le projet MOHICAN en ce qui concerne la récolte d'informations sur les acquis des étudiants qui entrent à l'université (Leclercq & al., 2001) :

- rassembler des données fiables (dépassant les on-dit et les rumeurs) ;
- fournir concrètement et de façon précoce aux étudiants d'abord, aux encadrants ensuite, des informations à visée formative (utilisables en vue d'améliorer les chances de réussite).

2. Les 10 tests MOHICAN

Dans leur ouvrage intitulé « *Le premier des MOHICANs – Une recherche-action de Monitoring Historique des Candidatures* » (Leclercq & al. 2001), les auteurs du projet s'expliquent sur les caractéristiques des 10 épreuves MOHICAN dont les données nous ont permis d'expérimenter les nouveaux indices d'analyse de la qualité spectrale qui seront présentés plus loin.

Les auteurs commencent par préciser ce que ne sont pas les 10 tests (intitulés « *check up* ») MOHICAN :

[51] D. Leclercq (Président, Ulg), J.-F. Charlier (FUCAM), J.M. De Ketele (UCL), M. Delhaxhe (ULg), P. Dupont (UMH), J.P. Lambert (FUSL), J.P. Lambotte (FUSAGx), E. Loeckx (CIUF), R. Peters (Secrétaire scientifique, ULg), A. Pilatte (FPMs), T. Reggers (Secrétaire scientifique, ULg), M. Romainville (FUNDP), J.L. Wolfs (ULB).

« *…l'ensemble des épreuves CHECK UP MOHICAN*
- ***n'est pas*** *un relevé exhaustif des compétences jugées pertinentes à l'entrée de diverses sections universitaires. Ainsi, par exemple, la statistique, l'orthographe, la littérature, les langues étrangères, etc ; sont absentes ;*
- ***n'est pas*** *un bilan des acquis de l'enseignement secondaire. Celui-ci a ses objectifs propres qui viennent d'ailleurs d'être spécifiés sous formes de « compétences terminales ». Ces spécifications n'étaient pas disponibles au moment de la conception des check up MOHICAN ;*
- ***n'est pas*** *une série d'épreuves parfaitement ciblées pour des sections particulières. Ainsi, ces épreuves sont les mêmes pour tous : l'épreuve de mathématique est la même pour les ingénieurs et pour les autres sections ; les épreuves de français sont les mêmes pour les romanistes et autres étudiants ; l'épreuve de vocabulaire n'est pas 'orientée' vers chaque discipline* ».

Ensuite l'équipe des 10 chercheurs[52] impliquée dans la réalisation des tests nous explique ce que sont réellement ces check up :

« *L es épreuves CHECK UP MOHICAN*
- ***sont*** *dix épreuves séparées, conçues par dix auteurs (ou groupes d'auteurs) différents, issus des 9 universités de la Communauté, sur la base de leurs observations (dans leurs institutions) sur des difficultés rencontrées par leurs étudiants de 1ère candidature. Souvent, ces observations sont étayées par des résultats à des épreuves ou des enquêtes auprès des enseignants universitaires ;*
- ***sont*** *des 'coups de sonde' dans un univers de connaissance plus large. C'est tout spécialement le cas du vocabulaire (seulement 45 questions !) ou des mathématiques (seulement 22 questions) ; sans parler des connaissances artistiques, repères historiques, socio-économique, de l'actualité. »*

Nous présentons en annexe (p. 482) les 10 check up '99 MOHICAN qui ont été soumis à une série de cohortes d'étudiants entrant en 1ère candidature lors de l'année académique 1999-2000.

Rappelons que les questions ont été créées par des experts provenant des institutions universitaires de la Communauté française de Belgique durant l'année académique 1998-1999. Le format des items était de type Questions à Choix Multiple (QCM). Pour 5 épreuves (voir tableau ci-dessous) il s'agissait de QCM-SG, c'est-à-dire de questions où deux Solutions Générales (SG) figuraient en plus des solutions habituellement proposées : « Aucune » (aucune des propositions est correcte) ou « Toutes » (toutes les solutions sont correctes). Voici les intitulés des épreuves, le nombre de questions (NQ), leurs auteurs et les étudiants auxquelles elles étaient destinées :

EPREUVE	NQ	AUTEUR(S)	Gr. ETUDIANTS
Vocabulaire	45	M. Monballin	Toutes le facultés
Syntaxe et articulation logique	12	J.-M. Defays	Toutes le facultés
Compréhension de texte	6	P. Hougardy	Toutes le facultés
Lecture de carte, graphiques, tableaux en géographie	10	F. Orban	Toutes le facultés
Mathématique	22	M. Lebrun & J. Lega	Sciences, médecine, ingénieurs
Physique	10	P. Chapelle	Sciences, médecine, ingénieurs
Chimie	8	A. Cornelis	Sciences, médecine, ingénieurs
Biologie	10	J.-C. Verhaeghe	Sciences, médecine, ingénieurs
Connaissance artistique	25	B. Noël & D. Leclercq	Sciences humaines
Repères historiques, économiques, actualité	25	D. Leclercq & F. Georges	Sciences humaines

[52] D., Leclercq, C. Conti, J.-M. De Ketele, M. Delhaxhe, P. Dupont, J.-P. Lambert, J.-P. Lambotte, B. Noël, M. Romainville et J.-L. Wolfs.

3. Administration des épreuves, traitement des données et feedbacks

Les check up ont été administrés entre le 4 et le 8 octobre 1999, sur une durée de trois heures, via des formuloms[53], dans 8 des 9 universités ou facultés universitaires de la Communauté française[54].

En octobre 1999 les données MOHICAN ont été lues et traitées par le SMART (voir description ci-avant, p. 57) en vue de permettre un envoi par le Service de Technologie de l'Education (STE) d'une enveloppe confidentielle vers chacun des quelque 4.000 étudiants testés. Les enveloppes contenaient une feuille de feedback par épreuve, donc 8 feuilles pour les étudiants de sciences, médecine, ingénieurs et 6 feuilles pour les étudiants de sciences humaines.

Voici un exemple de « feedback étudiant » conçu par le groupe de travail du STE pour l'épreuve « Arts, actualités, société » (Georges & al., 2001).

Ces informations destinées aux étudiants étaient :

- <u>personnalisées</u> (décrivant le plus diagnostiquement leur situation) ;
- <u>détaillées</u> car les feedbacks contenaient chacune de leurs réponses (confrontable aux réponses correctes et aux critères de correction rendus avec leurs questionnaires), leur graphique de réalisme et leur tableau spectral pour chaque test ;
- <u>normées</u> (leur permettant de se situer par rapport à la moyenne des étudiants de leur section) ;
- <u>précoces</u> (diffusées en octobre, à un moment permettant la réaction, c'est-à-dire des modifications, des réactions) ;
- <u>confidentielles</u> (aucune valeur individuelle nominative n'a été communiquée aux encadrants).

1	1	B	6	26	10	1999	1	25	4124
UCL	Ph & L	Class.	Nbre Etud.	Jour	Mois	Année	Version	Nbre Quest	Votre code

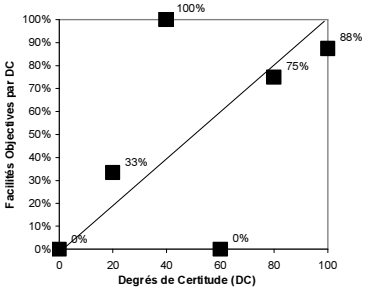
Diagnostic d'Auto-évaluation Cognitive (1) en

Arts, actualités, société

A. DESCRIPTIF

Cette épreuve comportait	25 Questions (NQ)
Vous avez donné	23 Réponses (NR)
Vous en avez réussi au total	16 Réponses Correctes (NRC)
Votre % de réponses correctes	**64** % ou Facilité Objective (FO)
Avec une Certitude Moyenne(2) de	71 % (CM)
Vous vous êtes surestimé de	**7** % d'Erreur Globale de Cert.(3) (EGC)
(2) sur TOUTES les réponses	(3) EGC=CM-FO

B. GRAPHIQUE DE REALISME



C. DETAILS DU CALCUL DE VOTRE REALISME

	NC		NU		DC		FOC		EC	NU*EC
Vous avez réussi	**0**	fois sur	**1**	la certitude	**0%**	soit à	**0,0%**	de réussite, donc avec	**0,0%** d'EC (total =	0)
Vous avez réussi	**1**	fois sur	**3**	la certitude	**20%**	soit à	**33,3%**	de réussite, donc avec	**-13,3%** d'EC (total =	40)
Vous avez réussi	**2**	fois sur	**2**	la certitude	**40%**	soit à	**100,0%**	de réussite, donc avec	**-60,0%** d'EC (total =	120)
Vous avez réussi	**0**	fois sur	**1**	la certitude	**60%**	soit à	**0,0%**	de réussite, donc avec	**60,0%** d'EC (total =	60)
Vous avez réussi	**6**	fois sur	**8**	la certitude	**80%**	soit à	**75,0%**	de réussite, donc avec	**5,0%** d'EC (total =	40)
Vous avez réussi	**7**	fois sur	**8**	la certitude	**100%**	soit à	**87,5%**	de réussite, donc avec	**12,5%** d'EC (total =	100)

Votre Total (pondéré) d'Erreurs de Certitude (TEC) est (en valeur Absolue, sans compter le sens des erreurs) = 360

Votre Erreur Moyenne (Absolue) de Certitude est **16** =TEC/NR=(360,0/23)=EMAC

Votre **réalisme** (100-EMAC) est donc de **84** (100-15,7)

Voici (chaque nombre gras) la **position spectrale** (de -100 à +100) de chacune de vos performances aux 10 questions de l'épreuve en fonction de la certitude ET de l'exactitude.

RC = la réponse correcte attendue

R = votre Réponse

Facilité Objective (FO) en % (ou taux d'exactitude en %) du groupe de **6** étudiants pour chacune des questions

D. TABLEAU SPECTRAL

RC	R		Vos Réponses Incorrectes (RI) -100	-80	-60	-40	-20	-0	OM	Vos Réponses Correctes (RC) 0	20	40	60	80	100	FO	CONTENUS DES QUESTIONS
3	3	Q1												**80**		50	Architecture : Notre Dame de Paris
3	3	Q2												**80**		67	Architecture : cathédrale d'Orléans
6	2	Q3			**-60**											67	Architecture : Parthénon
6	6	Q4												**80**		50	Architecture : Golden Gate
2	2	Q5										**40**				100	Peinture/Sculpture : impressionnisme
3	3	Q6												**80**		67	Peinture/Sculpture : peintres belges xx siècle
6	3	Q7		**-80**												33	Peinture/Sculpture : aveugles
1	1	Q8												**80**		67	Peinture/Sculpture : Liepchitz
1	1	Q9													**100**	100	Cinéma : SIDA - Philadelphia
5	5	Q10													**100**	83	Cinéma : autisme - Rain Man
7	7	Q11													**100**	100	Cinéma : Spielberg
5	5	Q12													**100**	83	Cinéma : Shoah
7	1	Q13	**-100**													67	Cinéma : Stanley Kubrik
3	4	Q14						**0**								33	Cinéma : décimation
7	1	Q15					**-20**									0	Théâtre : Berthold Brecht
6	5	Q16		**-80**												33	Théâtre : Dr Knock (J. Romains)
2	2	Q17										**40**				33	Littérature : Napoléon II
3	1	Q18					**-20**									50	Littérature : Baudelaire
3	3	Q19									**20**					67	Littérature : Kant
5	0	Q20							**0**							50	Littérature : Aldous Huxley
2	2	Q21													**100**	100	Littérature : autosémitisme
5	5	Q22													**100**	100	Musique : Queen
6	6	Q23												**80**		67	Musique : Beethoven
1	0	Q24							**0**							0	Danse : Jorge Donn
6	6	Q25													**100**	67	Danse : Boléro

Vos rép. Incorrectes = 7 soit 36% — Vos rép. Correctes = 16 soit 64%

Celles du groupe = 39% — Celles du groupe = 61%

Cert. Moy des R Incorrectes = 51% — Cert. Moy des R Correctes = 80%

Celles du groupe = 28% — Celles du groupe = 72%

Erreur — Exactitude

ignorée — reconnue — peu assurée — assurée

Méconnaissance — Ignorance — Connaissance

totale — partielle — partielle — totale

Savoirs

interprétations possibles des positions spectrales des réponses (avec certitudes)

[53] Formulaires destinés à la lecture optique de marques.

[54] L'Université Libre de Bruxelles pratique ce type de test depuis 1994 et a préféré ne pas en soumettre deux à ses étudiants.

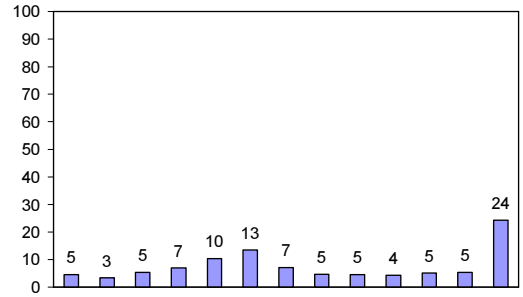
Des feedbacks furent également envoyés aux responsables des sections. En novembre 1999, MOHICAN a fourni à chacune des 51 sections concernées les résultats pour l'ensemble de ses étudiants. Voici un exemple de « feedback section » pour l'épreuve « Arts, actualités, société » (voir questions en annexe, p. 497) :

CIUF - MOHICAN - CHECK UP 1999

25 Nbre Quest — Nbre étudiants

Diagnostic d'Auto-évaluation Cognitive en

Arts, actualités, société



Graphique spectral de la distribution des réponses pour l'ensemble des 25 questions sur les étudiants.

Tableau Spectral des Réponses

RC			-100	-80	-60	-40	-20	-0	OM	0	20	40	60	80	100	CONTENUS DES QUESTIONS	FO	FS	FSC	FSI	DM	DN	MS	rpb	rpbS
3	100	Q1	3	3	6	9	11	9	4	6	11	12	11	8	9	Architecture : Notre Dame de Paris	56,8	43,3	51,4	-33	29,2	-13	16,3	0,31	0,24
3	100	Q2	2	3	6	8	11	10	5	8	12	11	11	7	7	Architecture : cathédrale d'Orléans	54,8	39,3	46,6	-30	25,6	-12	13,3	0,27	0,20
6	100	Q3	2	3	6	8	13	14	7	6	7	5	8	7	14	Architecture : Parthénon	47,8	41,8	59,2	-26	28,3	-12	16,5	0,51	0,42
6	100	Q4	10	7	11	12	12	8	5	3	1	3	4	5	19	Architecture : Golden Gate	35	56,2	76,3	-45	26,7	-27	-0,7	0,40	0,38
2	100	Q5	1	2	5	8	9	10	5	5	7	10	12	10	16	Peinture/Sculpture : impressionnisme	58,7	47,2	60,7	-28	35,6	-10	25,5	0,46	0,41
3	100	Q6	6	6	9	11	10	10	6	3	6	6	8	9	11	Peinture/Sculpture : peintres belges xx siècle	41,8	48,5	61,9	-39	25,9	-20	5,64	0,33	0,28
6	100	Q7	2	4	11	13	23	18	7	6	5	5	4	2	2	Peinture/Sculpture : aveugles	23,4	29,6	38	-27	8,92	-19	-9,9	0,30	0,13
1	100	Q8	4	6	9	8	8	9	6	6	8	9	10	9	8	Peinture/Sculpture : Liepchitz	49,4	45,7	53,1	-38	26,2	-17	9,25	0,32	0,14
1	100	Q9	1	1	1	1	1	2	1	2	1	2	2	7	80	Cinéma : SIDA - Philadelphia	93,6	89,4	93,6	-28	87,6	-1,4	86,2	0,36	0,53
5	100	Q10	2	1	1	1	3	4	3	2	1	2	2	6	71	Cinéma : autisme - Rain Man	85,2	82,5	91,8	-29	78,2	-3,4	74,8	0,41	0,60
7	100	Q11	13	4	3	2	1	2	4	2	2	2	5	12	50	Cinéma : Spielberg	73,1	81,9	87,6	-67	64	-15	48,5	0,49	0,16
5	100	Q12	1	1	1	1	1	2	3	2	1	2	3	8	75	Cinéma : Shoah	91,6	86,6	91,9	-28	84,2	-1,6	82,6	0,39	0,45
7	100	Q13	22	7	6	5	8	10	6	1	1	1	4	6	24	Cinéma : Stanley Kubrik	36,7	66,2	86,7	-54	31,9	-31	0,57	0,48	0,36
3	100	Q14	4	5	8	11	15	22	15	8	6	3	3	1	1	Cinéma : décimation	21,3	24,6	24,8	-25	5,28	-16	-10	0,09	-0,08
7	100	Q15	12	5	8	9	18	30	12	2	1	1	1	1	2	Théâtre : Berthold Brecht	6,22	31,1	54	-30	3,36	-24	-21	0,20	0,14
6	100	Q16	8	9	10	13	18	19	8	3	1	2	1	2	6	Théâtre : Dr Knock (J. Romains)	15,3	38,9	59	-35	9,02	-27	-18	0,29	0,21
2	100	Q17	2	2	6	9	19	25	11	6	7	5	4	2	3	Littérature : Napoléon II	26,8	24,1	38,6	-19	10,3	-12	-1,3	0,35	0,26
3	100	Q18	3	3	5	9	14	19	13	7	8	6	6	2	6	Littérature : Baudelaire	35,3	30	44,2	-22	15,6	-12	4,05	0,26	0,25
3	100	Q19	4	4	6	8	11	11	4	8	8	8	8	5	15	Littérature : Kant	51,8	44,9	55,8	-33	28,9	-15	14,3	0,34	0,27
5	100	Q20	1	1	2	4	10	21	13	8	6	3	4	4	24	Littérature : Aldous Huxley	48,2	38,4	65,2	-14	31,4	-5,3	26,1	0,39	0,57
2	100	Q21	0	1	1	2	4	9	6	5	4	3	5	7	54	Littérature : autosémitisme	77	67,4	83,7	-13	64,4	-2,2	62,2	0,46	0,65
5	100	Q22	1	0	1	1	1	2	3	2	1	1	4	9	73	Musique : Queen	91,2	85,8	91,5	-27	83,5	-1,6	81,8	0,38	0,50
6	100	Q23	7	6	10	11	14	15	7	4	3	3	4	4	13	Musique : Beethoven	30,2	45,2	66,7	-36	20,2	-23	-2,5	0,44	0,35
1	100	Q24	3	1	3	6	15	38	18	9	3	2	1	1	1	Danse : Jorge Donn	16,7	14,2	21,4	-13	3,57	-8,3	-4,8	0,16	0,08
6	100	Q25	2	2	3	6	11	20	9	5	4	3	5	6	24	Danse : Boléro	46,8	43,6	72,1	-19	33,8	-8,3	25,5	0,48	0,62

FO = Facilité Objective (% de Rép. Corr.) — FS = Facilité Subjective (Cert moy sur TOUTES les rép.) — FSC = M des certitudes des Rép. Correctes — FSI = Facilité Subjective des R Incorrectes — DM = Degré de Maîtrise (FSC*FO)/100 — DN = Degré de Non Maîtrise (FSI*DO)/100 — MS = Moyenne Spectrale (DM-DN) — rpb = Corrélation point bisériale Seuil=Rac(1/NQ) = 0,20 — rpbS = rpb Spectral (indépendant des autres questions)

Réponses Incorrectes (RI)	Omis	Réponses Correctes(RC)
44%	7%	49%
Cert. Moy. des R. I. = 30%		Cert. Moy. des R. C. = 63%

Erreur — **Exactitude**

ignorée — reconnue — peu assurée — assurée

Méconnaissance totale — partielle — Ignorance — partielle — **Connaissance** totale

Savoirs

dangereux — inutilisables — utilisables

4,56	3,35	5,45	7,02	10,4	13,5	7,15	4,69	4,6	4,3	5,21	5,48	24,3

interprétations possibles des positions spectrales des réponses (avec certitudes)

Ce feedback contient le graphique de la distribution spectrale des performances de la section à l'ensemble des questions (en haut à gauche). Dans l'exemple ci-dessus on constate que 24% des réponses ont été données correctement avec le pourcentage de certitude maximum (100%), que 5 % ont été données avec la 80% de certitude mais que 5% l'avaient été incorrectement avec la certitude 100%. On constate aussi 7% d'omission sur le total des réponses.

Le feedback pour la section contenait aussi le tableau spectral des réponses pour chaque question du test. Par exemple la 1ère question de ce test qui en comptait 25 était une question liée à l'architecture de la cathédrale Notre Dame de Paris. Dans le cadre de cette 1ère question du test, 9% des étudiants ont répondu correctement avec une certitude maximale (100%), 8% ont répondu correctement avec la certitude 80 (80%).

De gauche à droite, voici la signification des colonnes :

- colonne 1 : réponse correcte attendue (RC) ;
- colonne 2 : le total des taux de réponses répartis dans les colonnes 4 à 16 (en principe = 100%) ;

- colonne 3 : numéro de la question ;
- colonne 4 : taux de réponses incorrectes (RI) avec le degré de certitude 100 (codage « -100 ») ;
- colonnes 5 à 9 : suite des taux des RI (« -80 » à « -0 ») ;
- colonne 10 : taux d'omission (OM) ;
- colonnes 11 à 16 : taux des réponses correctes avec le degrés de certitude correspondants ;
- colonne 17 : synthèse du contenu de chaque question ;
- colonne 18 : Facilité Objective (FO) ou taux de réponses correctes ;
- colonne 19 : Facilité Subjective (FS) ou la certitude moyenne calculée sur toutes les réponses correctes et incorrectes mais pas sur les omissions ;
- colonne 20 : moyennes des certitudes des seules réponses correctes (FSC) ;
- colonne 21 : moyennes des certitudes des réponses incorrectes (sans le omissions) (FSI) ;
- colonne 22 : Degré de Maîtrise (DM) c'est-à-dire (FSC x FO)/100, donc le taux d'exactitude nuancé par la certitude moyenne des réponses correctes (DM est égal ou inférieur à FO) ;
- colonne 23 : Degré de Non maîtrise (DN), soit (FSI x DO)/100, DO = Difficulté Objective, soit 100 – FO ;
- colonne 24 : MS = Moyenne Spectrale, calculée par DM – DN. Elle ne tient pas compte des omissions ; MS est inférieur ou égal à DM ;
- colonne 25 : coefficient de corrélation point bisériale classique (*rpbis classique*) obtenu par la proposition correcte, la problématique de ce coefficient de discrimination est expliquée en détail plus loin (p. 171) ;
- colonne 26 : coefficient de corrélation point bisériale spectrale (*rpbis S*) obtenu par la proposition correcte, la problématique de ce coefficient d'analyse de la cohérence spectrale que nous avons mis au point (Gilles, 1999) sera expliquée en détail plus loin (voir p. 178).

Signalons ici brièvement (nous consacrons plus loin un chapitre entier à la problématique des *rpbis*) que les coefficients de corrélation point bisériale informent les enseignants sur la qualité des questions. Le *rpbis classique* est la corrélation entre les choix ou les rejets (0 ou 1) de la proposition correcte et les scores au total de l'épreuve (le nombre de réponses correctes). Nous verrons en détail plus loin que le *rpbis classique* nécessite le calcul d'une valeur repère pour contrer le problème du recouvrement entre le score de la question et le score total du test (voir p. 176). Le rpbis classique est un indice de discrimination qui permet d'évaluer dans quelle mesure les sujets qui ont choisi la proposition correcte (lorsqu'il est calculé sur celle-ci, mais on peut aussi le calculer pour les autres solutions incorrectes) récoltent, en moyenne, un nombre plus élevé de réponses correctes au total du test que les sujets qui se sont trompés à la question.

Le *rpbis spectral* est la corrélation entre les choix ou les rejets (0 ou 1) de la proposition correcte et les degrés de certitude qui ont accompagné ces choix/rejets, il s'agit d'un indice qui permet d'évaluer la cohérence spectrale, c'est-à-dire d'estimer dans quelle mesure les sujets qui ont choisi la proposition correcte ont accompagné celle-ci de degrés de certitude en moyenne plus élevés que les degrés de certitude utilisés par les sujets qui se sont trompés. Contrairement au *rpbis classique*, le *rpbis spectral* est calculé pour chaque question indépendamment des autres questions de l'épreuve. Avec le *rpbis spectral* il n'est donc pas nécessaire de calculer une valeur repère. Lorsque la question « fonctionne bien » du point de vue de la cohérence spectrale, on s'attend à ce que la corrélation soit positive et élevée. Nous quantifierons ces indices dans la suite de cette thèse et développerons les outils nécessaires à leur interprétation.

Chapitre II :

Introduction à l'analyse spectrale

Ce chapitre intitulé « Introduction à l'analyse spectrale » est tiré d'une communication cosignée par Dieudonné Leclercq et Jean-Luc Gilles et présentée par ce dernier au Colloque de l'Association pour le Développement des Méthodes d'Evaluation en Education (ADMEE) à Grenoble en septembre 1996. Le texte de la section B « Les techniques de recueil des certitudes » est publié sous la forme d'un article condensé dans l'ouvrage intitulé « L'activité évaluative réinterrogée » édité par Gérard Figari et Mohammed Achouche aux éditions De Boeck, Bruxelles (2001, pp. 134-146).

Sommaire

A. Les enjeux du recours aux pourcentages de certitude

1. L'incompétence est une situation normale de la vie

Les domaines dans lesquels chacun de nous est compétent sont bien moins nombreux que ceux où il est ignorant. L'apprentissage consiste à passer de certains niveaux d'incompétence à d'autres (moindres).

« *... j'ai formé une méthode, par laquelle il me semble que j'ai moyen d'augmenter par degrés ma connaissance, et de l'élever peu à peu ... j'en ai déjà recueilli de tels fruits qu'encore qu'au jugement que je fais de moi-même, je tâche toujours de pencher du côté de la défiance plutôt que vers celui de la présomption ... toutefois il se peut faire que je me trompe...* ». Descartes, 1636, Ed. 1952, p. 126-127, Nous soulignons)

2. L'ignorance (connue) n'est pas dangereuse

Ce n'est pas ce que nous ignorons qui nous cause des problèmes, mais ce que nous savons ... et qui est faux. Reconnaître (dans le sens "se rendre compte" et dans le sens "admettre publiquement") ses degrés d'incompétence est une habileté fondamentale, une compétence cruciale pour tout apprenant.

Sans faire du réalisme l'alpha et l'oméga de toute activité mentale (car son importance varie selon les domaines, les lieux, les circonstances, les individus), il faut poser le problème du statut intellectuel, affectif et surtout social du doute.

En fait, l'ignorance avouée ne devrait pas avoir de conséquence sociale négative. Vous ne connaissez pas la réponse à une question qu'un visiteur vous pose ? Il s'adresse à quelqu'un d'autre, ou vous prenez l'information là où elle se trouve. Mais les personnes qui manifestent "des prétentions de connaissance non fondées" (Ebel, 1968) constituent des dangers pour elles-mêmes et pour leur entourage.

Bertrand Russel le rappelait en termes humoristiques : *"L'ennui, dans ce monde, est que les fous sont sûrs d'eux-mêmes et les sages pleins de doute".*

Descartes (1636, Ed. 1952, p. 146) le disait déjà autrement : *"... en confessant plus ingénument ce que j'ignorais que n'ont coutume de faire ceux qui ont un peu étudié, et peut-être aussi en faisant voir les raisons que j'avais de douter de beaucoup de choses que les autres estiment certaines ...".*

3. Pourquoi cacher l'ignorance ?

Habituellement, il est honteux de ne pas savoir et bien des personnes (des institutions et des états) feraient n'importe quoi pour cacher leur ignorance. C'est là véhiculer des VALEURS FRELATEES. Il est bien plus précieux que chacun reconnaisse ses lacunes, surtout quand le sort des autres en dépend. Que penserions-nous d'une infirmière qui doute du contenu de sa seringue, mais qui, plutôt que de l'avouer en demandant confirmation à sa collègue, ferait l'injection ? Et du pilote d'avion qui décollerait sans être certain que toutes les vérifications ont été faites ?

Cette tendance à dissimuler l'ignorance n'est pas nouvelle; elle était déjà dénoncée par Descartes (1628, règle 2, Ed. 1952, p. 39) « *...ayant cru qu'il est indigne d'un homme docte d'avouer qu'il ignore quelque chose, ils se sont habitués à embellir leurs fausses raisons, ... et ... les ont ... données pour vraies ».*

Dans le même document, plus loin (p. 131) il montre l'ampleur sociale du problème : « *... pour les mauvaises doctrines, ... je pensais n'être plus sujet à être trompé ni par les promesses d'un alchimiste, ni par les prédictions d'un astrologue, ni par les impostures d'un magicien, ni par les artifices ou la vanterie d'aucun de ceux qui font profession de savoir plus qu'ils ne savent ».*

Tout formateur doit apprendre à savoir dire SIMPLEMENT devant les apprenants "je ne sais pas", afin que, eux aussi, osent avouer leurs doutes.

4. Le doute, c'est le moteur même de la connaissance

C'est parce qu'il prend conscience de son ignorance ou de son incertitude que l'apprenant cherche l'information, dans les livres, auprès de collègues, ou en "interrogeant" le milieu physique qui l'entoure.

On sait le rôle joué par les "ruptures d'équilibre" dans le développement intellectuel tel que l'a décrit Piaget (1956). Pendant les "paliers d'équilibre", le sujet ne remet rien en cause, est à l'aise dans ses (fausses) certitudes ... jusqu'à ce que se produise un bénéfique déséquilibre; source de "rééquilibration majorante" (c'est-à-dire à un niveau plus élevé).

L'aspect bénéfique de ces "réorganisations de la connaissance" était déjà signalé par Descartes (1636, Ed. 1952, p. 145) : « *... faisant ... réflexion, en chaque matière, sur ce qui la pouvait rendre suspecte et nous donner l'occasion de nous méprendre, je déracinais ... de mon esprit toutes les erreurs qui s'y étaient glissées auparavant [...]. Et, comme en abattant un vieux logis, on en réserve ordinairement les démolitions pour servir à en bâtir un nouveau, ainsi, en détruisant celles de mes opinions que je jugeais être mal fondées , je faisais diverses observations et acquérais plusieurs expériences, qui m'ont servi depuis à en établir de plus certaines.* »

5. Le sommet de la pyramide des objectifs cognitifs

Nous sommes ici au niveau le plus élevé de la taxonomie des objectifs cognitifs de Bloom *et al* (1956) : l'évaluation. Très peu de méthodes ont été consacrées à mesurer cette capacité. Elle est même souvent "évacuée" du domaine cognitif ... mais on ne l'étudie pas plus du côté affectif. Ce sont les économistes et les psychologues (et non les éducateurs), qui ont le plus développé les recherches sur cette capacité, si liée à la prise de décision. Or la décision est omniprésente dans les processus éducationnels.

Apprendre à apprendre, n'est-ce pas, d'abord se rendre compte de ce que l'on sait et de ce que l'on ne sait pas, ensuite mettre en œuvre les moyens qui paraissent les plus appropriés pour acquérir le savoir et enfin se faire une idée nette (la plus objective possible) du résultat atteint par l'opération ?

Prendre des initiatives en situation d'incertitude comporte, entre autres, une évaluation (forcément subjective) des chances de succès et d'échec des projets envisagés. Cela implique des choix permanents entre la réserve et l'audace, entre une attitude réceptive et une attitude "émettrice".

Se comporter de façon autonome suppose que l'individu s'en remette à son propre jugement, ait donc confiance en lui-même, se connaisse assez pour donner aux jugements des autres sur lui-même l'importance qu'ils méritent, n'être ni hermétique ni vulnérable. En retour, l'avis d'autrui peut être influencé positivement par la "sagesse" de l'auto-jugement de la personne jugée.

6. Vers un modèle épistémologique non manichéen

Quittons ces terrains affectifs, sociaux et moraux pour aborder l'épistémologie sous-jacente de la pratique scolaire. Bon nombre de personnes considèrent que la connaissance est affaire de tout ou rien : ou bien on "sait" les choses, ou bien on ne les "sait pas".

Il existe bien sûr des situations de connaissance parfaite (ou totale) : savoir que 2 + 2 font 4. Existent aussi des situations de connaissance nulle (quel est le second prénom d'Albert Einstein ?). Mais, le plus souvent, on est dans un état intermédiaire, de connaissance partielle.

Ainsi, si l'on demande à un adulte européen dans quel pays est la ville de QUITO, d'habitude il ne connaît pas la réponse exacte à cette question, mais il est capable de dire que ce n'est pas une ville de son pays, ni d'URSS, ni des USA, ... et il la placerait plutôt en Amérique du Sud. Cette approximation n'est pas mauvaise, puisqu'il s'agit de la capitale de l'Equateur. Cette connaissance partielle joue un grand rôle dans

les processus d'apprentissage; en effet, s'il a plusieurs cartes à sa disposition, notre adulte consultera d'abord celle d'Amérique du Sud.

Autre exemple : en quelle année est née la personne qui est en face de nous ? On peut ne pas "connaître" la réponse (précise) à cette question, mais être capable de bien des raisonnements à son sujet. Ainsi, si on connaît la durée de sa vie professionnelle, on peut l'ajouter aux années d'études (déduites à partir de sa profession). Comme il reste un doute, on lui demandera ingénument : "Vous avez débuté directement dans le métier ?" Enfin, on envisagera la possibilité que certaines années d'études aient été répétées.

On le voit, l'individu en situation de connaissance partielle est capable d'envisager une liste des solutions possibles, et d'éliminer certaines d'entre elles sur la base d'autres connaissances combinées à des raisonnements, bref par résolution de problèmes.

7. S'auto-évaluer s'apprend par l'expérience personnelle

Nous proposerons ci-après une procédure générale propre à poursuivre cet objectif dans l'enseignement. Il s'agit de mettre l'élève concrètement en face des *conséquences* (positives et négatives) *de ses actes*, afin qu'il apprenne, non pas par des règles enseignées, mais par des contingences de renforcement vécues selon le paradigme du conditionnement opérant décrit par B.F. Skinner, (1963). On est frappé de la clairvoyance de textes de Descartes (1636, Ed. 1952, p. 131) qui, avec le recul, ont une résonance "skinérienne" : « *Et me résolvant de ne chercher plus d'autre science que celle qui se pourrait trouver en moi-même, ou bien dans le grand livre du monde, j'employai le reste de ma jeunesse à voyager ... à recueillir diverses expériences, à m'éprouver moi-même dans les rencontres que la fortune me proposait ... Car il me semblait que je pourrais rencontrer beaucoup plus de vérité dans les raisonnements que chacun fait touchant les affaires qui lui importent, et dont l'événement le doit punir bientôt après s'il a mal jugé, que dans ceux que fait un homme de lettres dans son cabinet, touchant des spéculations qui ne produisent aucun effet, et qui ne lui sont d'aucune conséquence ...* (Nous soulignons). »

8. Des outils trop grossiers pour une matière subtile

Les chercheurs en éducation peuvent être comparés à des chimistes qui travailleraient avec une pelle à charbon. Or, on sait que la mesure précise des poids et des masses a amené des progrès décisifs en chimie. Si nous voulons comprendre un peu mieux les mécanismes d'apprentissage et de traitement de l'information, il est urgent que nous travaillions avec un plus grand nombre de nuances et de degrés de précision que les chercheurs classiques et que nous élaborions les loupes, les balances, les microscopes, … de la pédagogie. Les probabilités subjectives, ou les degrés de certitude qui en sont une forme dérivée, peuvent être l'un de ces outils. Bien entendu, se pose le problème de la validité, de la fidélité, de la sensibilité (ou acuité) des mesures permises avec ces nouveaux instruments. Chacun de ces points a reçu un examen approfondi ailleurs (Leclercq, 1993).

a) Vers une édumétrie instrumentée

Depuis des décennies, l'évaluation pédagogique a emprunté à la psychologie l'arsenal de ses méthodes, de ses instruments et des techniques statistiques. Même des apports aussi récents et importants que la théorie de la généralisabilité (Cronbach & al.,1963; Cardinet & Tourneur, 1975) ou que le modèle de Rasch (Rasch, 1966; Wright & Stone, 1979) ont d'abord été formulés dans un contexte psychométrique. Globalement, ces apports ont incontestablement été bénéfiques à l'éducation et continueront vraisemblablement à l'être. Sur certains points, cependant, ils présentent des dangers que ne manquent pas de signaler des pédagogues vigilants; par exemple, De Landsheere (1976) dénonce "le dangereux mythe de la courbe de Gauss" lorsqu'elle est appliquée sans discernement.

Il est grand temps qu'à côté de la psychométrie se constitue ce que Carver (1974) a appelé l'édumétrie, c'est-à-dire un ensemble d'approches, de concepts, d'instruments et de techniques spécifiques

aux problèmes pédagogiques. Nous avons un grand besoin de disposer d'instruments d'observation adaptés à l'objet même de l'évaluation des apprentissages : les modifications intra-individuelles.

Par exemple, au lieu de demander la date d'un événement, on peut en demander une estimation au moyen d'une fourchette (Hardy, 1981). Avant d'exiger le résultat exact d'une opération mathématique, on en demandera une (bonne) approximation (D'Hainaut, 1975). Plutôt que de dire aux étudiants : "Répondez uniquement si vous savez, et omettez si vous ignorez", ce qui est une consigne ambiguë (De Finetti, 1965), on leur permettra d'exprimer leurs doutes et leurs certitudes (Shufford & al., 1966; Adams & Adams, 1961). Au lieu de se limiter à demander une réponse à une question, on favorisera la production éventuelle de plusieurs réponses hiérarchisées selon leur vraisemblance (Leclercq, 1975; Bruno, 1990).

b) En finir avec la « *correction for guessing* » (pour divination)

Depuis trois quarts de siècle, les chercheurs en éducation ont essayé de "rendre équitable" la notation des étudiants à partir d'examens constitués de questions à choix multiple (QCM). Si l'on désigne par k le nombre de solutions proposées (dans notre exemple $k = 4$), par p la probabilité de désigner la solution correcte (supposée ici unique), par H le fait "donner une réponse au Hasard" et par | l'expression "étant donné que", on a : $p \mid H = 1/k$ (ici = 0,25)

Si l'on a prévu un tarif en cas de réponse correcte (*TC*) de + 1 point et un tarif en cas de réponse incorrecte (*TI*) de 0 point, alors l'étudiant complètement ignorant a intérêt à répondre au hasard, ce qui entraîne un score aux alentours de 25 points sur 100, si toutes les QCM ont $k = 4$. Ce "gonflement" du score (qui aurait dû être 0/100), dû à ce que certains appellent la probabilité automatique, rend les scores à des QCM incomparables aux scores obtenus à des questions ouvertes. On connaît la vieille parade à ce problème : la "correction pour divination" (*correction for guessing*) qui consiste à retirer des points en cas de réponse incorrecte, bref à fixer un tarif *TI* négatif.

Depuis Henmon (1911), on sait (voir Leclercq, 1986) que si l'on fixe le tarif en cas de réponse correcte (*TC*) à +1, et le tarif en cas de réponse incorrecte (*TI*) à -1/(k-1), le score d'une personne qui répondrait partout au hasard serait en moyenne zéro. C'est la *correction for guessing* classique qui a pour objectif de rendre nulle l'espérance mathématique. Il en découle qu'un apprenant qui ne sait rien a autant intérêt à omettre qu'à répondre au hasard.

Leclercq (1987) a montré que la *correction for guessing* est d'autant plus un instrument à bannir, qu'il peut être remplacé par les probabilités subjectives qui, elles,

- sont basées sur un modèle théorique plus pertinent;
- sont un principe de notation plus équitable;
- sont plus formatives (pour l'apprenant);
- sont plus informatives (pour l'apprenant et l'enseignant).

c) Une récente confusion scientifique à surmonter

Une large partie de la littérature (à très forte majorité anglo-saxonne) concernant les degrés de certitude (*confidence marking*) a été consacrée à trouver un système de notation des questions à choix multiple qui résolve de façon élégante et scientifique le vieux problème de la "correction pour divination".

La définition d'Ebel (1968) est très révélatrice de cette limitation de l'intérêt pour les degrés de certitude à ce seul problème technique : « *La pondération par degré de certitude (confidence weighting) est une façon spéciale de répondre aux questions d'un test [...], et une façon spéciale d'attribuer des points à ces réponses. En bref, on demande à l'étudiant d'indiquer non seulement quelle est la solution [...] qu'il croit correcte, mais en plus dans quelle mesure il est certain de l'exactitude de cette solution. Il recevra plus de points pour une réponse correcte donnée avec un degré de certitude élevé qu'avec un degré de certitude faible. Mais la pénalisation (points négatifs) pour une réponse incorrecte donnée avec un degré de certitude fort devra être suffisamment lourde pour décourager les prétentions de certitude non fondées* ».

Ce principe de notation a, en fait, une assez longue histoire (voir Henmon, 1911; Hollingworth, 1913; Trow, 1923; Hevner, 1932; Ahlgren, 1967, et Jacobs, 1968). Les recherches sur le sujet ont culminé, aux Etats-Unis, dans la période 1960-1970, puis ont été abandonnés. Quatre raisons expliquent la période de « creux » de ces recherches.

Raison 1 : On a cru que l'enjeu *principal* était docimologique (une technique d'attribution de points), alors qu'il est moral, épistémologique, lié à la façon dont nous concevons la connaissance et son utilisation sociale.

Raison 2 : *Les modèles* sous-jacents sont mal connus ou mal acceptés. Nombreux sont ceux qui, par crainte (mal justifiée) du subjectivisme, hésitent à s'engager dans cette voie. Ils ignorent qu'il existe des techniques permettant d'étudier objectivement la subjectivité. Or, l'influence de la personnalité sur l'auto-estimation a été étudiée depuis longtemps (Hevner, 1932; Wiley et Trimble, 1936; Swineford, 1938) et continue de l'être (Jacobs, 1971). Plutôt que de se plaindre de "l'intrusion" de l'affectif dans le cognitif, le formateur et le chercheur devraient y voir une voie ENFIN OBJECTIVE d'approcher cette facette individuelle, et, par la disponibilité d'UN NOUVEAU TYPE DE DONNEES sur l'apprenant, de nouvelles possibilités d'interventions et de compréhension.

Raison 3 : Les utilisateurs sont peu conscients de la nécessité, pour recueillir des données valides, de s'en tenir à des *méthodes* spécifiques et rigoureuses. La plupart des données fournies à ce jour par la littérature mondiale sont de peu d'intérêt, car récoltées dans de mauvaises conditions méthodologiques. Leclercq (1993, p. 213) signale à ce propos : « *...faced to inappropriate scales of tariffs (point awards), students discover that some strategies consisting to bias their intimate estimation of confidence pay more than telling the truth. In order to reinforce students to tell the truth, to make it optimal behaviour, tariffs must be computed according to decision theory (Von Neumann & Morgenstern, 1947). It must be noted that very few researchers have followed these methodological conditions, which has resulted in inconsistent data and finally ruined a fruitful approach which was almost abandoned during the late 1970s. Fortunately, a few scholars have continued to promote strict methodological requirements, what De Finetti (1956) calls 'Methods for discriminating levels of partial knowledge…' what Shufford & al. (1966 call 'Admissible probability measurement procedures', and what Van Naerssen (1962) calls 'A scale for the measurement of subjective probabilities'. To these pioneers must be added Darwin Hunt and James Bruno, the systems of whom are described in Leclercq and Bruno (1993)* ».

Raison 4 : Il existe une grande confusion quant aux *interprétations* à donner aux nouvelles informations recueillies. De très nombreuses recherches, notamment des traitements statistiques concernant la validité et la fidélité des nouvelles "mesures" obtenues, sont sans objet, car une grave équivoque pèse sur la notion même de "mesure". Leclercq (1993, p. 215) signale à ce propos : « *Numerous researches, most of which were published in the journal of Educational Measurement, raise the question : 'Are new (total) test scores (computed with new scales of tariffs taking confidence degrees into account) more valid and more reliable than classical ones (number of correct answers) ?' Results from these experimental results are confusing. Half of the studies show an increase in validity and a decrease in reliability whereas other studies find the contrary… Without being able to explain why. Actually, the problem itself is incorrectly stated since the new total score is not a measure, but the combination of two different measures : (1) the measure of the ability (number of correct answers) and (2) the measure of realism (quality of assessment). The new score can be more valid (i.e. reflect more accurately the learner's competency) only if the person is realistic!* ».

B. Les techniques de recueil des certitudes

Depuis des décennies, l'évaluation pédagogique a emprunté à la psychologie l'arsenal de ses méthodes, de ses instruments et des techniques statistiques. Même des apports aussi récents et importants que la théorie de la généralisabilité (Cronbach *& al.*,1972; Cardinet & *al.*, 1975) ou que le modèle de Rasch (Rasch, 1960; Wright & Stone, 1979) ont d'abord été formulés dans un contexte psychométrique. Globalement, ces apports ont incontestablement été bénéfiques à l'éducation et continueront à l'être. Sur certains points, cependant, ils présentent des dangers que ne manquent pas de signaler des pédagogues vigilants; par exemple, De Landsheere (1976) dénonce "*le dangereux mythe de la courbe de GAUSS*".

Il est grand temps qu'à côté de la psychométrie se constitue ce que Carver (1974) a appelé l'édumétrie, un ensemble d'approches, de concepts, d'instruments et de techniques spécifiques aux problèmes pédagogiques. Nous avons un grand besoin de disposer d'instruments d'observation adaptés à l'objet même de l'évaluation des apprentissages : les modifications intra-individuelles.

1. Les consignes dichotomiques (C1)

Van Naerssen et Van Beaumont (1965) dactylographiaient une QCM où $k = 3$ comme suit :

	Quelle est la capitale de la France ?
Z 1	Lyon
Z 2	Paris
Z 3	Marseille

Rappelons que Z est la première lettre de *zekerheid* (certitude en néerlandais). Les apprenants peuvent choisir une solution (2 par exemple) soit en entourant seulement le chiffre 2 (pas sûr =Ⓒ) ou en encerclant Z *et* 2 (sûr =ⓏⒸ).

Van Naerssen et Van Beaumont (1965) utilisent le barème suivant (où TC = tarif en cas de Réponse Correcte, et où TI = Tarif en cas de Réponse Incorrecte)

	TC	TI
sans Z	+ 0,5	0
avec Z	+ 1	-0,5

Sandbergen (1971) a utilisé ce même barème en doublant tous les tarifs.

	TC	TI
sans Z	1	0
avec Z	2	-1

Les consignes dichotomiques ont (en pire) les défauts des consignes ordinales (voir ci-après). Nous analyserons les barèmes (ou ensembles) de tarifs à la lumière de la théorie des décisions. Dans cette théorie, en effet, l'étudiant est supposé choisir le comportement (le degré de certitude) « qui lui rapporte la plus grande espérance de points », ou « qui a l'utilité attendue la plus élevée ». Les mots « espérance » et « attendue » signifient que l'on tient compte des probabilités de gagner, tandis que les mots « points » et « utilité » font référence aux tarifs en cas de réponse correcte (TC) et en cas de réponse incorrecte (TI). Le Score Attendu (ou S.A.) à une question est calculé par la formule S.A. = (p . TC) + (q . TI) où p = la probabilité (subjective) de réussir et q celle de se tromper (q = 1 – p). Pour 4 solutions ($k = 4$) proposées et une réponse au hasard, selon la formule de Sandbergen :

sans Z, S.A. = (1/4 . 1) + (3/4 . 0) = 0,25

avec Z, S.A. = (1/4 . 2) + (3/4 . -1) = -0,25

On voit donc que le barème encourage à répondre au hasard, mais sans Z. On voit aussi qu'il faut à peine dépasser p = 1/3 pour que l'étudiant ait intérêt à répondre :

sans Z, S.A. = (1/3 . 1) + (2/3 . 0) = 0,33

avec Z, S.A. = (1/3 . 2) + (2/3 . -1) = 0

Donc, un étudiant qui connaît avec un peu plus de 33 % de certitude a mathématiquement intérêt à répondre. En fait, les expériences de Cross et Frary (1977) montrent qu'en pratique ils y ont encore plus intérêt que mathématiquement (voir Leclercq, 1987, 128-130).

2. Les consignes ordinales (C2)

Par exemple (avec une échelle de certitude à quatre niveaux) :	Si vous êtes : pas sûr du tout, donnez le degré de certitude 0 peu sûr, donnez le degré de certitude 1 sûr, donnez le degré de certitude 2 très sûr, donnez le degré de certitude 3

De telles consignes sont vagues. Ce qui est « très sûr » pour un apprenant peut être « peu sûr » pour un autre. Puisque l'expérimentateur ne connaît pas l'interprétation exacte que différents apprenants ont donné de "sûr", deux réponses accompagnées du même degré de certitude ne peuvent être comparées, qu'elles aient été données par deux apprenants différents ou par la même personne. En effet, rien n'empêche l'apprenant de modifier, pendant le test, son interprétation de l'expression "sûr", puisque celle-ci n'a pas été définie. Des contrastes de facilité entre des questions successives peuvent amener l'apprenant à réviser son interprétation des divers degrés de certitude quand il passe d'une question à l'autre. Par conséquent, deux réponses avec le même degré de certitude peuvent difficilement être comparées, *même si elles sont données par un même apprenant*. Malgré ces faiblesses, ce type de consigne a été abondamment utilisé, avec diverses matrices de tarifs.

Jacobs (1971) utilise le barème suivant (échelle à 3 niveaux):

	TC	TI
Je devine	+1	0
Assez sûr	+2	-2
Certain	+ 3	-3

Leclercq (1973) utilisait un barème fort proche : (échelle à 3 niveaux plus omission)

	TC	TI
Omission	0	
Peu sûr	+1	-1
Moyennement sûr	+2	-2
Très sûr	+3	-3

Ces barèmes de tarifs sont commodes : les tarifs sont faciles à retenir (nombres entiers correspondant aux codes des degrés), et le calcul des scores peut se faire manuellement. Néanmoins, ces barèmes doivent être abandonnés car ils ne respectent pas les critères élémentaires de la théorie des décisions.

3. Les consignes par zones REGULIERES d'une échelle d'intervalles (C3)

Pour échapper aux critiques qui viennent d'être adressées aux échelles d'intervalles telles que celles qui ont été envisagées plus haut et qui peuvent être interprétées différemment par les répondants, l'expression du degré de certitude devrait se faire en termes de probabilités. Par ailleurs, la sensibilité (ou acuité) d'un apprenant dans l'estimation de ses chances de succès n'est pas fine au point qu'il puisse distinguer entre 0,372 et 0,373, ni même entre 0,37 et 0,38. C'est pourquoi, on se contentera de réponses probabilistes assez globales, correspondant à des *zones sur l'axe des probabilités*. Une des consignes utilisée en Belgique à partir de 1971 (Leclercq, 1973) est la suivante (avec les tarifs utilisés alors) :

Si vous attribuez à votre réponse une chance d'être correcte,	alors choisissez le degré de certitude :	**Barème A**		**Barème B**	
		TC	TI	TC	TI
... compris entre 0 et 25 %	0	0	0	0	0
... compris entre 25 et 50 %	1	+1	-1	+3	-1
... compris entre 50 et 75 %	2	+2	-2	+4	-2
... compris entre 75 et 100 %	3	+3	-3	+5	-5

Le barème A n'est pas conforme à la théorie des décisions (les étudiants n'ont pas intérêt à dire la vérité), c'est pourquoi, à partir de 1972, il a été remplacé par le barème B. Shufford, Albert et Massengil (1966) proposent d'utiliser les 10 digits (0, 1, 2, 3, 4, ... 9) pour désigner 10 intervalles (0 à 10 %, 10 à 20 %, 20 à 30 %, etc.) réguliers. Les points gagnés (0,1 ou 0,2 ou 0,3 ... jusque 1) et perdus (-0,1 ou -0,2 ou -0,3 ... jusque -1) sont faciles à calculer.

4. Les consignes par étoiles (C4)

De Finetti (1965) décrit un "système à cinq étoiles" où l'apprenant doit répartir cinq étoiles (ou astérisques) sur les diverses solutions proposées d'une QCM, chaque étoile équivalant à une probabilité de 0,20. Il existe seulement sept façons de distribuer cinq étoiles :

1	2	3	4	5	6	7
5	4	3	2	1	3	2
-	1	1	1	1	2	2
-	-	1	1	1	-	1
-	-	-	1	1	-	-
-	-	-	-	1	-	-

Michael (1968) a utilisé un "système à dix étoiles" où le score à la question vaut le nombre d'étoiles attribuées à la solution correcte, ce nombre étant divisé par 10. De ses données expérimentales, Michael conclut qu'un test noté selon des procédures classiques devrait être 1,7 fois plus long pour atteindre la même fidélité dans les mesures qu'avec la procédure des dix étoiles.

5. Les consignes par rapports (C5)

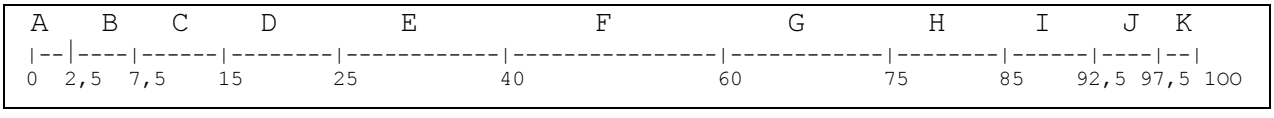
Edwards (1967), entre autres, a recommandé d'utiliser une échelle essentiellement logarithmique aux extrémités pour définir les degrés de certitude.

Voici un exemple d'une telle échelle :

A =	1 chance sur 1000
B =	1 chance sur 100
C =	1 chance sur 10
D =	1 chance sur 4
E =	1 chance sur 2
F =	3 chances sur 4
G =	9 chances sur 10
H =	99 chances sur 100
I =	999 chances sur 1000

Cette échelle à neuf degrés correspond assez bien aux propriétés souvent approximativement logarithmiques de la perception humaine. Il est probable que le "nombre magique 7" de Miller (1956) s'applique aussi aux probabilités subjectives, comme il s'applique à la vision, au toucher, à l'audition, au goût, etc. L'échelle d'Edwards paraît fondée, car nous savons que des différences aux extrémités de l'échelle sont plus perceptibles qu'au centre. Ainsi, un étudiant moyen peut faire la distinction entre 1 chance sur 100 et 10 chances sur 100 (9 % de différences) alors qu'il aura bien plus de difficultés à faire la différence entre 33 % et 42 % (la même différence de 9 %, mais au milieu de l'échelle des probabilités). Gardant ceci à l'esprit, on peut définir des zones de certitudes inégales, mais symétriques).

En voici un exemple :



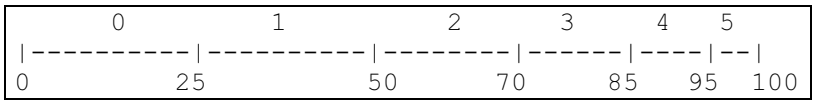
(Zones A et K = 2,5% - Zones D et H = 1O% - Zones B et J = 5% - Zones E et G = 15% - Zones C et I = 7,5% - Zone F = 2O%)

6. Les consignes par zones irrégulières et asymétriques (C6)

Dans les situations d'évaluation pédagogique, les zones de la portion droite de l'échelle des probabilités sont beaucoup plus utilisées que les zones de gauche, pour deux raisons :

a) Les tests sont le plus souvent utilisés comme post-tests plutôt que comme pré tests, la situation classique de mesure, en effet, consiste à vérifier dans quelle proportion les apprenants ont maîtrisé des matières qui leur ont été enseignées.
b) Il est fréquent que les étudiants aient, sur une matière, une réponse correcte, mais avec des doutes.

Il est donc raisonnable de concevoir une échelle dont la forme générale serait la suivante :



C'est ce type d'échelle que nous adopterons. Dans le cas d'un VRAI-FAUX Généralisé (VFG), chaque question étant binaire (vrai faux), les deux degrés inférieurs à 50 % sont, en pratique, inutiles. Même dans ce cas extrême, il reste à l'étudiant le choix entre 4 degrés de certitude !

7. Les consignes par échelles continues (C7)

Cette procédure, recommandée par De Finetti (1965) sous le nom *de "Continuous Confidence Marking"* autorise l'apprenant à exprimer son degré de certitude avec la précision qu'il veut : il peut aussi bien répondre par "30 %" que par "31,27 %". Ce mode de réponse est intéressant quand la continuité est nécessaire ... et possible (par exemple, il faut que l'étudiant n'attribue pas plus de 100 % à l'ensemble des solutions proposées si on lui garantit qu'une seule d'entre elles est correcte).

Baker (1969) a développé une technique qui garantit ce principe. Il présente sur l'écran vidéo d'un ordinateur, les quatre histogrammes correspondant aux probabilités (exprimées en pourcentages) attribuées à chacune des quatre solutions proposées d'une QCM. Quatre boutons permettent à l'apprenant de changer chaque valeur (la modification est immédiatement affichée sur l'écran). Les trois autres valeurs sont automatiquement adaptées (par une règle proportionnelle) de sorte que la somme des quatre valeurs soit toujours égale à 100. Dans cet exemple, les changements n'ont pas besoin d'être effectués pas à pas, mais peuvent être continus.

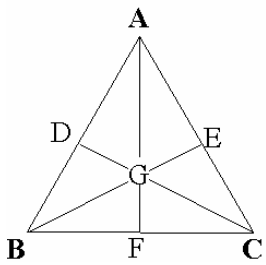
Du point de vue de la notation (l'attribution des points), on n'utilise plus, pour les consignes par échelle continue, une matrice de tarifs, mais une fonction mathématique. C'est aussi ce que font Shufford (1993) et Dirkzwager (1993).

8. Les consignes par technique du contour triangulaire (C8)

De Finetti (1965) a proposé des consignes où l'apprenant doit désigner le seule solution correcte parmi trois solutions proposées (A, B et C).

L'apprenant doit se positionner :

- soit sur un des sommets (A, B ou C) d'un triangle équilatéral ;
- soit exactement entre A et B (réponse D) ;
- soit exactement entre A et C (réponse E) ;
- soit exactement entre B et C (réponse F) ;
- soit au centre du triangle (réponse G).



Evidemment, de telles réponses doivent être interprétées comme suit :
« D = j'élimine C et j'hésite également entre A et B (ou je réponds A à 50 %, B à 50 % et à C à 0 %). »
« G = j'hésite entre les trois solutions proposées. »

Bruno et Baxter (1989) utilisent le même principe que De Finetti, mais en offrant trois positions intermédiaires (G, H et I) entre A et B au lieu d'une seule. Ils offrent de même trois positions, entre A et C (D, E et F) et trois autres positions entre B et C (J, K et L), plus un point central (M), ce qui fait, au total, 13 possibilités (au lieu de 7 précédemment).

Une telle procédure offre des possibilités de nuances considérables, puisque :	G signifie : A à 75 % B à 25 % C à 0 %

On pourrait continuer à raffiner la procédure en donnant des lettres codes aux intersections joignant les lettres déjà évoquées, de manière à permettre l'expression de subtilités telles que A à 66 %, B à 33 %, C à 0 % ou même A à 75 %, B à 12,5 %, C à 12,5 % en traçant les médianes du triangle. Une telle approche ne fonctionne *que* pour trois solutions proposées.

Cependant on pourrait imaginer d'utiliser les possibilités offertes par le dessin 3D sur ordinateur pour proposer des consignes à quatre solutions (pyramide à base triangulaire).

9. Les consignes par fractiles (ou fourchettes) (C9)

Quand la réponse à une question porte une valeur numérique repérable sur un continuum (par exemple, un poids, une surface, un prix, une date, une durée, une vitesse, etc.), la procédure des fractiles est une manière élégante et formative de faire face au problème de la connaissance partielle. On demande à l'apprenant de fournir, pour chaque question, deux réponses (c'est-à-dire deux fractiles) : la limite inférieure et la limite supérieure d'un intervalle (ce qui constitue une "fourchette". Plus l'intervalle est étroit, plus l'apprenant est sûr de lui et plus le risque est grand que la solution correcte "déborde" de l'intervalle... donc plus grand doivent être le tarif (positif) en cas de succès et le tarif (négatif) en cas d'échec. Quand l'intervalle n'inclut pas la valeur correcte, on parle de "surprise". On observe souvent un grand nombre (imprévu) de surprises. Par exemple, alors qu'on demande aux apprenants de fournir un intervalle tel qu'il inclue la réponse correcte dans 8O % des cas, on observe souvent un taux de succès inférieur à 5O %. Ce phénomène a été appelé par Pitz (1974) *hyper précision dans les intervalles*. Des consignes plus résignées (à l'hyper précision spontanée) demandent à l'apprenant de donner un intervalle de telle sorte qu'un tiers des réponses correctes se situent en dessous, un autre au-dessus et le troisième tiers à l'intérieur. Cette procédure est appelée "tertiles" par Pitz (1974). Exemple de consigne :

> En quelle année a eu lieu la bataille de HASTINGS ?
> *Fournissez une date limite inférieure et une date limite supérieure, de telle sorte que la réponse correcte ait 1/3 de chance d'être dans votre intervalle, 1/3 au-dessus et 1/3 en dessous.*

Si l'étudiant répond « Entre 1000 et 1050 », il y a là une surprise, la date correcte est 1066.

Albert et Raiffa (1982) utilisent cinq fractiles. Exemple de consigne :

> En quelle année a eu lieu la bataille de HASTINGS ?
> *Fournissez 4 dates limites pour qu'il y ait 20 % de chances que la réponse correcte « tombe » dans l'une des 5 zones.*

Si l'étudiant répond « 950, 1000, 1050 et 1100 », la réponse correcte est comprise dans la zone 4 (de 1050 à 1100), moins bonne que la zone 3, mais meilleure que la 5 ! Ce problème a été bien traité par Murphy et Winkler (1974), Lichtenstein *et al.* (1977), Hardy (1981).

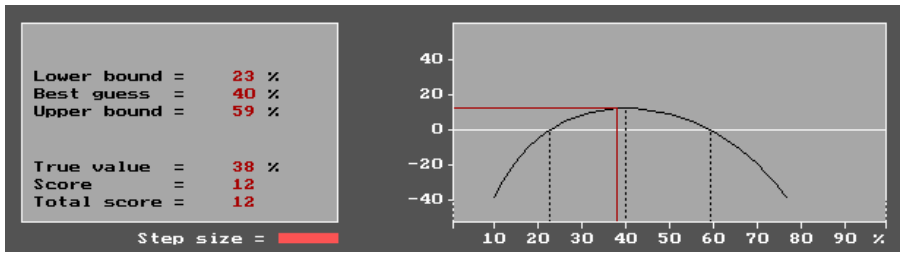
10. Les consignes par ajustement d'une distribution de probabilités (C10)

Jelle Van Lenthe (1993) permet de répondre à des questions appelant des réponses en pourcentages, par son logiciel ELI (les trois premières lettres du mot anglais ELICITATION).

« Quel est le pourcentage d'enfants de 11 ans qui passe plus de 2 heures par jour devant la TV en France, dans le sondage de mai 1966 ? »

Après avoir reçu la question (sur le haut de l'écran), l'étudiant voit s'afficher (sur le bas de l'écran) une distribution, centrée sur 50% et avec des « bords » éloignés de 15% (donc 35% et 65%).

L'étudiant peut (avec les touches → et ←) déplacer horizontalement cette courbe. La voici déplacée pour que son centre (la Moyenne) soit 40 %. Avec les touches ↑ et ↓, l'étudiant peut aussi la rendre plus pointue ou plus aplatie. Quand l'étudiant a confirmé sa réponse, la réponse correcte s'affiche (sous forme de barre verticale, ici à 38 %) ainsi que le score (sous forme de barre horizontale (ici 12 points).



La courbe représente ainsi les scores en ordonnée (Y) selon la position de la réponse correcte en abscisse (X) par rapport à cette courbe (cette situation est visualisée sur le graphique par les traits continus qui relient le point de la courbe à X et Y). On constate que le score est négatif quand la réponse correcte est en deçà de l'endroit où la partie gauche de la courbe croise le trait horizontal à zéro. Le score est également négatif quand la réponse correcte se situe au delà de l'intersection de la courbe avec l'axe horizontal à zéro.

C. Conclusions

Les différents états de connaissances partielles qui découlent de l'association d'un degré de certitude à une réponse autorisent un diagnostic plus subtil (et par là différents niveaux de remédiations) que lorsqu'on n'utilise pas de degré de certitude. Cette amélioration de la sensibilité de l'outil d'évaluation contribue également à une mesure plus fine de modifications intra-individuelles (par exemple entre un pré et un post-test).

1. Limite d'acceptabilité des consignes de recueil des degrés de certitude

Parmi les consignes énumérées ci-dessus, seules celles qui sont numérotées de C3 à C10 peuvent constituer ce que Shufford, Albert et Massengil (1966) appellent des "*procédures admissibles de mesure des probabilités (subjectives)*".

2. Choisir ce que l'on peut traiter

Le choix entre l'une d'elles dépend du but de la recherche ou de l'action pédagogique, des moyens disponibles, du degré de familiarité de l'apprenant et du formateur avec ce genre de pratique... Par exemple, il est inutile de permettre à l'apprenant de donner des degrés de certitude très détaillés (l'échelle continue des degrés de certitude) si le formateur n'a pas les moyens de les traiter et de donner une information appropriée en retour. Si un taux de certitude de 37,24 % est traité comme s'il valait 30 %, l'expression nuancée constitue une perte d'énergie. D'ailleurs, certains se sont interrogés sur la précision qu'il est possible d'atteindre des points de vue psychologiques, physiologiques et/ou pédagogiques pour de tels pourcentages (Miller, 1956).

3. Nécessaire progressivité

Il n'y a pas, à notre connaissance, de règles et principes d'auto-estimation de ses compétences qu'on puisse enseigner car cette performance varie avec les contenus, les enjeux, etc. Par contre l'apprentissage de cette habileté métacognitive se fait par l'ajustement de comportements d'auto-estimation après avoir été confronté aux conséquences de nos jugements. Les jeunes enfants ou les étudiants non entraînés devraient recevoir des consignes comportant peu de degrés de certitude et une procédure de notation limpide (qui puisse être appliquée rapidement à la main par l'étudiant lui-même).

Si un expérimentateur veut mesurer de subtiles modifications dans les connaissances, il devra bien sûr, utiliser un plus grand nombre de degrés de certitude.

Dans ce contexte de familiarisation, des feedbacks métacognitifs (indices de réalisme) sont cruciaux pour permettre à l'apprenant d'améliorer l'auto-estimation de ses compétences. Reliés à des normes (Gilles, 1996a; 1996b; 1997; 1998b) ces feedbacks lui permettront de se situer par rapport aux performances d'une population cible.

Chapitre III :

Indices classiques d'analyse de la qualité des épreuves

Sommaire

A. Introduction

Parmi la panoplie d'indices au service de l'évaluateur qui désire se pencher sur la qualité des épreuves, on distingue généralement quatre types d'outils : les indices (1) de facilité, (2) de fidélité, (3) de discrimination et (4) de validité.

Les indices de facilité, de fidélité et de discrimination seront présentés et discutés dans ce chapitre. Quant aux indices de validité nous ne les aborderons pas ici car nous ne les situons pas dans le champs de notre recherche sur les outils d'analyse de la qualité spectrale des épreuves.

Nous aborderons d'abord la problématique des indices de facilité selon trois approches :
- l'approche théorique ou la facilité « vue » du côté des experts ;
- l'approche introspective ou la facilité des items prédite par les étudiants ;
- l'approche expérimentale ou la facilité déterminée a posteriori sur la base des réponses des étudiants.

Nous présenterons ensuite les indices de fidélité qui seront utilisés plus loin dans le cadre de l'exploration de la qualité des épreuves MOHICAN :
- le coefficient de bipartition (avec correction de Spearman-Brown) ;
- le coefficient de bipartition de Guttman ;
- le coefficient alpha de Cronbach (avec calcul du nombre de questions (parallèles) à ajouter ou à retrancher au test pour obtenir un niveau alpha donné, et avec le calcul de l'alpha en cas de suppression d'une question donnée) ;
- la corrélation question-total.

Enfin, en ce qui concerne la problématique de la discrimination, Laveault & Gregoire (1997, p.231), différencient deux types d'approches dans l'analyse du pouvoir discriminatif d'une question.

Lorsqu'il s'agit de vérifier que les items contribuent à départager les étudiants qui ont obtenu un score élevé au test (et donc, on l'espère, les meilleurs étudiants), des autres qui ont obtenus un score plus faible, un bon item est alors « un item qui serait réussi par une plus grande proportion de sujets ayant obtenu un score élevé à l'examen que par des sujets ayant obtenus un score faible ».

Un test n'a cependant pas toujours pour but de discriminer les étudiants entre eux, notamment dans un contexte de pédagogie de maîtrise où le but est d'arriver à une situation où un maximum d'étudiants (au minimum 80%) maîtrisent un objectif d'apprentissage avant de passer au suivant. Dans une telle perspective d'évaluation, dite de 'mesure critériée', il s'agira moins de vérifier si l'item discrimine les « forts » des « faibles » que de vérifier si la question permet de faire la différence entre ceux qui maîtrisent l'objectif et ceux qui ne le maîtrisent pas.

La méthode de discrimination utilisée sera donc différente en fonction de l'approche envisagée. Laveault & Gregoire (1997) répertorient trois principales catégories de méthodes :
- les indices de discrimination pour la mesure critériée ;
- les indices de discrimination D ;
- les indices corrélationnels de discrimination.

Le rpbis Spectral Contrasté (*rpbis SC*) que nous présenterons dans la partie suivante, est à notre avis un outil d'analyse des items à ranger dans la catégorie des indices corrélationnels de discrimination et, comme son appellation l'indique, il est basé sur le principe du coefficient point bisérial classique. Nous présentons donc plus en détail les indices corrélationnels de discrimination en vue de permettre de situer le nouvel *rpbis SC* parmi ces autres indicateurs du pouvoir discriminatif des questions.

B. L'indice de facilité des questions (p)

Leclercq (1987) distingue trois grandes approches de la facilité des questions : (1) l'approche théorique où la facilité des questions est soit déduite d'un modèle, soit établie par un groupe d'experts, (2) l'approche introspective où les étudiants sont amenés à exprimer les chances de succès qu'ils accordent à leurs réponses et (3) l'approche expérimentale où la facilité des questions est observée à partir des réponses fournies par le groupe testé.

1. L'approche théorique ou la facilité des items « vue » du côté des experts

a) L'intervention d'un groupe d'experts

L'idée est de faire appel à un groupe d'experts du contenu et de leur demander de se prononcer sur la facilité/difficulté des questions. Notons que le fait de se prononcer sur le niveau de difficulté des questions d'une épreuve implique qu'on ait une idée préalable du niveau de compétence des sujets qui seront testés, compte tenu du contenu et de la difficulté du cours.

Nous avons utilisé ce type d'approche « expertise » depuis l'année académique 1997-1998 dans le cadre de l'épreuve de Compréhension d'un Texte de Vulgarisation Scientifique (test CTVS) soumise aux étudiants de la Faculté de Médecine de l'Université de Liège (Gilles, 1998a ; Gilles & Melon, 2000). Le test CTVS a lieu chaque année (avec un texte différent et de nouvelles questions) depuis 1997-1998 et constitue une des trois épreuves d'« *évaluation générale de la maîtrise du français* » en sciences médicales. Les résultats obtenus à ces tests de maîtrise du français interviennent à raison de 15% dans la cote de classement de chaque étudiant (les 85% restant proviennent des résultats aux examens) en fin de 3ème candidature. C'est suite à des mesures gouvernementales visant à limiter le nombre de médecins pratiquants que la Faculté de Médecine devrait opérer une sélection des étudiants en fin de 3ème candidature (si les actions de justice en cours ne viennent pas modifier ces mesures).

Lors de l'élaboration annuelle du test CTVS, un groupe d'experts formés d'enseignants de la Faculté de Médecine ainsi que d'encadrants accompagnant les étudiants dans leurs démarches d'apprentissage (cellule « Guidance étude » de l'ULg) est contacté en vue de donner un avis sur la qualité des questions qui seront proposées et sur le niveau de difficulté de celles-ci. L'implication d'encadrants indirectement engagés dans l'enseignement par leurs contacts individualisés avec les étudiants au niveau d'une guidance dans les méthodes d'apprentissage, nous paraît cruciale. Elle apporte un éclairage qualitatif important teinté de l'avis des étudiants sur la qualité des enseignements qu'ils suivent, mais aussi sur l'adéquation de la difficulté des items auxquels ils doivent répondre dans leurs épreuves.

b) Modèle hiérarchisé des difficultés

Leclercq (1987) propose une méthode de classement des items par ordre de facilité qui peut être réalisée à l'aide d'un groupe d'experts : *« ...on adopte alors pour indice de facilité de la question la moyenne des scores avancés par les divers experts. Par cette méthode, des indices numériques peuvent être obtenus rapidement, mais ils ne sont qu'hypothétiques et devront être confrontés aux approches introspectives et expérimentales.* ». L'auteur signale également que dans ce type d'approche il convient de préciser le degré d'accord entre experts en mentionnant l'écart type de la distribution des jugements ou les pourcentages d'accord.

Un des avantages de l'approche théorique qu'elle soit de type « expertise sur la facilité » avec ou sans « modèle hiérarchique des difficultés » est que dans la chronologie des étapes de réalisation d'une épreuve, cette approche se situe en amont de la passation et s'intègre ainsi dans une évaluation *a priori* de la qualité des questions.

2. L'approche introspective ou la facilité des items ressentie par les étudiants

Il s'agit ici de faire appel aux impressions des étudiants en leur demandant dans quelle mesure chaque question leur paraît facile/difficile. Une telle approche peut se concrétiser par l'intermédiaire des pourcentages de certitude associés aux réponses. Leclercq (1987) précise : « *La moyenne des indices introspectifs fournis par une population d'étudiants pour la réponse correcte peut être utilisée comme indice de facilité d'une question. De tels indices subjectifs (caractère attractif ou attractivité) peuvent être confrontés aux indices expérimentaux (ou popularité) et, bien entendu, aux indices théoriques (ou complexité). L'idéal, évidemment étant de combiner les trois approches.* »

La possibilité nous est offerte dans le cadre des tests MOHICAN, de mesurer pour chaque question le pourcentage moyen des certitudes. De plus, à l'aide des turbo analyses, nous pouvons mesurer cette facilité introspective par question (*piq*) au départ des données des sujets les plus réalistes, donc en assurant un meilleur niveau de validité introspective.

Nous présenterons dans la partie suivante (voir p. 251) les indices de facilité introspective par question calculés après turbo analyse dans le cadre de l'épreuve de physique (n = 2.497).

Nous présenterons aussi l'indice de Centration par question (*Cq*) où nous confrontons la facilité introspective d'une question à sa facilité objective (*poq*). Nous avons appliqué à l'indice *Cq* le principe de la turbo analyse (voir p. 254).

Les indices *piq* et *Cq* peuvent aussi être calculés pour un test entier ce qui donne les deux indices : *pit* (voir p. 268) et *Ct* (voir p. 270).

3. L'approche expérimentale

a) Principe de base

Lorsque les questions d'une épreuve sont notées à l'aide d'une même échelle, l'indice de difficulté d'une question est généralement noté *p*. Notons, ainsi que nous le discutons plus loin, qu'il aurait sans doute mieux valu appeler cet indice *p* « *indice de facilité* ». Par similitude avec l'indice de facilité/difficulté introspective d'une question nous proposons de noter l'indice de *facilité/difficulté* (classiquement *p*) *o*bjective d'une *q*uestion par « *poq* ». Cet indice *poq* est défini par :

$$p = \frac{\sum_{s=1}^{ns} x_{qs}}{ns} \quad (1)$$

avec :

s = l'indice des *s*ujets
x_{qs} = la note à la question *q* obtenue par le sujet *s*
ns = le nombre total de *s*ujets ayant répondu à la question *q*

Laveault & Gregoire (1997) proposent de diviser cette moyenne par l'étendue de la note lorsque les items d'un test sont notés sur des échelles différentes, ce qui produit un indice comparable variant entre 0 et 1 pour chaque question.

Leclercq (1987) signale à la suite de Wood (1977) et Ebel (1979) un aspect paradoxal de cet indice : à mesure que la fréquence des réponses correctes baisse, la difficulté s'élève. Ces auteurs recommandent d'utiliser l'appellation « *indice de facilité* » lorsqu'il est question de proportion de réponses correctes défini par : *nc*/*nr* (*nc* = nombre de réponses correctes et *nr* = le nombre total de réponses) et « *indice de difficulté* » lorsqu'on fait référence aux proportions d'échecs (omissions et réponses incorrectes) défini par : *ni*/*nr* ou 1-(*nc*/*nr*).

b) Les facteurs qui affectent la difficulté d'une question

(1) L'influence du temps imparti

Laveault & Gregoire (1997) font remarquer que lorsqu'un grand nombre de sujets n'ont pu répondre à un item faute de temps, l'indice de difficulté mesure deux choses : la difficulté de la question et la rapidité du répondant. Nuancer le calcul de la difficulté en tenant compte du nombre de sujets qui ont répondu à la question plutôt qu'au test ne permet pas de contrer le problème étant donné qu'il y a de fortes chances que ceux qui produiront une réponse à la question, les plus rapides, soient aussi les plus « forts » du point de vue de leur performance globale au test.

(2) La part des réponses dues au hasard

Nous avons déjà abordé le problème des réponses dues au hasard dans le contexte des questions à choix multiple et du « gonflement » des scores qui en résulte (voir p. 104). Nous avons signalé que les pénalités pour réponses dues au hasard, la « *correction for guessing* », fait l'objet de nombreuses critiques et ne constitue pas une solution efficace au problème. Par exemple, il découle de cette correction pour divination classique qu'un apprenant qui ne sait rien a autant intérêt à omettre qu'à répondre au hasard. De Landsheere (1979) signale également « *Elle repose notamment sur l'hypothèse gratuite que tous les sujets*

ont également deviné. De plus, on établit pas de distinction entre l'élimination de certains choix sur la base de connaissances réelles et la divination au pur hasard .».

Mentionnons ici avec toutes les réserves liées aux propos ci-dessus, la formule de correction pour l'indice de difficulté proposée par Laveault & Gregoire (1997) qui signalent qu'il est possible de corriger l'indice de difficulté pour l'effet du hasard « *…à chaque fois que l'on peut admettre que les leurres ont une chance à peu près égale d'être choisis.* ». L'indice corrigé pour l'effet du hasard devient alors :

$$p' = p - \left[\frac{1-p}{M-1}\right] \quad (2)$$

Avec
p' : l'indice de difficulté corrigé
p : l'indice de difficulté de départ
M : le nombre de propositions contenues dans la QCM

Notons que les auteurs ne recommandent cependant pas d'appliquer systématiquement cette correction, notamment parce qu'il est très peu plausible qu'un sujet réponde véritablement au hasard : « *…Celui-ci dispose toujours d'une connaissance partielle de la question qui lui permet d'éliminer des choix de réponses. Une question à cinq choix de réponses peut alors se ramener à trois ou deux choix*. ».

(3) Le niveau de compétence des répondants et la position de l'épreuve dans le contexte d'apprentissage

Le fait de passer une épreuve avant, pendant ou après un apprentissage aura des effets sur la difficulté. On ne s'attend par exemple pas à obtenir les mêmes niveaux de difficulté lors d'épreuves répétées. C'est le cas du test de lecture proposé par Boxus (1981). Ce test est présenté aux mêmes élèves (1ère primaire) à plusieurs reprises et on observe une évolution des résultats via des courbes caractéristiques de distribution des scores : d'abord une courbe « en i » avant l'apprentissage, ensuite une courbe gaussienne « en cloche » pendant, et finalement une courbe « en j » après l'acquisition des *skills* de lecture par les apprenants (Debry & al, 1998).

Le niveau de compétence des répondants souvent lié à la position de ces derniers dans le cursus des apprentissages joue sur la difficulté de l'épreuve. C'est aussi le cas des tests de progrès pratiqués à la Faculté de Médecine à Maastricht où est pratiquée la méthode du « *Problem Based Learning (PBL)* » (Barrows & Tamblyn, 1977, 1980; Van Der Vleuten & Wijnen, 1990; Van Der Vleuten, 1996; Leclercq, 1998). Chaque année à trois mois d'intervalle tous les étudiants des 6 années d'études sont soumis à une évaluation dont les résultats seront comptabilisés. Ces tests portent sur toute la matière de médecine (250 questions). Dans ce contexte d'évaluation il est normal qu'un étudiant de 1ère année ne puisse répondre à toutes les questions. Ainsi, le niveau de performance attendu et donc le niveau de difficulté de l'épreuve varie en fonction du niveau d'étude et de compétence des répondants. Les concepteurs de ces tests de progrès conçoivent leurs questions de façon à ce que les étudiants de 1ère année réussissent en moyenne environ 10% des questions, ceux de 2ème année environ 20%, etc.

4. Rapport entre facilité des items, symétrie des distributions et capacité à discriminer les sujets

Lorsque pour une épreuve donnée on est confronté à une série de questions particulièrement faciles la distribution des scores devient asymétriques et prend une forme « en j » dans le cas contraire d'une épreuve particulièrement difficile, la forme de la distribution est aussi asymétrique, mais « en i ». Dans d'autres épreuves présentant ni trop de questions faciles, ni trop de questions difficiles on observe un type de distribution de forme gaussienne et donc plus symétrique.

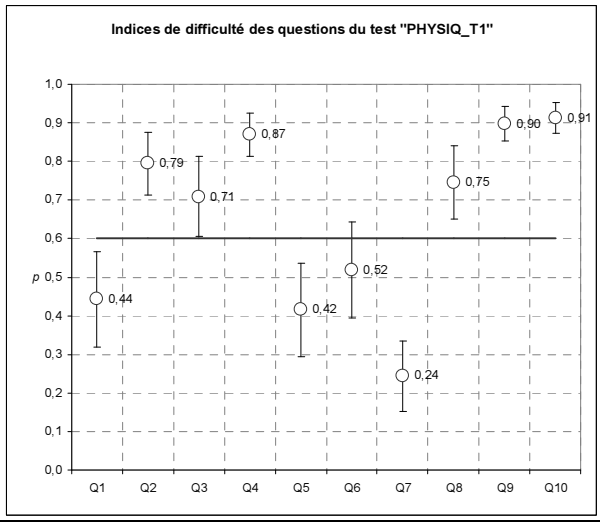
Il est aussi possible de présenter des distributions par question en fonction de l'aspect correct (1) ou incorrect (0) de la réponse fournie. Lorsque la question est particulièrement difficile, on obtient alors une distribution (0,1) de forme asymétrique dont le coefficient d'asymétrie sera positif (la colonne des 0 sera plus haute que la colonne des 1). Dans le cas d'une question particulièrement facile le coefficient d'asymétrie de la distribution sera négatif (la colonne des 1 sera plus haute que la colonne des 0). Lorsque les réponses correctes et incorrectes s'équilibrent, la distribution est symétrique et le coefficient d'asymétrie est proche de zéro (la colonne des 0 sera à peu près de même hauteur que la colonne des 1).

Habituellement, dans les situations d'évaluation où l'on cherche à différencier les sujets entre eux, les questions dont la distribution est symétrique et où réponses correctes et incorrectes s'équilibrent, sont recherchées car elles discriminent autant les « forts » que les « faibles ». En effet, une question trop facile ne discriminera en principe que parmi les sujets particulièrement « faibles », tous les autres ayant réussi la question (*en principe*, car cela dépend aussi du niveau de cohérence interne de la question, sa corrélation avec le score total, son *rpbis*). De même une question trop difficile ne discriminera en principe que parmi les sujets les plus « forts », tous les autres l'ayant ratée. Imaginons trois catégories d'items : (D) celle des questions Difficiles où *p* varie entre 0 et 0,33, (I) celle des questions de difficulté Intermédiaire où *p* varie entre 0,34 et 0, 66 et (F) celle des questions Faciles où *p* varie entre 0,67 et 1. Si dans une épreuve donnée nous ne rencontrons que des questions de type D et F, cette épreuve ne discriminera pas les sujets dont le niveau de compétence est moyen et qui sont ni particulièrement « forts », ni particulièrement « faibles ».

a) Lien entre difficulté et discrimination autour d'un seuil de réussite

En fait, si nous souhaitons nous assurer qu'un test discrimine bien les sujets qui ont atteint le seuil de réussite de ceux qui ne l'ont pas atteint, par exemple la note 12 de satisfaction à une épreuve universitaire (ou 0,6 sur une échelle de 0 à 1), nous devons veiller à ce que le test comporte un nombre suffisant de questions qui discriminent autour de ce seuil. Ainsi, comme que le signalent Laveault & Gregoire (1997) : « *A cause du grand nombre d'items dans le voisinage de la note de passage, de légères différences se traduiront par des changements importants au niveau du score total. De cette façon, le score total du sujet nous permettra de bien discriminer entre ceux qui ont atteint et ceux qui n'ont pas atteint la valeur seuil.* ».

Qu'en est-il de l'épreuve de physique (n = 2.497) ? Postulons qu'un seuil de réussite ait été fixé à 0,6 pour ce test. Notons qu'il existe des méthodes pour déterminer des scores seuils les plus valides possibles (V. de Landsheere, 1986; Kane, 1994; Laveault & Gregoire, 1997), nous ne les aborderons pas ici car la problématique des seuils mérite un chapitre en soi.



Pour que l'épreuve discrimine bien les sujets qui ont atteint ou non ce seuil, les valeurs *p* des questions devraient s'étaler autour de 0,6.

L'ordonnée reprend les valeurs de l'indice de difficulté *p*. Chaque question est représentée par un

cercle et, en référence avec l'ordonnée, indique dans quelle proportion l'item est réussi.

La dispersion des résultats autour de l'indice *p*, exprimée par la variance de l'item, est visualisée par des barres situées au-dessus et en dessous de chaque cercle.

Etant donné qu'il s'agit d'items corrigés de façon dichotomique (0, 1), la variance est calculée selon la formule simplifiée :

$$\boxed{\sigma_q^2 = p_q.q_q} \quad (3)$$

avec :

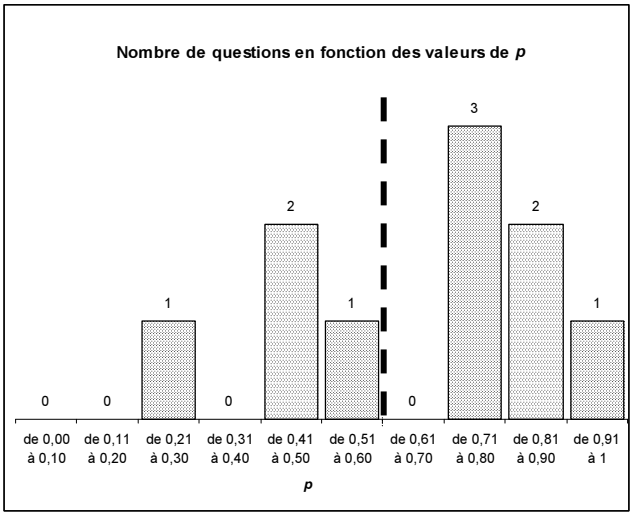
$_q$ = l'indice représentant les questions
p_q = la proportion de sujets ayant réussi la question q
q_q = la proportion de sujets ayant échoué la question q

La ligne horizontale sur le graphique désigne la valeur du seuil que nous avons fixé à 0,6. Six questions se situent au-dessus de ce seuil et quatre en dessous. La difficulté moyenne du test vaut 0,66 et est donc assez proche du seuil que nous avons fixé à 0,6.

	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	Moy.
p	0,44	0,79	0,71	0,87	0,42	0,52	0,24	0,75	0,90	0,91	0,655

Le graphique ci-dessous montre la distribution des 10 questions du test de physique (n = 2.497) en fonction des valeurs obtenues à l'indice de difficulté. Chaque bâton représente le nombre de questions pour un intervalle des valeurs *p* donné.

La ligne verticale en pointillés séparant les catégories « *0,51 à 0,60* » et « *0,61 à 0,70* » permet de visualiser le seuil (0,6).



Ce graphique montre mieux dans quelle mesure le test discrimine efficacement autour de la valeur du seuil. Nous constatons un « trou », c'est-à-dire l'absence de questions, au niveau *p* compris entre 0,61 et 0,70, juste à droite du seuil, et, juste à gauche, une seule question pour le niveau de difficulté compris entre 0,51 et 0,60. Ce test, dans l'hypothèse où un des objectifs des concepteurs aurait été de renvoyer un feedback stipulant « vous avez réussi… » ou « vous n'avez pas réussi… » avec un seuil de réussite fixé à 0,6, aurait discriminé avec plus d'efficacité les étudiants s'il avait compté des questions dont les *p* auraient été compris entre 0,61 et 0,70 et si il y avait eu plus de questions avec des valeurs *p* comprises entre 0,51 et 0,70.

Enfin, rappelons qu'une question de difficulté moyenne ne discrimine pas automatiquement de façon efficace les répondants. En effet, une question avec un *p* égal à 0,5 peut malheureusement aussi obtenir un mauvais coefficient de discrimination *rpbis*.

b) Rapport entre difficulté et discrimination lorsqu'on souhaite discriminer à différents niveaux de performances

Si l'objectif du concepteur du test est de sonder les compétences des sujets en vue par exemple de leur renvoyer un feedback du type « votre niveau de maîtrise se situe à… », l'enjeu de l'analyse du pouvoir discriminatif des questions n'est plus de vérifier si elles contribuent à discriminer à un seuil de réussite/difficulté donné, mais plutôt de contrôler si elles permettent une discrimination efficace de plusieurs niveaux de performances possibles chez les répondants.

Dans le cas où l'évaluateur souhaite discriminer à différents niveaux de performances, il devra veiller à utiliser des items dont les niveaux de difficulté permettront de discriminer aussi bien, par exemple, parmi les sujets « forts » que parmi les sujets « moyens » ou « faibles ».

Laveault & Gregoire (1997) estiment que dans de tels cas : « …*les items faciles ou difficiles jouent un rôle plus complexe et c'est au concepteur de s'interroger sur ce rôle en fonction des objectifs d'évaluation. L'analyse d'items peut l'aider à se poser les questions pertinentes quant au rôle joué par chaque item ainsi que sur les moyens pour améliorer la qualité de l'instrument de mesure… Un « bon » item nous permet d'atteindre notre objectif d'évaluation. Cet objectif sera atteint en choisissant des items de difficulté et de discrimination adéquates. ».*

Nous pensons que si l'objectif est de renseigner sur le niveau des performances des répondants, il faudra veiller à ce que des questions de tous niveaux de difficulté soient présentes dans l'épreuve. Dès lors, les concepteurs devront faire en sorte qu'aucun « trou » n'apparaisse dans un graphique de distribution des nombres de questions par niveau *p* tel que celui présenté au paragraphe précédent. Ils devront aussi équilibrer au mieux le nombre d'items pour chaque intervalle de difficulté.

Prévoir le niveau de difficulté d'une question avant que le test n'ait eu lieu relève à notre avis de la problématique de la validité de contenu qui est elle-même liée à la validité conceptuelle (liée à un modèle du contenu) des items d'une épreuve. Ces problèmes fondamentaux dans la construction d'un test doivent être pris en considération avant la phase de mise en œuvre. Nous n'aborderons pas ici la problématique de la validité car nous situons le champs de cette étude sur la mise au point d'outils d'aide à l'analyse spectrale des items dans une phase ultérieure du processus de réalisation d'un test, après qu'il ait été créé et soumis aux répondants, l'analyse spectrale de la qualité des questions s'effectuant lors de la phase de correction du test.

C. Les indices de fidélité

Dans quelle mesure un correcteur peut-il prétendre qu'un travail corrigé et classé dans la catégorie « excellent » bénéficierait de la même mention s'il était à corrigé dans d'autres conditions (autres correcteurs ou un mois plus tard) ? Les problématiques liées à la subjectivité de la correction ont été l'objet d'un courant de docimologie dite « négative » ou « critique » mené par Pieron (1963). Ce dernier et d'autres chercheurs à sa suite ont relevé une série de biais d'évaluation tels que l'inconstance d'un même évaluateur[55] et la discordance entre évaluateurs. Cette dernière est illustrée, entre autres, par AGAZZI (1967) qui observe à l'occasion de la correction des copies d'un baccalauréat par 6 correcteurs que 70 % des compositions françaises sont tantôt admises par les uns et tantôt refusées par les autres. Pieron & al. (1962) estiment qu'il faudrait 16 correcteurs pour stabiliser les notes en physique, 78 correcteurs en composition française et 127 en dissertation philosophique... (si ces correcteurs ne se concertent pas).

Des procédures d'évaluation automatisées visent à garantir la fidélité des mesures. Les QCM permettent d'échapper à la subjectivité des correcteurs et contribuent ainsi à augmenter la fidélité des évaluations; de plus, la simplicité de correction autorise un traitement informatisé rapide grâce à la lecture optique de marques. Ce fut le cas des épreuves MOHICAN qui ont été standardisées et ont fait l'objet d'une correction automatisée à l'aide du dispositif de lecture optique de marque du Système Méthodologique d'Aide à la Réalisation de Tests de l'Université de Liège (SMART-ULg) (voir p. 55).

Nous sommes bien sûr conscients qu'à côté des avantages qualitatifs et quantitatifs indéniables du questionnement à choix multiple (la capacité à couvrir une matière large, la simplicité et l'objectivité de la correction, la possibilité d'évaluer systématiquement et précisément des niveaux supérieurs d'activité mentale, …) il existe une série de limitations liées aux QCM comme par exemple l'incapacité des QCM à mesurer l'expression spontanée, l'aptitude à rédiger, l'invention de solutions nouvelles, la démonstration de raisonnements logiques d'une certaine complexité. Comme le souligne Leclercq (1986, p. 46) « *Les QCM ne peuvent résoudre tous le problèmes de l'évaluation pédagogique. Il s'agit d'une technique parmi d'autres, qui doit être choisie en fonction de son adéquation au contenu, au niveau de comportement sollicité, aux contraintes du moment* ».

1. Trois sources d'erreurs qui affectent la fidélité des mesures

Tout en contribuant à améliorer la constance des mesures, la standardisation des tests et l'automatisation de la correction n'empêchent pas que des erreurs puissent affecter la fidélité. La pratique enseigne que même dans le cas où nous pouvons administrer deux versions parallèles d'un test aux mêmes sujets et malgré de multiples précautions, des différences apparaîtront entre les deux séries de scores des répondants. La corrélation entre des résultats, en théorie devrait être parfaite, en pratique ne l'est pas. Nous pouvons tendre vers plus de fidélité dans les mesures mais atteindre une fidélité totale est pratiquement impossible. Leclercq (1987) distingue trois sources d'erreurs qui affectent la fidélité des mesures : l'instrument, le contexte de passation et l'individu.

De Landsheere signale à propos des erreurs liées aux instruments : « *Malgré le parallélisme des épreuves, il se peut que le contenu de l'une soit plus familier à un sujet que le contenu de l'autre. Idéalement, il faudrait administrer un grand nombre de formes parallèles afin de travailler sur un échantillon réellement représentatif de la tâche* ».

En ce qui concerne les circonstances de passation, il est impossible de reproduire parfaitement deux situations de testing, de nombreux de facteurs accidentels peuvent amener des différences dans les performances des sujets : bruits, disposition de salle, stylo cassé, éclairage, météo, …

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

Enfin, pour ce qui est de l'individu, les apprentissages réalisés entre deux tests, l'état de santé des étudiants, le niveau de fatigue et de stress, les motivations, … pourraient influencer les résultats.

2. Définition de la fidélité

Leclercq (1987) définit la fidélité d'un test comme étant « *sa capacité de fournir des mesures stables (reproductibles) si le test est appliqué de nouveau à un autre moment, ou dans d'autres circonstances*. ». Etant donné qu'il existe des sources d'erreurs qui affectent les mesures lors des passations, on peut exprimer le coefficient de fidélité de cette façon :

$$\boxed{r_{xx'} = 1 - E} \quad (4)$$

avec $r_{xx'}$: le coefficient de fidélité (coefficient de corrélation) de deux séries de mesures x et x'
E : une fonction numérique des erreurs [56] qui affectent les mesures et qui sont liées aux instruments, aux circonstances de passation et aux sujets eux-mêmes

La fidélité est un concept fondamental en évaluation, si d'une fois à l'autre un test ne mesure pas la même chose, il ne pourra pas constituer une mesure valide de ce qu'il prétend mesurer. Parmi les erreurs qui affectent les mesures, il convient de distinguer les erreurs systématiques des erreurs non-systématiques (ou aléatoires).

Laveault & Gregoire (1997, pp. 132-140), attirent notre attention sur le premier postulat de la théorie classique des tests (Spearman, 1907; Guliksen, 1950; Magnuson, 1967; Lord & Novick, 1968). Selon la théorie classique des tests, il est impossible d'obtenir une mesure complètement exempte d'erreurs.

D'après le premier postulat, le score observé (X) d'un sujet résulte de la somme de son score vrai (V) et de l'erreur aléatoire de mesure (E) associée à ce score :

$$\boxed{X = V + E} \quad (5)$$

Dès lors, pour estimer la fidélité d'une mesure, il faudra évaluer l'écart entre la note vraie et le score obtenu par le sujet. Démontrer la fidélité d'un test, reviendra donc à démontrer que la proportion d'erreur de mesure peut être négligée car en faible proportion par rapport à la "performance vraie" évaluée par le test.

D'une part le score vrai devrait être une constante et d'autre part l'erreur aléatoire de mesure est variable, ce qui entraîne que les scores observés sont variables aussi en fonction de l'erreur. Comme l'erreur aléatoire de mesure se distribue normalement, les scores observés se distribuent aussi normalement autour du score vrai.

Cependant l'idée que les scores se distribuent normalement autour du score vrai nous paraît discutable. En effet, nous savons que dans certaines situations de testing la moyenne des sujets peut être très proche du minimum (par exemple en situation de pré test) ou du maximum (lors d'un post-test), dans de telles situations les distributions des scores ne sont pas symétriques, mais asymétriques et en forme de « ***i*** » ou de « ***j*** ». Dès lors, les scores observés pourraient très bien ne pas se distribuer normalement autour du score vrai.

Laveault & Gregoire (1997) signalent par ailleurs : « *On peut affirmer que le postulat ne tient pas dès que le test mesure une caractéristique de l'individu qui exerce une influence directe ou indirecte sur sa façon de répondre au test, telle que la tendance à deviner, à tricher, à omettre certaines catégories de réponses, etc.* »

[56] « E » n'est pas numérique dans cette équation si on considère que la quantification des erreurs de mesure est dépendante des unités employées pour effectuer les mesures. Dès lors, E pourrait varier dans une plage différente de 0 à 1, c'est pourquoi nous préférons utiliser l'expression « fonction numérique des erreurs ».

Il existe dès lors des sources d'erreurs qui doivent aussi être prises en compte, il s'agit des erreurs systématiques : « *Par exemple, il y a erreur systématique lorsqu'un test est trop facile ou trop difficile. Deux sujets dont les scores vrais en mathématique sont différents, peuvent obtenir le même score vrai de 10/10 lorsque l'examen est trop facile… C'est pourquoi il faudrait réécrire l'équation de la manière suivante :*

$$\boxed{X = V + e_s + e_a} \quad (6)$$

Dans cette dernière expression, le score observé du sujet est la somme d'un score vrai, d'une erreur systématique et d'une erreur aléatoire. »

Une façon d'améliorer la fidélité consiste à augmenter le nombre de questions. « *Plus il y a d'items dans un test, plus l'erreur type d'estimation de cette moyenne sera faible et, par conséquent, plus l'erreur type sera réduite »,* à condition que les items mesurent la même chose.

Laveault & Gregoire (1997) proposent trois définitions générales de la fidélité à partir de la théorie classique des tests. Notons que ces auteurs utilisent le terme « fiabilité » plutôt que « fidélité ».

L'indice de fiabilité, corrélation entre les scores observés et les scores vrais.

D'un point de vue pratique cette définition n'apporte rien car elle ne nous donne pas les valeurs des scores vrais. L'intérêt réside dans le fait qu'elle nous permet de comprendre que « *meilleure est la fiabilité, meilleure sera la prédiction du score vrai à partir du score observé* ».

L'indice de fiabilité, proportion de variance des scores observés imputable aux scores vrais.

Donc, plus le test est fiable, plus la variance des scores observés est due aux scores vrais et non aux fluctuations aléatoires. Selon Laveault & Gregoire : « *Concrètement, le coefficient de fiabilité* $r_{xx'} = 0{,}81$ *signifie que 81% de la variance des scores observés est attribuable à la variance des scores vrais*. ».

L'indice de fiabilité, corrélation entre scores observés à deux formes parallèles (définition opérationnelle).

Il est démontré (la démonstration figure dans Laveault & Gregoire (1997), p.138) que la corrélation entre deux formes parallèles d'un test nous permet d'en estimer la fiabilité mais pour autant que les postulats de la théorie classique soient adéquats pour décrire les résultats. Notons que cette définition opérationnelle de la fidélité correspond à celle utilisée par Leclercq (1987) cité au début de cette section.

3. L'évaluation de la fidélité des épreuves

Une première façon d'évaluer la fidélité d'une épreuve consiste à administrer à deux reprises le même test à un même groupe d'étudiants après un délai plus ou moins long (procédure « *test-retest* »), puis à calculer la corrélation des deux séries de mesures (coefficient de stabilité). Théoriquement, dans le cas d'un test qui serait totalement fidèle, donc qui ne contiendrait pas d'erreurs de mesures, la corrélation devrait être égale à 1. Cependant, nous savons que dans la réalité cette corrélation parfaite ne peut être obtenue. En effet, de multiples causes seront à l'origine des différences de performances entre deux passations d'un test (environnement, stress, familiarité des items, … Voir plus haut « *Trois sources d'erreurs qui affectent les mesures* »).

Une façon d'essayer d'éviter le problème de la source d'erreurs liées à l'effet d'apprentissage provoqué par le fait d'être confronté une seconde fois au même test, consiste à créer deux formes équivalentes d'une même épreuve (méthode des formes parallèles). En postulant que les deux tests mesurent la même réalité, la corrélation entre les deux formes devrait permettre d'estimer la fidélité de l'épreuve. Malheureusement, nous savons que d'une part il est difficile de garantir le total parallélisme entre deux tests réputés équivalents et, que d'autre part, des fluctuations dans les performances des participants peuvent aussi intervenir dans l'explication des variations dans les scores des deux formes (par exemple une baisse du niveau d'anxiété lors de la seconde épreuve pourrait influencer les performances). Comment reconnaître les fluctuations liées à l'instrument de celles liées aux sujets ?

On le voit, les deux approches précédentes (« *test-retest* » et formes parallèles) présentent d'énormes problèmes. Dès lors, une troisième voie d'évaluation de la fidélité des épreuves consiste à utiliser les informations fournies en une seule passation. Il s'agit d'une approche basée sur l'évaluation de la consistance interne du test. Plusieurs techniques sont possibles, nous reprendrons ici les deux plus souvent citées : le coefficient de bipartition ou « *split-half* » et « l'alpha de Cronbach ».

En ce qui concerne le coefficient de bipartition, signalons brièvement ici (il sera détaillé plus loin, p. 130) que l'idée est de séparer une épreuve en deux moitiés, la première moitié étant constituée par les résultats obtenus par exemple aux questions paires et l'autre moitié aux questions impaires. Le coefficient de bipartition étant la corrélation entre les deux séries de résultats. Il subsiste cependant encore deux problèmes avec la méthode de bipartition. D'une part on obtiendra autant de mesures de la fidélité qu'il y a de possibilités de combiner les questions dans des moitiés différentes de l'épreuve. D'autre part, comme la corrélation implique la moitié des questions et non l'entièreté, la fidélité est systématiquement sous-estimée (nous verrons plus loin que le coefficient de correction de Spearman permet de contrecarrer cet effet).

Pour ce qui est du coefficient alpha de Cronbach (détaillé plus loin, p. 137), le problème des multiples moitiés possibles est résolu étant donné qu'il correspond à la moyenne des coefficients de fidélité que l'on aurait obtenus par « *split-half* » successifs en examinant toutes les combinaisons de questions en deux moitiés. Notons également que l'alpha est calculé à partir des résultats de l'entièreté des questions du test (et non plus des moitiés).

4. L'évaluation de la fidélité des mesures dans le contexte des épreuves MOHICAN

Une procédure « *test-retest* » aurait permis de mesurer la stabilité des résultats en soumettant le même test aux mêmes étudiants quelques temps après la première passation. Un des buts de l'opération était de fournir rapidement à chaque participant un feedback personnalisé contenant entre autres les réponses correctes aux différents tests, ce qui a pour effet de « brûler » les épreuves pour une seconde passation. A supposer que la première passation n'aie pas eu d'effet d'apprentissage sur les répondants, le contexte du projet MOHICAN ne nous permet quand même pas de mettre en place une telle procédure, ne serait-ce que parce que les abandons ou les changements d'orientation des étudiants font que quelques semaines après l'entrée en 1ère candidature, le groupe initial est déjà amputé d'une partie de ses répondants. Au sujet du problème des effets d'apprentissage liés aux procédures « *test-retest* », Laveault & Gregoire (1997, p.140) signalent que : « *Si, par exemple, les sujets les plus forts lors de la première passation sont aussi ceux qui, au moment du retest, se rappellent mieux des questions posées la première fois, il risque d'y avoir corrélation entre le score vrai de l'élève au premier test et l'erreur aléatoire de mesure au second, ce qui enfreint le postulat 5 du modèle de la théorie classique*. »

La mise en place de formes parallèles des tests MOHICAN n'a pas non plus été envisagée pour des raisons de coût et de timing des opérations. La création de deux épreuves jumelles aurait en effet été coûteuse en temps pour les créateurs des épreuves, la rédaction de questions qui soient différentes tout en restant équivalentes n'étant pas simple.

En édumétrie les évaluateurs sont souvent confrontés à ces problèmes de lourdeur de mise en œuvre des procédures de « *test-retest* » et de formes parallèles et une façon de contourner le problème consiste alors à faire appel à des méthodes d'estimation de la cohérence interne. Leclercq (1987, p.147) rappelle : « *Faute de pouvoir mesurer la fidélité, on en est réduit à l'estimer par divers artifices, tous basés sur la cohérence interne (internal consistency) ou homogénéité de l'épreuve.* ». Dans la suite de cette étude nous proposerons deux types de coefficients qui nous permettront de mesurer la cohérence interne des tests MOHICAN en vue d'en estimer dans une certaine mesure la fidélité (voir tableau ci-après), il s'agit du coefficient de bipartition (*split half coefficient*) et de l'alpha de Cronbach.

Leclercq (1987, p.147) propose un tableau qui permet de visualiser dans quelle mesure chaque méthode permet de tenir compte des sources d'erreurs qui affectent la fidélité des mesures. Nous reprenons ci-après ce tableau en y remplaçant les formules de Kuder Richardson KR20, KR21 et de Guilford par le coefficient alpha de Cronbach qui nous paraît équivalent du point de vue de l'information fournie et qui sera détaillé plus loin. De Landsheere (1979) signale à ce propos : « *α est une formule générale du coefficient d'équivalence ou de Kuder Richardson 21. Alpha est, en fait, la moyenne de tous les coefficients de bipartition (split-half coefficient) possibles pour un même test (Cronbach) dans le cas d'un parallélisme rigoureux entre les deux formes étudiées.* ». Les cases hachurées signalent l'absence de réponses à l'estimation de la source d'erreur correspondante. Nous indiquons MOHICAN ou barrons ~~MOHICAN~~ lorsque l'option, respectivement, est ou n'est pas envisagée dans le cadre de cette étude.

Méthodes de calcul :		Sources d'erreurs couvertes :		
		Les questions du test	Les variations dans l'administration	Les variations chez l'individu
Calcul de la fidélité	Coefficient de stabilité : procédure « *test-retest* » et calcul du *r* de Bravais-Pearson		~~MOHICAN~~	~~MOHICAN~~
	Coefficient d'équivalence : procédure des formes parallèles et calcul du *r* de Bravais-Pearson	~~MOHICAN~~		
Calcul de la cohérence interne	Coefficient de bipartition : découpage des résultats en deux moitiés et calcul du *r* de Bravais-Pearson	MOHICAN		
	Coefficient Alpha de Cronbach	MOHICAN		

5. Coefficient de bipartition ($r_{xx'}$)

a) Principe et inconvénients

Le principe de base du coefficient de bipartition repose sur le calcul de la corrélation *r* de Bravais-Pearson à partir de deux moitiés des résultats obtenus à une épreuve administrée en une fois. La méthode est aussi appelée méthode de bissection ou méthode de fiabilité par moitié ou encore en anglais « *split-half* ».

(1) Sous-estimation systématique due à la réduction de moitié des questions

Un inconvénient de la méthode de bipartition est qu'elle ne nous donne qu'une estimation de la fidélité qui repose sur la moitié seulement des questions, le coefficient de bipartition risque donc de sous-estimer systématiquement la fidélité d'un test. Nous savons en effet que plus un test comporte d'items, plus il est fiable à condition que les items mesurent bien la même chose.

(2) Problème de l'équivalence des moitiés

La corrélation entre les scores des deux moitiés d'un même test pose un second problème, celui de la constitution de chaque moitié. Il y a en effet de fortes chances qu'on obtienne des coefficients de fidélité différents selon la manière dont les deux moitiés ont été constituées. Corréler les items du début du test, de la question un à la question de la fin de la première moitié, avec les items de la seconde moitié qui terminent le test avec tout ce que cela sous-entend comme erreurs de mesures liées à la fatigue, au manque de temps, etc. donnera un coefficient très probablement différent d'une corrélation des deux parties créées sur la base d'un tirage aléatoire des questions.

On peut aussi créer les moitiés en tentant d'équilibrer dans chacune les items en rapport avec les mêmes contenus et en veillant à répartir les difficultés. On peut aussi considérer, mais sans aucune garantie, qu'en reprenant les questions paires dans une moitié et les questions impaires dans l'autre on obtient deux moitiés équivalentes. Dans le cadre de cette étude, nous avons préféré dans un premier temps sélectionner les items selon qu'ils étaient pairs ou impairs puis dans un second temps nous avons équilibré ces formes parallèles du point de vue de la difficulté des questions.

b) Méthodes de calcul

La formule de base est celle du coefficient de corrélation *r* de Bravais-Pearson (voir p. 150).

(1) Importance de l'équilibrage de la difficulté des sous-tests

Un test trop facile, où la proportion de réponses correctes est très élevée et où la distribution des scores est décentrée vers le maximum, obtiendra une asymétrie négative (courbe en ***j***) et un autre trop difficile, où la proportion de réponses incorrectes est très élevée, obtiendra une asymétrie positive (courbe en ***i***). Or, comme le rappellent Laveault & Gregoire (1997, p. 152) : « *…la corrélation r de Pearson ne peut atteindre sa valeur maximum 1 que lorsque les variables en corrélation sont symétriques ou possèdent le même type d'asymétrie*. ». Dès lors un mauvais équilibrage des moitiés peut entraîner le calcul d'une corrélation entre par exemple, une première moitié de test trop facile et une seconde trop difficile, donc entre une distribution asymétrique négative pour la première et une distribution asymétrique positive pour la deuxième, et donc, hors des limites statistiques du modèle de corrélation linéaire de Bravais-Pearson.

(2) Correction de la sous-estimation liée à la réduction du nombre d'items

Comme nous l'avons souligné plus haut, appliqué aux moitiés des résultats ce coefficient de corrélation sous-estime la fidélité, le nombre de questions ayant diminué de 50% dans chaque sous-test. Le coefficient de Spearman-Brown permet d'apporter une solution à ce problème de sous-estimation de la fidélité en apportant une correction au *r* de Bravais Pearson.

La formule de Spearman-Brown (*Spearman Brown prophecy formula*) s'exprime mathématiquement de cette façon :

$$rS = \frac{kr_{xx'}}{1 + (k-1)\,.\,r_{xx'}} \quad (7)$$

avec $r_{xx'}$: comme expression de la corrélation entre les deux moitiés de l'épreuve (voir définition du coefficient de corrélation de Bravais-Pearson, p. 150).

k : la proportion dans laquelle le test doit être allongé pour calculer le coefficient de fidélité corrigé pour sous-estimation

Dans le cas du coefficient de bipartition, les sous-tests sont proportionnellement deux fois moins long que le test entier, dès lors *k* vaut 2, ce qui entraîne :

$$rS = \frac{2r_{xx'}}{1 + r_{xx'}} \quad (8)$$

Laveault & Gregoire (1997, p.145) insistent sur le fait que cette correction n'est toutefois valable que si les moitiés du test correspondent à la définition de deux épreuves strictement parallèles : « *Lorsque les variances des deux moitiés sont forts différentes, l'estimation de la fiabilité du test entier risque d'être faussée.* ».

(3) Coefficient de fiabilité par moitié : lorsque les variances diffèrent

La fiabilité par moitié de Guttman est une autre méthode de calcul similaire à celle de Spearman-Brown, mais où cette fois on considère qu'il existe des différences dans les variances des résultats des deux moitiés. Le coefficient de Guttman s'exprime :

$$rG = \frac{2.(\sigma_t^2 - \sigma_{t1}^2 - \sigma_{t2}^2)}{\sigma_t^2} \quad (9)$$

avec rG : l'expression du coefficient par moitié de Guttman

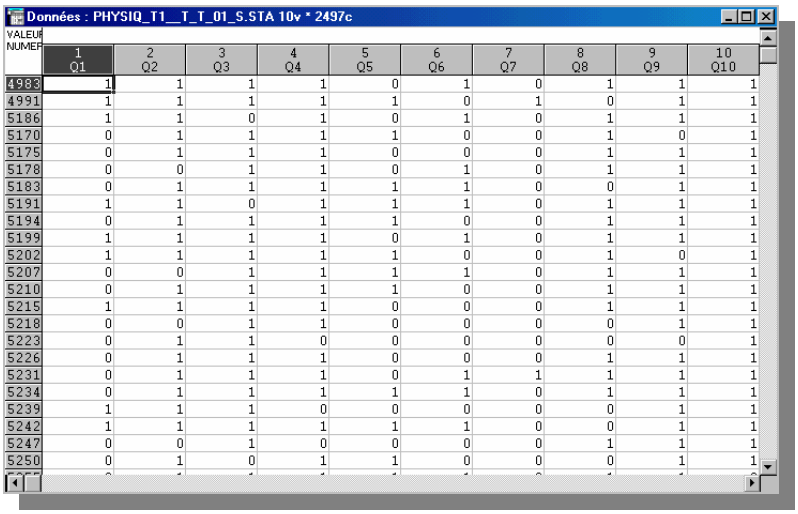
σ_t^2 : la variance de tous les items (les deux moitiés)

σ_{t1}^2 : la variance des items de la première moitié

σ_{t2}^2 : la variance des items de la deuxième moitié

c) Matrices « binaires » et « spectrales » des épreuves MOHICAN

Les scores des épreuves MOHICAN peuvent se présenter sous deux formes : l'une, binaire, ne tenant pas compte des informations métacognitives et l'autre, spectrale, intégrant les pourcentages de certitude en plus de l'information sur l'aspect correct ou non des réponses. La première consisterait à présenter les données sous la forme de 0 ou de 1, pour chaque réponse fournie par chaque sujet si la réponse est correcte on indique 1, si elle ne l'est pas, on note 0. Les résultats des tests se présenteraient alors sous la forme de matrices composées de 0 ou de 1. Les lignes représentant les sujets et les colonnes les questions

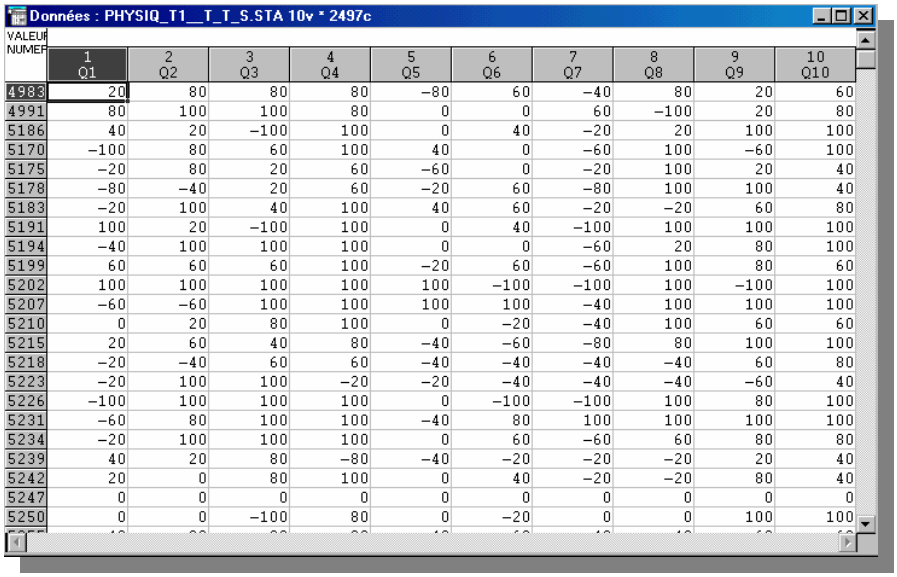

Données : PHYSIQ_T1__T_T_01_S.STA 10v * 2497c

VALEUR NUMER	1 Q1	2 Q2	3 Q3	4 Q4	5 Q5	6 Q6	7 Q7	8 Q8	9 Q9	10 Q10
4983	1	1	1	1	0	1	0	1	1	1
4991	1	1	1	1	1	0	1	0	1	1
5186	1	1	0	1	0	1	0	1	1	1
5170	0	1	1	1	1	0	0	1	0	1
5175	0	1	1	1	0	0	0	1	1	1
5178	0	0	1	1	0	1	0	1	1	1
5183	0	1	1	1	1	1	0	0	1	1
5191	1	1	0	1	1	1	0	1	1	1
5194	0	1	1	1	1	0	0	1	1	1
5199	1	1	1	1	0	1	0	1	1	1
5202	1	1	1	1	1	0	0	1	0	1
5207	0	0	1	1	1	1	0	1	1	1
5210	0	1	1	1	1	0	0	1	1	1
5215	1	1	1	1	0	0	0	1	1	1
5218	0	0	1	1	0	0	0	0	1	1
5223	0	1	1	0	0	0	0	0	0	1
5226	0	1	1	1	0	0	0	1	1	1
5231	0	1	1	1	0	1	1	1	1	1
5234	0	1	1	1	1	1	0	1	1	1
5239	1	1	1	0	0	0	0	0	1	1
5242	1	1	1	1	1	1	0	0	1	1
5247	0	0	1	0	0	0	0	1	1	1
5250	0	1	0	1	1	0	0	0	1	1

La matrice obtenue ne contient dès lors pas d'information sur la certitude avec laquelle l'étudiant a accompagné sa réponse. Cette matrice peut être qualifiée de « binaire ».

Une autre façon de présenter les données (Leclercq & Jans, 1999) consiste à inclure dans l'aspect correct ou incorrect la certitude des sujets en indiquant le pourcentage de certitude qui a accompagné la réponse. Le signe moins qui apparaît devant certains pourcentages de certitude montre que la réponse était incorrecte. Lorsqu'il n'y a pas de signe devant le pourcentage la réponse était correcte.

Ce type de matrice peut donc être qualifiée de « spectrale ». Sur base des informations contenues dans ce second type de matrice, nous pouvons calculer des indices de fidélité et les comparer avec les indices de fidélité obtenus à partir de la première matrice binaire. Nous avons inclus dans le logiciel *SCANTEST 2.0* un module de création de fichiers de matrices de résultats binaires « *Fichier Statistica avec réponses 1 ou 0* » ou de résultats spectraux « *Fichier Statistica avec réponses en %* ».


Données : PHYSIQ_T1__T_T_S.STA 10v * 2497c

VALEUR NUMER	1 Q1	2 Q2	3 Q3	4 Q4	5 Q5	6 Q6	7 Q7	8 Q8	9 Q9	10 Q10
4983	20	80	80	80	−80	60	−40	80	20	60
4991	80	100	100	80	0	0	60	−100	20	80
5186	40	20	−100	100	0	40	−20	20	100	100
5170	−100	80	60	100	40	0	−60	100	−60	100
5175	−20	80	20	60	−60	0	−20	100	20	40
5178	−80	−40	20	60	−20	60	−80	100	100	40
5183	−20	100	40	100	40	60	−20	−20	60	80
5191	100	20	−100	100	0	40	−100	100	100	100
5194	−40	100	100	100	0	0	−60	20	80	100
5199	60	60	60	100	−20	60	−60	100	80	60
5202	100	100	100	100	100	−100	−100	100	−100	100
5207	−60	−60	100	100	100	100	−40	100	100	100
5210	0	20	80	100	0	−20	−40	100	60	60
5215	20	60	40	80	−40	−60	−80	80	100	100
5218	−20	−40	60	60	−40	−40	−40	−40	60	80
5223	−20	100	100	−20	−20	−40	−40	−40	−60	40
5226	−100	100	100	100	0	−100	−100	100	80	100
5231	−60	80	100	100	−40	80	100	100	100	100
5234	−20	100	100	100	0	60	−60	60	80	80
5239	40	20	80	−80	−40	−20	−20	−20	20	40
5242	20	0	80	100	0	40	−20	−20	80	40
5247	0	0	0	0	0	0	0	0	0	0
5250	0	0	−100	80	0	−20	0	0	100	100

Ces fichiers sont destinés au logiciel *STATISTICA 5.1* produit par la firme *Statsoft.* Ces matrices peuvent être créés à partir d'une sélection d'étudiants établie à partir de la section ou de la faculté, ou à partir d'un niveau de réalisme donné. Dans ce dernier cas, nous avons alors, après transfert du fichier dans *STATISTICA 5.1*, la possibilité de calculer les indices de fidélité en relation avec les données issues de turbo analyses. L'analyse de la fidélité des fichiers binaires et spectraux est réalisée à l'aide du module « *Fiabilité d'échelle et analyse d'échelles* » du logiciel *STATISTICA 5.1*.

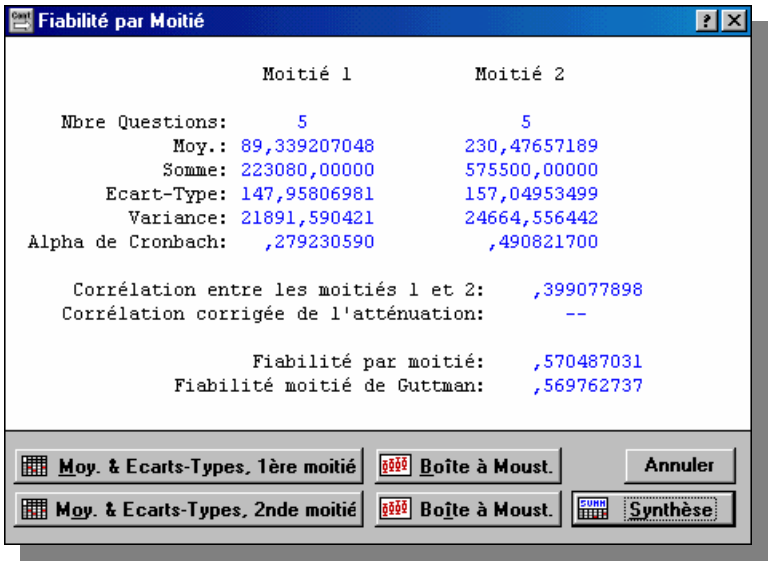
d) Utilisation des ressources du logiciel STATISTICA 5.1 produit par la firme Statsoft

Le module « *Fiabilité et analyse d'échelles* » du logiciel *STATISTICA 5.1* produit par la firme *Statsoft* permet de sélectionner les questions en vue de constituer les deux moitiés du test.

Dans l'exemple qui va suivre nous avons repris les 10 questions du test de physique soumis à 2.497 étudiants. Les items de chaque moitié ont été sélectionnés sur base du critère « pairs – impairs ». La première moitié contient donc les items impairs 1, 3, 5, 7 et 9, tandis que la seconde comprend les items pairs 2, 4, 6, 8 et 10.

Lorsque l'utilisateur a défini les sous-tests et après avoir sélectionné l'option *Fiabilité par moitié* dans les menus suivants, *STATISTICA 5.1* présente l'écran ci-dessous.

Les données sont en rapport avec la matrice spectrale (voir p. 132) créée à partir des résultats au test de physique (n = 2.497).



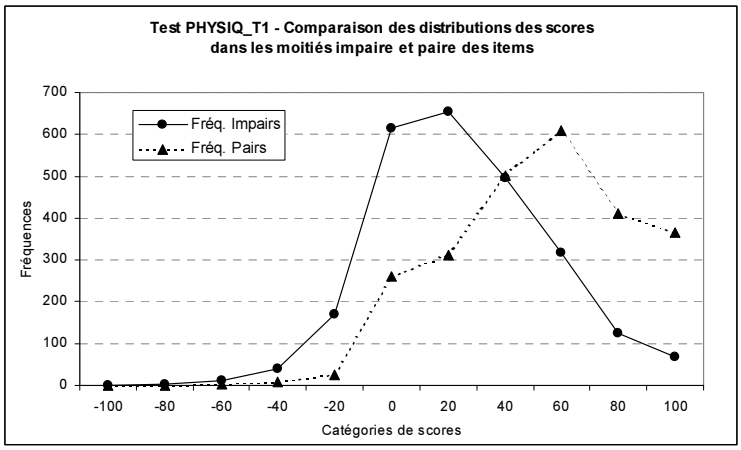
Voici la boîte de dialogue proposée dans *STATISTICA 5.1.* La moyenne de chaque moitié est calculée à partir de la somme des pourcentages des 5 items (pairs ou impairs en fonction de la moitié), il s'agit donc d'une moyenne à situer dans le contexte de cet exemple sur une plage variant de –500 (-100 * 5) à +500 (100 * 5). Remarquons les moyennes très différentes dans la moitié 1 (89,3) et dans la moitié 2 (230,5). Le logiciel donne également l'alpha de Cronbach pour chaque moitié (nous reviendrons plus loin sur cet indice de cohérence interne, p. 137) et nous observons une nette différence entre l'alpha de la 1ère moitié (0,28) et l'alpha de la 2ème moitié (0,49).

Le r de Bravais-Pearson vaut ici 0,399 et lorsqu'on lui applique la correction de Spearman-Brown, reprise sous l'appellation « *Fiabilité par moitié* », l'indice de fidélité vaut alors 0,570. L'indice de fiabilité par moitié de Guttman vaut quant à lui 0,569, valeur quasi identique à celle du Spearman-Brown.

e) Equilibrage des sous-tests en fonction de la facilité/difficulté des questions

La faible fidélité obtenue dans le cadre de cet exemple doit être interprétée avec prudence. Nous savons que les limites de validité du coefficient de corrélation *r* imposent des distributions symétriques ou qui possèdent le même type d'asymétrie.

Or nous constatons que les moyennes sont très différentes dans les deux moitiés, la 1ère moitié étant particulièrement difficile étant donné la moyenne de 89 (la plage variant entre – 500 et + 500) et la seconde moitié plus facile avec une moyenne de 230. Les coefficients d'asymétrie des distributions sont dans un cas positif et dans l'autre négatif : pour la 1ère moitié, celle des items impairs, il est égal à 0,30 et pour la seconde moitié, celle des items pairs, il vaut –0,18.



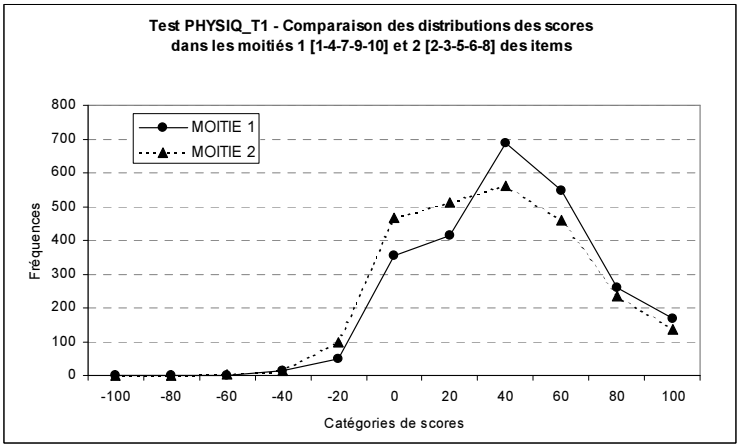
Le graphique des distributions des fréquences des moyennes des scores spectraux (minimum = -100 et maximum = 100) pour les deux moitiés constituées sur la base du critère « items pairs-impairs » illustre ce déséquilibre entre les deux moitiés.

	Moy.	Ec.Type
Q1	4,3	62,9
Q2	47,8	56,8
Q3	37,5	68,9
Q4	63,9	53,6
Q5	4,4	46,8
Q6	12	63,9
Q7	-20,3	61,4
Q8	43	52
Q9	63,4	48
Q10	63,6	45

Dès lors, nous devons reconsidérer la constitution des deux sous-tests en les équilibrant mieux du point de vue de la difficulté.

En nous basant sur les valeurs des moyennes et des écarts types des dix questions du test de physique, nous proposons une deuxième répartition des items en deux sous-tests. La 1ère moitié contenant les items 1-4-7-9-10 et la seconde moitié[57] les items 2-3-5-6-8.

Le graphique ci-contre des distributions des moyennes des scores aux deux moitiés constituées plus équilibrée du point de vue de la difficulté montre une amélioration de la situation. Les deux distributions sont plus proches (coefficients d'asymétrie : 0,06 pour la première moitié et 0,21 pour la seconde).

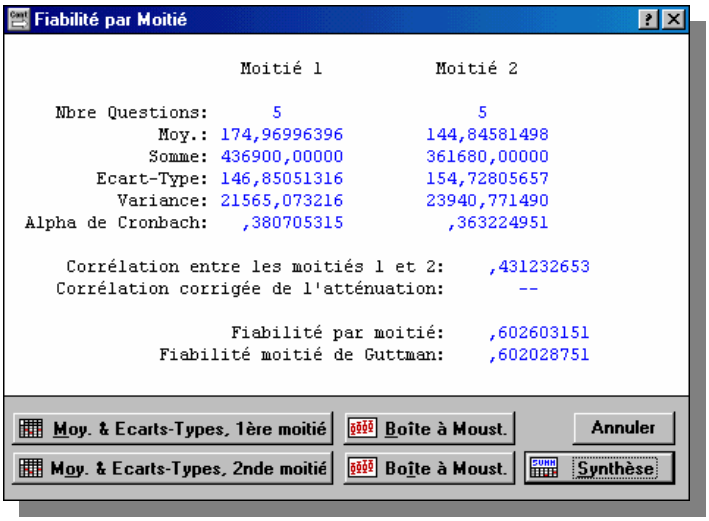


[57] Remarquons que d'autres sous-tests équilibrés en fonction de la facilité/difficulté auraient aussi pu être définis.

Les informations livrées par le module *Fiabilité et analyse d'échelles* de *STATISTICA* indiquent un rapprochement des moyennes : 174,9 pour la moitié 1 et 144,8 pour la moitié 2. De même qu'un rapprochement des écarts types et des variances : $\sigma_1 = 146{,}8$ et $\sigma_1^2 = 21565$ contre $\sigma_2 = 154{,}7$ et $\sigma_2^2 = 23940$.

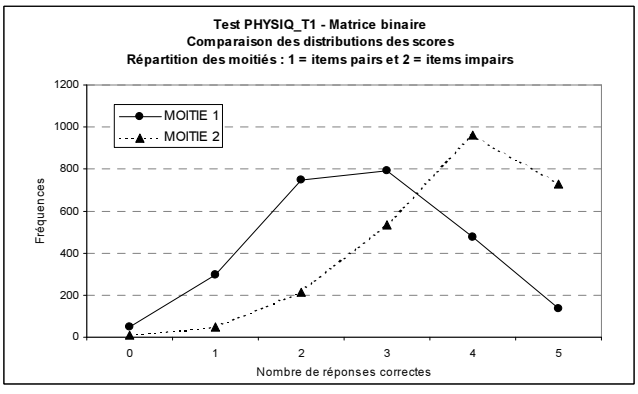
Les coefficients alpha de Cronbach se rapprochent également. Lorsqu'on compare avec les valeurs obtenues pour la répartition « pairs-impairs » des items, on observe une augmentation de l'alpha de la moitié 1 où il passe de 0,28 à 0,38 et une diminution pour la moitié 2 où il passe de 0,49 à 0,36.

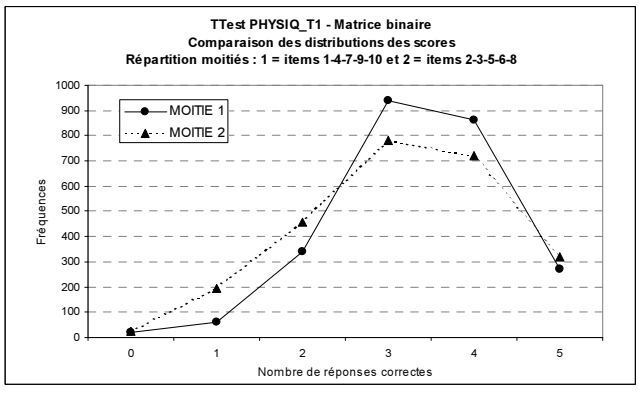
Le coefficient de corrélation augmente également de 0,40 à 0,43 ainsi que les indices de fidélité de Spearman-Brown qui et de Guttman qui passent tous les deux de 0,57 à 0,60.



f) Calcul des coefficients de bipartition sur les matrices binaires et spectrales

Les coefficients de fidélité présentés à partir de l'exemple du test de physique (n = 2.497) ont été calculés au départ des résultats spectraux. Nous pouvons aussi calculer ces coefficients de bipartition à partir des données binaires et comparer ensuite les valeurs obtenues à celles des données spectrales.





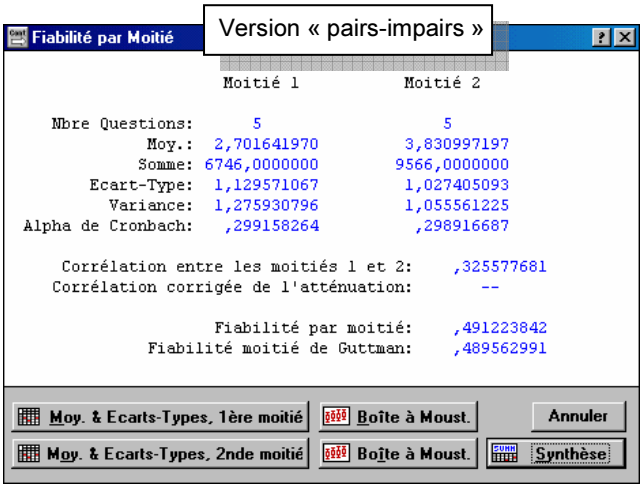
Avec une répartition « *pairs-impairs* » des items dans les deux moitiés, on observe le même type de problème que pour la matrice spectrale. Le graphique de gauche montre la distribution des nombres de réponses correctes pour les items impairs comparée à celle des items pairs. Le graphique de droite montre les distributions après équilibrage de chaque moitié en fonction de la difficulté des questions. La répartition est celle qui a déjà été employée pour la matrice spectrale, c'est-à-dire pour la moitié 1 : sélection des items 1-4-7-9-10 et pour la moitié 2 : items 2-3-5-6-8.

Le tableau ci-contre montre les différences de situation en fonction du type de répartition des items dans les moitiés. Tout comme pour la matrice spectrale, les moyennes et les coefficients d'asymétrie se rapprochent dans la version équilibrée.

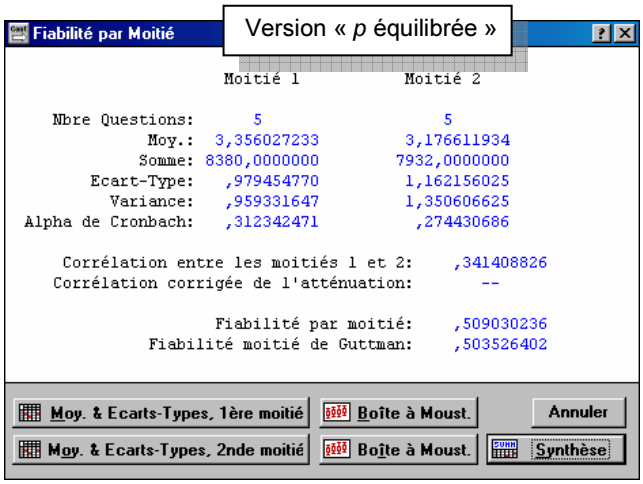
	Répartition pairs-impairs		Répartition équilibrée (*p*)	
	items 1-3-5-7-9	Items 2-4-6-8-10	items 1-4-7-9-10	Items 2-3-5-6-8
Moy.	2,7	3,8	3,4	3,2
E. t.	1,13	1,03	0,98	1,16
Asymétrie	0,00	-0,76	-0,41	-0,30

Les indices de fidélité s'améliorent légèrement : la corrélation entre les moitiés vaut 0,325 pour la version « *pairs-impairs* » et 0,341 pour la version « *p* équilibrée ». Avec la correction de Spearman-Brown, ces corrélations valent respectivement 0,491 et 0,509. Le coefficient de fidélité de Guttman vaut quant à lui 0,489 pour la répartition en items pairs et impairs et 0,503 pour la répartition selon les difficultés.

Version « pairs-impairs »



Version « *p* équilibrée »



g) Comparaison des coefficients de bipartition selon le type de matrice : binaire ou spectrale

Le tableau ci-dessous reprend les coefficients de bipartition obtenus au départ des matrices binaire et spectrale après répartition équilibrée des items en fonction de leur difficulté.

	MATRICE BINAIRE	MATRICE SPECTRALE
Coefficient de bipartition ($r_{xx'}$)	0,341	0,431
r avec correction Spearman-Brown (*rS*)	0,509	0,602
Coefficient de bipartition de Guttman (*rG*)	0,503	0,602

La comparaison des valeurs dans le cadre de l'exemple du test de physique ($n = 2.497$) montre de meilleurs coefficients de fidélité pour les données incluant les pourcentages de certitude des répondants.

6. Le coefficient alpha de Cronbach

Un des inconvénients liés à la méthode de bissection réside dans le fait qu'on obtient des coefficients de bipartition différents selon les items qui figurent dans chaque moitié. On obtient des valeurs différentes selon la répartition des items dans les deux moitiés du test.

Le coefficient alpha de Cronbach (Cronbach, 1951) repose sur le principe qu'il y a autant de parties dans un test que de questions, De Landsheere (1979, p. 42) rappelle à ce propos : « *Alpha est en fait la moyenne de tous les coefficients de bipartition (split-half coefficient) possibles pour un même test (Cronbach), dans le cas d'un parallélisme rigoureux entre les deux formes étudiées* ».

a) Méthode de calcul

La formule de calcul s'écrit :

$$\alpha = \frac{nq}{nq-1} . \left[1 - \frac{\sum_{q=1}^{nq} \sigma_q^2}{\sigma_t^2} \right] \quad (10)$$

avec q : l'indice des questions

nq : le nombre de questions dans le test

σ_q^2 : les variances des nq questions individuelles

σ_t^2 : la variance de la somme de toutes les questions du test

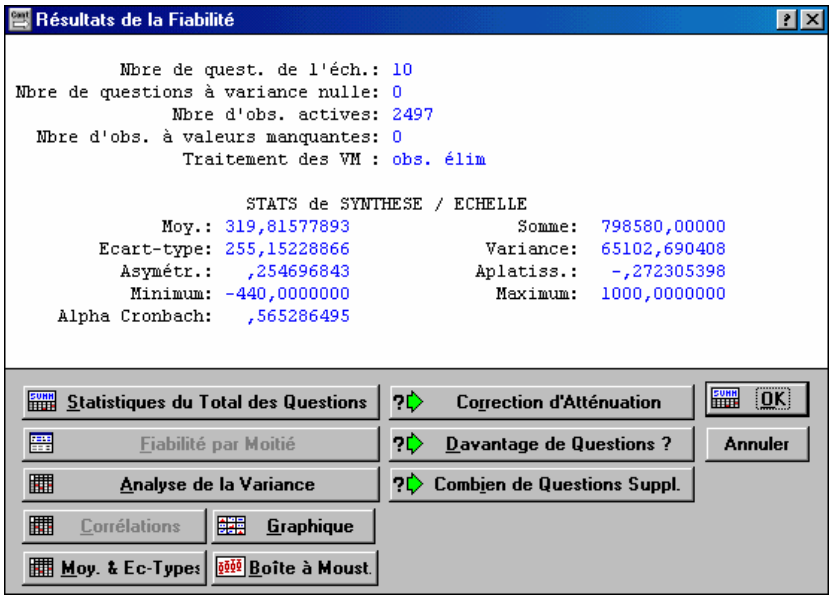
Dans le cas où aucune variance des scores observés est attribuable à la variance des scores vrais, où il n'y aurait donc aucune part de score vrai dans les scores observés mais uniquement de l'erreur, la variance de la somme de toutes les questions est alors identique à la somme des variances des questions individuelles et par conséquent l'alpha est égal à zéro

Dans l'hypothèse où toutes les questions seraient parfaitement fiables, où la cohérence interne du test serait parfaite, la part de score vrai dans les scores observés serait maximale et le coefficient alpha serait alors égal à 1. En fait, $1 - \frac{\sum_{q=1}^{nq} \sigma_q^2}{\sigma_t^2}$ deviendrait égal à $\frac{nq}{nq-1}$ et nous obtiendrions 1 pour alpha.

Laveault & Gregoire (1997, p. 146) rappellent que l'alpha sous-estime la fidélité : « *Le α de Cronbach repose sur le postulat fort que chaque item est parallèle aux autres (même degré de difficulté, même variance). Comme c'est rarement le cas dans la pratique, la valeur de fiabilité fournie par α sous-estime la fiabilité du score total au test. On peut donc affirmer que α est une valeur conservatrice de la cohérence interne du score total puisque $\alpha \leqslant r_{xx'}$* ».

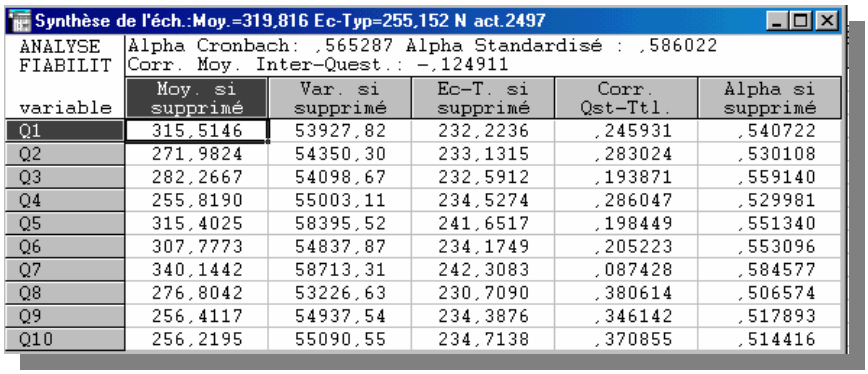
b) Indices calculés par *STATISTICA 5.1*

Le Module « *Fiabilité et analyse d'échelles* » du logiciel *STATISTICA 5.1 produit par la firme Statsoft* permet de calculer rapidement l'alpha de Cronbach ainsi qu'une série de statistiques en lien avec cet indice de cohérence interne. Il offre par exemple, la possibilité du calcul du nombre de questions supplémentaires (similaires à celle déjà contenues dans le test) nécessaires pour obtenir un niveau alpha donné. Les statistiques ci-contre concernent les données spectrales du test de physique (n = 2.497). Les moyenne, somme, écart type, variance, coefficients d'asymétrie et d'aplatissement, minimum et maximum sont calculés à partir des scores totaux de tous les items confondus.



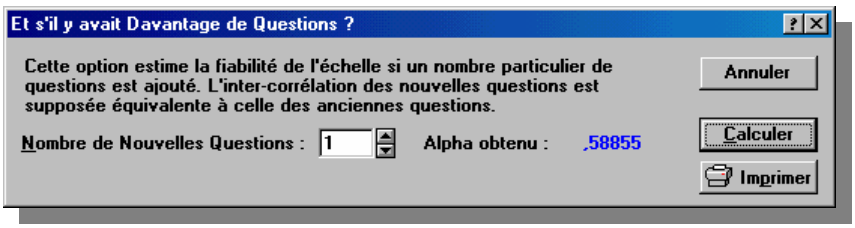
L'alpha est égal à 0,56 ce qui est peu élevé, on considère habituellement qu'un niveau alpha correct se situe au minimum à 0,80.

STATISTICA 5.1 permet aussi de simuler les résultats qu'obtiendrait l'épreuve si une question donnée était supprimée. On voit par exemple que pour ce test, et avec les scores calculés sur base des pourcentages de certitude, si la question 7 était supprimée, le coefficient alpha remonterait à 0,58.



variable	Moy. si supprimé	Var. si supprimé	Ec-T. si supprimé	Corr. Qst-Ttl.	Alpha si supprimé
Q1	315,5146	53927,82	232,2236	,245931	,540722
Q2	271,9824	54350,30	233,1315	,283024	,530108
Q3	282,2667	54098,67	232,5912	,193871	,559140
Q4	255,8190	55003,11	234,5274	,286047	,529981
Q5	315,4025	58395,52	241,6517	,198449	,551340
Q6	307,7773	54837,87	234,1749	,205223	,553096
Q7	340,1442	58713,31	242,3083	,087428	,584577
Q8	276,8042	53226,63	230,7090	,380614	,506574
Q9	256,4117	54937,54	234,3876	,346142	,517893
Q10	256,2195	55090,55	234,7138	,370855	,514416

Une autre option permet de calculer la valeur de l'alpha si « x » questions parallèles étaient ajoutées à celles déjà présentes dans l'épreuve. On voit dans l'exemple ci-contre qu'une question parallèle supplémentaire entraînerait pour le test de physique une amélioration de l'alpha qui augmente légèrement en passant de 0,565 à 0,588.



Le nouvel alpha est calculé en utilisant la formule de Spearman-Brown déjà utilisée dans le cadre de la correction pour sous-estimation du coefficient de bipartition (p. 131) :

$$\alpha S = \frac{k\alpha}{1+(k-1).\alpha} \quad (11)$$

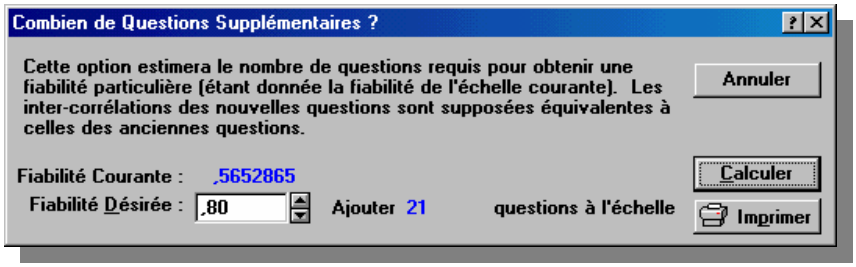
avec α : l'alpha de Cronbach
k : la proportion d'allongement du test

En attribuant à k la proportion d'allongement du test lorsqu'on y ajoute une question parallèle, on obtient la valeur du nouvel indice alpha. Dans l'exemple du test de physique on désire connaître l'alpha en simulant l'ajout d'une question parallèle dans un test qui en contient 10. Pour cette simulation l'épreuve contient ainsi 11 items. Cette allongement représente en terme de proportion : 11/10 = 1,1, dès lors le calcul du nouvel alpha devient :

$$\alpha S = \frac{1{,}1.0{,}5652865}{1+(1{,}1-1).0{,}5652865} = \frac{0{,}6218152}{1+0{,}1.0{,}5652865} = \frac{0{,}6218152}{1+0{,}0565287} = \frac{0{,}6218152}{1{,}0565287} = 0{,}5885454$$

c) Coefficient d'allongement du test pour une fidélité désirée

Une autre option du module « *Fiabilité et analyse d'échelle* » du logiciel *STATISTICA 5.1* permet le calcul du nombre de questions supplémentaires (similaires à celles qui existent déjà) qu'il faudrait prévoir pour atteindre un seuil de fiabilité désiré. On voit que dans le cadre de notre exemple il faudrait 21 questions supplémentaires similaires à celles qui existent déjà pour atteindre un alpha de 0,80.



Ce coefficient d'allongement du test pour une fidélité désirée est obtenu après transformation de la formule de Spearman-Brown. On isole k de façon à déterminer la proportion représentant le nombre d'items nécessaires, la formule devient :

$$k = \frac{r_{xx'}(1-r_{yy'})}{r_{yy'}(1-r_{xx'})} \quad (12)$$

avec :

k = la proportion de questions nécessaires pour atteindre le coefficient de fidélité désiré
$r_{xx'}$ = le coefficient de fidélité désiré
$r_{yy'}$ = le coefficient de fidélité actuel

Pour l'exemple du test physique cela donne :

$$k = \frac{0{,}8.(1-0{,}56528)}{0{,}56528.(1-0{,}8)} = \frac{0{,}8.0{,}43472}{0{,}56528.0{,}2} = \frac{0{,}347776}{0{,}113056} = 3{,}0761$$

Cela signifie que le test devrait être trois fois plus long pour obtenir un coefficient de fidélité de 0,80. Les 10 items de la version de départ doivent être multipliés par 3,0761, ce qui donne 30,7 donc 31 questions au total et donc 21 items parallèles à ajouter aux 10 qui existent déjà, comme nous l'indique le logiciel *STATISTICA 5.1*.

Dès lors le nombre de questions à ajouter au test pour atteindre un alpha de 0,8 ($kq[\alpha 0,8]$) se calcule :

$$\boxed{kq\,[\alpha 0,8] = (nq\,.\,k) - nq} \quad (13)$$

avec :

k = la proportion de questions nécessaires pour atteindre un alpha de 0,8

nq = le nombre de questions qui figurent dans le test

Ce qui donne dans le cadre de notre exemple :

$$kq[\alpha 0,8] = (10\,.\,3,0761) - 10 = 30,761 - 10 = 20,761 \approx 21$$

Remarquons que nous pouvons aussi utiliser cette formule pour calculer combien de questions il faudrait retrancher lorsque l'alpha est plus élevé que 0,8. Par exemple, si le test de 10 questions obtenait un alpha de 0,85, nous pourrions connaître le nombre de questions parallèles à retirer du test pour obtenir un alpha inférieur de 0,8 :

$$k = \frac{0,8.(1-0,85)}{0,85.(1-0,8)} = \frac{0,8.0,15}{0,85.0,2} = \frac{0,12}{0,17} = 0,706$$

et

$$kq[\alpha 0,8] = (10\,.\,0,706) - 10 = 7,06 - 10 = -2,94 \approx -3$$

Dans ce cadre de cet exemple il faudrait donc retirer 3 questions parallèles pour obtenir un alpha égal à 0,8.

d) Comparaison des alpha de Cronbach calculés sur matrices spectrale et binaire

Les informations liées à l'alpha de Cronbach peuvent être calculées sur base des données spectrales ou des données binaires. Voici un comparatif des valeurs obtenues dans le cadre du test de physique (n = 2497) :

	Données binaires	**Données spectrales**
Moyenne	6,53	320
Somme	16312	798580
Ecart type	1,76	255
Variance	3,08	65103
Asymétrie	-0,29	0,25
Aplatissement	-0,35	-0,27
Minimum	0	-440
Maximum	10	1000
Alpha de Cronbach	**0,47**	**0,57**
Alpha si ajout d'une question	**0,50**	**0,59**
Combien de questions supplémentaires pour α = 0,8	**35**	**21**

La valeur du coefficient alpha de Cronbach est plus élevée lorsqu'il a été calculé à l'aide des données spectrales (α = 0,59) plutôt qu'en utilisant les données binaires (α = 0,47). Dans le cas des données spectrales l'ajout d'une question parallèle entraîne une plus forte augmentation de l'alpha (0,59 contre 0,50) et il faut moins de questions supplémentaires pour atteindre un alpha de 0,8 (21 items en plus contre 35 dans la cas d'une estimation à partir des données binaires). Nous avons comparé les valeurs des coefficients alpha et des autres indices de fidélité des 10 épreuves MOHICAN plus loin dans cette étude (voir p. 309).

e) Alpha obtenu par le test lorsqu'on retire la question *q*

Nous pouvons aussi comparer la corrélation de chaque question avec le total du test et les valeurs alpha obtenues par l'épreuve si la question envisagée était supprimée :

	Corrélation Qst-Total		Alpha si supprimé	
	BINAIRE	SPECTRAL	BINAIRE	SPECTRAL
q1	,26	,25	,420	,541
q2	,17	,28	,452	,530
q3	,14	,19	,464	,559
q4	,19	,29	,450	,530
q5	,18	,20	,451	,551
q6	,19	,21	,450	,553
q7	,20	,09	,444	,585
q8	,24	,38	,430	,507
q9	,18	,35	,453	,518
q10	,19	,37	,451	,514

La dernière colonne du tableau montre que l'alpha le plus élevé pourrait être obtenu après suppression de la question 7 dans le cadre d'une estimation à partir des données spectrales (α spectral Q7 = 0,585). On voit que ce manque de cohérence interne de la question 7 établie à partir des données spectrales ne se retrouve pas dans les valeurs de l'alpha calculées à partir des données binaires (α binaire Q7 = 0,444). Dans le cas des données binaires, l'alpha le plus élevé serait obtenu après suppression de la question 3 (α binaire Q3 = 0,464). Les conclusions auxquelles nous pourrions aboutir quant à la suppression de certaines questions ne coïncident donc pas et dépendent du type de résultats à partir desquels nous calculons l'alpha de Cronbach.

f) Corrélation question *vs* total sans la question envisagée

Dans le tableau ci-dessus, les colonnes « Corrélation Qst-Total » donnent les corrélations entre la question respective et le score de la somme totale (sans la question envisagée). Nous remarquons que dans le cas de ce test de physique soumis à 2.497 étudiants les corrélations sont peu élevées. Les corrélations sont plus élevées lorsqu'elles sont calculées au départ des données spectrales sauf pour la question 1 (quoique dans cet item elles sont pratiquement identiques) et 7.

Remarquons que lorsqu'il est calculé à partir des données binaires, l'item 7 est corrélé à 0,20 alors qu'avec les données spectrales la corrélation descend à 0,09.

7. Matrices des covariances et des corrélations

a) Principe de construction de la matrice des covariances

Rappelons la formule de calcul des covariances :

Données spectrales – Matrice des covariances

	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10
q1	3962									
q2	261	3230								
q3	269	471	4759							
q4	461	447	402	2878						
q5	472	234	157	161	2192					
q6	728	329	117	489	221	4093				
q7	127	354	195	-119	277	57	3765			
q8	465	695	629	470	337	288	455	2710		
q9	435	447	389	707	241	547	-117	502	2322	
q10	380	511	482	581	146	299	70	732	760	2069

$$\operatorname{cov} q_x q_y = \frac{1}{ns}\sum_{s=1}^{ns}(q_{x\,s} - \overline{q}_x)(q_{y\,s} - \overline{q}_y) \quad (14)$$

avec : s = l'indice des sujets
ns = le nombre de sujets
q_{x_s} = le score du sujet s à la QCM x
q_{y_s} = le score du sujet s à la QCM y
$\overline{q}_x$ = moyenne des scores QCM x
$\overline{q}_y$ = moyenne des scores QCM y

La covariance est la moyenne du produit des écarts des données par rapport à leurs moyennes respectives. La covariance est donc une mesure de la relation existant entre deux séries de données. De Landsheere (1979, p.63) citant Dagnelie, rappelle : « *La covariance de deux séries d'observations est positive ou négative « selon que la relation entre les deux séries de données est croissante ou décroissante, c'est-à-dire selon que les valeurs élevées d'une série correspondent, dans l'ensemble, aux valeurs élevées ou aux valeurs peu élevées de l'autre* ». Ajoutons que si les deux plages de données sont indépendantes, la valeur de la covariance sera proche de zéro.

La matrice ci-dessus contient les covariances des scores spectraux qui intègrent les pourcentages de certitude en plus de l'information sur l'aspect correct ou non des réponses (voir p. 132). La diagonale (3962, 3230, 4759, …) reprend les valeurs des variances de chacune des 10 questions du test. Etant donné que les covariances « $q_x q_y$ » sont identiques aux covariances « $q_y q_x$ », les mêmes valeurs se répètent symétriquement par rapport à la diagonale de la matrice et nous ne présentons que les valeurs qui figurent sous la diagonale.

Les matrices de covariances des tests MOHICAN s'interprètent relativement facilement étant donné que les items ont tous la même étendue (dans le cas des matrices spectrales la marge de variation des scores est de –100 à 100). Le module « *Fiabilité et analyse d'échelles* » de *STATISTICA 5.1* nous donne la matrice des covariances.

Nous observons sur cette matrice liée au test de physique (n = 2.497) et calculée à l'aide des scores spectraux (avec pourcentages de certitude), deux covariances négatives, l'une entre l'item 4 et l'item 7 et l'autre entre l'item 9 et l'item 7, ces items contribuent moins à la cohérence interne de l'épreuve.

b) Principe de construction de la matrice des corrélations

La matrice des corrélations (coefficients *r* de Bravais-Pearson, p. 150) nous donne aussi de précieuses indications sur la relation entre les questions de l'épreuve.

Données spectrales – Matrice des corrélations

	q1	q2	q3	q4	q5	q6	q7	q8	q9
q2	,07								
q3	,06	,12							
q4	,14	,15	,11						
q5	,16	,09	,05	,06					
q6	,18	,09	,03	,14	,07				
q7	,03	,10	,05	-,04	,10	,01			
q8	,14	,24	,18	,17	,14	,09	,14		
q9	,14	,16	,12	,27	,11	,18	-,04	,20	
q10	,13	,20	,15	,24	,07	,10	,02	,31	,35

Nous observons dans cette matrice des corrélations calculées à partir des scores spectraux, des corrélations qui ne sont guère élevées. Les corrélations entre les questions 4 et 7 et entre les questions 7 et 9 sont négatives, mais en fait assez proches de zéro. La matrice des corrélations est également fournie par le module « *Fiabilité et analyse d'échelles* » de *STATISTICA 5.1.*

c) matrices des covariances et de corrélations des scores binaires

Nous avons introduit les matrices des covariances et des corrélations en utilisant les résultats spectraux. Nous allons maintenant présenter les matrices obtenues à partir des résultats binaires (p. 132).

Données binaires – Matrice des covariances

	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10
q1	,25									
q2	,02	,16								
q3	,01	,01	,21							
q4	,02	,01	,01	,12						
q5	,04	,01	,01	,01	,24					
q6	,04	,02	,01	,02	,02	,25				
q7	,02	,02	,01	,00	,02	,01	,18			
q8	,03	,02	,02	,01	,02	,01	,03	,19		
q9	,01	,01	,01	,01	,01	,01	,00	,01	,09	
q10	,01	,00	,01	,01	,00	,01	,01	,02	,01	,08

Lorsqu'on calcule le tableau carré des covariances des 10 questions du test de physique (n = 2.497) à partir des données fournies par une correction dichotomique (réponse correcte = 1 ou incorrecte = 0), on aboutit à une matrice telle que celle qui est présentée ci-contre.

Les covariances entre les questions sont peu élevées et lorsqu'on compare avec la matrice obtenue à partir des données spectrales (voir point précédent), on constate que les covariances négatives n'apparaissent plus sur cette matrice binaire.

En ce qui concerne la matrice des corrélations calculées à partir des données binaires, nous observons des corrélations peu élevées.

Données binaires – Matrice des corrélations

	q1	q2	q3	q4	q5	q6	q7	q8	q9
q2	,08								
q3	,06	,06							
q4	,14	,07	,09						
q5	,15	,06	,05	,05					
q6	,16	,09	,03	,10	,07				
q7	,09	,12	,07	,02	,10	,07			
q8	,12	,10	,09	,06	,11	,06	,17		
q9	,09	,04	,06	,12	,08	,10	,03	,07	
q10	,09	,04	,09	,09	,03	,07	,09	,14	,12

Un avantage de la matrice des corrélations par rapport à la matrice des covariances réside dans le fait qu'elle facilite la comparaison des valeurs binaires et spectrales étant donné l'étendue des scores de -100 à 100 pour la matrice spectrale et les scores dichotomiques (0 ou 1) pour la matrice binaire.

La comparaison des corrélations de la matrice binaire avec celles de la matrice spectrale montre des valeurs en moyenne un peu plus faibles pour la matrice binaire. Nous constatons aussi que les corrélations négatives n'apparaissent plus sur cette dernière matrice binaire.

8. Conventions pour la notation des indices de fidélité dans le cadre de cette recherche

Etant donné les particularités liées à l'emploi de matrices binaires et spectrales dans le cadre de l'évaluation de la fidélité des épreuves MOHICAN, nous proposons une notation particulière pour les indices présentés dans ce chapitre.

D'une part, nous ajouterons en indice « *mb* » aux indices classiques pour signifier que les calculs ont été réalisés au départ de *m*atrices *b*inaires.

D'autre part, nous utiliserons « *ms* » pour notifier l'emploi de *m*atrices *s*pectrales dans le calcul des indices de fidélité.

Voici le récapitulatif des notations qui seront utilisées à partir d'ici (nous indiquons entre parenthèses les pages où ces indices sont exposés) :

- $r_{xx'}$ = coefficient de bipartition d'un *t*est (p. 130) ;
- $rS_{xx'}$ = correction de Spearman-Brown appliquée au coefficient de bipartition (p. 131) ;
- rG = coefficient de bipartition de Guttman (p. 131) ;
- α = coefficient alpha de Cronbach (p. 137) ;
- $kq[\alpha 0,8]$ = nombre de questions à ajouter ou retrancher au test pour obtenir un alpha égal à 0,8 (p. 139) ;
- α-q= alpha obtenu par le test lorsqu'on retire la question q (p. 141) ;
- r_{qt} = corrélation question-total (p. 141) ;
- r_{qxqy} = corrélation des scores d'une question x avec les scores d'une question y (p. 143) ;
- cov_{qxqy} = covariance des scores d'une question x avec les scores d'une question y (p. 142).

D. Principaux indices de discrimination des items

1. Les indices de discrimination pour la mesure critériée

Dans une perspective de pédagogie de maîtrise, l'évaluateur sera soucieux de vérifier que les questions d'un test de maîtrise permettent de différencier les étudiants qui maîtrisent un objectif de ceux qui ne le maîtrisent pas. Laveault & Gregoire (1997, p. 245) soulignent : « *Les items les plus utiles en mesure critériée sont ceux qui sont les plus sensibles à l'enseignement. Si l'enseignement a été profitable, le degré de difficulté de ces items devrait changer considérablement. De plus, lorsque nous devons nous prononcer sur la maîtrise d'un objectif, ces items devraient nous permettre de prendre des décisions appropriées. Enfin, si les items en question proviennent d'un même domaine d'items, ils devraient être réussis ou échoués conjointement.* »

a) L'indice de sensibilité à l'enseignement

L'indice proposé par Cox et Vargas (1966) mesure les gains bruts (voire les pertes) obtenus après un apprentissage et permet de mettre en évidence les items les plus influencés par l'enseignement.

(1) Principe

L'indice de Sensibilité *S* est calculé en faisant la différence entre le score de difficulté d'un item, la proportion *p* de sujets qui ont répondu correctement à l'item (donc entre 0 et 1) avant et après la situation d'apprentissage.

$$S = p_{post} - p_{pré} \quad (15)$$

Avec

p_{post} : la mesure de la difficulté de l'item après l'apprentissage (post-test) ;

$p_{\text{pré}}$: la mesure de la difficulté de l'item avant l'apprentissage (pré test).

(2) Interprétation

Selon Laveault & Gregoire (1997, p. 246), « *Plus l'écart S est élevé, plus la mesure porte sur des items permettant de mesurer l'effet de l'enseignement. Moins S est élevé, moins l'item est utile car il a porté sur une question qui était tout aussi bien réussie avant qu'après. Un tel item ne permet pas de discriminer l'effet de l'enseignement*. ». Cette interprétation doit cependant être fortement nuancée car, comme le postulent ces auteurs, l'enseignement influence le score de difficulté d'un item après l'apprentissage et un *S* peu élevé peut parfois autant être expliqué par un enseignement déficient que par un item peu discriminant… et ces auteurs d'ajouter plus loin : « *Une valeur négative de S ou une valeur de 0 peuvent être interprétées de deux façons : (1) L'item ne convient pas, car il ne porte pas sur l'enseignement. (2) L'enseignement n'a eu aucun effet sur la réussite des élèves* ».

(3) Affinements possibles

Un rapport entre l'amélioration (moyennant les nuances apportées plus haut) de l'item et l'amélioration maximale possible pourrait aussi être calculé en se basant sur la formule de Mac Guigan (1967). L'indice de sensibilité deviendrait alors un indice de Sensibilité Relative que nous proposons de noter *SR*.

$$SR_1 = \frac{p_{post} - p_{pré}}{1 - p_{pré}} \quad (16)$$

Avec

p_{post} : la mesure de la difficulté de l'item après l'apprentissage (post-test) ;

$p_{pré}$: la mesure de la difficulté de l'item avant l'apprentissage (pré test).

Remarquons aussi que dans le cas de figure d'un indice de sensibilité relative, en cas de perte, la formule devrait être légèrement modifiée comme le préconise D'Hainaut (1973) dans la problématique des procédures de calcul des gains relatifs d'apprentissage :

$$SR_2 = \frac{p_{post} - p_{pré}}{p_{pré}} \quad (17)$$

b) L'indice de discrimination au seuil de maîtrise ou l'indice *B*

Brennan (1972) propose un indice similaire à l'indice *D* (voir *infra*), mais où la constitution du groupe « fort » et du groupe « faible », se fait sur base de l'atteinte ou non d'un seuil de maîtrise au score total du test. La valeur de *B* est calculée en soustrayant le score de difficulté de l'item (somme des scores à l'item divisée par le nombre de sujets) obtenu à partir des résultats des étudiants qui n'ont pas atteint le seuil de maîtrise au test par le score de difficulté calculé à l'aide des données de ceux qui ont atteint ce seuil de maîtrise.

(1) Principe

La formule de calcul de l'indice *B* est simple :

$$B = p_{M+} - p_{M-} \quad (18)$$

Avec

p_{M+} : proportion de sujets qui ont atteint le seuil de Maîtrise pour l'ensemble des questions du test et qui ont répondu correctement à l'item ;

p_{M-} : proportion de sujets qui n'ont pas atteint le seuil de Maîtrise et qui ont répondu correctement à l'item.

(2) Interprétation

L'indice, *B* peut varier de –1 à 1. Plus l'écart *B* est grand, plus l'item discrimine entre les étudiants qui atteignent le seuil de maîtrise à l'ensemble des items du test (groupe supérieur M+) et ceux qui n'atteignent pas ce seuil (groupe inférieur M-). Une valeur négative indique une situation paradoxale où l'item a été réussi par plus d'étudiants sous le seuil de maîtrise que par d'étudiants qui ont atteint ce seuil. Une valeur 0 indique que l'item ne discrimine pas entre les étudiants du groupe supérieur et du groupe inférieur.

2. Les indices de discrimination D

a) La méthode des deux groupes extrêmes (*D, D net, Upper Lower*)

Lorsqu'il s'agit d'évaluer le pouvoir discriminatif d'une question en se référant à sa capacité à distinguer les scores élevés des scores faibles au test, la méthode de l'indice *D* se révèle particulièrement facile à mettre en œuvre. Leclercq (1987), signale que l'indice *D* est aussi appelé *D net* (de l'anglais *Net Discrimination Index*) et qu'il repose sur les travaux de Kelley (1939), Flanagan (1939) et de Findley (1956).

(1) Principe

Le principe de calcul des indices *D* et *B* sont similaires. L'indice *D* consiste à calculer la différence entre l'indice de difficulté de l'item (la proportion de sujets qui ont répondu correctement) obtenu par un groupe d'étudiants particulièrement « forts » à l'ensemble des questions du test, le Groupe Supérieur (GS) et un autre groupe d'étudiants particulièrement « faibles », qui obtiennent les notes les moins élevées au total du test, le Groupe Inférieur (GI). Kelley (1939) propose de prendre les 27% des sujets les meilleurs au test pour la constitution du GS et pour la constitution du GI, les 27% des sujets qui obtiennent les scores les plus faibles. Wiersma et Jurs (1990) rappellent : « *27% is used because it has shown that this value will maximise differences in normal distributions while providing enough cases of analysis* ». Leclercq (1987, p. 70) suggère d'utiliser 25%.

Findley (1956), propose de calculer l'indice *D* de la façon suivante :

$$D = p_+ - p_- \quad (19)$$

Avec
p_+ : la proportion de sujets qui appartiennent à GS et qui ont répondu correctement à l'item ;
p_- : la proportion de sujets qui appartiennent à GI et qui ont répondu correctement à l'item.
Les effectifs de GS et de GI sont équivalents (27% de l'effectif total pour Kelley).

(2) Interprétation

L'indice varie entre 1 et –1. Plus l'écart *D* est grand, plus l'item discrimine entre les « forts » et les « faibles ». Une valeur négative indique une situation paradoxale où l'item a été réussi par plus d'étudiants « faibles » que par d'étudiants « forts ». Une valeur 0 indique que l'item ne discrimine pas les étudiants du GS et du GI.

Ebel (1965) propose des valeurs repères pour interpréter l'indice *D* :
- $D \geqslant 0{,}4$: l'item discrimine très bien ;
- $0{,}30 \leqslant D \leqslant 0{,}39$: l'item discrimine bien ;
- $0{,}20 \leqslant D \leqslant 0{,}29$: l'item discrimine peu ;
- $0{,}10 \leqslant D \leqslant 0{,}19$: l'item est à améliorer ;
- $D < 0{,}10$: l'item est sans utilité pour l'examen.

Selon Leclercq (1987), l'indice *D net* calculé sur peu d'étudiants constitue une mesure grossière de la discrimination et que plus le nombre de sujets augmente, plus l'indice se rapproche de la valeur du rpbis.

L'indice *D* est particulièrement utile lorsqu'on ne dispose pas d'ordinateur et que l'on est amené à calculer manuellement le pouvoir discriminatif des questions d'un test. Laveault & Gregoire (1997) font remarquer que « *L'indice D … ne porte que sur la moitié des données (54%), ce qui diminue le travail de calcul* ».

b) La méthode des quatre groupes

(1) Principe

Cette méthode est proposée par De Landsheere (1980). Les quatre groupes sont constitués en fonction du rendement scolaire global des élèves, de A les plus « forts » à D les plus « faibles ». Pour chaque item du test on détermine alors le nombre de sujets par groupe qui répondent correctement. On calcule ensuite la moyenne obtenue dans chaque groupe pour chaque question et enfin, on synthétise ces données dans un tableau en vue de comparer les moyennes pour chaque question dans les quatre groupes.

Voici un exemple de tableau récapitulatif proposé par De Landsheere (1980, p. 94) :

Question	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
Maximum	6	6	6	6	10	8	8	50
Groupe A	4,9	5,1	5,3	5,2	4,3	5,8	4,3	34,9
Groupe B	3,2	4,9	5,8	4,1	4,7	5,9	3,8	32,4
Groupe C	3,4	4,6	5,4	3	3,4	3,6	3,9	27,3
Groupe D	2	3,4	5	2,8	2,3	4,1	2,3	21,9
Ordre des scores moyens	A C B D	A B C D	B C A D	A B C D	B A C D	B A C D	A C B D	A B C D
Discrimine	NON	OUI	NON	OUI	NON	NON	NON	OUI

(2) Interprétation

Les quatre moyennes doivent s'ordonner du groupe A au groupe D sinon on considère que l'item ne discrimine pas.

Par exemple, pour la question 1, le groupe B obtient une moins bonne moyenne (3,2) que le groupe C (3,4), dès lors l'ordre moyen des scores pour cet item devient A-C-B-D et la question devient suspecte. Par contre, pour la question 2, les moyennes s'ordonnent en fonction des quatre groupes et on considère alors que l'item contribue à la discrimination totale du test.

Cependant nous pensons qu'il faut nuancer la dernière ligne du tableau car si deux groupes obtiennent des scores moyens très proches il devient difficile de les discriminer, dès lors, une solution pourrait consister à proposer trois catégories au lieu de quatre.

(3) Affinements

De Landsheere (1980) propose une solution graphique simple qui permet de nuancer les résultats obtenus avec cette méthode. Elle consiste à prendre en compte les scores individuels des élèves de chaque groupe et à les pointer dans un tableau reprenant en abscisses les groupes et en ordonnées la cote obtenue par chaque sujet. Pour une question cela donne :

Score obtenu	Groupe A	Groupe B	Groupe C	Groupe D
6	IIIIIIIIII	IIIIII	II	IIIIIIII
5	IIIIIIIII	IIII	II	
4	II	II	IIII	II
3	II	II	II	
2		II	II	II
1				II
0		II	IIII	IIIIIIII

Ensuite, il faut indiquer dans chaque groupe le quartile supérieur « QS » (la note au milieu de la moitié supérieure de la série de notes ordonnées), le médian « M » (la note située au milieu de la série) et le quartile inférieur « QI » (la note du milieu de la moitié inférieure). Ce qui donne pour l'exemple de la question :

Score obtenu	Groupe A	Groupe B	Groupe C	Groupe D
6	QS	QS		QS
5	M	M	QS	
4	QI		M	
3		QI		M
2			QI	
1				QI
0				

Idéalement nous devrions assister à une descente des QS, M et QI. Selon De Landsheere (1980, p. 96), « *La descente est plus ou moins accusée selon le pouvoir discriminatif. Une montée indique qu'un groupe inférieur (d'après le résultat de l'ensemble de l'examen) obtient de meilleurs points que le groupe supérieur.* ». Dans l'exemple on observe une montée pour le groupe D (le plus « faible ») et on peut en déduire que les correcteurs ont été « trop généreux » pour les meilleurs (QS) du groupe « faible ».

Leclercq (1987, p. 69) signale « *Si de telles analyses étaient pratiquées à partir d'un nombre suffisant de sujets (30 dans chaque groupe), elles permettraient de repérer comme le font les rpbis, les questions ambiguës ou mesurant une autre variable que l'ensemble des autres questions* ».

3. Les indices corrélationnels de discrimination

L'évaluation de la tendance de l'item à discriminer les sujets qui, en moyenne, obtiennent un nombre total de réponses correctes plus élevés que les autres sujets peut s'effectuer à l'aide d'une corrélation entre le score à l'item et le score total. Le coefficient de Bravais-Pearson[58] (défini ci-après) pourrait être utilisé pour calculer ces corrélations, mais il requiert en principe des échelles de mesure continues et d'après Laveault & Gregoire (1997, p. 234), « *Lorsque l'item est corrigé de manière dichotomique (0,1) ou encore de manière ordinale (A, B, C, D, E ou encore 0, 1, 2, 3, 4 points), le r de Pearson ne fournit pas une valeur exacte de la corrélation entre deux variables* ». Des alternatives existent : le coefficient de corrélation ρ (*rho*) de Spearman, les corrélations bisériales (rbis) et point bisériales (rpbis), le coefficient ϕ (*phi*) et la corrélation tétrachorique. Leur emploi dépend des postulats qui sont fait sur l'échelle de mesure utilisée.

Dans cette partie nous aborderons succinctement ces différents indices de discrimination des items. Nous reviendrons plus en détail sur la corrélation point bisériale dans la partie suivante car le rpbis Spectral est bâti sur le principe de cet indice de discrimination.

a) Variables métriques : le coefficient *r* de Bravais-Pearson

(1) Champ d'application

Le coefficient de corrélation *r* de Bravais-Pearson permet de corréler des mesures qui n'ont pas obligatoirement les mêmes dimensions et qui peuvent être distribuées différemment (Landercy, 1981, 1983). Il est utilisé lorsque les deux variables dont il faut évaluer la corrélation sont continues (quand théoriquement une valeur est toujours possible entre deux autres).

(2) Formule de calcul

Considérons deux séries de scores X_j et Y_j (j : 1,…, n) de deux variables X et Y. Soient $\overline{X}$ et $\overline{Y}$ leurs valeurs moyennes et σ_x et σ_y leurs variances, la formule de base pour la corrélation s'écrit :

$$r = \frac{\sum_{j=1}^{n} \left(\frac{X_j - \overline{X}}{\sigma_x}\right).\left(\frac{Y_j - \overline{Y}}{\sigma_y}\right)}{n} \quad (20)$$

On retrouve dans la formule précédente l'expression des notes centrées réduites des variables X et Y ($Z_{X_j} = \frac{X_j - \overline{X}}{\sigma_x}$). Dès lors, le *r* de Bravais-Pearson étant le résultat du produit moyen des couples de notes centrées réduites correspondant aux mesures, la formule devient :

[58] Bien que le coefficient de corrélation soit largement associé au nom de Karl Pearson (1857-1936), on attribue généralement l'idée fondamentale et les premières approches numériques à Francis Galton (1822-1911). Raymond Brisbois (1967, p. 38) précise : « *Quand Pearson, de 1890 à 1893, prit contact avec les travaux très prometteurs de Galton, il se passionna pour les corrélations et pour tout ce qui touche à la statistique. Il apporta à l'œuvre de son grand prédecesseur les approfondissements mathématiques dont ce dernier se disait incapable. Il en résultat d'abord le coefficient de corrélation, … puis tout une série de formules secondaires se développant comme un halo autour du noyau fondamental selon les situations particulières des problèmes posés.* ». Quand à l'apport de de Bravais, il semble beaucoup plus limité, ainsi Brisbois (1967, p. 29) rapporte : « *…il ne paraît même pas sûr que Bravais … ait pensé à une interdépendance entre variables observées. On peut continuer à appeler de Bravais-Pearson le coefficient de corrélation … le mérite de Bravais est celui d'un précursseur seulement ; son équation ne pénétra en statistique que par Karl Pearson.* ».

$$r = \frac{\sum_{j=1}^{n} Z_{X_j} Z_{Y_j}}{n}$$

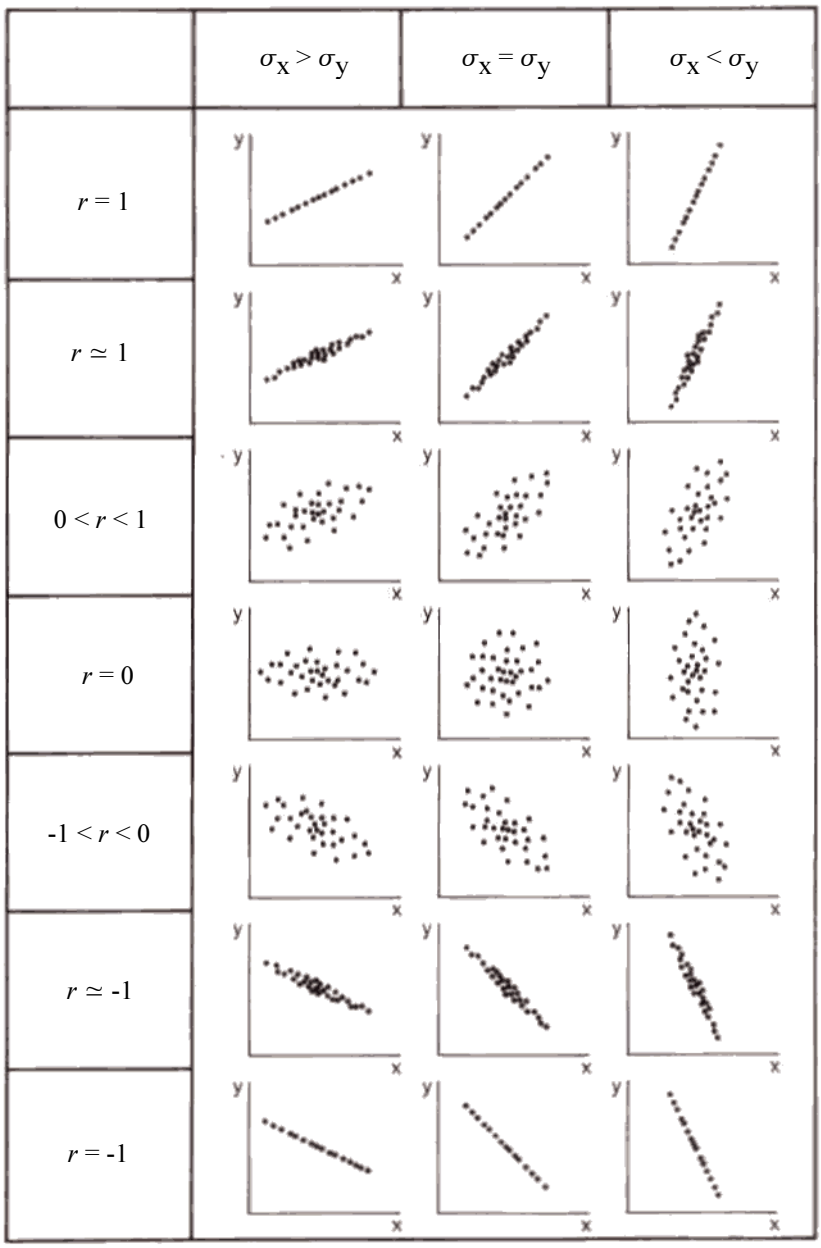
On appellera r^2 proportion de liaison ou plus simplement liaison.

Landercy (1981, p. 109) nous renseigne sur les limites de validité du coefficient *r* de Bravais-Pearson, ce coefficient implique que : « *(1) les couples de points soient indépendants entre eux (pas d'ensembles hybrides), (2) les variables X et Y soient continues et métriques ; la forme des distributions peut varier pour autant qu'elles restent uni modales et assez symétriques, (3) la relation entre les variables X et Y soit linéaire.* ».

Diagrammes de dispersion et liaison avec les valeurs du coefficient de corrélation

Dagnelie (1992, p. 123) signale que la valeur maximum « 1 » ne peut être obtenue que si les points observés se trouvent tous sur une même droite non parallèle aux axes de coordonnées. L'interprétation des valeurs de *r* est facilitée par la figure ci-dessous tirée de Dagnelie (1992). On voit bien sur cette figure la relation entre une dispersion linéaire des données et une liaison forte entre deux variables (*r* proche de ou égal à 1).

Nous ne nous étendrons pas ici sur la procédure de calcul du *r* de Bravais-Pearson car la plupart des logiciels de calcul ou de statistique intègrent ce coefficient de corrélation.



(3) Interprétation

L'interprétation du *r* dépend du contexte de son utilisation. Ainsi, si il est utilisé en tant que coefficient de liaison entre deux formes parallèles d'un même test appliqué dans des conditions équivalentes, les valeurs obtenues pour *r* devront être élevées et atteindre 0,8 ou 64% de liaison et même 0,9 (81% de liaison) pour établir l'équivalence des épreuves. Par contre, il est aussi généralement admis qu'une liaison entre un test d'aptitude et les résultats obtenus en cours de cursus scolaire n'atteignent que rarement 0,6 (36% de liaison) pour un test unique et 0,8 (64% de liaison) si l'on combine plusieurs tests (Landercy, 1981).

En recherche corrélationnelle, un coefficient de corrélation élevé (plus précisément proche de 1) montre une liaison forte entre deux variables et on peut en déduire (avec prudence) que ces dernières sont très peu influencées par des facteurs extérieurs à l'expérience. A l'inverse des recherches corrélationnelles, dans les recherches expérimentales des coefficients de corrélation moins élevés peuvent potentiellement fournir de l'information quant à une liaison certaine entre deux variables, la faible valeur pouvant être liée à des différences incontrôlées et non constantes de la situation expérimentale (attention, fatigue,…). On peut alors penser que si ces facteurs incontrôlés étaient maintenus constants, la corrélation augmenterait.

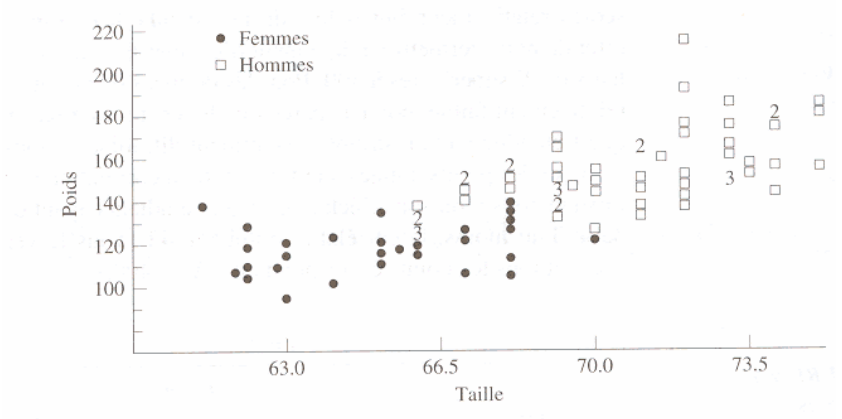
Très utilisé, le coefficient de corrélation *r* de Bravais-Pearson doit être interprété avec prudence. Par exemple, comme Landercy (1981, p. 109) le rappelle : « *…le coefficient de corrélation linéaire est un indice, une estimation de la tendance qu'ont deux variables à varier ensemble et non pas une mesure absolue de liaison sur une échelle linéaire ou d'intervalle… un coefficient r = .5 n'indique pas une relation deux fois plus forte qu'un coefficient r = .25 et de même, une augmentation de r = .4 à r = .6 n'est pas équivalente à une augmentation de r = .5 à r = .7* ». Si le coefficient *r* est utilisé pour des variables non continues l'interprétation en est évidemment encore plus délicate bien que parfois instructive.

Corrélation n'implique pas causalité

Il convient d'insister sur ce point : un coefficient de corrélation ne peut absolument pas nous indiquer pourquoi il existe une corrélation plus ou moins forte entre deux variables. Droesbeke (2001, p. 398) rappelle à ce sujet : « *L'existence d'une dépendance entre variables doit se traiter indépendamment de la détermination de r. C'est sur base de démonstrations, de convictions ou d'hypothèses, que l'interprétation de la valeur de r se fera en terme de mesure d'intensité de dépendance ou de simple association de comportements.* ».

L'effet de sous-groupes hétérogènes

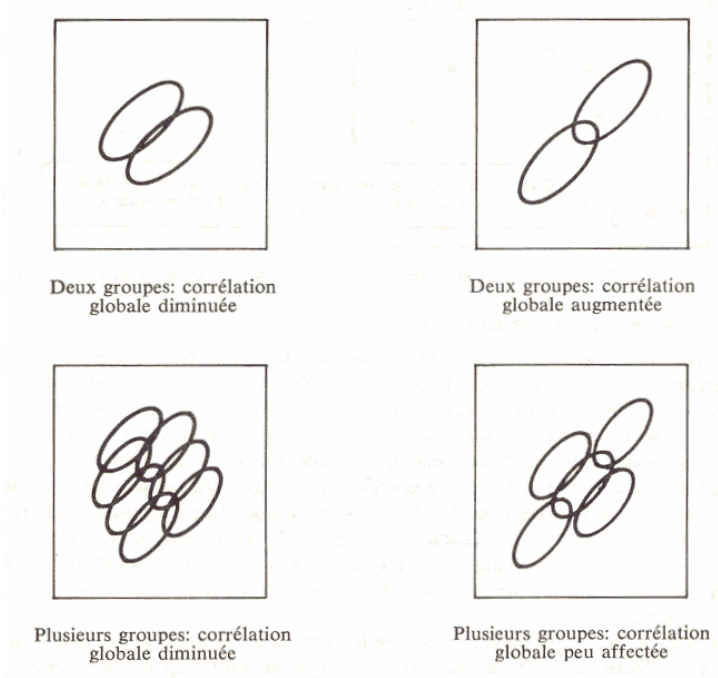
Lorsqu'on interprète les résultats d'analyse de corrélation, la présence éventuelle de sous-groupes hétérogènes dans les données est à prendre en compte. A titre d'exemple, Howell (1998, p.304) reprend les données d'une étude portant sur sur 92 étudiants universitaires américains à qui l'on demande de noter leur taille (en pouces), leur poids (en livres), leur sexe ainsi que plusieurs autres variables. On obtient ainsi le graphique suivant de relation entre la taille et le poids pour les hommes et les femmes combinés.



Howell a pris soin de montrer sur le graphique où étaient les points liés au hommes et ceux qui concernent les femmes.

Il commente de la façon suivante : « *Lorsqu'on combine les données des hommes et des femmes, la relation est étonnamment bonne : la corrélation est de .78. Par contre, si l'on examine les données séparément pour les deux sexes, la corrélation retombe à .60 pour les hommes et à .49 pour les femmes. (…) L'important est de constater que la corrélation élevée lorsque les sexes étaient combinés n'est pas simplement due à la relation entre la taille et le poids. Elle s'explique aussi en grande partie par le fait que les hommes sont en général plus grands et plus gros que les femmes. (…) les expérimentateurs doivent faire preuve de prudence lorsqu'ils combinent des données provenant de différentes sources. Il se peut que la relation entre deux variables soit masquée par une troisième.* ».

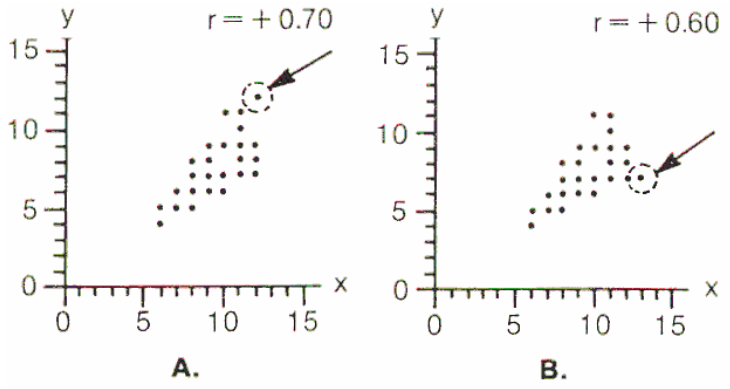
Brisbois (1967, p. 69) présente une série de schémas[59] qui montrent comment des sous-groupes peuvent influencer les valeurs des corrélations globales :



L'effet de points particuliers sur la valeur du coefficient de corrélation dans les échantillons de petite taille

A l'aide des deux graphiques en nuages de points ci-dessous, Nitko (1996, p. 53) montre à quel point un changement dans une seule paire de mesures peut affecter la corrélation de l'ensemble, en particulier lorsque le nombre de couples de mesures est peu élevé (ici $N = 25$ paires de scores).

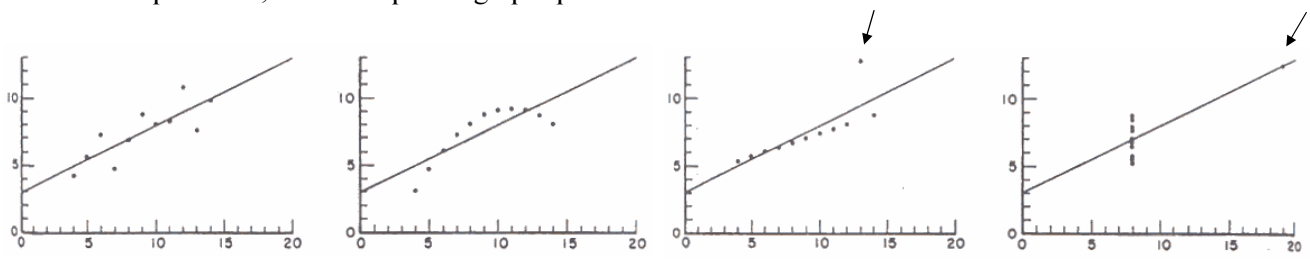
La corrélation de .70 calculée à partir des données représentées dans le graphique en uage de points A descend à .60 dans le graphique B lorsque le sujet qui obtient X = 12 et Y = 12 est remplacé par un autre sujet qui obtient X = 13 et Y = 7.



[59] inspirés de dessins analogues de J.P. Guilford, *Fundamental Statistics in Psychology and Education*, New York, McGraw-Hill Book Co., Inc., 1956, 3e édition, page 324.

Des nuages de points très différents peuvent aboutir à des coefficients de corrélation identiques

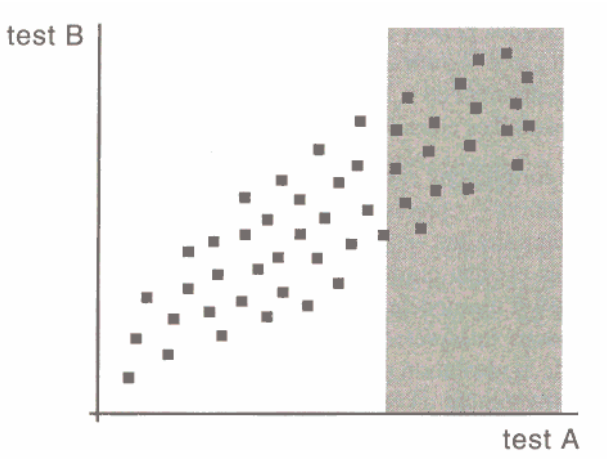
Droesbeke (2001, p. 398) présente les quatre ensembles de données imaginées par Anscombe[60] où *r* vaut à chaque fois 0,816 alors que les graphiques montrent des situations clairement différentes :



La présence de quelques individus à l'une ou l'autre extrémité du groupe risque de fausser l'interprétation de *r*. On remarque dans les deux derniers graphiques l'influence de points isolés (nous les signalons par les flèches) qui expliquent les valeurs similaires de *r* alors que les données se présentent très différemment.

L'effet de la réduction de l'étendue

Cet effet ce présente lorsque les résultats d'un sous-ensemble particulier se concentrent sur une zone étroite de l'étendue possible des scores. Laveault et Grégoire (1997, p. 72) illustrent le phénomène à l'aide du schéma ci-contre.



A titre d'exemple, les auteurs citent une étude célèbre de Thorndike[61] commanditée par l'U.S. Air Force : « *Une batterie de tests avait été constituée pour prédire le succès de l'apprentissage du pilotage. Sur base des résultats à ces tests, seuls 13% des candidats étaient suffisamment qualifiés pour être admis dans le programme d'apprentissage. Toutefois, dans un but expérimental, on décida d'admettre tous les candidats. A la fin de la période d'entraînement, on évalua les qualités de pilote de chacun et l'on calcula les corrélations entre ce critère et les résultats aux différents tests. Ces corrélations furent calculées pour l'ensemble du groupe (N = 1036) et pour le groupe des meilleurs candidats (N=136). On constata ainsi que la corrélation entre le critère et le test de coordination complexe était de 0,40 pour l'ensemble du groupe et de –0,03 pour le groupe restreint. (…) La valeur des prédictions de la batterie de tests était donc très faible si l'on se basait sur les seuls résultats des candidats les plus brillants. Par contre, cette même qualité des prédictions était satisfaisante lorsque l'on évitait la réduction de l'étendue des scores en en calculant les coefficients de corrélations à partir des résultats de l'ensemble du groupe*. »

Selon Howell (1998, p. 303), lorsqu'on est confronté à cet effet de la réduction d'étendue, suivant la nature des données la valeur du *r* peut soit s'accroître, soit diminuer, mais le plus souvent *r* diminue.

[60] F.J. Anscombe, Graphs in Statistical Analysis, *The American Statistician*, 27, 1973, pp. 17-21.
[61] R.L. Thorndike, *Personnel selection : test and measurement techniques*. New York : Wiley, 1949.

b) Variables ordinales : le coefficient ρ (*rho*) de Spearman

(1) Champ d'application

Le coefficient ρ (*rho*) de Spearman s'applique à des échantillons appareillés exprimés en variables ordinales. Dès lors lorsque l'une des deux variables est ordinale et que l'autre est métrique, le ρ de Spearman peut être appliqué après transformation de la variable métrique en variable ordinale. Selon Landercy (1981), « *Si la relation entre les données est linéaire, la différence entre ρ et r est très petite. Si la relation n'est pas linéaire mais qu'elle ne présente pas d'extremum, les différences de rang ne sont pas affectées : ρ est plus grand que r mais sera une meilleure approximation que r en ce qui concerne la relation entre les deux variables. Le coefficient ρ peut être calculé à partir d'un petit nombre de données.* »

(2) Formule de calcul

Selon la démonstration exposée par Landercy (1981), le coefficient ρ peut se calculer à l'aide de la formule suivante :

$$\boxed{\rho = 1 - \frac{6.\sum_j D_j^2}{n^3 - n}} \quad (21)$$

Avec

$D_j^2 =$ la *D*ifférence au carré des rangs pour chaque couple de données *j*

$n =$ le *n*ombre de couples de données

(3) Interprétation

D'après Hoteling & Pabst (1936), rapportés par Laveault & Gregoire (1998), la corrélation ρ de Spearman possède une efficacité relative à 91% par rapport à la corrélation *r* de Pearson. Ce qui signifie que le ρ calculé sur un échantillon de 100 sujets a la même précision qu'une corrélation de Pearson portant sur 91 sujets lorsque les conditions pour le calcul d'une corrélation *r* de Pearson sont respectées. Laveault & Gregoire recommandent l'utilisation du ρ de Spearman *« ...chaque fois que l'une des deux variables ne se distribue pas normalement ou encore ne rencontre pas les conditions d'une échelle à intervalles égaux* ».

c) Variable réellement dichotomique : la corrélation point bisériale (rpbis)

(1) Champ d'application

La corrélation point bisériale (rpbis) est utilisée lorsqu'une des deux données est métrique et que l'autre est dichotomique. Leclercq (1987, p. 50) rappelle que le point signifie « *le fait que la corrélation bisériale s'applique aux dichotomies vraies ou situations dichotomiques, par exemple la réussite ou l'échec à une question, le choix (ou non) d'une solution à une QCM, le sexe (m ou f) d'un étudiant, etc. ».* Dans l'analyse des items, le rpbis permet de corréler les scores obtenus à l'item et les scores au total du test. Nous reviendrons en détail sur cet indice rpbis dans la partie suivante car il est à la base du principe de fonctionnement du rpbis Spectral.

(2) Formules de calcul

Leclercq (1987) reprend deux formules équivalentes :

$$\text{formule 1: } \boxed{rpbis = \frac{Mx - Ma}{\sigma}\sqrt{pq}} \quad (22)$$

et

$$\text{formule 2 : } \boxed{rpbis = \frac{Mx - Mt}{\sigma}\sqrt{\frac{p}{q}}} \quad (23)$$

Voici une troisième formule possible:

$$\text{formule 3 : } \boxed{rpbis = \frac{Mt - Ma}{\sigma}\sqrt{\frac{q}{p}}} \quad (24)$$

Avec

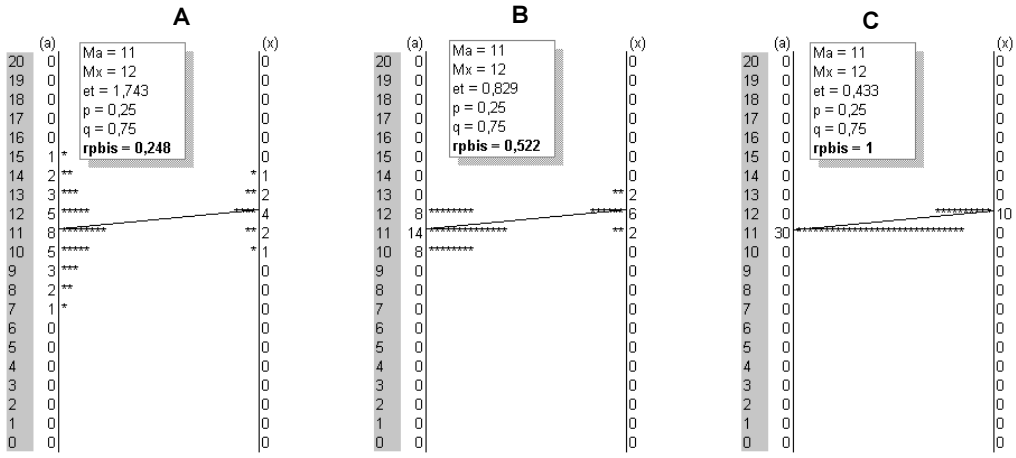
- p = proportion des étudiants qui ont choisi la proposition pour laquelle on calcule le rpbis
- q = $1 - p$ = proportion des étudiants qui n'ont pas choisi la proposition pour laquelle on calcule le rpbis
- Mx = la moyenne des nombres de réponses correctes qui ont été fournies à l'ensemble du test par les étudiants qui ont choisi la proposition analysée
- Ma = la moyenne des nombres de réponses correctes fournies par les étudiants qui n'ont pas choisi la proposition analysée
- Mt = $p \,.\, Mx + q \,.\, Ma$ = la moyenne des nombres de réponses correctes de tous les étudiants au test
- σ = l'écart type des nombres de réponses correctes de tous les étudiants au test

(3) Interprétation

La première série de trois graphiques (A, B et C) qui suit montre des simulations de fréquences des nombres de réponses correctes obtenues par 40 sujets à l'ensemble des questions d'un test comprenant 20 QCM.

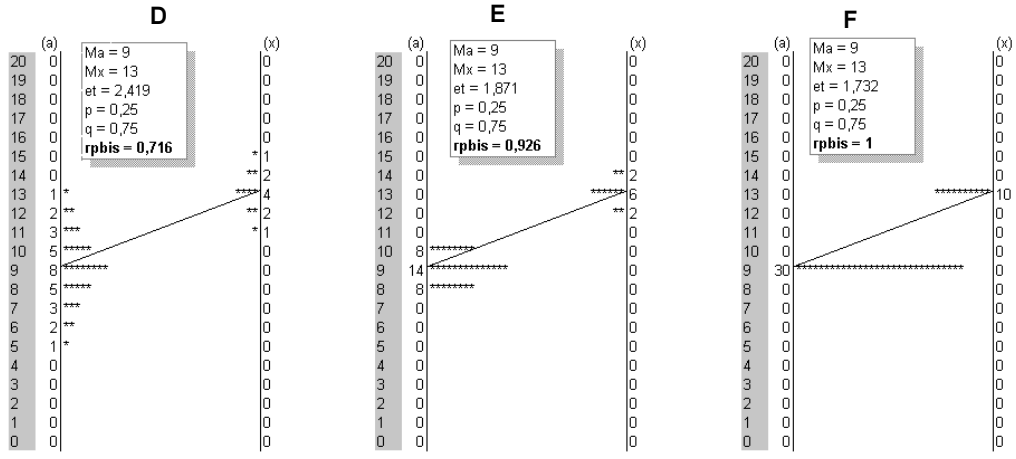
Sur chacun des graphiques, l'axe « (x) » à droite reprend les fréquences des cotes de 10 sujets sur 40 qui ont choisi une proposition « x » à une QCM (p = 0,25). L'axe « (a) » à gauche reprend les fréquences de 30 autres sujets (q = 0,75) qui n'ont pas choisi la proposition.

Sur les graphiques A, B et C, la moyenne à l'ensemble des questions du test pour les sujets qui ont choisi la solution ne varie pas et vaut 12 (Mx = 12), la moyenne des autres ne varie pas non plus et vaut 11 (Ma = 11). La pente des droites obliques qui relient ces moyennes Ma et Mx est identique pour A, B et C.



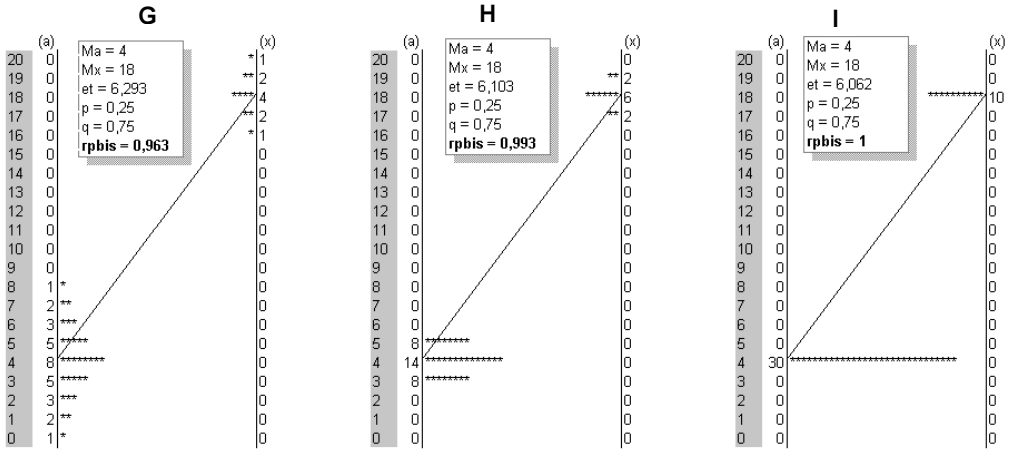
En ce qui concerne la situation A, la dispersion (et) des nombres de réponses correctes vaut 1,748. Pour B la dispersion des données autour de Ma et Mx est plus faible, ce qui influence la dispersion totale (et = 0,829). Le troisième graphique C montre quant à lui une concentration des données sur Mx et Ma (les 10 sujets qui ont choisi la solution obtiennent tous 12 réponses correctes à l'ensemble des questions du test et les 30 autres en obtiennent tous 11). L'écart type des nombres de réponses correctes lié à C vaut 0,433. Etant donné que p et q ne varient pas ainsi que Ma et Mx, il est logique que plus l'écart type des nombres de réponses correctes pour l'ensemble des données diminue, plus le rpbis augmente : rpbis A (0,248) < rpbis B (0,522) < rpbis C (1).

Les trois graphiques suivants (D, E et F) montrent des situations analogues mais cependant différentes dans la mesure où l'écart entre Ma et Mx est plus élevé. Précédemment cet écart était égal à 1 alors qu'ici il vaut 4 (Mx = 13 et Ma =9). Il en résulte des droites obliques plus pentues.



Lorsqu'on compare les rpbis des graphiques D et E avec ceux des graphiques A et B, on constate que rpbis A (0,248) < rpbis D (0,716) et que rpbis B (0,248) < rpbis E (0,926). Quant au rpbis lié au graphique F, il obtient la valeur 1 tout comme le rpbis lié au graphique C.

Enfin, en ce qui concerne les graphiques G, H et I ci-contre, l'écart entre Mx et Ma est encore plus prononcé (Mx – Ma = 14). On constate que les rpbis des graphiques G(0,963) et H(0,993), sont plus élevés que les rpbis des graphiques des



séries précédentes (A, B et D, E). Le rpbis lié au troisième graphique I où la dispersion des données est nulle autour de Ma et de Mx (comme sur C et F) obtient la valeur maximum 1.

La comparaison « horizontale » de chacune des trois séries de graphiques montre à p, q, Ma et Mx constants une augmentation des valeurs du rpbis en parallèle avec une diminution de la dispersion des données.

La comparaison « verticale » des situations A-D-G, B-E-H et C-F-I montre que l'écart entre Mx et Ma influence aussi la valeur du rpbis à condition que la dispersion des données autour de Ma et Mx ne soit pas nulle.

Le rpbis, un indice de « fracture »

Notons que la seule différence entre Ma et Mx pourrait être indicative de l'état d'une « fracture » entre le groupe de sujets qui a choisi une proposition et le groupe des autres sujets. Landercy (1981, p. 122) signale à ce propos : « *Si l'on regarde la formule du coefficient rpbis, on constate qu'elle dépend directement de la différence entre les moyennes Mx et Ma et que plus Mx – Ma sera grand, plus r sera grand.».* Comme nous venons de l'observer, d'une part ce raisonnement s'applique aux situations où la dispersion des données autour des moyennes Mx et Ma n'est pas nulle, et, d'autre part, il faut souligner que la dispersion des données intervient aussi dans l'explication des valeurs des rpbis. La prise en compte de la dispersion des données nous paraît constituer un avantage de la formule du rpbis sur le simple calcul de la différence entre Mx et Ma car comme le montrent par exemple les graphiques D et E, des différences de moyennes Mx et Ma égales (Mx-Ma = 4 dans le cas des situations D et E) peuvent traduire des « fractures » dont la « netteté » diffère dans la mesure où les données sont plus ou moins dispersées (rpbis D = 0,716 et rpbis E = 0,926).

Alors que la simple différence entre Mx et Ma donne une indication à propos de « l'ampleur » de la « fracture » mais sans en montrer la « netteté », le rpbis en tient compte en intégrant la dispersion des données dans la formule.

Trois conditions pour qu'un rpbis récolte la valeur maximum 1

Le graphique C montre que le rpbis d'une solution à une QCM peut récolter la valeur maximum 1 alors que la différence entre Mx et Ma n'est pas élevée. En fait, le rpbis vaut « 1 » lorsqu'on observe simultanément les trois conditions suivantes :

[1] les sujets qui ont choisi la solution obtiennent tous un nombre identique de réponses correctes au total du test (la dispersion de ces scores est nulle);

[2] les sujets qui n'ont pas choisi cette solution récoltent aussi un même nombre total de réponses correctes (la dispersion de ces scores est nulle aussi) ;

[3] le nombre de réponses correctes obtenu par les sujets qui ont choisi la solution est supérieur au nombre de réponses correctes observé chez ceux qui ne l'ont pas choisie.

Ces trois conditions étaient remplies pour les situations présentées dans les graphiques C, F et I où le rpbis est égal à « 1 ». Soulignons qu'il extrêmement rare de voir ces trois conditions remplies dans les données issues de tests où on mesure les acquis des étudiants. Dans le cadre des épreuves MOHICAN nous n'avons pas rencontré ce type de configuration des données.

Indications sur l'adéquation des solutions proposées

Notons que lorsque, en moyenne, les sujets qui ont choisi la solution obtiennent un nombre de réponses correctes pour l'ensemble des questions du test moins élevé que ceux qui n'ont pas choisi cette solution, alors Mx est inférieur à Ma et il en résulte un rpbis négatif.

De Landsheere (1979, p. 44) signale : « *Le rpbis est positif si l'item est réussi, en moyenne, par les sujets qui obtiennent un bon score au test ; un coefficient négatif correspond à la situation opposée… Le rpbis fournit aussi de précieuses informations sur l'adéquation des distracteurs.* ».

Notons qu'un rpbis élevé pour une proposition n'indique pas forcément que ceux qui n'ont pas choisi cette solution obtiennent en moyenne des scores faibles. Dans une autre variante de l'exemple du graphique E présenté précédemment, Mx pourrait être égal à 19/20 et Ma égal à 15/20, dans ce cas de figure Ma est loin d'être faible.

Dans un autre cas de figure analogue à la situation représentée par le graphique E mais où Mx vaudrait 6 et Ma vaudrait 2, le rpbis serait positif (et élevé) pour un item réussi par des sujets qui obtiennent un nombre de réponses correctes peu élevé au test (Mx vaut 6/20). La valeur positive et proche de 1 d'un rpbis ne montre donc pas forcément que ceux qui ont choisi la solution sont ceux qui obtiennent de « bons » scores.

Dès lors, nous préférons interpréter l'indice de la façon suivante : le rpbis d'une proposition est positif lorsque les sujets qui choisissent celle-ci obtiennent, en moyenne, un nombre total de réponses correctes plus élevé que les autres sujets. Le rpbis est négatif lorsque les sujets qui ont choisi la proposition obtiennent, en moyenne, un nombre total de réponses correctes moins élevé que les autres sujets.

De Landsheere (*op. cit.*) propose aussi d'interpréter les valeurs de la façon suivante : « *En gros, un rpbis supérieur à .20 est considéré comme satisfaisant pour la réponse correcte ; le rpbis d'un distracteur devrait être négatif.* ». Ce raisonnement de De Landsheere s'applique à des épreuves qui comportent entre 20 et 30 questions. Nous verrons, dans la section suivante, que l'interprétation du rpbis d'une question pose le problème du recouvrement de l'item dans le score total au test, ce problème peut être contré en calculant un seuil rpbis pour la réponse correcte (voir *infra* p. 176). Entre 20 et 30 questions, ce seuil rpbis varie entre 0,22 et 0,18 pour la réponse correcte.

d) Variable dichotomisée : la corrélation bisériale (rbis)

(1) Champ d'application

Le coefficient de corrélation bisérial (rbis) est utilisé lorsque les deux variables sont distribuables de manière continue (théoriquement une valeur est toujours possible entre deux autres) mais que l'une des deux, dont la distribution métrique serait normale, a été dichotomisée. Par exemple le découpage bisérial d'une population en sujets dont la taille serait supérieure à 1,80 m (1ère série) et en sujets qui mesurent moins de 1,80 m (2ème série).

(2) Formules de calcul

Le principe est similaire à celui du coefficient point bisérial (rpbis) mais on tient compte dans le calcul de la hauteur de la courbe normale centrée réduite au point *z* correspondant à une densité de probabilité égale à *p*. Landercy (1981) rappelle : « *…comme la variable dichotomisée doit être distribuée normalement et continûment, on tient compte dans le calcul de la densité de probabilité Y de cette distribution à la valeur de la coupure…* ». La formule devient :

$$rbis = \frac{Mx - Ma}{\sigma} . \frac{pq}{Y} \quad (25)$$

ou

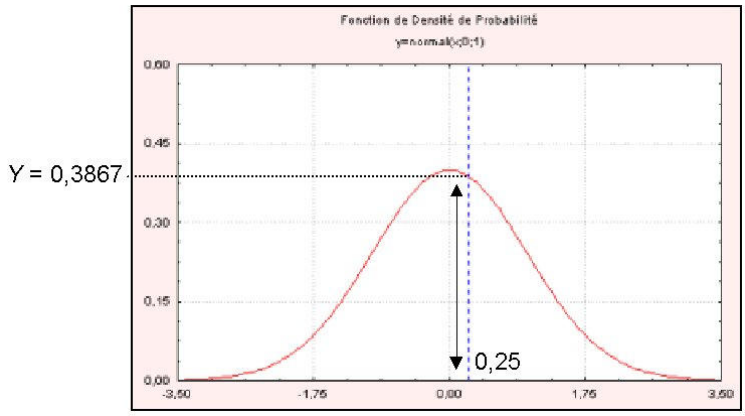
$$rbis = \frac{Mx - Mt}{\sigma} . \frac{p}{Y} \quad (26)$$

Avec Y = ordonnée de la courbe normale centrée réduite en fonction de l'abscisse *z* (voir ci-dessous comment on obtient *z* à partir d'une table ou d'un logiciel calculateur de probabilités)

La formule de calcul peut être établie à partir de celle du rpbis (Landercy, 1981; Leclercq, 1987; Laveault & Gregoire, 1997) :

$$rbis = \frac{\sqrt{pq}}{Y} rpbis \quad (27)$$

En ce qui concerne le calcul de la valeur *Y*, par exemple pour une valeur *p* de 0,6 la densité de probabilité dans une distribution normale centrée réduite correspond à un score *z* de 0,253. Cette probabilité peut être fournie par une table des valeurs de la loi normale ou être obtenue à l'aide d'un calculateur de probabilité, tel que celui que l'on trouve dans le logiciel *STATISTICA* produit par la firme *StaSoft*. A partir de ce score *z* on peut alors obtenir la valeur de l'ordonnée *Y* à partir d'une table des « *Ordonnées Y de la courbe normale centrée réduite en fonction de l'abscisse z* ». Dans le cas de notre exemple, *Y* vaut 0,3867.



(3) Interprétation

Selon Landercy (1981), quatre conditions doivent être remplies pour que le rbis fournisse une bonne estimation de la corrélation de Pearson : « *(a) la relation est linéaire, (b) la variable dichotomisée est extraite d'une population normale, (c) la coupure s'effectue vers le centre de la distribution (p et q ne peuvent jamais dépasser .90) et (d) la variable métrique est extraite d'une population distribuée unimodalement de manière pas trop dissymétrique* ».

Laveault & Gregoire (1997, p. 238) signalent que « *Lord et Novick (1968) ont démontré que la corrélation bisériale obtenue est 20 % supérieure au coefficient de corrélation point bisériale* ». Il semble aussi que le coefficient de corrélation bisériale soit plus performant lorsque *p* ou *q* sont élevés, les mêmes auteurs rapportent : « *Dans le cas de valeurs extrêmes de p et q, Magnuson (1967) a démontré que la corrélation bisériale pouvait être jusqu'à quatre fois supérieure à la corrélation point bisériale. Ceci est dû au fait que la faible variance des items affecte grandement la valeur maximum que peut prendre la corrélation point bisériale, qui est un équivalent algébrique du r de Pearson* ».

e) Deux variables réellement dichotomiques : le coefficient de corrélation ϕ (*phi*)

(1) Champ d'application

Le coefficient de corrélation ϕ (*phi*) est utilisé pour évaluer la relation entre deux variables réellement dichotomiques. Avec ϕ on peut par exemple calculer la corrélation entre deux items vrai/faux et établir ainsi la valeur prédictive de réussite à une des deux questions en fonction de l'autre. En fait, le coefficient ϕ s'applique à des types de situations qui peuvent se traduire par des tables de contingence deux fois deux.

(2) Formule de calcul

Landercy (1981) propose le calcul du coefficient à l'aide d'une table de contingence. Pour deux items *J* et *K* « vrai/faux » que l'on comparerait, la table de contingence s'exprimerait comme suit,

		Item *J* 1	0	
Item	1	A	B	A+B
K	0	C	D	C+D
		A+C	B+D	*n = A+B+C+D*

« A » représentant le nombre de sujets ayant répondus « vrai » aux items *K* et *J*, « D » le nombre de sujets ayant répondu « faux » aux deux items, « C » le nombre des sujets ayant répondu « vrai » à *J* et « faux » à *K,* et enfin B, le nombre de sujets ayant répondu « vrai » à *K* et « faux » à *J*.

Selon Landercy (1981) la formule de calcul du ϕ peut s'écrire (en développant le *r* de Bravais-Pearson) :

$$\phi = \frac{AD - BC}{\sqrt{(A+B)(A+C)(D+B)(D+C)}} \quad (28)$$

Laveault & Gregoire (1998) proposent une formule basée sur les proportions :

$$\phi = \frac{\left|(p_{JK} - p_J \cdot p_K)\right|}{\sqrt{p_J \cdot q_J \cdot p_K \cdot q_K}} \quad (29)$$

Avec
p_{JK} = A/*n* = la proportion des sujets ayant répondu « vrai » aux deux items
p_J = (A+C)/*n* = la proportion des sujets ayant répondu « vrai » à l'item *J*
p_K = (A+B)/*n* = la proportion des sujets ayant répondu « vrai » à l'item *K*
q_J = 1- p_J = la proportion des sujets ayant répondu « faux » à l'item *J*
q_K = 1- p_K = la proportion des sujets ayant répondu « faux » à l'item *K*

(3) interprétation

Le coefficient ϕ sera :

- grand et positif si les effectifs des cases « A » et « D » sont grands et ceux des cases « B » et « C » petits, ce qui montre une corrélation importante et positive, si B = C = 0 alors ϕ = 1 ;
- grand et négatif si « A » et « D » sont petits et « B » et « C » sont grands, ce qui montre une corrélation négative importante, si A = D = 0 alors ϕ = -1 ;
- proche de zéro si « A . B » et « C . D » sont plus ou moins équivalents, on a une absence de corrélation. Si « A = B = C = D », la répartition uniforme des effectifs montre aussi une absence de corrélation.

Landercy (1981) signale que le ϕ est construit à partir d'un χ^2 (*khi carré*) : *« ...c'est la racine carrée du χ^2 moyen »*. La formule de calcul du ϕ peut aussi s'écrire (De Landsheere, 1979; Landercy, 1981) :

$$\phi = \sqrt{\frac{\chi^2}{N}} \quad (30)$$

f) Deux variables dichotomisées : le coefficient de corrélation tétrachorique (r_t)

(1) Champ d'application

Le coefficient r_t tétrachorique est une bonne approximation du r de Bravais-Pearson lorsque les deux variables au départ continues, distribuées normalement et reliées linéairement ont été dichotomisées. C'est par exemple le type de situation que l'on rencontre lorsqu'on souhaite évaluer la relation entre l'échec et la réussite dans une branche et une autre, à deux items d'un test avant et après un apprentissage, … Ces situations peuvent aussi être traduites dans une table de contingence :

		Item *J*		
		1	0	
Item	1	A	B	A+B
K	0	C	D	C+D
		A+C	B+D	*n = A+B+C+D*

(2) Formule de calcul

Selon Landercy (1981), calculer le coefficient r_t revient à résoudre l'équation :

$$r_t + r_t^2.\frac{Z_x.Z_y}{2} + r_t^3\frac{(Z_x^2-1).(Z_y^2-1)}{6} + ... = \frac{AD-BC}{Y_x.Y_y.n^2} \quad (31)$$

Avec

Z_x et Z_y = les notes centrées réduites correspondant aux valeurs x et y où s'effectue la dichotomisation.

Y_x et Y_y = les densités d'effectifs de la loi normale à Z_x et Z_y .

A, B, C et *D* = les effectifs de chaque case de la table de contingence.

Il s'agit d'une équation complexe qu'il est difficile de résoudre « à la main » comme le rappelle Landercy (1981) : « *Il est bien évident que la résolution générale de l'équation est très compliquée et ne peut s'effectuer que sur ordinateur* ».

(3) Interprétation

Tout comme pour le coefficient ϕ, d'après la répartition des effectifs dans la table des contingences on peut dire que :

- en cas de corrélation positive parfaite, les effectifs se répartiront dans les cases « A » et « D » ;
- en cas de corrélation négative parfaite, les effectifs se répartiront en « B » et « C » ;
- en cas de corrélation nulle, le produit AB sera équivalent au produit CD, c'est le cas en particulier lorsque les effectifs sont équivalents dans les cases « A », « B », « C » et « D ».

D'après Landercy (1981), « r_t *est moins fiable que r. Cependant* r_t *se rapproche d'autant plus de r que : (1) n est grand, (2) la corrélation linéaire est forte et (3) la division en catégories est près des médianes* ». Laveault & Gregoire (1997) évoquent la possibilité d'utiliser dans certaines conditions le coefficient ϕ en lieu et place du r_t : « *Dans le cas des items de difficulté moyenne, la corrélation* ϕ *et la corrélation tétrachorique fournissent les mêmes résultats. La différence est plus importante dans les cas extrêmes où les items très faciles ou très difficiles sont mis en corrélation. Le calcul des corrélations tétrachoriques est particulièrement recommandé si l'on souhaite réaliser une analyse factorielle sur une matrice des inter corrélations entre les items. Mis à part ce cas particulier, il semble qu'à défaut de pouvoir employer les corrélations tétrachoriques, les corrélations* ϕ *peuvent constituer une alternative pratique quoique imparfaite.* »

g) Tests de signification pour les indices corrélationnels de discrimination

Lorsque la valeur d'un indice corrélationnel de discrimination est proche de zéro, il convient de s'interroger sur son niveau de significativité. En d'autres termes, est-ce par hasard que cette valeur de l'indice est différente de zéro, ou au contraire, cette différence est-elle significative ?

La forme de la distribution d'échantillonnage des *r* de Bravais-Pearson dépend de la taille *n* de l'échantillon et du paramètre moyen $\bar{r}$ de la population théorique. Ceci étant dû au fait que la distribution d'échantillonnage est délimitée à ses extrémités par les valeurs –1 et 1, dès lors la forme de la distribution sera d'autant plus biaisée que la valeur de $\bar{r}$ se rapprochera de la valeur minimum ou maximum. La distribution d'échantillonnage sera symétrique uniquement si $\bar{r} = 0$.

(1) Test de signification lorsque le nombre de sujets dans l'échantillon est supérieur à 50

Magnuson (1967) a montré que lorsque *N* est plus grand que 50 et $\bar{r}$ est proche de zéro, l'écart type de la distribution des *r* de Bravais-Pearson, pouvait être estimé par :

$$S_r = \frac{1}{\sqrt{n-1}} \quad (32)$$

Avec
S_r = l'écart type de la distribution des *r*.
n = le nombre de sujets de la corrélation.

Dès lors, comme le soulignent Laveault & Gregoire (1997, p. 243), « *...plus l'échantillon est petit et plus grande devra être la corrélation entre deux variables avant que celle-ci ne puisse être considérée comme significativement différente de 0. Plus le nombre de répondants à un test est petit, plus l'indice de discrimination devra être grand avant que l'on considère qu'un item contribue à différencier les sujets quant à leur score total.* ».

C'est la même formule qui est employée pour évaluer le degré de signification des coefficients de corrélation point bisériale et ϕ. Par contre, Laveault & Gregoire (1998) recommandent d'utiliser la formule de Kurtz & Mayo (1979) pour le cas de la corrélation bisériale.

$$S_{rbis} = \frac{\sqrt{\frac{pq}{n-1}}}{Y} \quad (33)$$

Avec *Y* = ordonnée de la courbe normale centrée réduite en fonction de l'abscisse *z* (voir à la section concernant le calcul de la corrélation bisériale comment on obtient *z* à partir d'une table ou d'un logiciel calculateur de probabilités)

Que ce soit par la formule de Magnuson ou la formule de Kurtz et Mayo, l'écart type calculé sert à déterminer un intervalle de confiance. Si la corrélation se situe à l'intérieur de l'intervalle au seuil de signification choisi (0,05 ou 0,01), on en déduit que la corrélation n'est pas significativement différente de zéro.

Laveault & Gregoire (1997), prenant l'exemple d'une corrélation de 0,34 établie sur 82 sujets (S_r = 0,111) rappellent : « *Les valeurs comprises entre ± 1,96S_r déterminent un intervalle de confiance à l'intérieur duquel se situent 95% des valeurs de corrélations qui peuvent se produire au hasard entre 82 couples de données pour lesquels il n'y a pas de corrélation…Dans le cas qui nous intéresse, cet intervalle est compris entre ±0,22. Une corrélation de 0,34 serait donc considérée comme significativement différente de 0.* »

(2) Test de signification indépendant du nombre de sujets dans l'échantillon

Landercy (1981), propose une méthode d'évaluation du niveau de significativité du *r* à partir de la droite de régression et qui implique la transformation du *r* en *t* :

« *Si le coefficient de corrélation $\bar{r}$ dans la population est nul, alors dans la population, la droite de régression de y en x est parallèle à l'axe des x et a pour équation $y = \bar{y}$, c'est-à-dire que la pente est nulle. Fisher a montré que pour des échantillons appareillés de N couples de données, la distribution d'échantillonnage des pentes a autour de la pente nulle était une distribution d'échantillonnage en t de Student à N-2 degrés de liberté, l'erreur type de cette distribution étant la racine carrée du quotient de l'erreur moyenne estimée de variation :* $\frac{\sum (y_i - \bar{y}_i)^2}{N-2}$,

par la variation totale des x_i *:* $\sum (x_i - \bar{x}_i)^2$.

On a : $S = \sqrt{\frac{\sum (Y_i - \bar{y}_i)^2}{(N-2)\sum (x_i - \bar{x})^2}} = \sqrt{\frac{\sum N\sigma_y^2 (1-r^2)}{(N-2)N\sigma_x^2}} = \frac{\sigma_y}{\sigma_x}\sqrt{\frac{\sum 1-r^2}{N-2}}$

On peut calculer : $t = \frac{|a-o|}{S} = \frac{r\frac{\sigma_y}{\sigma_x}}{\frac{\sigma_y}{\sigma_x}\sqrt{\frac{\sum 1-r^2}{N-2}}} = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$

et comparer ainsi la valeur trouvée aux valeurs critiques de la table à N-2 degrés de liberté. Si la valeur de t calculée est supérieure à la valeur de la table pour un niveau de signification donné, cela signifie que la pente de la droite de régression est différente de 0 à ce taux de signification et qu'il y a donc une relation entre les données x_i *et* y_i*. ... Cette méthode, très simple et très rapide ne pose pas de conditions préalables quant à la taille de l'échantillon.*»

Donc, après calcul de la valeur du *t* à l'aide de la formule de Fisher de transformation de la valeur *r* en *t* de Student :

$$t_c = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} \quad (34)$$

avec : *n* = le nombre de couples de mesures appareillées
r = le coefficient de corrélation sur lequel est appliqué le test de signification $r \neq 0$

Nous obtenons ainsi la valeur du *t* calculé (t_c) que nous pouvons comparer aux valeurs théoriques (t_t) de la table du *t* de Student à *n*-2 degrés de liberté : si la différence est significative, on peut affirmer que la corrélation est différente de zéro.

PARTIE II

Instrumentation de l'analyse de la qualité spectrale des examens standardisés universitaires

Chapitre IV :

Application de la problématique du coefficient de corrélation point bisériale à l'analyse spectrale des QCM

Sommaire

A. Problématique du coefficient de corrélation point bisériale classique

1. Formules de calcul du coefficient de corrélation bisériale de point classique pour une question

Dans cette partie nous utiliserons l'appellation coefficient de corrélation bisériale de point (*rpbis*) en la complétant le plus souvent du terme « classique » pour une distinction claire avec le coefficient de corrélation bisériale de point Spectral Contrasté (*rpbis SC*) calculé à partir des informations liées aux degrés de certitude et dont la problématique sera détaillée dans la partie suivante.

Les formules de coefficient de corrélation bisériale de point classique pour une question ont été décrites précédemment à la page 156. Rappelons ici les deux formules reprise par Leclercq (1987, p. 50) :

$$\text{formule 1: } \boxed{rpbis = \frac{Mx - Ma}{\sigma}\sqrt{pq}}$$

et

$$\text{formule 2 : } \boxed{rpbis = \frac{Mx - Mt}{\sigma}\sqrt{\frac{p}{q}}}$$

ainsi qu'une troisième formule possible:

$$\text{formule 3 : } \boxed{rpbis = \frac{Mt - Ma}{\sigma}\sqrt{\frac{q}{p}}}$$

Mx étant, lorsqu'on évalue la performance d'une question, la moyenne des scores à l'ensemble des questions du test pour les étudiants qui ont choisi la réponse correcte à la question envisagée. *Ma* étant la moyenne des scores à l'ensemble du test pour les étudiants qui n'ont pas choisi la réponse correcte à la question. *Mt*, dans le cas des formules 2 et 3, représente la moyenne de tous les scores au test. Enfin, l'écart type (σ) dans les 3 formules est calculé à partir de tous les scores à l'ensemble des questions du test.

Remarquons à la suite de Leclercq (1987, p. 50) que le *rpbis* pour une question est « *la combinaison en une seule formule des aspects quantitatif (pouvoir séparateur =* $\sqrt{pq}$ *) et qualitatif (*$\frac{Mx - Ma}{\sigma}$*) du caractère discriminatif d'une question ».*

a) Aspect quantitatif : le pouvoir séparateur d'une question

En ce qui concerne le pouvoir séparateur d'une question, c'est-à-dire sa capacité à distinguer les étudiants qui réussissent la question et les étudiants qui échouent, Leclercq (*op. cit.* p. 44) rappelle qu'il dépend de la moyenne des fréquences des résultats en termes d'échecs (0) ou réussites (1) des étudiants à la question considérée. Le tableau ci-dessous montre que le pouvoir séparateur d'une question est fonction de *p* (proportion de réponses correctes) :

Si	p = 0,5	alors pq =	0,5 x 0,5	= 0,25	et $\sqrt{pq}$	= 0,5
	p = 0,4 ou 0,6	alors pq =	0,4 x 0,6	= 0,24	et $\sqrt{pq}$	= 0,489
	p = 0,3 ou 0,7	alors pq =	0,3 x 0,7	= 0,21	et $\sqrt{pq}$	= 0,458
	p = 0,2 ou 0,8	alors pq =	0,2 x 0,8	= 0,16	et $\sqrt{pq}$	= 0,4
	p = 0,1 ou 0,9	alors pq =	0,1 x 0,9	= 0,09	et $\sqrt{pq}$	= 0,3
	p = 0,05 ou 0,95	alors pq =	0,05 x 0,95	= 0,0475	et $\sqrt{pq}$	= 0,217

La question qui est réussie par tous les étudiants (p = 1), ne permet pas de distinguer les étudiants entre eux, il n'est pas possible de les répartir en deux groupes distincts en fonction du critère « réussite/échec » à la question. Son pouvoir séparateur est nul. Il en va de même si tous les étudiants échouent à la question (p = 0).

Leclercq (op. cit. p.44) signale que pendant bon nombre d'années on n'a accordé d'importance qu'au seul pouvoir séparateur avec une tendance à supprimer les questions trop bien ou trop mal réussies. Or, qu'une question d'une épreuve pédagogique ait été réussie par tous les étudiants ou par aucun d'eux n'est pas en soi une raison suffisante pour éliminer cette question.

En effet, il suffit de se reporter à des situations de type pré test où il n'est pas rare que des questions obtiennent des taux de réussite proches de 0%, ou à des situations de type post-test où, au contraire, il est possible d'observer des taux de réussite élevés proches de 100%. Malheureusement, on rencontre encore actuellement cette tendance à vouloir systématiquement supprimer les questions réussies par moins de 30% des étudiants ou par plus de 80% où on fait passer le but de discrimination (légitime en psychologie différentielle) avant le but de mesurer les acquis (légitime en éducation). Voici un exemple de ce type de considération que nous avons rencontré dans Kehoe (1995) : « *The proportion of students answering an item correctly also affects its discrimination power. This point may be summarised by saying that items answered correctly (or incorrectly) by a large proportion of examinees (more than 85%) have markedly reduced power to discriminate. On a good test, most items will be answered correctly by 30% to 80% of the examinees* ». Selon l'auteur, les questions d'un « bon » test devraient obtenir des taux de réussite qui oscillent entre 30% et 80%. Répétons que nous ne pouvons adhérer à ce principe que dans une perspective de psychométrie différentielle et non dans une perspective édumétrique, le pouvoir discriminateur n'étant à notre avis en éducation pas le seul argument à tenir en compte pour décider de la suppression d'une question dans une situation de testing pédagogique.

b) Aspect qualitatif : le caractère discriminatif d'une question

En ce qui concerne l'aspect plus qualitatif ($\frac{Mx - Ma}{\sigma}$) (de $\frac{Mx - Mt}{\sigma}$ ou de $\frac{Mt - Ma}{\sigma}$) du caractère discriminatif d'une question, dans la formule 1, une différence « *Mx – Ma* » (ou « *Mx – Mt* » ou « *Mt – Ma* ») positive montre que ceux qui ont réussi la question ont une moyenne au test critère plus élevée (et ceux qui ont raté, une moyenne moins élevée), habituellement c'est ce à quoi on s'attend. Cette différence est influencée par la dispersion des scores. Leclercq (*op. cit.* p. 49) souligne « *Quand l'écart type est grand, il est possible d'observer entre Mx et Ma une différence plus importante que si les résultats avaient été moins dispersés. C'est pourquoi la valeur numérique de l'indice de discrimination tient compte de l'écart type, en exprimant la différence Mx – Ma par rapport à l'écart type* ».

De Landsheere (1979) signale que le rpbis joue un rôle important dans l'analyse des items des tests : « *…il est utilisé comme indice de corrélation entre chaque item d'un test et le score total… Le rpbis renseigne donc sur la cohérence interne du test.* »

2. Procédure de calcul du *rpbis classique* pour chaque proposition d'une question à choix multiple

La matrice ci-après reprend les choix d'un groupe de 20 étudiants pour une Question à Choix Multiple (QCM) comprenant 4 propositions, de P1 à P4 avec une possibilité d'omission notée « OM ». Dans l'exemple, P3 correspond à la réponse correcte pour cette question (données en gras sur fond grisé). Pour chaque étudiant, le choix d'une proposition est exprimé par un 1 et l'absence de choix des autres possibilités par des 0. Ainsi, l'étudiant « etu1 » a choisi la proposition P4, l'étudiant « etu2 » la proposition P1, « etu3 » la proposition P3 qui est correcte, etc.

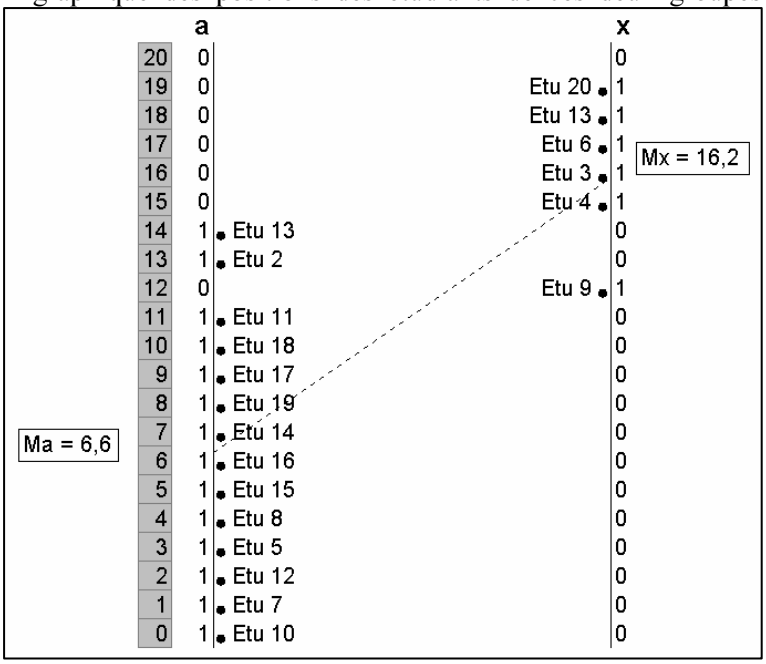
La dernière colonne reprend pour chaque étudiant le nombre de réponses correctes qu'il a fourni pour l'ensemble des questions du test (20 questions au total dans notre exemple). Il s'agit de la deuxième série de données qui interviendront dans le calcul des corrélations et qui explique l'appellation « bisériale ».

L'avant dernière ligne de la matrice reprend les moyennes pour les colonnes OM, P1 à P4 et NRCT et la dernière ligne l'écart type des NRCT.

	OM	P1	P2	**P3**	P4	NRCT Nombre de Réponses Correctes au Test
etu1	0	0	0	**0**	1	14
etu2	0	1	0	**0**	0	13
etu3	0	0	0	**1**	0	16
etu4	0	0	0	**1**	0	15
etu5	1	0	0	**0**	0	3
etu6	0	0	0	**1**	0	17
etu7	1	0	0	**0**	0	1
etu8	0	0	0	**0**	1	4
etu9	0	0	0	**1**	0	12
etu10	0	0	1	**0**	0	0
etu11	0	0	0	**0**	1	11
etu12	0	1	0	**0**	0	2
etu13	0	0	0	**1**	0	18
etu14	0	1	0	**0**	0	7
etu15	0	0	1	**0**	0	5
etu16	0	0	1	**0**	0	6
etu17	0	0	1	**0**	0	9
etu18	0	0	0	**0**	1	10
etu19	0	1	0	**0**	0	8
etu20	0	0	0	**1**	0	19
Moyennes	0,1	0,2	0,2	**0,3**	0,2	Mt = 9,5
Ecart type						5,8

Prenons le cas du calcul du *rpbis classique* de la réponse correcte (P3) pour la question considérée. Dans le cas de l'analyse des propositions d'une QCM à l'aide du *rpbis classique*, c'est le total des réponses correctes au test pour chaque étudiant qui nous servira de mesure critère. Deux groupes d'étudiants peuvent être envisagés : d'une part ceux qui ont choisi la solution P3 et d'autre part les autres qui l'ont rejetée et qui ont donc opté pour un des distracteurs (P1, P2 ou P4) ou qui ont omis (OM).

Ci-contre figure une représentation graphique des positions des étudiants de ces deux groupes lorsqu'on considère à la fois le rejet choix (axe a) ou le choix (axe x) de P3 et le nombre de réponses correctes pour l'ensemble des questions du test (nombres de réponses correctes dans la zone grisée). Chaque étudiant du premier groupe (ceux qui ont rejeté P3) est placé sur un l'axe « a », les autres sont placés sur l'axe « x ».

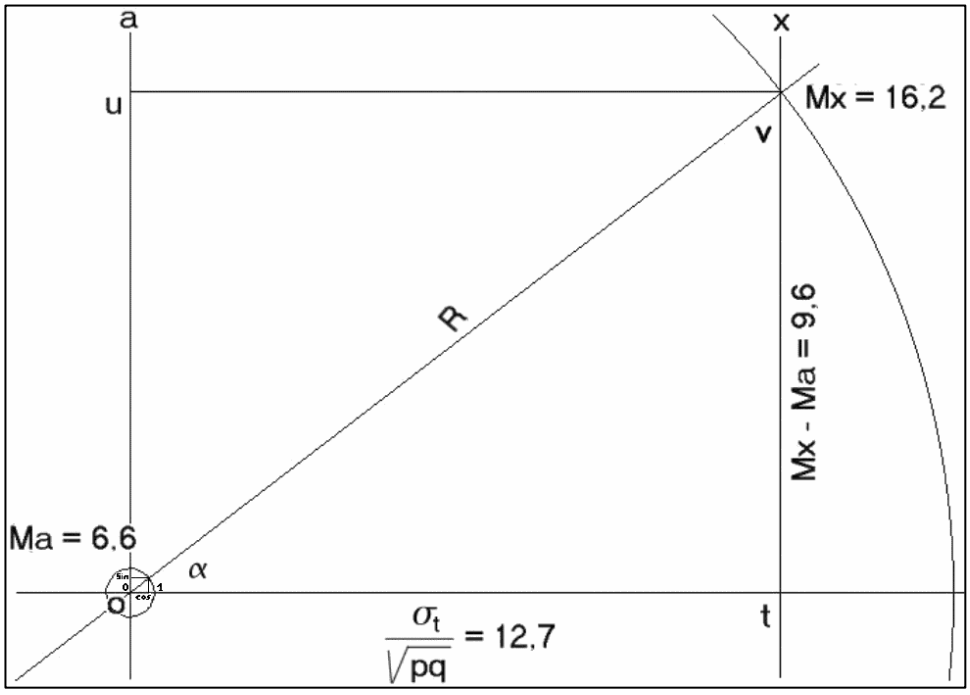


Dans cet exemple le groupe des 6 étudiants qui ont choisi P3 obtient une moyenne de 16,2 réponses correctes à l'ensemble des questions du test. Les 14 autres étudiants obtiennent une moyenne de 6,6 réponses correctes. Le pouvoir séparateur de P3 (et donc de la question puisque ici P3 est la réponse correcte) est de 0,3 (la moyenne des choix « 1 » et rejets « 0 » de la proposition P3), 30 % des étudiants ont choisi cette proposition.

En ce qui concerne l'aspect qualitatif du pouvoir discriminatif de la proposition P3, *Mx – Ma* vaut 9,6. La valeur positive de cette différence indique qu'il y a concordance entre le choix de P3 et les scores à l'ensemble des questions du test. Rapportée à l'écart type, cette différence vaut 9,6/5,8 = 1,66.

On observe sur le graphique précédent une oblique reliant *Ma* à *Mx*, la pente de cette oblique est d'autant plus forte que la différence entre *Ma* et *Mx* est élevée.

Dans le contexte de la proposition P3, *pq* vaut 0,3 * 0,7 = 0,21 et racine de *pq* = 0,46. Le *rpbis* de P3 calculé selon la formule 1 vaut : *rpbis* = 1,66 ∗ 0,46 = 0,76. La valeur positive et élevée de ce coefficient de corrélation indique que les sujets qui ont choisi P3, en l'occurrence la réponse correcte, sont des sujets qui, globalement, obtiennent des scores plus élevés au test que ceux qui n'ont pas choisi P3.



Leclercq (op. cit. p.53) propose également une représentation graphique du *rpbis classique*. En voici une présentation adaptée à notre exemple. On retrouve sur ce graphique les axes a et x du graphique précédent ainsi que les positions des moyennes *Ma* (6,6) et *Mx* (16,2).

L'hypoténuse du triangle rectangle « t-o-v » est aussi le rayon (R sur le graphique) du cercle dont *Ma* est le centre et *Mx* un point de la circonférence. A l'aide du théorème de Pythagore nous pouvons calculer $R = \sqrt{(Mx - Ma)^2 + \frac{\sigma^2}{pq}} = \sqrt{92{,}16 + \frac{33{,}64}{0{,}21}} = 15{,}9$.

Connaissant R, nous pouvons ensuite calculer les sinus et cosinus de l'angle α (inscrits dans un cercle trigonométrique dont le rayon vaut 1). Sin_α = (*Mx* – *Ma*) / R = 9,6 / 15,9 = 0,603. $Cos_\alpha = \left(\frac{\sigma}{\sqrt{pq}} \right)$ / R = 12,7 / 15,9 = 0,798.

Enfin, nous obtenons la valeur du rpbis en calculant la tangente de l'angle α à partir de ces deux dernières valeurs : $Tan_\alpha = Sin_\alpha / Cos_\alpha$ = 0,603 / 0,798 = 0,756.

On voit que le *rpbis classique* correspond au quotient du sinus par le cosinus de l'angle α, la valeur de la tangente de cet angle. L'angle α nous donne aussi le coefficient angulaire (la pente) de la droite qui passe par les points *Ma* et *Mx* et qui vaut 37,08°.

Le principe de calcul du *rpbis classique* est le même pour chacune des propositions de la question et peut aussi être appliqué aux omissions.

On s'attend dans le cas des réponses incorrectes à des valeurs négatives ou proches de 0 pour le coefficient *rpbis classique*. On peut en effet logiquement s'attendre à ce que le choix d'une proposition incorrecte soit effectué par une proportion élevée d'étudiants qui récoltent les moins bons scores à l'ensemble des questions du test.

Voici le tableau des valeurs des *rpbis classiques* pour la question de notre exemple :

	OM	P1	P2	**P3**	P4
rpbis classique	-0,43	-0,17	-0,39	**0,76**	0,02

3. Inconvénients liés aux *rpbis classiques*

a) Problème du recouvrement entre le score de la question et le score total du test

Le *rpbis classique* pose un premier problème lié au recouvrement des performances des étudiants à la question dans le nombre de réponses correctes à l'ensemble du test (en anglais « *overlap* »).

Prenons le cas extrême d'un test comprenant 2 questions. Le taux de réussite de chaque question interviendrait alors pour la moitié de l'explication du taux de réussite global de l'épreuve. Dans ce cas de figure, le poids de la question dont on analyse les *rpbis classiques* serait tel que la corrélation bisériale de point en deviendrait anormalement surfaite.

Pour contrer ce problème de recouvrement entre le score de la question et le score total, Leclercq (*op. cit.*, p. 54) envisage trois solutions.

La première consiste à sortir le score de la question dans le calcul du score total de l'épreuve. La mesure critère devient donc le total sans la performance à la question considérée des réponses correctes au test pour chaque étudiant.

Une seconde solution consiste à calculer une valeur repère pour le rpbis de la réponse correcte, la valeur de ce repère dépend du nombre de questions (*nq*) posées dans le test. Leclercq (1987, p. 54) rappelle « *Dans le cas, purement théorique, où les questions d'un test n'auraient aucune corrélation entre elles (r = 0), la corrélation entre la question et le total du test vaut automatiquement* $1/\sqrt{nq}$ *, ainsi pour nq = 9, rpbis vaut 1/3 ou 0,33. C'est ce qu'on appelle la valeur repère du rpbis d'une question* ».

Hardy (1983) a élaboré une méthode de calcul de valeurs repères valables pour chaque solution d'une QCM, étant donné les proportions de choix.

Enfin, une troisième solution, proposée par Henrysson (1963), consiste à corriger vers le bas la valeur du *rpbis classique* à l'aide de la formule suivante :

$$C\,rpbis = \left(\sqrt{\frac{n}{n-1}}\right)\left(\frac{rpbis_i \;\; \sigma - \sqrt{pq}}{\sqrt{\sigma^2 - pq}}\right) \quad (35)$$

avec $C\,rpbis$ = corrélation point bisériale corrigée pour « recouvrement »
σ = écart type des résultats au total de l'épreuve

Dans le cadre des épreuves que nous corrigeons à l'aide du Système Méthodologique d'Aide à la Réalisation de Tests (SMART) nous avons choisi la seconde solution qui consiste à calculer une valeur repère pour la réponse correcte selon la formule :

$$valeur\ repère\ rpbis = 1/\sqrt{nq} \quad (36)$$

avec nq = nombre de questions dans le test

Dans le cas de notre exemple qui comportait 20 questions, cette valeur repère serait de $1 / \sqrt{20} = 0{,}22$, pour 25 questions ce serait 0,20, pour 36 questions ce serait 0,16. On voit que pour des épreuves académiques habituelles (où *nq* « *tourne autour de 20 à 30 questions* »), la valeur repère est proche de 0,20. C'est ce qui explique ce que disait De Landsheere (1979) (voir *supra* p. 159).

b) Problème de la non comparabilité des *rpbis classiques* d'une épreuve à l'autre

La valeur du *rpbis* classique d'une question est influencée par le score total au test qui, dans le cas de nos analyses des propositions d'une QCM, nous sert de mesure critère. Ce score total dépend des scores obtenus aux autres questions du test.

En général, les évaluateurs veillent à ne pas utiliser les mêmes questions d'une année à l'autre, les questions d'un test standardisé d'évaluation des connaissances des étudiants ne sont pas les mêmes dans les différents tests mêmes voulus « parallèles ». Après quelques sessions d'examen, les évaluateurs se retrouvent face à un nombre important de questions et tôt ou tard se pose le problème de la gestion d'une banque de questions.

Lorsqu'on gère une banque de questions il est tentant d'utiliser les *rpbis* classiques des tests antérieurs en vue de décisions liées à la sélection de questions qui feront partie ou non d'un test ultérieur. Cependant, le rpbis classique est avant tout un indice de cohérence par rapport au score total obtenu à l'ensemble des questions du test lui-même. D'une part, les questions sont différentes d'un test à l'autre et la valeur du *rpbis* classique pour une question qui se retrouve dans deux tests différents subit cette variation de contexte. D'autre part, d'une année académique à l'autre, les étudiants ne sont plus les mêmes (dans les situations de testing académique habituelles) et les rpbis classiques sont donc calculés à partir des performances des étudiants qui pourraient être très différentes d'un test à l'autre. Pour ces deux raisons, l'utilisation du *rpbis* classique comme indice absolu de cohérence interne est inopérante.

Le *rpbis classique* n'est donc pas un indice « *test free* ». Nous verrons plus loin que dans le cas du *rpbis Spectral Contrasté*, à la différence du *rpbis classique*, l'influence des autres questions contenues dans le test ne joue plus, les choix ou non choix d'une proposition d'une QCM n'étant plus corrélés avec les scores totaux à l'ensemble de l'épreuve mais avec les degrés de certitude qui ont accompagné ces choix au sein de la question. Ceci dit, comme pour le *rpbis classique*, dans le contexte d'examens universitaires, les étudiants sont évidemment différents d'une année académique à l'autre et on ne peut malheureusement pas être certain que la façon d'utiliser les degrés de certitude soit similaire dans deux groupes de répondants différents.

Cependant, nous verrons aussi que l'application de la turbo analyse qui débouche sur le calcul des rpbis Spectraux Contrastés Turbo (*rpbis SCT*), permet de sélectionner dans les matrices de résultats par question, les données des étudiants les plus réalistes, c'est-à-dire des répondants qui ont utilisé les degrés de certitude sans trop de sous-estimations ni trop de surestimations. Les données de ces étudiants réalistes sont donc d'une part plus valides et d'autre part nous donnent des points de « comparaison spectrale » pour calculer des *rpbis SCT* de questions identiques placées dans des tests soumis à des groupes d'étudiants différents mais dont les performances des sujets les plus réalistes peuvent être comparées du point de vue de leur gestion des degrés de certitude (grâce à l'indice de réalisme des sujets *Rs*).

B. Problématique du rpbis Spectral avec traitement Contrasté (rpbis SC)

1. Principe

Comme nous le verrons en détail, le recours aux degrés de certitude offre l'avantage de permettre le calcul d'un rpbis Spectral Contrasté (*rpbis SC*) qui au lieu de prendre comme mesure critère le nombre de réponses correctes obtenues à l'ensemble des autres questions du test, utilise les degrés de certitude qui ont accompagné les réponses au sein d'une question.

Le *rpbis SC* est calculé en corrélant les degrés de certitude (de 0 à 5 dans le cas des données de la recherche MOHICAN) qui ont accompagné les réponses à une QCM avec les choix/rejets (1 ou 0) des propositions de cette QCM.

Pour la réponse correcte, le *rpbis SC* positif, généralement obtenu, montre dans quelle mesure les sujets qui ont choisi la proposition correcte ont accompagné celle-ci de degrés de certitude en moyenne plus élevés que les degrés de certitude utilisés par les sujets qui se sont trompés.

Lorsqu'on calcule le *rpbis SC* d'un distracteur, nous prenons en compte les degrés de certitude des étudiants qui ont choisi ce distracteur et les degrés de certitude des étudiants qui ont choisi la réponse correcte, nous contrastons les certitudes liées au distracteur considéré avec les certitudes liées aux réponses correctes en excluant les certitudes liées aux autres distracteurs, d'où l'appellation « contrasté ».

Pour une proposition incorrecte, la corrélation des choix/rejets avec les degrés de certitude est en principe négative. Ces valeurs *rpbis SC* négatives auxquelles on s'attend montrent dans quelle mesure les choix du distracteur ont, en moyenne, été accompagnés de degrés de certitude moins élevés lorsqu'on compare avec les degrés de certitude utilisés par ceux qui ont choisi la réponse correcte.

2. Matrice des données

Voici la matrice reprenant les choix d'un groupe de 20 étudiants pour une question à choix multiple comprenant 4 propositions, de P1 à P4 avec une possibilité d'omission notée « OM ». Rappelons que dans l'exemple, P3 correspond à la réponse correcte pour cette question (données en gras sur fond grisé). Pour chaque étudiant, le choix d'une proposition est exprimé par un 1 et le non choix des autres possibilités par des 0. Ainsi, l'étudiant « etu1 » a choisi la proposition P4, l'étudiant « etu2 » la proposition P1, « etu3 » la proposition P3 qui est correcte, etc.

Nous rappelons ci-contre la correspondance entre les degrés de certitude (DC) cochés sur les formulaires de réponses spéciaux pour la lecture optique (formuloms) et les pourcentages de certitude (%C) employés dans le cadre des épreuves MOHICAN.

DC	%C
0	0%
1	20%
2	40%
3	60%
4	80%
5	100%

La colonne DC de la matrice ci-dessous reprend pour chaque étudiant le degré de certitude (ici entre 0 et 5, un nombre plus élevé indique donc une certitude plus grande) qui a accompagné la réponse sélectionnée. Par exemple, « etu1 » a accompagné la proposition P4 d'un degré de certitude 0 et l'étudiant « etu2 » la proposition P1 d'un degré de certitude 1.

	OM	P1	P2	**P3**	P4	DC
etu1	0	0	0	**0**	1	0
etu2	0	1	0	**0**	0	1
etu3	0	0	0	**1**	0	4
etu4	0	0	0	**1**	0	5
etu5	1	0	0	**0**	0	0
etu6	0	0	0	**1**	0	3
etu7	1	0	0	**0**	0	0
etu8	0	0	0	**0**	1	3
etu9	0	0	0	**1**	0	2
etu10	0	0	1	**0**	0	0
etu11	0	0	0	**0**	1	1
etu12	0	1	0	**0**	0	2
etu13	0	0	0	**1**	0	5
etu14	0	1	0	**0**	0	1
etu15	0	0	1	**0**	0	0
etu16	0	0	1	**0**	0	3
etu17	0	0	1	**0**	0	2
etu18	0	0	0	**0**	1	2
etu19	0	1	0	**0**	0	0
etu20	0	0	0	**1**	0	4
Moyennes	0,1	0,2	0,2	**0,3**	0,2	1,9
Ecart type						1,67

3. Méthode de calcul du rpbis Spectral Contrasté (*rpbis SC*)

a) Calcul du *rpbis SC* dans le cas d'une réponse correcte

Nous proposons l'adaptation suivante de la formule 1 du *rpbis* classique lorsqu'il s'agit de calculer le rpbis Spectral Contrasté d'une réponse correcte (*rpbis SC*^{c}) :

$$rpbis\ SC^c = \frac{Dx^c - Da^c}{\sigma}\sqrt{pq} \quad (37)$$

Avec Dx^c = la moyenne des Degrés de certitude des étudiants qui ont choisi la réponse correcte

Da^c = la moyenne des Degrés de certitude des étudiants qui n'ont pas choisi la réponse correcte

σ = l'écart type des degrés de certitude de tous les étudiants

p = la proportion d'étudiants qui ont choisi la réponse correcte de la question

q = la proportion d'étudiants qui n'ont pas choisi la réponse correcte de la question

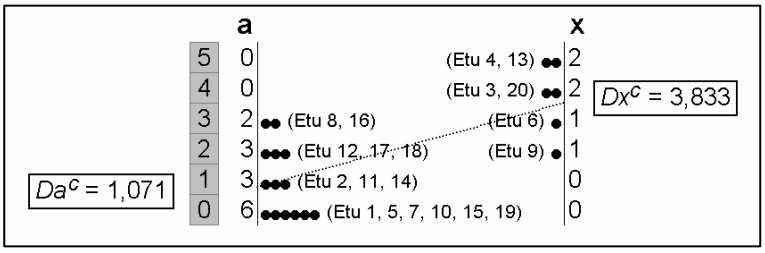
Pour la lisibilité de la formule nous avons ajouté un c en exposant car nous généraliserons ensuite dans la section suivante le calcul du *rpbis SC* aux réponses incorrectes.

Dans le cas de la proposition P3, le *rpbis SC c* est calculé en utilisant cette partie de la matrice :

	P3	DC
etu1	**0**	0
etu2	**0**	1
etu3	**1**	4
etu4	**1**	5
etu5	**0**	0
etu6	**1**	3
etu7	**0**	0
etu8	**0**	3
etu9	**1**	2
etu10	**0**	0
etu11	**0**	1
etu12	**0**	2
etu13	**1**	5
etu14	**0**	1
etu15	**0**	0
etu16	**0**	3
etu17	**0**	2
etu18	**0**	2
etu19	**0**	0
etu20	**1**	4

Reprenons le graphique présenté dans le cadre du *rpbis* classique mais en remplacant cette fois les nombres de réponses correctes par les degrés de certitude (de 0 à 5 dans la zone grisée). L'axe « a » correspond au rejet de la réponse correcte P3 (les « 0 » dans la colonne P3 de la matrice ci-dessus). La position de chaque sujet représenté par un rond noir correspond au degré de certitude choisi (par exemple, Etu 8 et Etu 16 ont accompagné leur réponse incorrecte du degré de certitude 3). L'axe « x » reprend quant à lui les positions des sujets qui ont choisi P3 (les « 1 » dans la colonne P3 de la matrice ci-dessus). Les positions des ronds noirs sur l'axe « x » montrent que les sujets qui ont choisi P3 ont, en moyenne, accompagné cette réponse correcte de degrés de certitude plus élevés que les autres sujets.

Soient Dx^c et Da^c les moyennes des Degrés de certitude respectivement pour les étudiants qui ont choisi la réponse correcte (P3) et pour les étudiants qui n'ont pas choisi cette solution (qui ont donc répondu P1, P2 ou P4 ou omis OM).



La différence $Dx^c - Da^c$ pour la proposition P3 vaut 3,833-1,071 = 2,762. Rapportée à l'écart type, cette différence vaut 2,762/1,67 = 1,653. La valeur positive de cette différence indique qu'il y a concordance entre le choix de P3 et l'utilisation de degrés de certitude élevés.

Dans le contexte de notre exemple, le *rpbis SC c* de la proposition P3 calculé selon la formule 1 vaut 1,653 ∗ racine de 0,21 = 0,76. La valeur positive et élevée de ce coefficient de corrélation indique que les étudiants qui ont choisi P3, en l'occurrence la réponse correcte, sont aussi les étudiants qui ont accompagné le choix de cette proposition P3 d'un degré de certitude en moyenne plus élevé que le groupe des étudiants qui ont opté pour une autre réponse.

En d'autres termes, le *rpbis SC c* de la réponse correcte est calculé en corrélant les données de la colonne des choix/rejets de la proposition correspondant à la réponse correcte avec la colonne des degrés de certitude qui ont accompagné, selon les cas, soit la réponse correcte, soit une réponse incorrecte.

Dans notre exemple cette corrélation est de 0,76. Nous dirons que le *rpbis SC* ^{c} pour la réponse correcte vaut 0,76 (nous consacrerons plus loin un paragraphe à l'interprétation des valeurs obtenues).

Nous allons maintenant généraliser le calcul du *rpbis SC* aux propositions incorrectes.

b) Calcul du *rpbis SC* dans le cas des propositions incorrectes

Nous proposons de définir le *rpbis SC* ^{i} dans le cas des propositions incorrectes par :

$$rpbis\ SC^{\,i} = \frac{Dx^{i} - Da^{i}}{\sigma}\sqrt{pq} \qquad (38)$$

Avec Dx^{i} = la moyenne des Degrés de certitude des étudiants qui ont choisi la réponse incorrecte pour laquelle nous calculons le *rpbis SC*^{i}

Da^{i} = la moyenne des Degrés de certitude des étudiants qui ont répondu correctement

σ = l'écart type des degrés de certitude des étudiants sélectionnés, c'est-à-dire de ceux qui ont choisi la réponse incorrecte envisagée et ceux qui ont répondu correctement

p = parmi les étudiants sélectionnés, la proportion de ceux qui ont choisi la réponse incorrecte pour laquelle nous calculons le *rpbis SC*^{i}

q = parmi les étudiants sélectionnés, la proportion d'étudiants qui ont choisi la proposition correcte

Envisageons la procédure de calcul. Prenons la proposition incorrecte P1. Dans un premier temps nous procéderons à une sélection des étudiants dont les données interviendront dans le calcul.

En sélectionnant d'une part les étudiants qui ont choisi la proposition P1 (celle pour laquelle nous calculons le *rpbis SC*^{i}) et d'autre part les étudiants qui ont choisi la réponse correcte P3 nous contrastons les résultats en évitant de faire intervenir les autres réponses incorrectes (P2, P4 et OM) dans le calcul de la corrélation. Ceci nous amène à éliminer une série de sujets dans la matrice de départ :

	OM	P1	P2	**P3**	P4	DC
~~*etu1*~~	~~*0*~~	~~*0*~~	~~*0*~~	~~***0***~~	~~*1*~~	~~*0*~~
etu2	0	1	0	**0**	0	1
etu3	0	0	0	**1**	0	4
etu4	0	0	0	**1**	0	5
~~*etu5*~~	~~*1*~~	~~*0*~~	~~*0*~~	~~***0***~~	~~*0*~~	~~*0*~~
etu6	0	0	0	**1**	0	3
~~*etu7*~~	~~*1*~~	~~*0*~~	~~*0*~~	~~***0***~~	~~*0*~~	~~*0*~~
~~*etu8*~~	~~*0*~~	~~*0*~~	~~*0*~~	~~***0***~~	~~*1*~~	~~*3*~~
etu9	0	0	0	**1**	0	2
~~*etu10*~~	~~*0*~~	~~*0*~~	~~*1*~~	~~***0***~~	~~*0*~~	~~*0*~~
~~*etu11*~~	~~*0*~~	~~*0*~~	~~*0*~~	~~***0***~~	~~*1*~~	~~*1*~~
etu12	0	1	0	**0**	0	2
etu13	0	0	0	**1**	0	5
etu14	0	1	0	**0**	0	1
~~*etu15*~~	~~*0*~~	~~*0*~~	~~*1*~~	~~***0***~~	~~*0*~~	~~*0*~~
~~*etu16*~~	~~*0*~~	~~*0*~~	~~*1*~~	~~***0***~~	~~*0*~~	~~*3*~~
~~*etu17*~~	~~*0*~~	~~*0*~~	~~*1*~~	~~***0***~~	~~*0*~~	~~*2*~~
~~*etu18*~~	~~*0*~~	~~*0*~~	~~*0*~~	~~***0***~~	~~*1*~~	~~*2*~~
etu19	0	1	0	**0**	0	0
etu20	0	0	0	**1**	0	4

Nous obtenons la sélection suivante :

	OM	P1	P2	**P3**	P4		DC
etu2	0	1	0	**0**	0		1
etu3	0	0	0	**1**	0		4
etu4	0	0	0	**1**	0		5
etu6	0	0	0	**1**	0		3
etu9	0	0	0	**1**	0		2
etu12	0	1	0	**0**	0		2
etu13	0	0	0	**1**	0		5
etu14	0	1	0	**0**	0		1
etu19	0	1	0	**0**	0		0
etu20	0	0	0	**1**	0		4

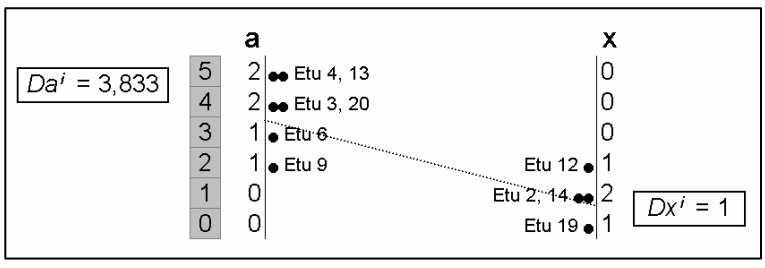
Les 10 étudiants sélectionnés ont choisi soit la proposition incorrecte P1 (les 1 dans la colonne P1), soit la réponse correcte (les 1 dans la colonne P3). Si nous éliminons ensuite les colonnes OM, P2, P3, P4 et supprimons les lignes vides, nous obtenons la matrice « bisériale » ci-après.

	P1		DC
etu2	1		1
etu3	0		4
etu4	0		5
etu6	0		3
etu9	0		2
etu12	1		2
etu13	0		5
etu14	1		1
etu19	1		0
etu20	0		4

Ecart type : 1,676

Donc, parmi ces 10 étudiants, 4 ont choisi la proposition P1, les 6 autres ont répondu correctement.

Sur la représentation graphique ci-contre l'axe « a » montre quels degrés de certitude ont été choisi par les étudiants qui ont répondu correctement P3 (les 0 dans la colonne P1 de la dernière matrice ci-dessus). L'axe « x » correspond aux degrés de certitude qui ont accompagné le choix de la proposition incorrecte P1.



La différence $Dx^i - Da^i$ vaut –2,833. Rapportée à l'écart type, cette différence vaut -2,833/1,676 = -1,69. Le signe négatif de cette différence indique que les sujets qui ont choisi le distracteur P1 ont, en moyenne, accompagné leur choix incorrect de degrés de certitude moins élevés que les degrés de certitude utilisés par les sujets qui ont répondu correctement.

Il est à remarquer que le Da^i d'une réponse incorrecte vaut le Dx^c de la réponse correcte (p. 180).

Dans le contexte du calcul du *rpbis* SC^i pour cette proposition P1, *pq* vaut 0,4 * 0,6 = 0,24 et racine de *pq* = 0,489. Le *rpbis* SC^i de P1 calculé selon la formule proposée au début de cette section vaut : -1,69 ∗ 0,489 = -0,83. La valeur négative de ce coefficient de corrélation indique que les étudiants qui ont choisi P1, une réponse incorrecte, sont aussi les étudiants qui ont accompagné le choix de cette proposition d'un degré de certitude en moyenne moins élevé que le groupe des étudiants qui ont opté pour la réponse correcte P3.

Nous dirons que le *rpbis* SC^i pour la proposition incorrecte P3 vaut -0,83.

Ce principe de calcul est appliqué aux autres propositions incorrectes ainsi qu'aux omissions.

c) Interprétation des valeurs des *rpbis SC* des propositions de la question envisagée dans notre exemple

Une fois calculés, les *rpbis SC* des propositions d'une question sont synthétisés dans un tableau. Pour faciliter la lecture des résultats nous indiquons la réponse correcte en gras sur fond grisé et rassemblons sous l'appellation *rpbis SC* le *rpbis* SC^c (celui de la proposition P3) et les *rpbis* SC^i (de P1, P2, P4 et de l'omission) de la question envisagée dans notre exemple.

	OM	P1	P2	**P3**	P4
rpbis SC	-0,87	-0,83	-0,74	**0,76**	-0,72

Le *rpbis SC* positif pour la réponse correcte (0,76) signifie que lorsqu'on compare les valeurs des degrés de certitude des étudiants qui ont choisi la réponse correcte avec les valeurs de ceux qui ont répondu incorrectement, on observe des degrés de certitude en moyenne plus élevés chez ceux qui ont répondu correctement à la question. C'est la situation à laquelle on peut logiquement s'attendre lorsqu'une réponse correcte fonctionne bien.

Si le *rpbis SC* de la réponse correcte était négatif cela signifierait que les étudiants qui ont choisi la bonne réponse l'ont fait, en moyenne, de façon moins assurée que ceux qui ont choisi les propositions incorrectes, et qui ont eux, en moyenne, accompagné leurs erreurs avec des degrés de certitude plus élevés. Notre hypothèse est qu'alors la qualité de la question pourrait être remise en cause. Nous pensons que lorsque ce cas de figure se présente, il faut prendre des informations complémentaires auprès du concepteur de la question, de l'enseignant et des étudiants pour ensuite procéder à des vérifications et éventuellement mettre en place des mesures correctives (supprimer la question, s'interroger à propos de la réponse correcte, valoriser une proposition incorrecte).

Il convient aussi de vérifier les valeurs du *rpbis SC* pour les propositions incorrectes et l'omission. Dans notre exemple, les *rpbis SC* négatifs des réponses incorrectes (omission [-0,87], P1 [-0,81], P2 [-0,74] et P4 [-0,72]) montrent que les étudiants qui n'ont pas fourni la réponse correcte ont systématiquement accompagnés leurs choix de degrés de certitude moins élevés que le groupe des étudiants qui ont choisi la réponse correcte. C'est la situation à laquelle on peut logiquement s'attendre pour les propositions incorrectes d'une QCM bien rédigée.

Enfin, terminons cette section relative au *rpbis SC* en notant qu'il serait tout à fait possible d'appliquer le principe du traitement contrasté dans le contexte du calcul des *rpbis classiques*. En effet, nous pourrions aussi dans le cas du calcul des *rpbis classiques* des propositions incorrectes décider de ne prendre en compte que les données des étudiants qui ont répondu correctement et les données de ceux qui ont choisi la proposition envisagée. Nous comptons appliquer ce principe du traitement contrasté aux *rpbis classiques* lors de travaux de recherche ultérieurs.

C. Problématique du coefficient de corrélation bisériale de point Spectral Contrasté avec turbo analyse (rpbis SCT)

1. Principe

L'idée de départ est d'intégrer le réalisme des étudiants dans le calcul du rpbis Spectral Contrasté. Tenir compte du réalisme des étudiants dans le *rpbis SC* permet d'augmenter le crédit que nous pouvons accorder aux valeurs obtenues. En effet, si nous calculons le *rpbis SC* à l'aide des données provenant des seuls étudiants qui s'auto-estiment particulièrement bien, sans trop de sous-estimations ni trop de surestimations, nous augmentons la fiabilité des valeurs du *rpbis SC*. A l'inverse, si nous nous basons sur toutes les données et que bon nombre d'étudiants sont peu réalistes, nos indices *rpbis SC* seraient biaisés, entachés par l'incohérence d'utilisation des degrés de certitude.

a) Calcul du *R*éalisme des sujets (*Rs*)

Le réalisme d'un étudiant pour un test donné se calcule sur l'ensemble des certitudes que l'étudiant à fournies à ce test. Voici la formule utilisée par les concepteurs du projet MOHICAN (Leclercq & al, 2000) pour calculer l'indice de *R*éalisme des *s*ujets (*Rs*) :

$$\boxed{Rs = 100 - EMAC} \quad (39)$$

Avec

$EMAC = \frac{\sum_i (|\, C_i - TE_i \,|).NU_i}{NR_t}$ = l'Erreur Moyenne Absolue de Certitude

i = indice des degrés de certitude

C_i = Valeur de la certitude i (en pourcents)

NC_i = Nombre de réponses correctes pour la certitude i

NU_i = Nombre d'utilisations de la certitude i *(si* $NU_i = 0$ alors l'indice i est ignoré)

TE_i = Taux d'Exactitude de la certitude i (en pourcents) = 100 x NC_i / NU_i

NR_t = Nombre total de réponses par test ($\Sigma_i NU_i$)

NC = Nombre total de réponses correctes ($\Sigma_i NC_i$)

Voici un exemple de tableau récapitulatif des réponses et certitudes fournies par un étudiant :

Degrés de Certitude	0	1	2	3	4	5	
C_i	0	20	40	60	80	100	Σ
NU_i	1	1	4	3	3	0	12
NC_i	0	0	3	3	3	0	9
TE_i	0	0	75	100	100	XXX	
$\|C_i - TE_i\|$	0	20	35	40	20	XXX	
$(\|\, C_i - TE_i \,|).\, NU_i$	0	20	140	120	60	XXX	340

$$EMAC = \frac{340}{12} = 28{,}3$$

$$Rs = 100 - 28{,}3 = 71{,}7$$

Les cases marquées XXX correspondent aux degrés de certitude non utilisés ($NU_i = 0$).

Dans notre exemple, l'étudiant a fourni une fois la certitude « 0% » pour une réponse qui s'est avérée incorrecte, sa prédiction s'est réalisée et on peut dire dans ce cas que son auto-estimation était parfaite car (1) sa prédiction « *aucune chance d'être correcte* » correspond à la réalité du fait « *la réponse est incorrecte* » et (2) il n'a pas commis d'Erreur de Certitude ($|C_i - TE_i| = 0$) avec la certitude 0%, son

taux d'exactitude pour cette Certitude (TE_i) vaut aussi 0. Le total pondéré des erreurs lié à cette certitude i = 0% (($|TE_i - C_i|$). NU_i) pondère la valeur de l'erreur de certitude en la multipliant par le nombre d'utilisations de la certitude (NU_i), dans le cas de cette certitude 0%, le total pondéré des erreurs vaut 0% (1 * 0% = 0%).

On voit également dans l'exemple que l'étudiant n'a pas fourni de certitude « 100% », dans ce cas, les autres valeurs ne peuvent être calculées et ceci explique les cases contenant XXX dans le tableau.

La dernière colonne du tableau contient les sommes des NU_i et ($| C_i - TE_i |$). NU_i. Dans notre exemple l'étudiant a répondu aux 12 questions en accompagnant ses réponses de certitudes, donc ΣNU_i vaut 12. La somme des totaux pondérés des erreurs de certitude vaut 340%. Dès lors, on peut calculer l'Erreur Moyenne Absolue de Certitude (*EMAC*) en divisant la somme des totaux pondérés des erreurs de certitude par le nombre de réponses.

Dans notre exemple, *EMAC* vaut 28,3% (340% / 12). Cela signifie qu'en moyenne pour les 1es questions du test, le sujet de notre exemple commet 28,3% d'erreurs de certitude.

A partir de l'indice de l'Erreur Moyenne Absolue de Certitude (*EMAC*) il est possible de calculer le Réalisme des sujets (*Rs*) en soustrayant la valeur de *EMAC* à 100 :

$$Rs = 100 - EMAC$$

Ce qui signifie que l'étudiant qui commet en moyenne 28,3% d'erreurs de certitude sur l'ensemble de ses réponses au test, manifeste un réalisme de 71,7 (100 - 28,3).

Cette formule de calcul du réalisme résulte d'une simplification de la consigne de recueil des degrés de certitude auprès des étudiants entrant à l'université en 1ère candidature et qui forment le public visé par les épreuves MOHICAN. C'est une autre formule qui est habituellement utilisée avec les étudiants de la Faculté de Psychologie et des Sciences de l'Education (FAPSE) de l'Université de Liège qui sont préalablement entraînés à l'auto-estimation de leurs réponses avant les examens et qui dans ce contexte sont confronté à une échelle des degrés de certitude différente. Pour une comparaison des échelles et des formules de réalisme utilisées dans MOHICAN et à la FAPSE, voir *infra* p. 272).

b) Principe de la « turbo analyse »

Reprenons la matrice des choix du groupe de 20 étudiants que nous avons utilisée dans le cadre du calcul des rpbis Spectraux Contrastés. Nous avons ajouté une colonne supplémentaire contenant les valeurs des réalismes des sujets calculés selon la méthode exposée précédemment (p. 184).

	OM	P1	P2	**P3**	P4	DC	Réalisme
etu1	0	0	0	**0**	1	0	55
etu2	0	1	0	**0**	0	1	86
etu3	0	0	0	**1**	0	4	81
etu4	0	0	0	**1**	0	5	92
etu5	1	0	0	**0**	0	0	83
etu6	0	0	0	**1**	0	3	74
etu7	1	0	0	**0**	0	0	43
etu8	0	0	0	**0**	1	3	76
etu9	0	0	0	**1**	0	2	43
etu10	0	0	1	**0**	0	0	93
etu11	0	0	0	**0**	1	1	82
etu12	0	1	0	**0**	0	2	66
etu13	0	0	0	**1**	0	5	89
etu14	0	1	0	**0**	0	1	82
etu15	0	0	1	**0**	0	0	94
etu16	0	0	1	**0**	0	3	34
etu17	0	0	1	**0**	0	2	72
etu18	0	0	0	**0**	1	2	87
etu19	0	1	0	**0**	0	0	86
etu20	0	0	0	**1**	0	4	84
Moyennes	0,1	0,2	0,2	**0,3**	0,2	1,9	
Ecart type						1,67	

Le principe de la turbo analyse appliquée au rpbis Spectral Contrasté est de calculer cet indice sur la base des étudiants les plus réalistes, ceux dont les données sont les plus fiables du point de vue de gestion des degrés de certitude. Nous avons donc opéré une sélection parmi ces étudiants en ne prenant en compte que les sujets dont le réalisme est supérieur à une valeur que nous fixons arbitrairement dans cet exemple à 80. C'est donc en nous basant sur les scores de réalisme (dernière colonne) que nous éliminons les données d'une série d'étudiants dont le réalisme est inférieur à 80 :

	OM	P1	P2	**P3**	P4	DC	Réalisme
~~etu1~~	~~0~~	~~0~~	~~0~~	~~**0**~~	~~1~~	~~0~~	~~55~~
etu2	0	1	0	**0**	0	1	86
etu3	0	0	0	**1**	0	4	81
etu4	0	0	0	**1**	0	5	92
etu5	1	0	0	**0**	0	0	83
~~etu6~~	~~0~~	~~0~~	~~0~~	~~**1**~~	~~0~~	~~3~~	~~74~~
~~etu7~~	~~1~~	~~0~~	~~0~~	~~**0**~~	~~0~~	~~0~~	~~43~~
~~etu8~~	~~0~~	~~0~~	~~0~~	~~**0**~~	~~1~~	~~3~~	~~76~~
~~etu9~~	~~0~~	~~0~~	~~0~~	~~**1**~~	~~0~~	~~2~~	~~43~~
etu10	0	0	1	**0**	0	0	93
etu11	0	0	0	**0**	1	1	82
~~etu12~~	~~0~~	~~1~~	~~0~~	~~**0**~~	~~0~~	~~2~~	~~66~~
etu13	0	0	0	**1**	0	5	89
etu14	0	1	0	**0**	0	1	82
etu15	0	0	1	**0**	0	0	94
~~etu16~~	~~0~~	~~0~~	~~1~~	~~**0**~~	~~0~~	~~3~~	~~34~~
~~etu17~~	~~0~~	~~0~~	~~1~~	~~**0**~~	~~0~~	~~2~~	~~72~~
etu18	0	0	0	**0**	1	2	87
etu19	0	1	0	**0**	0	0	86
etu20	0	0	0	**1**	0	4	84

Cette opération de sélection de données aboutit à une nouvelle matrice :

	OM	P1	P2	**P3**	P4	DC	Réalisme
etu2	0	1	0	**0**	0	1	86
etu3	0	0	0	**1**	0	4	81
etu4	0	0	0	**1**	0	5	92
etu5	1	0	0	**0**	0	0	83
etu10	0	0	1	**0**	0	0	93
etu11	0	0	0	**0**	1	1	82
etu13	0	0	0	**1**	0	5	89
etu14	0	1	0	**0**	0	1	82
etu15	0	0	1	**0**	0	0	94
etu18	0	0	0	**0**	1	2	87
etu19	0	1	0	**0**	0	0	86
etu20	0	0	0	**1**	0	4	84
Moyennes	0,08	0,25	0,17	0,33	0,17	1,92	
						Ecart type 1,93	

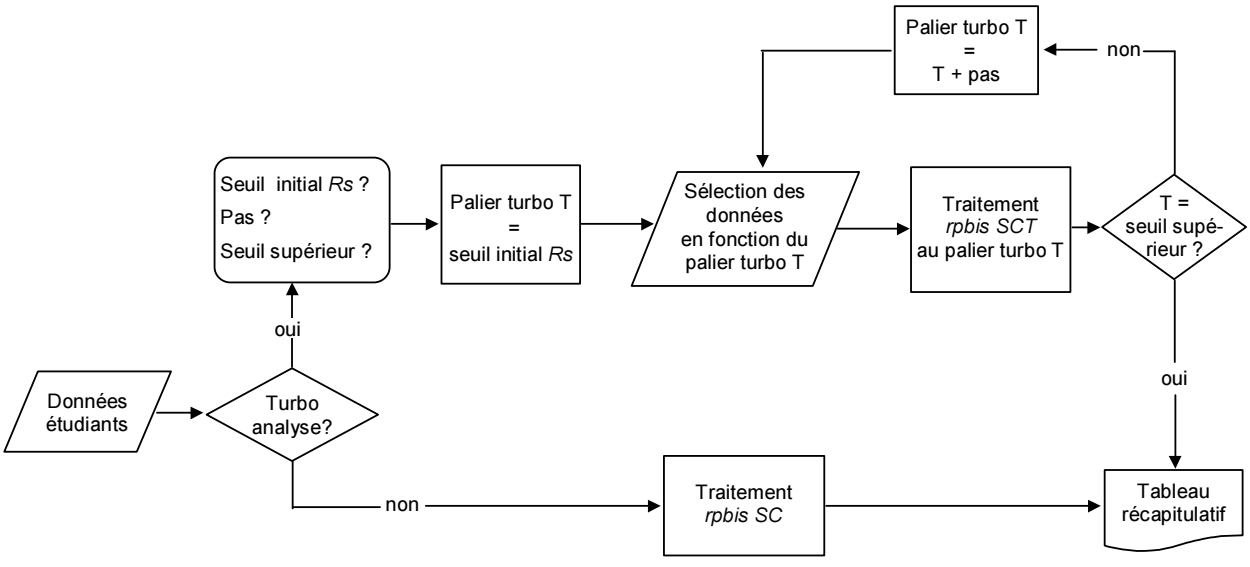
Nous calculons ensuite les *rpbis SC* à partir des données de ces étudiants dont le score en réalisme est supérieur ou égal à 80. Remarquons que nous aurions pu réaliser une autre turbo analyse (et nous le ferons dans la suite) en fixant le seuil exigé de réalisme à 90 (ou toute autre valeur comprise entre 0 et 100). Pour faciliter l'interprétation des *rpbis SC* issus de turbo analyses nous ajouterons à l'appellation *rpbis SC* un *T* suivi du seuil de réalisme, dans cet exemple 80 (exprimé en pourcents) qui a été exigé pour sélectionner les données (donc *rpbis SCT80*).

Calculés à partir de la nouvelle matrice de notre exemple, les *rpbis SCT80* des propositions et de l'omission valent :

	OM	P1	P2	**P3**	P4
rpbis SCT80	-0,97	-0,97	-0,98	**0,94**	-0,94

Dans le cadre de ce processus, l'appellation « *turbo analyse* » est à mettre en relation avec la possibilité d'enchaîner successivement des traitements *rpbis SCT* en exigeant à chaque fois un seuil de réalisme plus élevé pour la sélection des données. Nous avons programmé un logiciel intitulé *SCANTEST 2.0 pour épreuves MOHICAN Check up '99* (voir p. 195) qui nous permet de réaliser de tels enchaînements de traitements. Avec ce logiciel, au fur et à mesure des sélections des données des étudiants de plus en plus réalistes, nous obtenons des *rpbis SCT* de plus en plus fiables.

En résumé, dans le contexte de notre exemple on peut schématiser la turbo analyse de la façon suivante :



Dans notre exemple, nous avons sélectionné les étudiants dont le réalisme était supérieur ou égal à 80, nous parlerons alors de palier de Turbo analyse « T80 », ce qui a aboutit à la sélection des données de 12 étudiants dont le score *Rs* est supérieur ou égal à 80 sur les 20 de la matrice initiale. Nous pouvons encore augmenter le seuil de réalisme exigé et le porter à 90 (palier « T90 ») mais il ne nous reste alors plus que les données de 3 étudiants pour effectuer le traitement *rpbis SCT* :

	OM	P1	P2	**P3**	P4
etu4	0	0	0	**1**	0
etu10	0	0	1	**0**	0
etu15	0	0	1	**0**	0

DC
5
0
0

Réalisme
92
93
94

Ceci entraîne l'impossibilité de calculer le *rpbis SCT* de propositions qui n'ont pas été choisies par au moins un des trois étudiants (ici, les propositions P1 et P4 ainsi que l'omission OM) et explique les « XXXX » dans le tableau récapitulatif des *rpbis SCT90* ci-dessous.

	OM	P1	P2	**P3**	P4
rpbis SCT90	XXXX	XXXX	-1	**1**	XXXX

Au palier de turbo analyse « T90 », un seul sujet a répondu correctement en accompagnant sa réponse d'un degré de certitude 5 et les deux autres sujets ont choisi la proposition P2 en l'accompagnant d'un même degré de certitude 0. Les trois conditions évoquées précédemment (p. 158) sont donc réunies pour que le *rpbis SCT90* de P3 soit égal à 1 et celui de P2 égal à –1.

En effet, un *rpbis spectral* récolte la valeur maximum « 1 » lorsqu'on observe simultanément les trois conditions suivantes dans les résultats des sujets sélectionnés en fonction d'une turbo analyse ou/et d'un traitement contrasté :

[1] les sujets qui ont choisi la solution accompagnent tous leur réponse d'un même degré de certitude ;

[2] les sujets qui n'ont pas choisi cette solution accompagnent tous leur réponse d'un même degré de certitude ;

[3] le degré de certitude des sujets qui ont opté pour la solution est supérieur au degré de certitude de ceux qui ne l'ont pas choisie.

Ces trois conditions étaient réunies pour que le *rpbis SCT90* de P3 récolte la valeur maximum « 1 ».

Un *rpbis spectral* récolte la valeur « -1 » lorsque les conditions [1] et [2] sont réunies et, en ce qui concerne la condition [3], lorsque le degré de certitude des sujets qui ont opté pour la solution est inférieur au degré de certitude de ceux qui ne l'ont pas choisie. Ces trois dernières conditions étaient réunies pour que le *rpbis SCT90* de P2 récolte la valeur « -1 ».

***

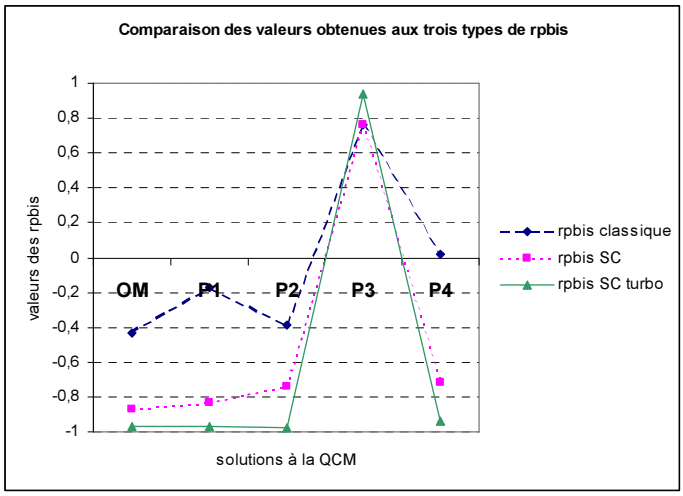
Signalons ici qu'il existe très probablement d'autres méthodes qui permettraient de tirer parti des performances en réalisme des sujets pour améliorer la fiabilité des *rpbis SC*. Il en existe au moins une, suggérée par M. Daniel Defays, qui consisterait à donner plus de poids aux données des étudiants dont le réalisme est élevé. L'avantage réside alors dans le fait qu'une fois les résultats pondérés en fonction du niveau de réalisme il devient possible de prendre en compte les données de tous les sujets. Nous comptons exploiter cette voie méthodologique prometteuse dans nos prochains travaux de recherche, notamment lorsque nous serons confronté à des situations où le nombre de sujets impliqués est moins élevé que dans le cadre des épreuves MOHICAN. Nous verrons plus loin que dans le cadre de « *check up '99* » les nombres très élevés d'étudiants permettent des effectifs encore importants au palier de turbo analyse T80.

2. Comparaison des valeurs obtenues aux *rpbis classique*, *rpbis SC* et *rpbis SCT80* dans le contexte de notre exemple

a) Récapitulatif des valeurs obtenues

Voici le tableau :

	OM	P1	P2	**P3**	P4
rpbis classique	-0,43	-0,17	-0,39	**0,76**	0,02
rpbis SC	-0,87	-0,83	-0,74	**0,76**	-0,72
rpbis SCT80	-0,97	-0,97	-0,98	**0,94**	-0,94



Le graphique ci-contre reprend ces valeurs. En ce qui concerne les rpbis des solutions incorrectes on observe une augmentation négative des valeurs en fonction du type de rpbis :

rpbis classique > *rpbis SC* > *rpbis SCT80*,

avec des différences très fortes entre d'une part les *rpbis* classiques et d'autre part les deux types de rpbis spectraux. Quant aux rpbis de la solution correcte, les différences sont moins marquées lorsqu'on compare les trois types. Dans notre exemple, la valeur du *rpbis* classique pour la réponse correcte (0,76) est égale à celle obtenue par le *rpbis SC* (0,76). La valeur du *rpbis SCT80* (0,94) dépasse quant à elle celles des *rpbis* classique et *rpbis SC*. Les lignes sur le graphique ne sont là que pour guider l'œil du lecteur.

b) Différences entre la moyenne pondérée des rpbis pour les propositions incorrectes et le rpbis de la proposition correcte

Il est utile de calculer la moyenne pondérée des valeurs des rpbis obtenue par les propositions incorrectes d'une QCM. Nous définissons la moyenne pondérée des rpbis des propositions incorrectes par :

$$\overline{rpbis^i} = \frac{\sum_{j=1}^{nj} (rpbis_j^i * n_j^i)}{NR^i} \quad (40)$$

avec :

j = indice des propositions incorrectes pour une question donnée

nj = *n*ombre de propositions incorrectes j pour une question donnée

n_j^i = nombre d'utilisations de la proposition *i*ncorrecte j

$rpbis_j^i$ = la valeur du *rpbis* de la proposition incorrecte j

NR^i = le *N*ombre total de *R*éponses *i*ncorrectes à la question (= $\sum_j n_j^i$)

Pour des raisons de lisibilité, nous ajoutons un « *i* » en exposant pour désigner le fait qu'il s'agit de la réponse incorrecte et une barre au dessus de l'indice $rpbis^i$ lorsque la valeur de ce dernier a été calculée en effectuant la moyenne pondérée des rpbis des propositions incorrectes.

Dans le contexte de notre exemple, voici ce que donne cette méthode de calcul de la moyenne pondérée des rpbis des propositions incorrectes de la question :

- pour le rpbis classique

	OM	P1	P2	P4	Σ
$rpbis^i_j$	-0,43	-0,17	-0,39	0,02	
n^i_j	2	4	4	4	$NR^i = 14$
$rpbis^i_j * n^i_j$	-0,87	-0,69	-1,56	0,09	$\overline{rpbis^i} =$ -3,03

$\overline{rpbis^i}$ *classique* = -3,03 / 14 = -0,22

- pour le rpbis Spectral Contrasté

	OM	P1	P2	P4	Σ
$rpbis^i_j$	-0,87	-0,83	-0,74	-0,72	
n^i_j	2	4	4	4	$NR^i = 14$
$rpbis^i_j * n^i_j$	-1,747	-3,312	-2,942	-2,898	$\overline{rpbis^i} =$ -10,9

$\overline{rpbis^i}$ *SC* = -10,9 / 14 = -0,78

- pour le rpbis Spectral Contrasté avec turbo analyse

	OM	P1	P2	P4	Σ
$rpbis^i_j$	-0,97	-0,97	-0,98	-0,94	
n^i_j	1	3	2	2	$NR^i = 8$
$rpbis^i_j * n^i_j$	-0,97	-2,905	-1,963	-1,885	$\overline{rpbis^i} =$ -7,725

$\overline{rpbis^i}$ *SCT80* = -7,725 / 8 = -0,97

Voici le tableau comparatif des valeurs obtenues pour les différents types de $\overline{rpbis^i}$. Nous y avons ajouté deux colonnes « $rpbis^c$ » correspondant au rpbis obtenu par la réponse correcte (ici P3) et « $rpbis^c$ - $\overline{rpbis^i}$ » correspondant à la différence entre le rpbis de la réponse correcte et la moyenne pondérée des rpbis des réponses incorrectes de façon à mettre en évidence pour chaque type de rpbis l'amplitude de cette différence. Cette différence peut théoriquement varier entre 0 et 2.

	$\overline{rpbis^i}$	$rpbis^c$	$rpbis^c$ - $\overline{rpbis^i}$
rpbis classique	-0,22	0,76	**0,98**
rpbis SC	-0,78	0,76	**1,54**
rpbis SCT80	-0,97	0,94	**1,91**

La différence obtenue pour le *rpbis SC* avec une turbo analyse dont le seuil de réalisme est fixé à 80 (1,91) est pratiquement le double de celle obtenue pour le *rpbis* classique (0,98). La différence « $rpbis^c$ - $\overline{rpbis^i}$ » du *rpbis SC* (1,54) se situe entre les valeurs obtenues pour les deux autres types de rpbis.

Nous utiliserons plus loin cette méthode de calcul des différences « $rpbis^c$ - $\overline{rpbis^i}$ » pour exprimer de façon précise les Niveaux de Cohérence Spectrale des questions (*NCSq*) à l'aide des *rpbis spectraux* (voir p. 231) ainsi que les Niveaux de Cohérence Interne des questions (*NCIq*) à l'aide des *rpbis classiques* (voir p. 233).

3. En synthèse, ce que mesurent les *rpbis spectraux*, leurs valeurs attendues, les enjeux

Rappelons notre intuition de départ pour la construction des indices *rpbis spectraux.* Logiquement les étudiants qui répondent correctement à une question devraient fournir des pourcentages de certitude plus élevés que les étudiants qui répondent incorrectement. Ainsi, pour une QCM qui fonctionne normalement du point de vue de l'utilisation des pourcentages de certitude, nous devrions observer une tendance à fournir, en moyenne, des certitudes plus élevées chez les sujets qui choisissent la réponse correcte que chez ceux qui choisissent un distracteur. Dans ce cas de figure nous dirons qu'il y a « cohérence spectrale ».

Dès lors que cette situation ne se présente pas, lorsque les certitudes des sujets qui ont choisi la réponse correcte sont, en moyenne, moins élevées que les certitudes fournies par les sujets qui ont choisi une proposition incorrecte, nous nous trouvons face à un problème « d'incohérence spectrale » dans l'utilisation des pourcentages de certitude.

En corrélant les choix ou les rejets (1 ou 0) de la réponse correcte d'une QCM avec les pourcentages de certitude qui les ont accompagnés, nous mesurons la tendance à utiliser des certitudes plus élevées chez les étudiants qui ont répondu correctement en comparaison avec les certitudes de ceux qui répondent incorrectement. C'est le principe de base des *rpbis spectraux* qui nous permettent de chiffrer la « cohérence spectrale » (dans ce cas ci pour une proposition correcte).

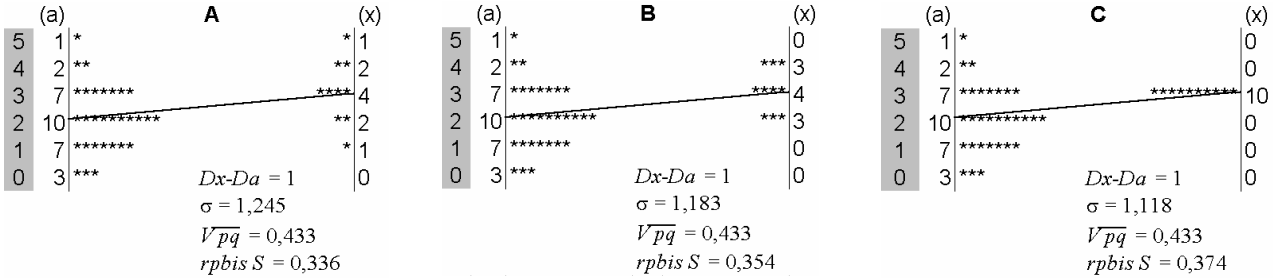
Lorsqu'il s'agit de mesurer la cohérence spectrale d'un distracteur la procédure est différente. Nous faisons alors intervenir dans le calcul de la corrélation, les données des étudiants qui ont choisi cette proposition incorrecte en contraste avec les seules données des sujets qui ont choisi la proposition correcte. Ceci évite d'introduire dans la mesure de la cohérence spectrale du distracteur envisagé, le « bruit » qu'engendreraient les données des autres propositions incorrectes. C'est le principe du « traitement Contrasté », d'où l'appellation *rpbis Spectral Contrasté* (*rpbis SC*).

Enfin, nous avons vu qu'il est possible de calculer des *rpbis SC* plus valides en nous basant sur les données des étudiants qui commettent moins d'erreurs dans leurs auto-estimations. C'est la raison pour laquelle nous proposons d'opérer une sélection dans les données utilisées pour le calcul des *rpbis SC* sur la base du critère du niveau de Réalisme atteint par les sujets (*Rs*). C'est le principe de la « Turbo analyse », d'où l'appellation *rpbis Spectral Contrasté Turbo* (*rpbis SCT*). Le terme « Turbo » fait référence à la montée en puissance de l'instrument en terme de qualité d'information fournie au fur et à mesure que l'on prend en compte les données des étudiants qui commettent de moins en moins d'erreurs dans leurs auto-estimations.

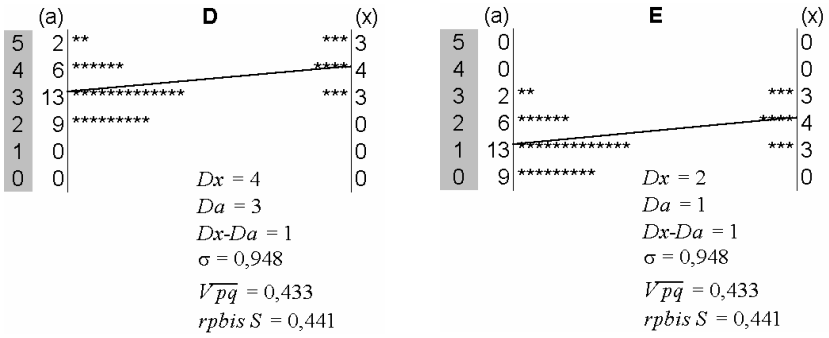
Comme il s'agit de corrélations, les *rpbis SC* et *rpbis SCT* varient dans une plage de valeurs comprises entre –1 et 1. La « cohérence spectrale » implique pour une proposition correcte une valeur positive et significativement différente de zéro. Pour les distracteurs les valeurs doivent idéalement être négatives et significativement différentes de zéro.

Précédemment, nous avons rappelé que le *rpbis classique* peut être considéré comme un indice de « fracture » (voir p. 158). La seule différence entre la moyenne des nombres de réponses correctes récoltées par ceux qui ont choisi la proposition (*Mx*) et la moyenne des autres (*Ma*) nous informe déjà sur l'ampleur de cette fracture. Par rapport à cette simple différence de moyennes, le *rpbis classique* présente cependant l'avantage d'une prise en compte de la dispersion des données (voir p. 157). En ce qui concerne les *rpbis spectraux* nous pouvons les assimiler à des indices de « fracture subjective » entre le groupe des sujets qui ont choisi une proposition et le groupe des autres (tous les autres sujets lorsqu'on calcule le rpbis de la réponse correcte et seulement le groupe de ceux qui ont répondu correctement lorsqu'on calcule le rpbis spectral contrasté d'un distracteur). La différence entre le pourcentage moyen de certitude dans chacun des groupes nous informe donc aussi sur l'ampleur de cette « fracture subjective ». Cependant, comme dans le cas du *rpbis classique*, cette simple différence de moyennes ne permet pas de prendre en compte la dispersion des données, ce que font par contre les *rpbis spectraux*.

Voici trois situations (A, B et C) qui illustrent l'intérêt d'une prise en compte de la dispersion des données dans le cadre des *rpbis spectraux*. Sur chacun des graphiques, l'axe « (x) » à droite reprend les fréquences des degrés de certitude (voir échelle des degrés de certitude, p. 178) de 10 sujets sur 40 qui ont choisi une proposition « x » à une QCM ($p = 0{,}25$). L'axe « (a) » à gauche sur chacun des graphiques reprend les fréquences de 30 autres sujets ($q = 0{,}75$) qui n'ont pas choisi la proposition. Sur chacun des trois graphques, la moyenne des degrés de certitude des sujets qui ont choisi la proposition x ne varie pas ($Dx = 3$), la moyenne des degrés de certitude des autres sujets ne varie pas non plus ($Da = 2$). Dans ces trois situations la différence des moyennes des certitudes vaut donc 1 ($Dx - Da = 1$). Cependant, les trois graphiques présentent des situations où les fractures subjectives, bien que même ampleur, ne présentent pas la même « netteté ». On constate que pour la situation A les certitudes sont assez dispersées autour de la moyenne *Dx*, pour la situation B les choix des degrés de certitudes sont moins dispersés autour de *Dx* et pour C les certitudes des sujets qui ont choisi la proposition x sont toutes concentrées sur le même degré 3 (les 10 étudiants qui ont choisi la proposition x ont tous accompagné leur choix du degré de certitude 3). Les rpbis spectraux traduisent ces différences par des valeurs à chaque fois un peu plus élevées : *rpbis spectral* A = 0,336, *rpbis spectral* B = 0,354 et *rpbis spectral* C = 0,374 alors que la simple différence des moyennes reste constante (*Dx* - *Da* = 1).



Notons qu'il existe des « fractures subjectives hautes » et des « fractures subjectives basses ». Les deux exemples ci-dessous montrent de tels cas de figure. Dans les deux exemples D et E, la différence *Dx* - *Da* vaut 1 et le *rpbis spectral* vaut 0,441 ($p = 0{,}75$).



Dans la situation D, le degré de certitude moyen des sujets qui ont choisi la solution x (*Dx*) vaut 4 (ce qui correspond à un pourcentage de certitude 80%). Pour le groupe des sujets qui n'ont pas choisi x, le degré de certitude moyen (*Da*) vaut 3 (ce qui correspond à un pourcentage moyen de 60%). Dans le cas de E, *Dx* vaut 2 (pourcentage de certitude moyen 40%) et *Da* vaut 1 (pourcentage de certitude moyen 20%). La certitude moyenne des sujets dans D est plus élevée (pourcentage de certitude moyen total = 65,5%) que dans E (pourcentage de certitude moyen total = 25,5%). L'exemple D montre donc une situation où les sujets sont globalement plus sûrs de leur réponses que ceux de la situation E, mais du point de vue de la discrimination spectrale, l'ampleur et la netteté de la fracture subjective entre les deux groupes (x) et (a) est la même : *rpbis spectral* de D et *rpbis spectral* de E = 0,441.

Enfin signalons que, parallèlement aux trois conditions requises pour qu'un *rpbis classique* soit égal à 1 (voir p. 158), les trois conditions suivantes doivent être remplies pour que le *rpbis spectral* atteigne sa valeur maximum 1 :

[1] les sujets qui ont choisi la solution accompagnent tous leur réponse du même pourcentage de certitude (la dispersion des données autour de ce pourcentage de certitude est nulle) ;

[2] les sujets qui n'ont pas choisi cette solution accompagnent aussi tous leur réponse du même pourcentage de certitude (la dispersion des données autour de ce pourcentage de certitude est aussi nulle) ;

[3] le pourcentage de certitude des sujets qui ont choisi la solution est supérieur au pourcentage de certitude des autres sujets (tous les autres sujets lorsqu'on calcule le rpbis de la réponse correcte et seulement le groupe de ceux qui ont répondu correctement lorsqu'on calcule le rpbis spectral contrasté d'un distracteur).

Le *rpbis spectral* est égal à –1 lorsque les deux premières conditions sont remplies et que, en ce qui concerne la troisième, le pourcentage de certitude des sujets qui ont choisi la solution est inférieur au pourcentage de certitude des autres sujets.

Notons que ces cas de figure où les *rpbis spectraux* atteignent leur valeur maximum supérieure « 1 » ou inférieure « –1 » sont exceptionnels. Dans le cadre des épreuves MOHICAN nous n'avons pas rencontré ce type de configuration des données où les trois conditions sont remplies.

Quels sont les enjeux liés aux *rpbis spectraux* ? Comme pour la production d'aliments (animaux ou végétaux), l'évaluation des acquis des étudiants peut engendrer ses propres nuissances. En particulier, parmi les questions des tests, certaines peuvent produire des mesures non valides, des résultats inadéquats. D'où la nécessité d'analyses quant à la qualité de ces questions, préalablement à leur ingestion par le système de distribution des notes aux consommateurs (professeurs et étudiants). Ces analyses doivent porter sur de grandes quantités de questions, et ce rapidement (tout spécialement lors des sessions de juin et de septembre, les délibérations constituant des « deadlines » impératives). Leur fiabilité est cruciale. Les nouveaux indices *rpbis spectraux* peuvent-ils augmenter cette fiabilité dans le cadre des épreuves standardisées ? Oui s'ils diminuent les « non-détections » de propositions problématiques au sein des QCM et les « fausses alertes » (la mise en évidence d'une valeur anormale récoltée par une proposition alors que les experts du contenu n'y décèlent pas d'anomalie). Nous verrons dans la troisième partie de cette thèse ce que donne cette approche dans le contexte des tests MOHICAN et comparerons les performances des *rpbis classiques* avec celles des nouveaux *rpbis spectraux*.

Chapitre V :

Informatisation des procédures de calcul : SCANTEST 2.0, un logiciel pour l'analyse de la qualité spectrale des épreuves MOHICAN

Sommaire

Cette recherche sur les rpbis Spectraux Contrastés avec Turbo analyse (*rpbis SCT*) nécessite des outils informatiques qui n'existent actuellement pas sur le marché. En vue de faciliter les traitements nous avons programmé une série de procédures en *Visual basic pour Microsoft Excel*. Ces procédures nous permettent, à partir des données MOHICAN, d'extraire les informations nécessaires au calcul des rpbis et de calculer les rpbis SC avec turbo analyse. Nous avons intitulé l'ensemble des programmes et l'interface utilisateur : « *SCANTEST 2.0* » en référence au logiciel de calcul des rpbis Spectraux « *SCANTEST 1.0* » que nous avions conçu et programmé dans le cadre d'une recherche précédente (Gilles, 1998a).

A. Les étapes du traitement dans SCANTEST 2.0

Dans la version 1.0 de *SCANTEST* seuls les *rpbis SC* liés à la proposition correcte des QCM étaient calculés. La nouvelle version *SCANTEST 2.0* permet non seulement le calcul des *rpbis SC* pour toutes les propositions des QCM des tests MOHICAN mais aussi d'effectuer des turbo analyses sur ces *rpbis SC*, donc de calculer les *rpbis SCT*.

1. Etape 1 : filtrage des données à traiter

Les fichiers de données MOHICAN sont disponibles sur le web[62]. Nous les avons téléchargés et convertis en feuilles de données *Microsoft Excel* : une feuille par test. Chaque feuille de données contient les réponses, les certitudes ainsi que les codes des étudiants, des facultés et des sections. Ces feuilles ont ensuite été incorporées au programme *SCANTEST 2.0*. La 1ère étape d'un traitement consiste à filtrer les informations liées à un test en fonction :

1. d'une (ou de toutes les) faculté(s) où les données ont été récoltées ;
2. d'une (ou de toutes les) section(s) au sein de cette (ces) faculté(s).

2. Etape 2 : traitement des données et calcul des rpbis classiques et des rpbis Spectraux Contrastés

La seconde étape aboutit au calcul des rpbis classiques et spectraux. Les données utilisées lors des traitements sont celles qui ont été préalablement filtrées à l'étape précédente. Le programme génère une feuille de données *Microsoft Excel* par question contenant une matrice de données suivie des *rpbis* classiques, *rpbis SC* et d'une série d'informations utiles relatives à chaque proposition (certitudes moyennes, pourcentage de choix, nombre d'étudiants qui ont choisi la proposition, … voir *infra* pp. 213 et 214). Une nouvelle feuille *Microsoft Excel* contenant les scores de réalisme des sujets est aussi créée lors de cette étape, elle permettra ultérieurement de procéder aux turbo analyses.

3. Etape 3 : calcul des rpbis Spectraux Contrastés avec turbo analyse

Le programme autorise le calcul des rpbis Spectraux Contrastés avec Turbo analyse (*rpbis SCT*). Dans une section précédente (voir *supra* p. 184) nous avons vu qu'il est possible de calculer les *rpbis SCT* à différents seuils de Réalisme (*Rs*) en sélectionnant les données des étudiants en fonction de leurs scores *Rs* (exprimés en pourcents). Dans le cadre d'une turbo analyse, il est possible avec *SCANTEST 2.0* de définir le seuil inférieur ainsi que le seuil supérieur de réalisme qui sera exigé pour inclure les données dans les traitements. L'utilisateur peut alors déterminer à partir de quel niveau minimum et jusqu'à quel niveau maximum de réalisme il faudra prendre en compte les données des sujets pour calculer les *rpbis SCT*, et ce, avec un pas qui lui aussi peut être fixé par l'utilisateur.

De manière générale, le programme sélectionnera les données des étudiants dont le seuil de réalisme est compris entre le niveau inférieur et le niveau supérieur. De manière plus précise, une première analyse sera faite en tenant compte de tous les étudiants qui ont un niveau de réalisme compris entre ce niveau inférieur et le niveau supérieur. Ensuite une seconde analyse sera faite en tenant compte des

[62] Adresse Internet : « http://139.165.55.56/mohicand/ ».

étudiants dont le réalisme est compris entre un nouveau niveau inférieur (égal au niveau antérieur augmenté du pas) et le niveau supérieur. Les analyses ultérieures prennent en compte les étudiants qui ont un réalisme compris entre le niveau inférieur initial augmenté d'un nombre entier de fois le pas et le niveau supérieur, et ce, jusqu'à ce que le niveau inférieur ait atteint le niveau supérieur.

Par défaut, les trois valeurs : (1) le seuil inférieur initial de réalisme, (2) le seuil supérieur de réalisme et (3) le pas sont fixés respectivement à 80, 100 et 10. Cela signifie qu'une turbo analyse peut alors avoir lieu au départ des données des étudiants dont le réalisme est supérieur ou égal à 80 et qu'elle génèrera deux rapports étant donné le pas de 10 : un premier contenant les *rpbis SCT* pour un seuil de réalisme 80, un second contenant les *rpbis SCT* pour un seuil de réalisme 90 (80 + pas de 10). Nous dirons dans le cadre de cet exemple que deux paliers de Turbo analyse ont été franchi, l'un à T80 et l'autre à T90.

4. Etape 4 : Traitements liés aux autres indices d'analyse de la qualité spectrale et classique des épreuves et mise en forme des informations

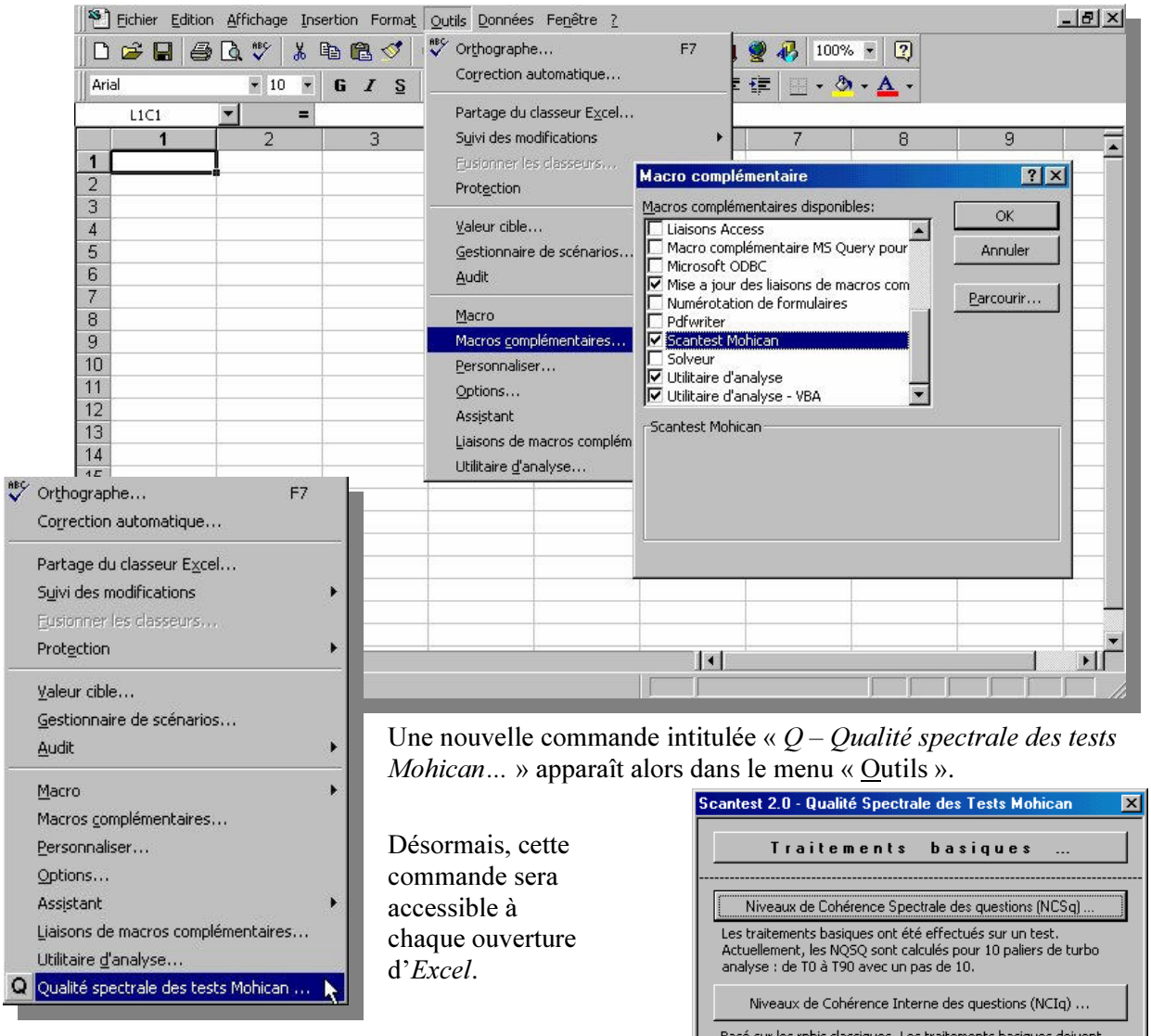
D'autres indices qui seront exposés plus loin sont également calculés à l'aide de *SCANTEST 2.0*. Il s'agit des :

- Niveaux de Cohérence Spectrale des questions (*NCSq*) (voir explications détaillées, p. 231) ;
- Niveaux de Cohérence Interne des questions (*NCIq*) (p. 233) ;
- Indices de Centration par question (*Cq*) (p. 254) ;
- Fréquences et statistiques descriptives des scores de Réalisme des sujets (*Rs*) (pp. 272, 274) et de Centration des sujets (*Cs*) (p. 277) ;
- Profils Spectraux des questions (*PSq*) (p. 235) et Indices de Réalisation des prédictions par question (*Rq*) (p. 242) ;

Enfin, lors de cette quatrième étape, nous utilisons des procédures qui reprennent les données calculées aux étapes précédentes, le but étant de les mettre en forme afin d'en faciliter l'interprétation. C'est notamment lors de cette dernière étape que nous mettons en forme les protocoles d'analyse de la qualité des épreuves MOHICAN en fonction de trois niveaux d'analyse : (1) test, (2) QCM et (3) propositions.

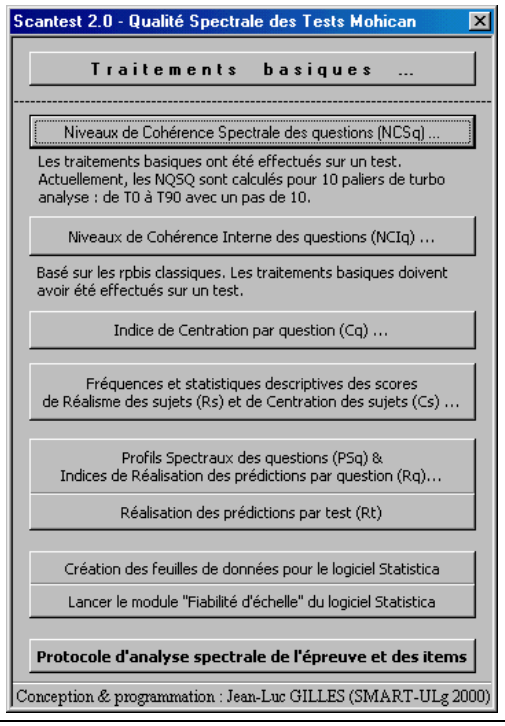
B. Interface utilisateur

Le programme de traitement *SCANTEST 2.0 pour les tests MOHICAN* a été conçu sous la forme d'une macro complémentaire destinée au tableur *Microsoft Excel.* Après avoir placé le fichier « *Scantest Mohican.xla* » dans le sous-répertoire « *Macrolib* » du répertoire « *Microsoft Office* », l'utilisateur active une première fois le programme à l'aide du menu « *Outils* » et de la commande « *Macros complémentaires…* ».



Une nouvelle commande intitulée « *Q – Qualité spectrale des tests Mohican…* » apparaît alors dans le menu « Outils ».

Désormais, cette commande sera accessible à chaque ouverture d'*Excel*.



Un clic sur la commande « *Q – Qualité spectrale des tests Mohican…* » active le menu principal qui permet d'accéder aux 10 modules de *SCANTEST 2.0 pour les tests MOHICAN*.

Nous allons maintenant décrire les modules de traitements [1] à [10] accessibles via l'interface utilisateur.

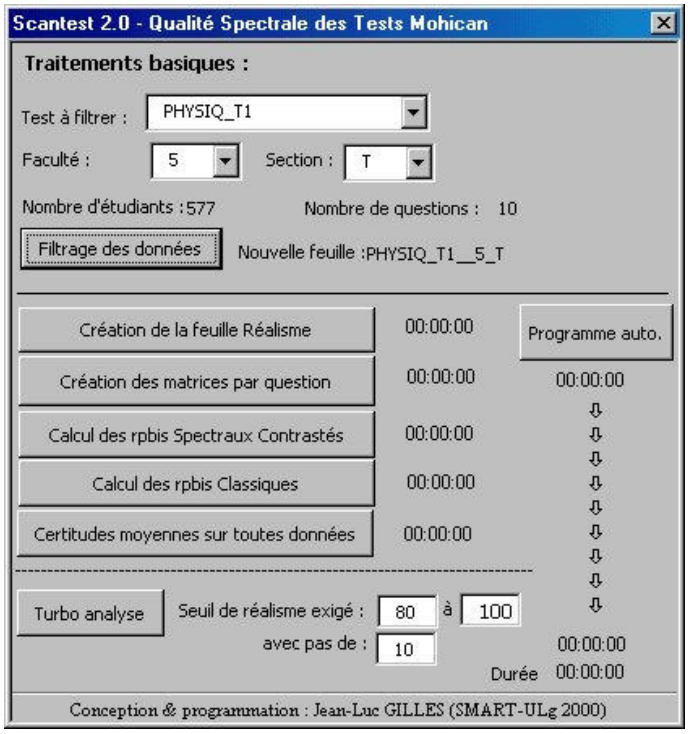
Avant de commencer les traitements, soit on crée un nouveau fichier *Microsoft Excel*, soit on utilise un fichier contenant déjà une série de traitements, mais dans ce deuxième cas l'utilisateur sera attentif à la taille que peut prendre un fichier en fin d'analyse lorsque les ressources des 10 modules ont été utilisées sur un grand nombre de données. A titre d'exemple, le fichier complet généré par *SCANTEST 2.0* après utilisation des modules [1] à [10] pour l'analyse de l'épreuve de vocabulaire qui comprend 45 questions soumises à un total de 3.846 étudiants, occupe 35.842 Ko sur le disque dur d'un PC.

C. Modules de traitements

SCANTEST 2.0 génère des tableaux de données et des graphiques qui seront exploités plus loin dans cette recherche. Ces résultats issus des traitements sont exposés dans les chapitres suivants où ils font l'objet d'une série d'analyses, de comparaisons et de généralisations. Nous n'interprèterons donc pas ici les sorties générées par les modules de traitements et parfois présentées à titre d'exemple. Pour le lecteur qui souhaiterait immédiatement en savoir plus sur les résultats des analyses spectrales et/ou classiques envisagées dans les modules, nous mentionnons systématiquement les références des sections où ils sont exposés en détail.

1. Module « Traitements basiques » [1]

L'appellation « Traitements basiques » est à prendre ici dans le sens de « *traitements qui sont à la base de ceux des modules [2] à [10] de SCANTEST* ». En effet, nous verrons plus loin que les traitements basiques produisent des résultats (voir *infra* p. 213) qui serviront de point de départ à d'autres traitements qui permettront d'envisager des synthèses et des diagnostics pour différents niveaux d'analyse de la qualité spectrale des épreuves MOHICAN (voir p. 207). Ces traitements en aval des traitements basiques sont réalisés à l'aide des modules [2] à [10].



L'interface ci-contre, montre que ce sont les données du test de physique intitulé « PHYSIQ_T1 » (T1 parce que les données sont de type 1, c'est-à-dire réponses accompagnées des degrés de certitude) qui seront filtrées en fonction :

(1) du code de la faculté, « 5 » correspondant aux facultés des Sciences, on ne sait pas de quelles universités, par principe déontologique, les responsables du projet MOHICAN n'ont pas voulu permettre de comparaisons inter-universitaires ;

(2) du code de la section, ici « T » signifie que toutes les sections sont sélectionnées.

Ensuite, un clic sur le bouton « Filtrage des données » provoque la création d'une nouvelle feuille contenant les données filtrées dans le classeur *Microsoft Excel*. Le panneau indique que les données de 577 étudiants ont été copiées dans une nouvelle feuille intitulée « PHYSIQ_T1__5_T » et que le test contient 10 questions.

Ensuite, une série de boutons permettent de lancer les procédures de traitement :

- « Création de la feuille de réalisme » qui calcule le réalisme (voir procédure de calcul en *infra* p. 184) de chacun des sujets de la sélection ;
- « Création des matrices par question » qui reprend les données relatives à chaque question (réponses et certitudes fournies) sur une feuille séparée (une matrice par question) ;
- « Calcul des rpbis Spectraux » qui effectue les traitements nécessaires en vue du calcul du rpbis Spectral Contrasté pour chaque proposition d'une question (*Dx*, *Da*, *Dx-Da*, (*Dx-Da*)/écart type, *p*, *q*, *pq*, racine de *pq*, et *rpbis SC*) ;
- « Calcul des rpbis classiques », traitements nécessaires en vue du calcul du rpbis pour chaque proposition d'une question (*Mx*, *Ma*, *Mx-Ma*, *Mx-Ma*/écart type et *rpbis*, les *p*, *q*, *pq* et racine de *pq* ayant déjà été calculés dans le cadre du *rpbis SC*) ;
- « Certitudes moyennes sur toutes données » ;
- « Calcul des rpbis Spectraux Turbo » avec les valeurs par défaut : seuil de réalisme minimum exigé à 80, seuil maximum à 100 et pas fixé à 10.

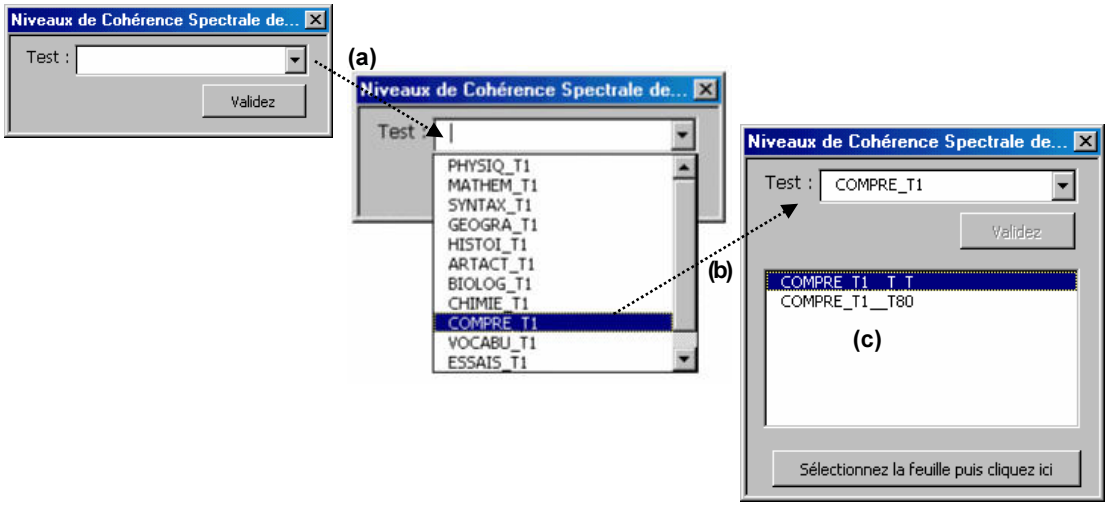
A côté des boutons liés à ces différentes procédures de calcul figurent des étiquettes qui affichent le temps écoulé après la réalisation du traitement (une turbo analyse « *80 à 100 avec pas de 10* » pratiquée à partir des données d'un test MOHICAN avec 2.355 sujets prend à elle seule plus d'une heure de traitement sur un PC de type pentium II cadencé à 366 Mhz).

Enfin, le bouton « Programme auto. » permet d'enchaîner automatiquement les différentes procédures les unes après les autres, y compris la turbo analyse. Pour une analyse détaillée des résultats des traitements basiques nous renvoyons le lecteur à la section « Résultats des traitements basiques effectués à l'aide de *SCANTEST 2.0* », page 213).

2. Module « Niveaux de Cohérence Spectrale des questions (*NCSq*) » [2]

Ce module de traitement calcule à partir des résultats des traitements basiques, les valeurs des indices de Niveaux de Cohérence Spectrale des questions (*NCSq*) qui seront détaillés plus loin (voir p. 231).
Après avoir cliqué sur le bouton « Niveaux de Cohérence Spectrale des questions (NCSq) », l'utilisateur obtient une nouvelle boîte de dialogue.

En cliquant sur la flèche du menu déroulant « Test : » **(a)**, l'utilisateur déploie ce menu et fait apparaître la liste des noms de code des 10 épreuves MOHICAN. Après avoir sélectionné une épreuve **(b)**, ici « COMPRE_T1 », la boîte de dialogue s'allonge et une nouvelle zone **(c)** apparaît. Cette zone contient la liste des feuilles de données issues des filtrages qui ont été effectués à l'aide du module « Traitements basiques » (voir p. 200). C'est sur les données filtrées sélectionnées (ici « COMPRE_T1__T_T ») que seront effectués les traitements *NCSq* de ce module. Cette procédure de sélection d'une feuille de données est aussi utilisée dans les autres modules.



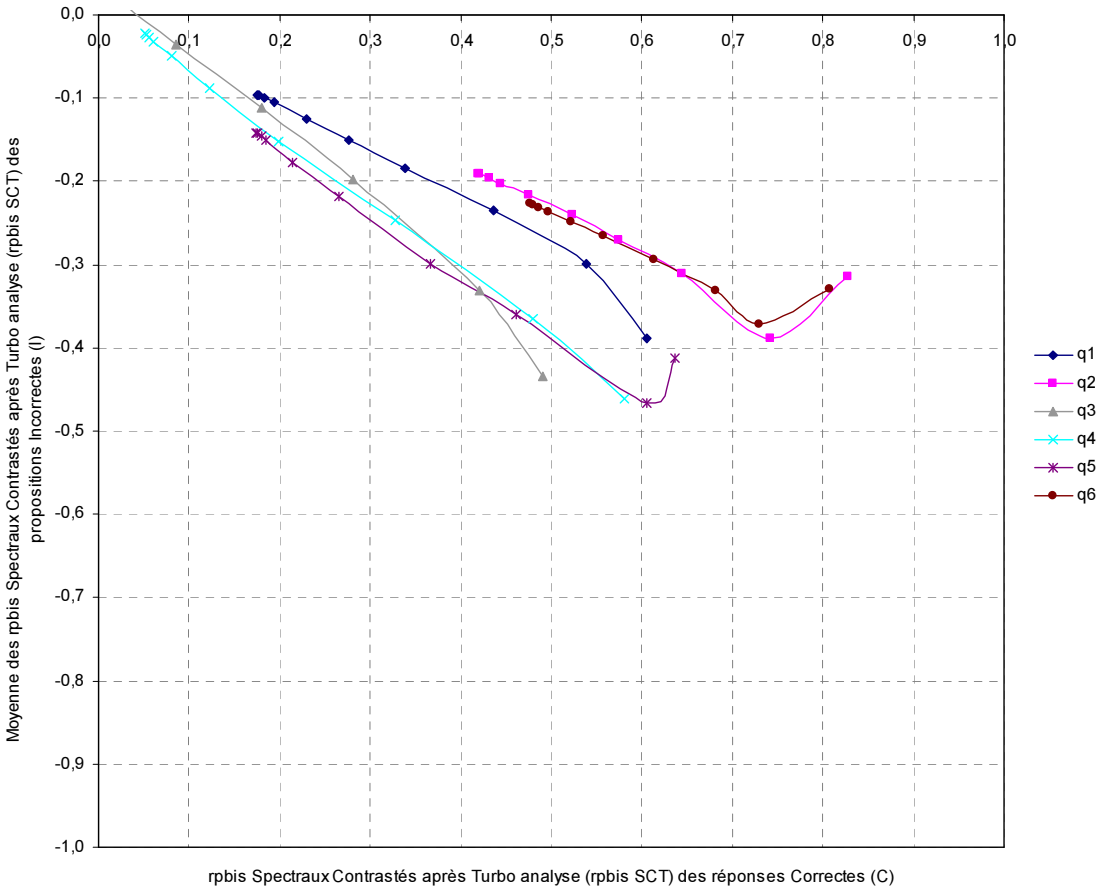
Lorsque l'utilisateur clique sur le bouton « *Sélectionnez la feuille puis cliquez ici* », une nouvelle feuille de données *Microsoft Excel* est créée et placée dans le fichier qui renferme déjà les résultats des traitements basiques. Le nom de cette nouvelle feuille débute par l'intitulé de la feuille des données filtrées à partir desquelles les calculs sont effectués (dans notre exemple « COMPRE_T1__T_T ») et se termine par « _NCSQ » (cela donne une nouvelle feuille « COMPRE_T1__T_T_NCSQ »).

Sans entrer ici dans une explication détaillée de l'indice *NCSq*, disons que le Niveau de Cohérence Spectrale de chaque question peut être représenté par un point dont la projection orthogonale est déterminée en référence à un plan horizontal par la valeur du *rpbis SC* de la réponse correcte et à un second plan perpendiculaire au premier par la moyenne pondérée des *rpbis SC* des propositions incorrectes (la procédure de calcul de cette moyenne pondérée est exposée en *infra* p. 189).

Ce module fonctionne après un traitement basique préalablement effectué avec une turbo analyse comportant 10 paliers de T0 à T90 et dont le pas est fixé à 10. Ceci permet de calculer les *NCSq* à partir des données des étudiants de plus en plus réalistes, au fur et à mesure de la progression du traitement dans les 10 paliers. On obtient alors pour chaque question 10 points correspondant aux valeurs des *NCSq* à chaque palier de turbo analyse. Ces points peuvent être reliés par des traits formant un « *Brin Spectral par question (BSq)* ». Nous appelons l'ensemble de ces brins une « *Gerbe Spectrale des questions (GSq)* ».

A titre d'exemple, voici la représentation graphique des *NCSq* du test de compréhension. L'interprétation et l'analyse des informations liées aux *NCSq* et à ce type de graphique sont expliquées au chapitre suivant (p. 228).

Zoom sur la zone de qualité du graphe de la Gerbe Spectrale des questions (GSq) après Turbo analyse (de T0 à T90 avec pas de 10)



La nouvelle feuille de donnée générée dans le cadre de ce module contient les valeurs des indices *NCSq* aux 10 paliers turbo ainsi que les synthèses graphiques. Ces indices et synthèses graphiques sont par

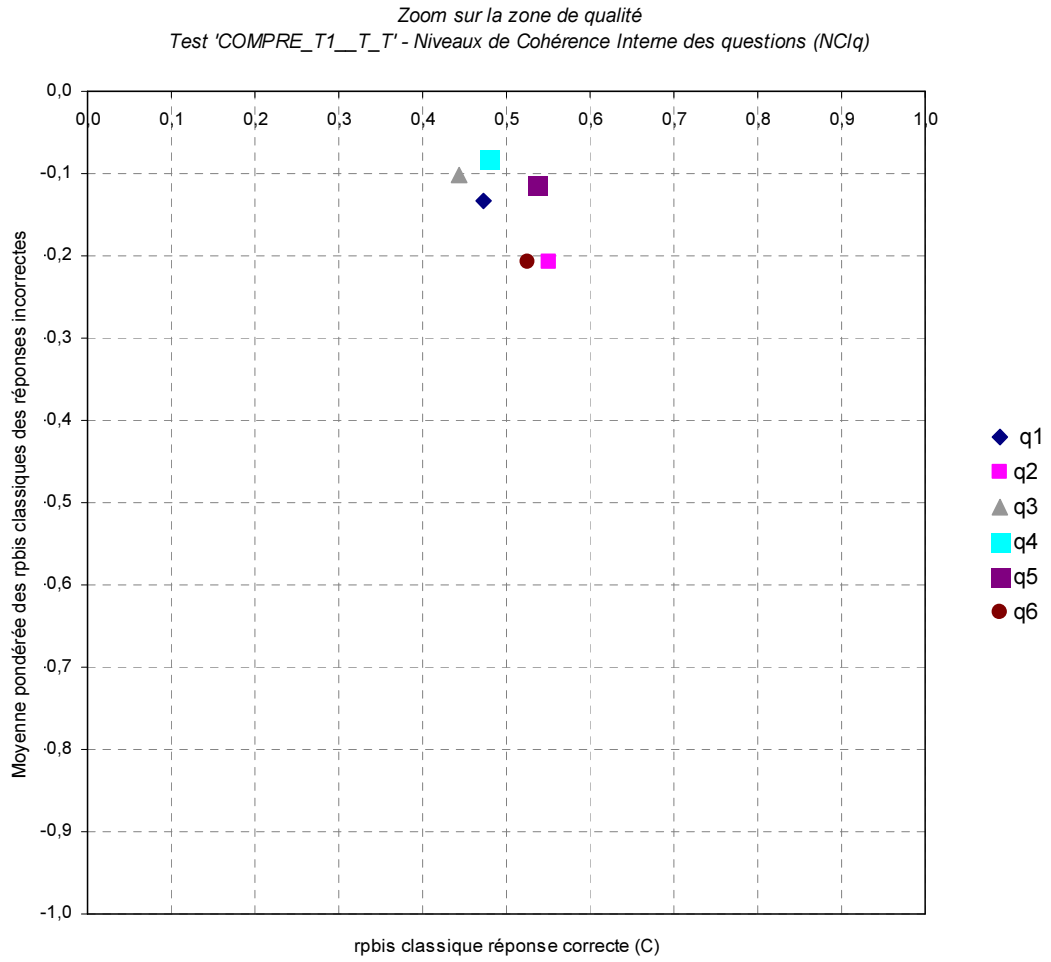
la suite réutilisées dans le cadre du module [10] pour l'établissement des protocoles d'analyse de la qualité spectrale des épreuves.

A partir des *NCSq*, *SCANTEST 2.0* calcule les Niveaux de Cohérence Spectrale du test (*NCSt*) aux 10 paliers de turbo analyse et en livre aussi une représentation graphique (voir détails p. 261).

3. Module « Niveaux de Cohérence Interne des questions (*NCIq*) » [3]

Dans le cadre de cette recherche nous souhaitons comparer les résultats des analyses spectrales avec les résultats d'analyses classiques. Nous avons vu précédemment que le *rpbis* classique est un indice corrélationnel de discrimination des sujets où la variable critère utilisée est le nombre de réponses correctes fournies à l'ensemble des questions du test (voir en *infra* pp. 150, 156 et 171). A l'aide de ce module nous appliquerons au *rpbis classique* le même type de traitement que celui que nous avons utilisé pour calculer les *NCSq*, mais cette fois en calculant la moyenne pondérée pour les propositions incorrectes (procédure de calcul p. 189) en ayant recours aux *rpbis classiques*. En soustrayant ensuite cette moyenne pondérée au *rpbis classique* de la réponse correcte, nous obtenons un Niveau de Cohérence Interne des questions (*NCIq*).

A titre d'exemple, voici la représentation graphique des *NCIq* du test de compréhension. Plus un point représentant une QCM se rapproche du coin inférieur droit du graphique, plus cette question participe à la cohérence interne de l'épreuve. Ce type de graphique est repris plus loin lors de la réalisation des protocoles d'analyse des épreuves MOHICAN.



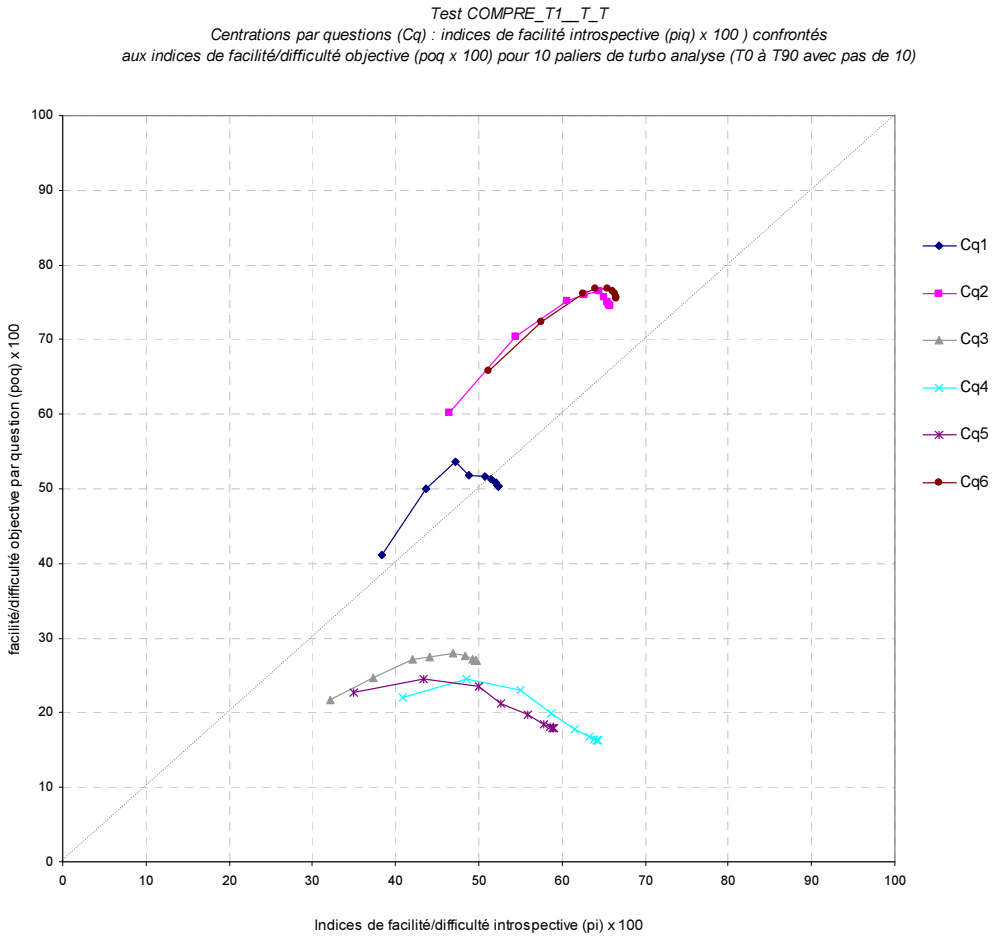
La procédure de sélection des données préalablement filtrées avec le module « Traitements basiques » et qui serviront à réaliser les traitements *NCIq* est la même que celle utilisée pour les traitements

NCSq (même boîte de dialogue mais avec un titre différent : « ...*NCIq* »). L'utilisation de ce module de traitement *NCIq* entraîne la création d'une nouvelle feuille de données *Microsoft Excel* placée dans le fichier qui renferme les résultats des traitements basiques. Le nom de cette nouvelle feuille emprunte celui de la feuille des données filtrées à partir desquelles les traitements sont effectués et se termine par « _NCIQ ». Cette nouvelle feuille contient les données de base, les valeurs des indices *NCIq* ainsi que les représentations graphiques.

4. Module « Indice de Centration par question (*Cq*) » [4]

L'indice de Centration par question (*Cq*) est un indice spectral qui combine la difficulté/facilité objective d'une question et sa difficulté/facilité introspective. *SCANTEST 2.0* fournit la valeur de l'indice *Cq* à chacun des 10 paliers de turbo analyse. Pour un palier de turbo analyse, l'indice *Cq* se calcule en soustrayant l'indice de difficulté/facilité objective de la question (*poq*) à l'indice de difficulté/facilité introspective obtenu à cette même question (*piq*). Les explications relatives à l'indice *Cq* et à sa procédure de calcul figurent page 254.

A titre d'exemple voici la représentation graphique des valeurs de l'indice *Cq* aux 10 paliers de turbo analyse pour les 6 questions du test de compréhension.



En calculant la moyenne des *Cq* pour une épreuve, on obtient la valeur de l'indice de Centration par test (*Ct*) (voir p. 270).

La procédure de sélection des données filtrées avec « Traitements basiques » est la même que celle utilisée pour les traitements *NCSq* (même boîte de dialogue mais avec un titre différent : « ...*Cq* »). Comme pour les deux modules précédents, une nouvelle feuille de données *Microsoft Excel* est générée. Le nom de

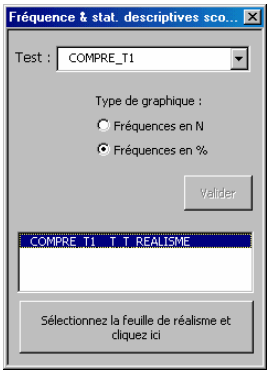
cette nouvelle feuille emprunte aussi celui de la feuille des données filtrées à partir desquelles les traitements sont effectués et se termine par « _CQ ». Elle contient les données, les valeurs des indices *Cq* et *Ct* ainsi que les représentations graphiques .

5. Module « Fréquences et statistiques des scores de Réalisme (*Rs*) et de Centration (*Cs*) des sujets » [5]

Nous avons vu dans le cadre du module des « Traitements basiques » qu'une option de l'interface permet la création d'une feuille de données qui contient les scores en Réalisme (*Rs*) des sujets. Ces scores *Rs* sont calculés selon la formule exposée page 184. A l'aide des données de cette feuille, nous calculerons les fréquences des *Rs* ainsi qu'une série de statistiques : moyenne, écart type, minimum, maximum, N, Kurtosis, asymétrie, médiane et mode. Un graphique des fréquences est également créé.

Nous avons aussi calculé les Centrations (*Cs*) des sujets en soustrayant à leurs certitudes moyennes les taux d'exactitude de leurs réponses. L'indice *Cs* et sa procédure de calcul sont détaillés plus loin, page 277. Comme pour l'indice *Rs*, la procédure de traitement inclus le calcul des fréquences, moyenne, écart type, minimum, maximum, N, Kurtosis, asymétrie, médiane et mode. Un graphique des fréquences est également créé.

La procédure de sélection des données filtrées (voir « Traitements basiques », p. 200) est similaire à celle utilisée pour les *NCSq* (p. 201) à ceci près que la boîte de dialogue offre une option supplémentaire « Type de graphique » qui permet de créer soit un graphique des fréquences en nombres de sujets, soit un graphique des fréquences en pourcentages.

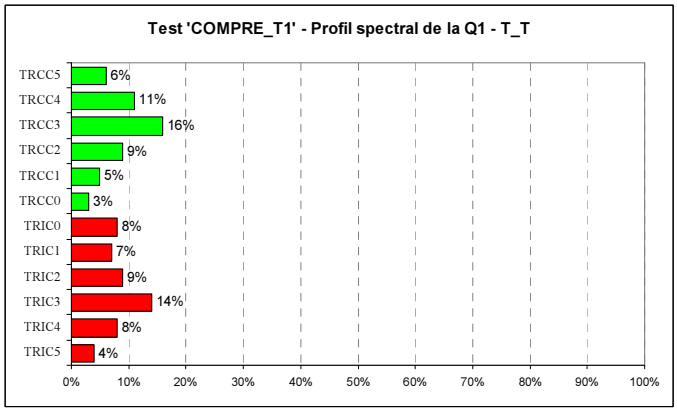


En ce qui concerne le stockage des informations, tous les résultats des traitements sont placées dans la feuille des scores de réalisme créée lors des traitements basiques.

Ces informations seront utilisées par la suite dans la réalisation des protocoles d'analyse spectrale des tests.

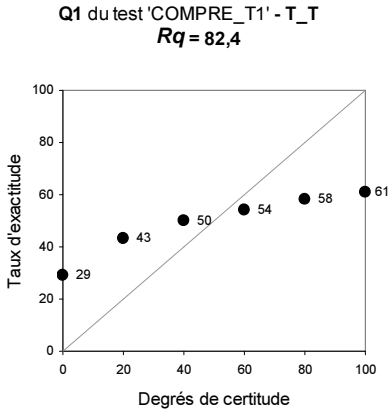
6. Module « Profils Spectraux des questions (*PSq*) & Indices de Réalisation des prédictions par question (*Rq*) » [6]

L'utilisation des degrés de certitude lors des épreuves MOHICAN offre la possibilité de calculer par question les Taux de Réponses Correctes (TRC) ou Incorrectes (TRI) pour chaque pourcentage de certitude (C0 = 0%, C1 = 20%, C2 = 40%, C3 = 60%, C4 = 80% et C5 = 100%). C'est ce que fait ce module de traitement pour aboutir aux Profils Spectraux des questions (*PSq*). Un exposé détaillé des *PSq* est proposé à la page 235.



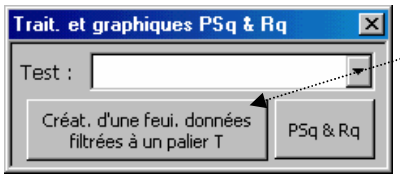
A l'aide des TRC et TRI de chaque certitude, *SCANTEST* élabore ensuite des graphiques tel que celui-ci : le *PSq* de la première question du test de compréhension (la somme des TRC et TRI vaut 100%).

Il est possible d'évaluer comment les taux d'exactitude des réponses fournies à une question par l'ensemble du groupe s'ajustent aux pourcentages de certitudes qui accompagnent ces réponses. Ce type de traitement aboutit à l'indice de Réalisation des prédictions par question (*Rq*). La réalisation des prédictions pour une question peut être visualisée à l'aide d'un graphique tel que ci-contre.

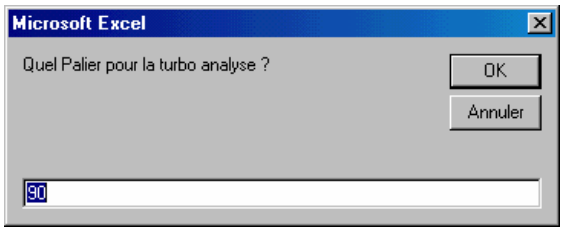


Il s'agit de l'indice *Rq* et du graphique liés à la 1ère QCM du test de compréhension. Sur ce graphique généré par *SCANTEST 2.0*, on constate que 29% des étudiants qui ont utilisé le pourcentage de certitude « 0 » ont répondu correctement. Ces étudiants se sont donc sous-estimés avec le degré de certitude « 0 ». Idéalement, aucun (0%) étudiant qui choisit le pourcentage de certitude « 0 » ne devrait répondre correctement, la concordance est alors parfaite entre les choix du degré de certitude « 0 » (les prédictions) et le pourcentage de réponses correctes (0%) obtenu dans les faits (les taux d'exactitude). L'indice et le graphique *Rq* sont détaillés plus loin (p. 242).

La procédure de sélection des données filtrées utilisées pour les traitements est similaire à celle qui est utilisée pour les *NCSq* (p. 201), à ceci près que nous y avons ajouté une option « Création d'une feuille de données filtrées à un palier T ».



Cette option permet de générer à partir des données filtrées une nouvelle feuille *Microsoft Excel* qui ne contient que les résultats des étudiants dont les scores de réalisme sont supérieurs ou égaux à un seuil déterminé par l'utilisateur (90 dans la boîte de dialogue ci-contre).

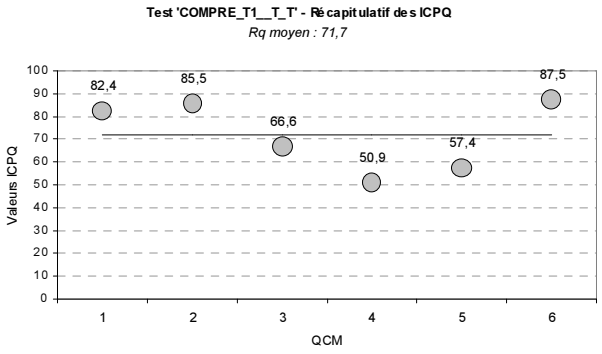


Nous pouvons alors dans un second temps réaliser les traitements *PSq* et *Rq* à partir des données de cette nouvelle feuille (voir pp. 240 et 246) en la sélectionnant selon la méthode décrite précédemment (voir p. 201).

En ce qui concerne le stockage des informations *PSq* et *Rq*, une nouvelle feuille de données *Microsoft Excel* est générée pour les accueillir. Le nom de cette feuille commence par le code du test et se termine par « PROFILS_S_ » suivi des caractères d'identification de la faculté et de la section séparé par « _ ». Dans le cas de notre exemple où les données utilisées pour les traitements sont celles des étudiants appartenant à Toutes « T » les facultés et Toutes « T » les sections cela donne le nom de feuille suivant : « COMPRE_T1_PROFILS_S_T_T ». Ces informations seront utilisées par la suite dans la réalisation les protocoles d'analyse spectrale des tests.

7. Module « Moyenne *Rq* par test » [7]

A l'aide de ce module, nous pouvons créer sur la feuille de données générée pour les *PSq* et *Rq* le graphique ci-contre.



Ce graphique « Récapitulatif des *Rq* » reprend les valeurs *Rq* de chacune des questions représentées par des cercles grisés. La valeur moyenne des *Rq* calculée sur toutes les questions du test (6 questions) est représentée par un trait horizontal (dans le contexte de l'exemple du test de compréhension, cette moyenne vaut 71,7).

8. Module « Création des feuilles de données pour *Statistica 5.1* » [8]

Nous verrons plus loin qu'une série d'indices de fidélité classiques peuvent être calculés à l'aide du logiciel *STATISTICA 5.1* produit par la firme *Statsoft.* Pour permettre l'importation des données des épreuves MOHICAN dans *STATISTICA,* nous avons programmé une procédure qui transforme la feuille des données filtrées (voir Traitements basiques, p. 200) en un fichier qui peut être utilisé dans ce logiciel de traitements statistiques. Ces fichiers sont sauvegardés dans un répertoire « result_statistica » lui-même subdivisé en sous-répertoires aux noms des différents tests MOHICA N (dans l'exemple du test de compréhension cela donne le chemin de sauvegarde «…/result_statistica/compre_t1/ »).

Deux types de fichiers sont produits, d'une part des fichiers « matrice binaire » et d'autre part des fichiers « matrice spectrale ». Dans ces deux types de matrices, les lignes représentent les sujets et les colonnes les questions.

Les « matrices binaires » sont composées de 0 ou de 1. Pour chaque réponse fournie par chacun des étudiants, si la réponse est correcte on indique 1, si elle ne l'est pas, on note 0 (voir p. 132).

Les « matrices spectrales » présentent les résultats des étudiants sous la forme des pourcentages de réussite qui ont accompagné les réponses. Ces pourcentages sont précédés du signe moins lorsqu'ils correspondent à des réponses incorrectes. Lorsqu'il n'y a pas de signe devant le pourcentage, la réponse est correcte.

Ces matrices nous permettent de réaliser des traitements à l'aide du module « *Fiabilité d'échelle* » du logiciel *STATISTICA 5.1*. Ces analyses sont décrites précedemment, page 132.

9. Module « Lancer le module *Fiabilité d'échelle* du logiciel *Statistica* » [9]

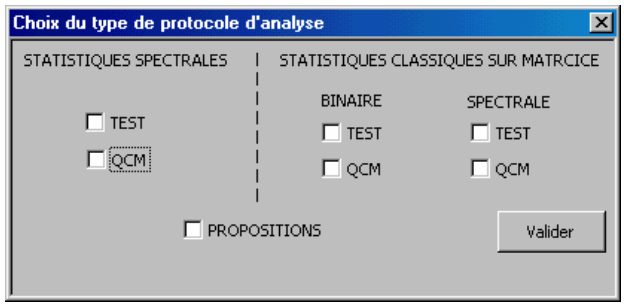
Si le logiciel *STATISTICA 5.1* est installé sur le PC de l'utilisateur dans le répertoire « c:\statistica\ » (répertoire par défaut proposé lors de l'installation du programme), un clic sur ce bouton de l'interface permet de lancer le module « *Fiabilité et analyse d'échelle* » de *STATISTICA 5.1* produit par la firme *Statsoft*. Les fichiers contenant les matrices de données créées à l'aide du module précédent peuvent alors être importées dans ce logiciel de traitements statistiques. L'utilisation des ressources de *STATISTICA 5.1* dans le cadre de cette recherche est décrite page 133.

10. Module « Protocole d'analyse spectrale de l'épreuve et des items » [10]

Ce dernier module vise à produire des synthèses à partir des informations qui ont été générées à l'aide des 9 modules précédents.

Lorsque l'utilisateur active ce 10ème module il reçoit la même procédure de sélection d'une feuille de données filtrées que celle qui est utilisée au 2ème module « *NCSq* » (voir p.201).

Mais ici, après avoir cliqué sur le bouton « *Sélectionnez la feuille puis cliquez ici* », il reçoit la boîte de dialogue ci-contre.



Cette interface permet d'obtenir des synthèses pour deux catégories de statistiques : spectrales et classiques, et, dans ce dernier cas, pour deux types de matrices : binaire d'une part et spectrale d'autre part.

Ces différentes statistiques peuvent être envisagées à trois niveaux d'analyse (voir p. 211) : test, QCM et propositions (les solutions proposées au sein des QCM). En ce qui concerne le niveau d'analyse « propositions », nous avons regroupé les statistiques spectrales et classiques dans la même synthèse.

Avant la création de ces synthèses, deux nouvelles feuilles de données *Microsoft Excel* sont créées pour la catégorie des statistiques classiques. L'une est en rapport avec les traitements sur la matrice binaire, l'autre concerne la matrice spectrale. *SCANTEST 2.0* calcule sur ces feuilles de données une série d'indices de fidélité : le coefficient de bipartition (voir p. 130), le coefficient de bipartition avec correction de Spearman-Brown (p. 131), le coefficient de fiabilité de Guttman (p. 131), le coefficient alpha de Cronbach (p. 137) et le nombre de questions à ajouter pour obtenir un alpha de 0,8 (p. 139). D'autres informations sont aussi produites : distribution et graphique des scores, indices de facilité/difficulté du test, courbes de fréquences des scores des deux moitiés de tests constituées pour le calcul du coefficient de bipartition.

C'est au départ des informations figurant sur ces deux nouvelles feuilles et sur toutes les autres qui ont été créées dans le cadre des traitements réalisés lors des 9 modules précédents que *SCANTEST 2.0* élabore les synthèses.

Nous avons intitulé l'ensemble de ces synthèses : « *Protocole SCANTEST 2.0 pour épreuves MOHICAN* ». Après traitement d'une sélection de données filtrées, elles sont collationnées dans un fichier protocole dont le nom commence par celui qui fut attribué à la feuille de données filtrées lors des traitements basiques (voir module [1], p.200) suivi de « _P » et de l'extension de fichier *Microsoft Excel* « .xls ».

Les fichiers protocoles sont sauvegardés par *SCANTEST* et placés dans un répertoire « …/protocoles/… » subdivisé en sous-répertoires qui portent les noms de code des épreuves MOHICAN. Dans le cas de l'exemple du test de compréhension cela donne le chemin d'accès et le nom de fichier protocole suivants : « …/protocoles/COMPRE_T1/COMPRE_T1__T_T_P.xls »).

Un des avantages liés à cette méthode de stockage par synthèses spécifiques rassemblées dans les feuilles de données d'un *fichier Microsoft Excel* est de nous permettre d'affiner nos analyses en retravaillant des résultats sans devoir recommencer toute la procédure de calcul.

Chapitre VI :

Analyses spectrales des propositions au sein d'une QCM

Sommaire

A. Matrices de résultats

Nous présentons ici les résultats obtenus après sélection de la commande « *Traitements basiques…* » (voir p. 200). Ce module de *SCANTEST 2.0* produit des matrices de résultats par question qui permettent de calculer les *rpbis* classiques, les *rpbis SC*, les *rpbis SCT*, les effectifs en nombres et en pourcentages ainsi que les certitudes moyennes pour chaque proposition au sein d'une question. Ces informations brutes sont ensuite d'une part mises en formes (voir p. 216) pour une présentation synthétique dans les protocoles d'analyse et d'autre part utilisées dans les traitements liés aux indices d'évaluation de la qualité spectrale globale de chaque question (p. 227) et du test (p. 256).

1. Matrice de résultats d'un traitement « Calcul des rpbis Spectraux Contrastés»

Les rpbis Spectraux Contrastés (*rpbis SC*) sont calculés à l'aide des procédures activées par la commande « Calcul des rpbis Spectraux » de l'interface utilisateur *SCANTEST 2.0* (p. 199). Rappelons que ce traitement a préalablement nécessité un filtrage des données (voir p. 200). Dans le cadre de l'exemple de l'épreuve de physique, tous les sujets de toutes les sections de toutes les facultés impliquées dans ce test[63] ont été sélectionnés (ce qui représente 2.497 étudiants).

Dans une première phase du traitement, le programme crée une série de nouvelles feuilles *Microsoft Excel*, une par question, qui contiennent les résultats bruts à la suite des données. Voici ces résultats bruts produits par *SCANTEST 2.0* pour les *rpbis SC*. Nous verrons plus loin (p. 216) qu'ils subiront une mise en forme à l'aide du module « Protocole d'analyse spectrale de l'épreuve et des items » décrit au chapitre précédent (p. 207).

	A	B	C	D	E	F	G	H
1	OM	P1	P2	P3	P4	P5	P6	P7
2501	0,47	3,26	1,87	2,29	2,30	2,29	2,26	1,00
2502	3,26	2,19	3,26	3,26	3,26	3,26	3,26	3,26
2503	-2,79	1,07	-1,39	-0,97	-0,96	-0,97	-1,00	-2,26
2504	-1,65	0,63	-0,82	-0,57	-0,57	-0,57	-0,59	-1,34
2505	0,02	0,44	0,01	0,24	0,03	0,13	0,11	0,00
2506	0,98	0,56	0,99	0,76	0,97	0,87	0,89	1,00
2507	0,02	0,25	0,01	0,18	0,03	0,12	0,10	0,00
2508	0,15	0,50	0,11	0,43	0,17	0,34	0,32	0,06
2509	-0,25	0,31	-0,09	-0,25	-0,10	-0,20	-0,19	-0,08
2510	59	1102	31	606	76	334	280	9

Dx
Da
Dx-Da
Dx-Da)/ecart type
p
q
pq
racine pq
((Dx-Da)/ecart type) ∗ (racine pq) = rpbis SC
n

Les colonnes A à H reprennent pour une QCM (ici la 1ère) du test de physique les valeurs des informations nécessaires au calcul des rpbis Spectraux (*Dx*, *Da*, *Dx-Da*, *Dx-Da*/écart type, *p*, *q*, *pq*, et racine de *pq*) des différentes possibilités de réponse : OM (omission) et P1 à P7 (les propositions). L'écart type des degrés de certitude utilisés (de 0 à 5) vaut dans le cadre de notre exemple 1,69 (cette valeur n'est pas reprise dans cette partie du tableau). L'avant dernière ligne (n° 2509) contient les valeurs des *rpbis SC* pour l'omission et les 7 propositions. Enfin, la dernière ligne (n° 2510) contient à titre indicatif le nombre de sujets (*n*) concernés par chaque possibilité de réponse.

P1 qui est la réponse correcte obtient un *rpbis SC* positif à 0,31. Les *rpbis SC* des réponses incorrectes sont négatifs pour l'omission (-0,25) et les propositions 3 (-0,25), 5 (-0,20) et 6 (-0,19). Ils sont aussi négatifs mais cependant proches de 0 pour les propositions 2 (-0,09), 4 (-0,1) et 7 (-0,08).

[63] Il s'agit des facultés de Médecine, Sciences, Sciences Appliquées, Sciences Agronomiques et Médecine Vétérinaire. Par contre, ne sont pas concernés par ce test : les étudiants des facultés de Droit, des Sciences Economiques, de Philosophie et Lettres et de Psychologie et des Sciences de l'Education.

2. Matrice de résultats d'un traitement « Calcul des rpbis Classiques »

SCANTEST calcule aussi les valeurs des *rpbis* classiques. Voici le tableau des résultats fournis par le programme dans le cadre de notre exemple.

	A	B	C	D	E	F	G	H
1	OM	P1	P2	P3	P4	P5	P6	P7
2513	5,25	7,55	6,00	5,69	5,51	5,66	6,03	5,22
2514	6,56	5,73	6,54	6,80	6,56	6,67	6,60	6,54
2515	-1,31	1,82	-0,54	-1,11	-1,05	-1,00	-0,57	-1,32
2516	-0,75	1,04	-0,31	-0,63	-0,60	-0,57	-0,32	-0,75
2517	-0,11	0,51	-0,03	-0,27	-0,10	-0,19	-0,10	-0,04

Mx
Ma
Mx-Ma
(Mx-Ma)/écart type
((Mx-Ma)/écart type)(racine pq) = rpbis Classique*

Remarquons qu'on ne retrouve plus les lignes *p*, *q*, *pq* et racine de *pq* car ces valeurs ont déjà été calculées dans le cadre du *rpbis SC* et nous les réutilisons ici afin d'alléger le traitement.

Comme pour les *rpbis SC*, nous obtenons des valeurs négatives pour les réponses incorrectes P3 (-0,27), P5 (-0,19) et pour l'omission (-0,11). Les autres propositions incorrectes obtiennent aussi des *rpbis* classiques négatifs mais proches de zéro (P2 = -0,03, P4 = -0,10, P6 = -0,1 et P7 = -0,04).

La réponse correcte P1 obtient un *rpbis* classique égal à 0,51. Pour contrer le problème du recouvrement des scores de chaque QCM dans le score total de l'épreuve nous calculons un seuil selon la formule $1/\sqrt{nq}$ (*nq* étant le nombre de questions figurant dans le test). Il y avait 10 QCM dans le test de physique, ce qui donne un seuil calculé valant 0,32. Le *rpbis* classique de la réponse correcte se situe à 19 points au-dessus de ce seuil.

3. Matrice de résultats d'un traitement «Calcul des rpbis Spectraux Turbo»

Dans cette section nous présenterons les matrices de résultats bruts liés aux *rpbis SCT80* et *rpbis SCT90* pour la 1ère question du test de physique. Ces résultats, comme l'indiquent les références aux paliers *T80* et *T90* (voir principe de la turbo analyse, p. 186), ont été calculés dans le 1er cas (pour *rpbis SCT80*) à partir des données des étudiants dont le Réalisme (*Rs*) se situe entre 80 et 100, et, dans le 2ème cas (*rpbis SCT90*), à partir des données des étudiants dont *Rs* se situe entre 90 et 100, donc un sous-groupe plus réaliste (voir procédure de calcul du réalisme p. 184). La procédure de paramétrage de *SCANTEST 2.0* pour obtenir ces *rpbis SCT* est exposée p. 200).

	A	B	C	D	E	F	G	H
1	OM	P1	P2	P3	P4	P5	P6	P7
2637	0,02	0,51	0,02	0,21	0,03	0,11	0,11	0,00
2638								
2639	0,22	3,90	2,31	2,42	2,32	2,33	2,19	0,50
2640	3,90	2,26	3,90	3,90	3,90	3,90	3,90	3,90
2641	-3,67	1,64	-1,58	-1,47	-1,57	-1,56	-1,70	-3,40
2642	-2,43	1,08	-1,05	-0,97	-1,04	-1,03	-1,13	-2,24
2643	0,02	0,51	0,02	0,21	0,03	0,11	0,11	0,00
2644	0,98	0,49	0,98	0,79	0,97	0,89	0,89	1,00
2645	0,02	0,25	0,01	0,17	0,03	0,10	0,10	0,00
2646	0,13	0,50	0,12	0,41	0,16	0,31	0,31	0,04
2647	-0,31	0,54	-0,13	-0,40	-0,17	-0,32	-0,35	-0,10
2648	18	540	16	222	28	114	114	2
2649	2,67	78,42	46,25	48,91	48,15	46,92	44,73	10,00
2650								
2651	0,02	0,64	0,01	0,15	0,02	0,06	0,10	0,00
2652								
2653	0,40	4,21	3,00	2,52	3,14	2,55	2,23	0,00
2654	4,21	2,41	4,21	4,21	4,21	4,21	4,21	4,21
2655	-3,81	1,80	-1,21	-1,68	-1,06	-1,66	-1,98	-4,21
2656	-2,65	1,25	-0,84	-1,17	-0,74	-1,15	-1,38	-2,93
2657	0,02	0,64	0,01	0,15	0,02	0,06	0,10	0,00
2658	0,98	0,36	0,99	0,85	0,98	0,94	0,90	1,00
2659	0,02	0,23	0,01	0,13	0,02	0,06	0,09	0,00
2660	0,13	0,48	0,11	0,35	0,15	0,24	0,30	0,00
2661	-0,33	0,60	-0,09	-0,42	-0,11	-0,28	-0,41	xxxx
2662	5	199	4	46	7	20	31	0
2663	8,00	85,40	60,00	51,00	56,67	48,89	42,86	xxxx

Résultats sur base données sujets avec R entre 80 et 100

Pourcentages de sujets
Da
Dx-Da
(Dx-Da)/ecart type
p
q
pq
racine pq
rpbis SCT80
n
Certitudes moyennes

Résultats sur base données sujets avec R entre 90 et 100

Pourcentages de sujets
Dx
Da
Dx-Da
(Dx-Da)/ecart type
p
q
pq
racine pq
rpbis SCT90
n
Certitudes moyennes

Les paramètres utilisés ici pour la turbo analyse sont : « Seuil de réalisme exigé : *80* à *100* avec pas de : *10* ». Ce paramétrage signifie que nous calculerons dans un premier temps les *rpbis SCT* sur les

données des étudiants dont le réalisme est supérieur ou égal au seuil initial fixé ici à 80 et inférieur ou égal au seuil supérieur 100. Ensuite, dans un second temps, après un pas de 10 ajouté au seuil inférieur initial de réalisme, nous recommencerons le traitement *rpbis SCT* sur les données des étudiants dont le réalisme est cette fois supérieur ou égal à 90 (80 + pas de 10) et inférieur ou égal au seuil supérieur 100. Il convient donc de distinguer dans la matrice générée par *SCANTEST 2.0*, deux séries de résultats similaires mais dont les valeurs changent en fonction des deux groupes d'étudiants envisagés : (1) ceux dont le score de réalisme est compris entre 80 et 100 et (2) ceux dont le réalisme est compris entre 90 et 100.

Par rapport aux matrices précédentes, deux types de données supplémentaires sont fournies : d'une part le pourcentage de sujets (lignes N°2637 et N°2651) et d'autre part les certitudes moyennes (lignes N° 2649 et N°2663) des sujets pour chaque possibilité de réponse. Dans l'exemple, lorsqu'on compare le nombre de sujets (*n*) (lignes N°2648 et N°2662), on observe une diminution du nombre d'étudiants dans la seconde série de résultats étant donné l'exigence en réalisme plus élevée (pour la réponse correcte, le nombre de sujets passe de 540 à 199). En parallèle à cette diminution du nombre de sujets, on observe pour cette question un pourcentage plus élevé de sujets ayant répondu correctement dans le groupe dont le réalisme est supérieur ou égal à 90 (pourcentage de réponses correctes : T80 = 51% et T90 = 64%).

En ce qui concerne les *rpbis SCT* (lignes N°2647 et N°2661), pour la réponse correcte P3 le *rpbis SCT90* (0,60) est plus élevé que le *rpbis SCT80* (0,54). Pour les réponses incorrectes, le *rpbis SCT90* de la proposition P7 n'est pas calculable car cette proposition n'a été choisie par aucun étudiant dont le réalisme était supérieur ou égal à 90. Les *rpbis SCT90* des propositions P3 et P6 sont plus négatifs que les *rpbis SCT80* de ces propositions, même observation pour l'omission (OM). Pour les propositions P2, P4 et P5 c'est le contraire, les *rpbis SCT80* sont plus négatifs que les *rpbis SCT90*.

B. Protocoles d'analyse des PROPOSITIONS au sein d'une QCM

Pour faciliter la lecture des statistiques spectrales relatives aux propositions d'une QCM, nous avons programmé dans *SCANTEST 2.0* un module « Protocole d'analyse spectrale de l'épreuve et des items » (voir p. 207) qui met en page de façon plus lisible les informations exposées aux paragraphes précédents.

Nous présentons ici les extraits du *protocole SCANTEST 2.0 pour épreuves MOHICAN* en rapport avec l'exemple de la première QCM du test de physique, plus précisément, avec le niveau d'analyse « PROPOSITIONS – STATISTIQUES CLASSIQUES ET SPECTRALES » de cette question.

1. Mise en forme des statistiques classiques

En ce qui concerne les statistiques classiques liées aux propositions, nous avons gardé trois types d'informations présentées dans les résultats bruts (p. 214) : les nombres de réponses (N Rép.), les pourcentages de réponses (% Rép.) et les *rpbis* classiques (rpbis). La réponse correcte est signalée par des statistiques en caractères gras.

Voici l'extrait du *protocole SCANTEST 2.0 pour épreuves MOHICAN* en lien avec les statistiques classiques du test de physique :

QUESTION N° 1

a) Statistiques classiques

	OM	**P1**	P2	P3	P4	P5	P6	P7	P8	P9
1) N Rép.	59	**1102**	31	606	76	334	280	9	0	0
2) % Rép.	2%	**44%**	1%	24%	3%	13%	11%	0%	0%	0%
3) rpbis	-0,11	**0,51**	-0,03	-0,27	-0,10	-0,19	-0,10	-0,04	xxxx	xxxx

Il existe dans certaines épreuves MOHICAN, des questions comportant 9 propositions de réponses, ce qui explique la présence des colonnes P8 et P9 dans cette présentation standardisée des résultats des questions. Etant donné l'absence de réponses pour P8 et P9 (N Rép. = 0 et % Rép. = 0) le rpbis est incalculable et ce qui explique les « xxxx » dans les cases.

2. Mise en forme des rpbis Spectraux Contrastés

Nous avons repris les certitudes moyennes (C Moy.) et les *rpbis SC* qui figurent dans les résultats bruts calculés à l'aide après traitement « Calcul des rpbis Spectraux Contrastés » (p. 213) du module « Traitements basiques ». Voici l'extrait du *protocole SCANTEST 2.0 pour épreuves MOHICAN* en lien avec ces statistiques spectrales du test de physique :

b) Statistiques spectrales

	OM	**P1**	P2	P3	P4	P5	P6	P7	P8	P9
1) C Moy.	9%	**65%**	37%	46%	46%	46%	45%	20%	xxxx	xxxx
2) rpbis SC	-0,25	**0,31**	-0,09	-0,25	-0,10	-0,20	-0,19	-0,08	xxxx	xxxx

3. Mise en forme des rpbis Spectraux Contrastés après Turbo analyse

Pour des raisons de place nous nous limitons ici à l'exemple des résultats bruts repris ci-avant (p. 214). Ceci explique la numérotation de l'extrait du protocole ci-dessous qui commence à « *8.* » étant donné que nous ne montrons pas ici les statistiques des *rpbis SCT* aux paliers de turbo analyse T10 à T70.

8. Palier de Turbo analyse : T80

	OM	**P1**	P2	P3	P4	P5	P6	P7	P8	P9
8.1 N Rép. T80	18	**540**	16	222	28	114	114	2	0	0
8.2 % Rép. T80	2%	**51%**	2%	21%	3%	11%	11%	0%	0%	0%
8.3 C. Moy. T80	3%	**78%**	46%	49%	48%	47%	45%	10%	xxxx	xxxx
8.4 rpbis SC T80	-0,31	**0,54**	-0,13	-0,40	-0,17	-0,32	-0,35	-0,10	xxxx	xxxx

9. Palier de Turbo analyse : T90

	OM	**P1**	P2	P3	P4	P5	P6	P7	P8	P9
9.1 N Rép. T90	5	**199**	4	46	7	20	31	0	0	0
9.2 % Rép. T90	2%	**64%**	1%	15%	2%	6%	10%	0%	0%	0%
9.3 C. Moy. T90	8%	**85%**	60%	51%	57%	49%	43%	xxxx	xxxx	xxxx
9.4 rpbis SC T90	-0,33	**0,60**	-0,09	-0,42	-0,11	-0,28	-0,41	xxxx	xxxx	xxxx

Pour chaque proposition, aux paliers de turbo analyse envisagés, *SCANTEST 2.0* sélectionne dans le fichier *Microsoft Excel* contenant les résultats bruts les informations concernant le nombre de réponses (N Rép. T…), le pourcentage de réponses (% Rép. T…), la certitude moyenne (C. Moy. T…), les *rpbis SCT* (rpbis SCT…) et les introduit dans le fichier *Microsoft Excel* contenant le protocole d'analyse.

4. Caractéristiques de la turbo analyse appliquée au test de physique pour établir nos premières observations

Concernant le principe de la turbo analyse, nous avons vu qu'il est possible, et nous l'avons programmé dans *SCANTEST 2.0* (voir pp. 197 et 200), de définir à partir de quel niveau minimum et jusqu'à quel niveau maximum de réalisme on peut prendre en compte les données des sujets pour calculer les *rpbis SCT*, et ce, avec un pas qui lui aussi est à définir. Dans les traitements qui ont abouti aux premiers résultats qui vont suivre, nous avons choisi de fixer le seuil inférieur initial à 0, le seuil supérieur à 100 et le pas à 10. Ce paramétrage de la turbo analyse entraîne 10 analyses successives.

La première analyse est faite en tenant compte des sujets dont le réalisme (*Rs*) est compris entre 0 (seuil inférieur initial) et 100 (seuil supérieur), c'est-à-dire tous les étudiants (ce qui revient à calculer les *rpbis SC*).

Ensuite, une seconde analyse est faite en tenant compte des données des étudiants dont le réalisme est compris entre 10 (seuil inférieur initial augmenté du pas égal à 10) et 100 (niveau supérieur), cette analyse aboutit aux *rpbis SCT10*.

Après cela une troisième analyse est faite où les données sont celles des étudiants dont le réalisme est compris entre 20 (seuil inférieur de l'analyse précédente augmenté du pas) et 100 (niveau supérieur) ce qui permet le calcul des *rpbis SCT20*. Les analyses se succèdent ainsi franchissant à chaque fois un nouveau palier de turbo analyse jusqu'à ce que le seuil inférieur soit égal au seuil supérieur. En tout, ce paramétrage aura permis de calculer les *rpbis SCT* à 10 paliers de Turbo analyse : T0, T10, T20, T30, T40, T50, T60, T70, T80 et T90.

C. Constats et questions à propos des analyses spectrales des PROPOSITIONS de la 1ère QCM du test de physique

Dans cette section, nous dressons une série de constats liés à l'analyse des propositions de la première QCM du test de physique après turbo analyse en 10 paliers. Ces constats soulèvent des questions dont nous chercherons les réponses dans les protocoles d'analyse des propositions des questions de l'ensemble des 10 épreuves MOHICAN.

Nous signalerons par un point d'interrogation « ? » suivi d'un numéro d'ordre, le tout entre crochets, les premières questions que nous nous posons au départ des constats effectués. Ces questions seront reprises plus loin où, après analyse des données de l'ensemble des questions des 10 épreuves MOHICAN, nous tenterons d'y apporter les réponses.

1. Seuil d'infléchissement des effectifs des propositions lors d'une turbo analyse

Voici le tableau de ventilation des sujets aux différents seuils exigés en réalisme dans le cadre de la turbo analyse décrite au paragraphe précédent. Ces données concernent la 1ère question du test de physique.

		OM	%OM	P1	%P1	P2	%P2	P3	%P3	P4	%P4	P5	%P5	P6	%P6	P7	%P7	Total	%Total
T0	$R \geqslant 0$	59	2,4%	1102	44,1%	31	1,2%	606	24,3%	76	3,0%	334	13,4%	280	11,2%	9	0,4%	2497	100%
T10	$R \geqslant 10$	57	2,3%	1101	44,1%	31	1,2%	606	24,3%	76	3,0%	334	13,4%	280	11,2%	9	0,4%	2494	99,9%
T20	$R \geqslant 20$	56	2,3%	1090	44,0%	31	1,3%	605	24,4%	76	3,1%	333	13,4%	279	11,3%	9	0,4%	2479	99,3%
T30	$R \geqslant 30$	54	2,2%	1080	43,9%	31	1,3%	600	24,4%	76	3,1%	330	13,4%	278	11,3%	9	0,4%	2458	98,4%
T40	$R \geqslant 40$	53	2,2%	1064	43,9%	31	1,3%	594	24,5%	75	3,1%	323	13,3%	272	11,2%	9	0,4%	2421	97,0%
T50	$R \geqslant 50$	47	2,0%	1040	44,3%	30	1,3%	573	24,4%	71	3,0%	312	13,3%	266	11,3%	8	0,3%	2347	94,0%
T60	$R \geqslant 60$	44	2,0%	968	44,8%	30	1,4%	522	24,2%	62	2,9%	283	13,1%	243	11,3%	7	0,3%	2159	86,5%
T70	$R \geqslant 70$	34	1,9%	825	46,9%	25	1,4%	404	23,0%	44	2,5%	219	12,5%	204	11,6%	3	0,2%	1758	70,4%
T80	$R \geqslant 80$	18	1,7%	540	51,2%	16	1,5%	222	21,1%	28	2,7%	114	10,8%	114	10,8%	2	0,2%	1054	42,2%
T90	$R \geqslant 90$	5	1,6%	199	63,8%	4	1,3%	46	14,7%	7	2,2%	20	6,4%	31	9,9%	0	0,0%	312	12,5%

Ventilation des effectifs de sujets aux différents paliers de turbo analyse (de T0 à T90) - test de physique (n = 2.497) - QCM 1

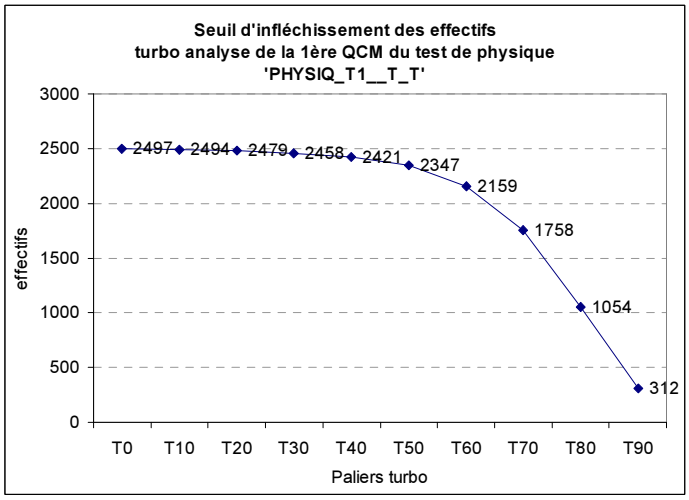
La colonne « Total » du tableau ci-dessus montre combien de sujets sont concernés à chaque palier de turbo analyse de T0 à T90. Par exemple, la ligne relative au niveau T0 ($Rs \geqslant 0$) indique que 59 (2,4%) des étudiants ont omis (OM), 1.102 (44,1%) ont choisi la réponse correcte P1 (fond gris), 31 (1,2%) ont coché la P2, etc.

Lorsque $Rs \geqslant 0$, c'est l'entièreté du groupe qui est concerné. La plus mauvaise performance en réalisme qu'un étudiant puisse obtenir est égale à 0 (100% d'erreurs moyennes de certitude) et, évidemment, tous les sujets obtiennent un $Rs \geqslant 0$, d'où un total qui correspond à 2.497 étudiants (100% dans la colonne « %Total »).

En ce qui concerne le niveau de turbo analyse $Rs \geqslant 10$, les nombres de sujets concernés sont quasi identiques. C'est seulement à partir du niveau de turbo analyse $Rs \geqslant 20$ que le nombre de sujets commence à diminuer légèrement pour la proposition P1. A ce seuil exigé de réalisme, 8 étudiants ayant choisi la P1 (1.102 – 1.090 = 12) ont un réalisme inférieur à 20. Au total, pour $Rs \geqslant 20$, il reste 2.479 étudiants soit 99,3% des 2.497 sujets de départ.

Lorsqu'on observe les chiffres du dernier niveau de turbo analyse $Rs \geqslant 90$, on constate que peu de sujets sont concernés par l'omission ($n = 5$), par la P2 ($n = 4$) et par la P4 ($n = 7$). En ce qui concerne la proposition P7, aucun étudiant dont le réalisme était supérieur à 90 ne l'a choisie. Le nombre total de sujets concerné descend à 314, soit 12,5% du nombre de départ (2.497).

Le graphique ci-dessous, montre un seuil d'infléchissement des effectifs à partir de T50 et T60.



[1.1] Observe-t-on ce type d'infléchissement des effectifs dans les autres questions des épreuves MOHICAN ?

[1.2] En pourcentages, les effectifs aux différents paliers de turbo analyse des autres épreuves MOHICAN sont-ils équivalents ?

Signalons que nous présenterons plus loin (p. 275) les distributions des scores de *R*éalisme des *s*ujets (*Rs*) en la comparant avec les résultats d'une autre sur l'indice *Rs* portant sur 28 examens avec degrés de certitude ayant eu lieu à la Faculté de Psychologie et des Sciences de l'Education de l'Université de Liège entre 1994 et 1996 (Gilles, 1996a).

2. Pourcentages de réponses correctes et réalisme des étudiants

Dans le tableau précédent de ventilation des effectifs de sujets aux différents paliers de turbo analyse, nous constatons que le pourcentage de sujets qui ont répondu correctement (P1) ne varie pratiquement pas jusqu'au niveau de turbo analyse $Rs \geqslant 70$ où il passe de 44,8% à 46,9% (+ 2,1%). Le pourcentage de réponses correctes continue ensuite à augmenter : pour $Rs \geqslant 80$ il vaut 51,2% (+ 4,3%) et pour $Rs \geqslant 90$ il est à 63,8% (+ 12,6%).

[1.3] Observe-t-on dans les autres questions des épreuves MOHICAN des pourcentages de réponses correctes plus élevés pour les groupes d'étudiants qui présentent les meilleurs scores à l'indice *Rs* (ceux dont les données sont sélectionnées aux paliers de turbo analyse les plus élevés) ?

Nous répondrons à cette question plus loin dans cette recherche après analyse des données de toutes les épreuves.

3. Configurations des rpbis

Les tableaux récapitulatifs ci-dessous reprennent les valeurs des différents types de rpbis pour la 1ère question du test de physique.

Voici le tableau récapitulatif des *rpbis* classiques :

	OM	**P1**	P2	P3	P4	P5	P6	P7
rpbis classique	-0,11	**0,51**	-0,03	-0,27	-0,10	-0,19	-0,10	-0,04

Valeurs rpbis classiques des 7 propositions et de l'omission de la 1ère QCM du test de physique

Nous observons un rpbis classique positif, au-dessus du seuil calculé pour contrer le problème du recouvrement entre les scores de la question et les score totaux au test (voir p. 176), qui vaut dans le cas de ce test comportant 10 questions : $1/\sqrt{10} = 0{,}32$. Quant aux réponses incorrectes, les *rpbis* classiques sont tous négatifs ou proches de 0. Cette configuration de *rpbis* classiques montre un fonctionnement correct de la QCM du point de vue de sa cohérence avec l'ensemble des questions du test.

Voici maintenant le tableau récapitulatif des rpbis Spectraux Contrastés :

	OM	**P1**	P2	P3	P4	P5	P6	P7
rpbis SC	-0,25	**0,31**	-0,09	-0,25	-0,10	-0,20	-0,19	-0,08

Valeurs des rpbis SC des sept propositions et de l'omission de la 1ère QCM du test de physique

Dans le contexte de cette 1ère QCM, les valeurs obtenues au *rpbis SC* nous permettent de dire qu'on observe chez les étudiants qui ont correctement répondu, une tendance à donner des degrés de certitude plus élevés que chez les étudiants qui ont échoué à la question. Le *rpbis SC* de la réponse correcte P1 est positif (0,31) et ceux des propositions incorrectes sont tous négatifs ou proches de zéro. C'est la situation à laquelle on peut logiquement s'attendre quand la question ne pose pas de problème particulier du point de vue de sa cohérence spectrale.

Pour les deux types d'indice on observe donc des configurations de valeurs dans lesquelles le rpbis de la réponse correcte est positif et les rpbis des réponses incorrectes sont négatifs.

On observe aussi ce type de configuration pour les valeurs obtenues après turbo analyse. Rappelons que les *rpbis SCT* sont alors calculés à partir de données de plus en plus fiables au fur et à mesure que l'on monte dans les paliers de turbo analyse, ces données provenant alors des étudiants dont le réalisme est de plus en plus élevé. Voici le tableau récapitulatif des *rpbis SCT* :

	OM	**P1**	P2	P3	P4	P5	P6	P7
rpbis SCT10	-0,25	**0,31**	-0,09	-0,25	-0,10	-0,20	-0,19	-0,08
rpbis SCT20	-0,25	**0,32**	-0,09	-0,26	-0,10	-0,20	-0,19	-0,08
rpbis SCT30	-0,25	**0,33**	-0,10	-0,26	-0,11	-0,21	-0,20	-0,08
rpbis SCT40	-0,25	**0,34**	-0,10	-0,27	-0,12	-0,22	-0,21	-0,09
rpbis SCT50	-0,24	**0,35**	-0,11	-0,28	-0,12	-0,22	-0,21	-0,08
rpbis SCT60	-0,26	**0,39**	-0,12	-0,31	-0,12	-0,23	-0,25	-0,10
rpbis SCT70	-0,28	**0,46**	-0,13	-0,35	-0,15	-0,27	-0,30	-0,08
rpbis SCT80	-0,31	**0,54**	-0,13	-0,40	-0,17	-0,32	-0,35	-0,10
rpbis SCT90	-0,33	**0,60**	-0,09	-0,42	-0,11	-0,28	-0,41	xxxx

Valeurs rpbis SCT aux sept propositions et à l'omission de la 1ère QCM du test de physique

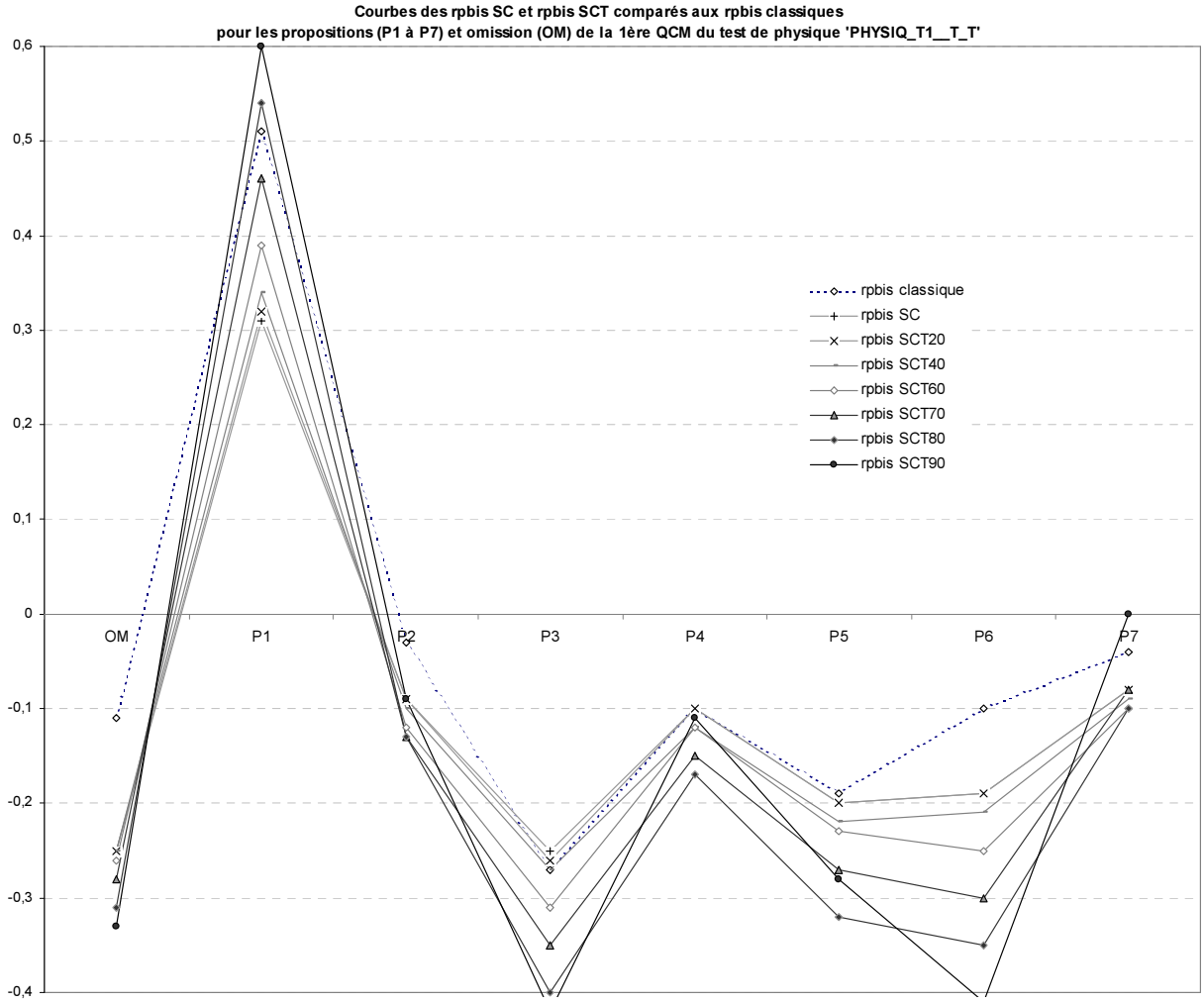
On constate que les *rpbis SCT* obtenus à des seuils élevés de réalisme (à partir de *rpbis SCT60*) ont tendance à être plus tranchés. En effet, les *rpbis SCT* de la réponse correcte sont de plus en plus élevés et les *rpbis SCT* des propositions incorrectes, dans l'ensemble, de plus en plus négatifs. Au palier T90 pour trois propositions incorrectes P2, P4 et P5 nous observons des valeurs négatives mais en hausse par rapport au(x) palier(s) précédent(s). Remarquons que les *rpbis SCT* de la proposition P7 sont plus stables et qu'il devient impossible de calculer le *rpbis SCT* à T90 étant donné l'effectif des sujets réduit à zéro (voir le tableau de ventilation des effectifs ci-avant, p. 218), d'où les « xxxx » dans la case.

Dès lors,

[1.4] Retrouve-t-on dans les autres questions des épreuves MOHICAN cette analogie des configurations des *rpbis* classiques, *rpbis SC* et *rpbis SCT* (valeurs positives à la réponse correcte et valeurs négatives aux réponses incorrectes) ? Autrement dit, lorsque la question est cohérente avec l'ensemble du test (mesure à l'aide du rpbis classique), est-elle systématiquement cohérente d'un point de vue spectral (mesure à l'aide des *rpbis SC* et *SCT*) ? Existe-t-il des cas où les configurations des différents rpbis se contredisent ?

4. *rpbis SCT* comparés aux *rpbis classiques*

Voici la représentation graphique des valeurs contenues dans les tableaux récapitulatifs précédents (voir pp. 219 et 220). Afin d'alléger le graphique nous n'avons pas repris les *rpbis SCT10*, *rpbis SCT30* et *rpbis SCT50.* Pour chacun de ces *rpbis SCT* non repris, les valeurs se situent respectivement entre le *rpbis SCT* qui le précède et le *rpbis SCT* qui le suit. Ce graphique permet de visualiser les valeurs des rpbis classiques et des rpbis spectraux pour l'omission et pour chacune des propositions de la 1ère question du test de physique (les traits reliant les valeurs n'ont pour but que de guider l'œil du lecteur).



Pour les propositions incorrectes (P2 à P7) ainsi que pour l'omission, lorsqu'on compare la courbe du *rpbis* classique avec les courbes des *rpbis* spectraux, on observe une tendance à voir les valeurs du *rpbis* classique se positionner au-dessus de celles des *rpbis SCT*.

En ce qui concerne la réponse correcte (P1), celle-ci obtient un *rpbis* classique de 0,51 et cette valeur n'est dépassée qu'à partir du *rpbis SCT80*.

Clarifions maintenant ces relations d'ordre des valeurs que nous observons pour les différents types de rpbis.

Lors de la présentation de la problématique des différents types de *rpbis* (voir p. 189), nous avions relevé dans le contexte de l'exemple théorique la relation suivante :

rpbis classique > *rpbis SC* > *rpbis SCT80.*

A l'aide des résultats de la turbo analyse nous pouvons maintenant observer ce qu'il en est (1) dans un cas réel et (2) en incluant les *rpbis SCT* calculés aux paliers de turbo analyse T10 à T90.

Rappelons aussi que nous avions exposé la possibilité de calculer pour les différents types de rpbis les moyennes pondérées des propositions incorrectes ($\overline{rpbis^i}$) (voir p. 189) pour ensuite soustraire ces $\overline{rpbis^i}$ aux rpbis des réponses correctes ($rpbis^c$) afin de calculer les différences [$rpbis^c$ - $\overline{rpbis^i}$].

Voici les valeurs obtenues pour la 1ère question du test de physique lorsqu'on applique cette procédure :

	$\overline{rpbis^i}$	$rpbis^c$	[$rpbis^c$ - $\overline{rpbis^i}$]
rpbis classique	-0,20	0,51	0,71
rpbis SC	-0,21	0,31	0,52
rpbis SCT10	-0,21	0,31	0,52
rpbis SCT20	-0,22	0,32	0,54
rpbis SCT30	-0,22	0,33	0,55
rpbis SCT40	-0,23	0,34	0,57
rpbis SCT50	-0,24	0,35	0,59
rpbis SCT60	-0,26	0,39	0,65
rpbis SCT70	-0,30	0,46	0,76
rpbis SCT80	-0,34	0,54	0,88
rpbis SCT90	-0,36	0,60	0,96

D'une part, on voit que dans le contexte de cette turbo analyse en 10 paliers de T10 à T90, la relation d'ordre pour $\overline{rpbis^i}$, $rpbis^c$ et [$rpbis^c$ - $\overline{rpbis^i}$] devient :

- $rpbis^c$: *SC* < *SCT10* < *SCT20* < *SCT30* < *SCT40* < *SCT50* < *SCT60* < *SCT70* < classique < *SCT80* < *SCT90* ;
- $\overline{rpbis^i}$: classique > *SC* > *SCT10* > *SCT20* > *SCT30* > *SCT40* > *SCT50* > *SCT60* > *SCT70* > *SCT80* > *SCT90* ;
- [$rpbis^c$ - $\overline{rpbis^i}$]: *SC* < *SCT10* < *SCT20* < *SCT30* < *SCT40* < *SCT50* < *SCT60* < classique < *SCT70* < *SCT80* < *SCT90* .

D'autre part, nous observons pour les *rpbis SCT* des différences [$rpbis^c$ - $\overline{rpbis^i}$] de plus en plus grandes au fur et à mesure que l'on monte dans les paliers de turbo analyse.

Chapitre VII :

Analyses spectrales des QCM

Sommaire

A. Outils d'aide à l'identification des Niveaux de Cohérence Spectrale d'une question (NCSq)

Rappelons que les valeurs des *rpbis SCT* ont été calculées à partir des données des 2.497 étudiants qui ont participé au test de physique[64]. Pour chaque QCM, les valeurs des *rpbis SCT* ont été obtenues après turbo analyse, celle-ci a été effectuée en 10 paliers selon les modalités suivantes : le seuil exigé de réalisme des sujets (*Rs*) de départ a été fixé à 0 ($Rs \geqslant 0$), les 9 autres seuils ont ensuite à nouveau été fixés de $Rs \geqslant 10$ à $Rs \geqslant 90$ en procédant par pas de 10 points de réalisme.

1. Qu'entendons-nous par cohérence spectrale d'une QCM ?

Lorsque nous avons exposé la problématique des rpbis spectraux (p. 178), nous avons signalé que le *rpbis SC* est calculé en corrélant les degrés de certitude (de 0 à 5 dans le cas des données de la recherche MOHICAN) qui ont accompagné les réponses à une QCM avec les choix/rejets (1 ou 0) des propositions de cette QCM.

Rappelons que pour la réponse correcte, le *rpbis SC* positif, généralement obtenu, montre dans quelle mesure les sujets qui ont choisi la proposition correcte ont accompagné celle-ci de degrés de certitude en moyenne plus élevés que les degrés de certitude utilisés par les sujets qui se sont trompés.

Rappelons aussi que lorsqu'on calcule le *rpbis SC* d'un distracteur, nous prenons en compte les degrés de certitude des étudiants qui ont choisi ce distracteur et les degrés de certitude des étudiants qui ont choisi la réponse correcte. Nous contrastons ainsi les certitudes liées au distracteur considéré avec les certitudes liées aux réponses correctes en excluant les certitudes liées aux autres distracteurs, d'où l'appellation « contrasté ». Pour une proposition incorrecte, la corrélation des choix/rejets avec les degrés de certitude est en principe négative. Ces valeurs *rpbis SC* négatives auxquelles on s'attend montrent dans quelle mesure les choix du distracteur ont, en moyenne, été accompagnés de degrés de certitude moins élevés lorsqu'on compare avec les degrés de certitude utilisés par ceux qui ont choisi la réponse correcte.

Dès lors, nous définissons la cohérence spectrale d'une QCM comme la propension à obtenir pour la réponse correcte des degrés de certitude en moyenne plus élevés que les degrés de certitudes qui accompagnent les réponses incorrectes. Une QCM dont le Niveau de Cohérence Spectrale (*NCSq*) est maximal obtient pour la réponse correcte un *rpbis SC* qui vaut 1 et pour chaque solution incorrecte des *rpbis SC* qui valent –1.

A contrario, lorsqu'une QCM obtient pour la réponse correcte des degrés de certitude en moyenne moins élevés que les degrés de certitudes qui accompagnent les réponses incorrectes, nous parlons d'incohérence spectrale. Une QCM dont le niveau d'incohérence spectrale est maximal récolte pour la réponse correcte un *rpbis SC* qui vaut -1 et pour chaque solution incorrecte des *rpbis SC* qui valent 1.

Nous verrons ci-après qu'il est possible de situer sur un graphique en nuage de points les niveaux de cohérence spectrale, voire d'incohérence spectrale, atteints par les questions d'un test, et ce, aux différents paliers de turbo analyse. Nous présenterons ensuite l'indice *NCSq* qui nous permet de chiffrer la cohérence spectrale, voire l'incohérence spectrale, d'une QCM aux différents paliers de turbo analyse (p. 231).

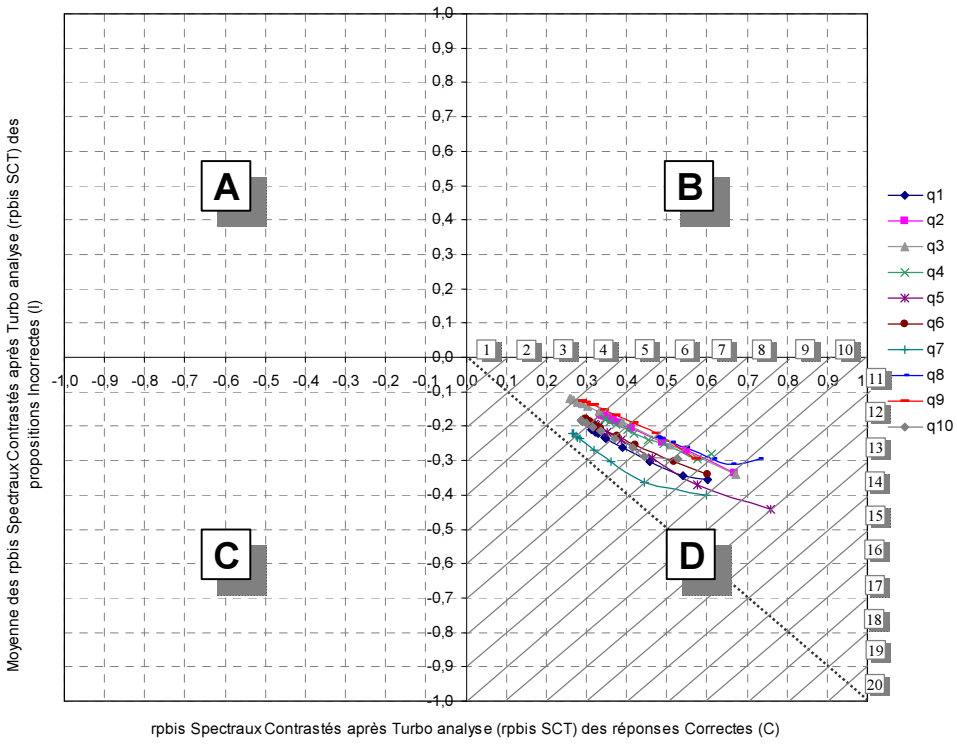
[64] C'est ce qui explique l'appellation « PHYSIQUE_T1__T_T » parfois associée au test de physique dans nos exemples et dans les titres de certains graphiques générés par *SCANTEST 2.0*. L'explication du codage « _T1__T_T » qui suit le nom du test « PHYSIQUE » figure à la page 6.

2. Brin Spectral d'une question (*BSq*) et Gerbe Spectrale d'un test (*GSt*) après turbo analyse

Les graphiques en nuage de points qui suivent permettent de visualiser les niveaux de cohérence spectrale des questions d'un test. Chacun reprend en ordonnée la moyenne des *rpbis SCT* des propositions incorrectes (voir méthode de calcul, p. 189) des QCM du test. Les valeurs des *rpbis SCT* des réponses correctes ont été placées en abscisse. On obtient ainsi 10 points pour chaque QCM, chaque point représentant pour un palier donné de turbo analyse, la valeur de la moyenne pondérée des *rpbis SCT* des propositions incorrectes d'une QCM en correspondance avec la valeur du *rpbis SCT* de la proposition correcte. Ces 10 points peuvent être reliés par des traits et former un « Brin Spectral par question » (*BSq*). L'épreuve de physique dont nous avons utilisé les données pour introduire l'outil est ainsi constituée de 10 *BSq*, formant une « Gerbe Spectrale du test » (*GSt*).

Le graphique ci-dessous est assez grossier et sera agrandi à la page suivante. Nous le discutons ici d'abord globalement. Les échelles des axes couvrent toute la plage des valeurs possibles pour les *rpbis SCT* : de –1 à +1.

Graphe de la Gerbe Spectrale des questions (GSq) du test de physique 'PHYSIQ_T1__T_T' (n = 2.497)



Dans le cas de figure d'une question qui fonctionne correctement, logiquement, les valeurs des *rpbis SCT* des propositions correctes devraient être positives, par contre, les propositions incorrectes, elles, devraient être négatives. Dès lors nous avons découpé l'aire du graphique en quatre zones A, B, C et D.

Si pour une QCM donnée, des points venaient à apparaître dans la **zone A**, ils seraient représentatifs d'une situation problématique à la fois pour les distracteurs et pour la réponse correcte : la moyenne pondérée des *rpbis SCT* des propositions incorrectes serait positive (les étudiants auraient tendance à accompagner de degrés de certitude plus élevés une ou plusieurs solutions incorrectes) et la valeur du *rpbis SCT* de la proposition correcte serait négative (les étudiants auraient tendance à accompagner de degrés de certitude moins élevés leur réponse correcte). Dans le cas de notre exemple du test de physique, aucune QCM n'entre dans cette zone A de double incohérence spectrale.

La **zone B** permet quant à elle de mettre en évidence des QCM qui présenteraient des problèmes liés à une ou plusieurs réponses incorrectes. La moyenne pondérée des *rpbis SCT* des solutions incorrectes est positive et le *rpbis SCT* de la réponse correcte est positif.

A l'inverse de la précédente, la **zone C** permet de visualiser les QCM qui présentent des problèmes liés à la proposition correcte. Le *rpbis SCT* de la réponse correcte est négatif et la moyenne pondérée des

rpbis SCT des réponses incorrectes est négative. Aucune question du test de physique n'entre dans les zones B ou C.

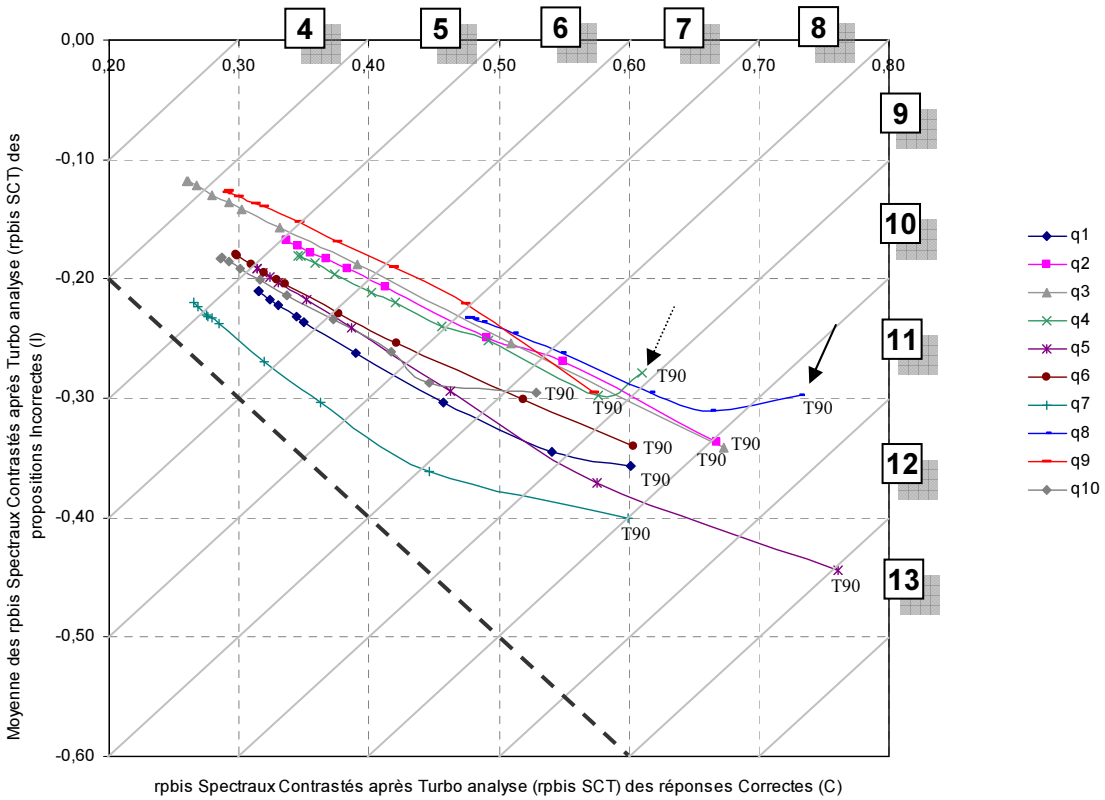
Enfin, la **zone D** peut être qualifiée de zone de double cohérence spectrale dans la mesure où un point qui représente une QCM dans cette zone signifie que, à un niveau de turbo analyse donné, la réponse correcte récolte un *rpbis SCT* positif tandis que la moyenne pondérée des *rpbis SCT* des propositions incorrectes est négatif. Toutes les QCM du test de physique figurent dans cette zone de double cohérence spectrale.

La ligne anti-diagonale en pointillés qui traverse la zone D du graphique désigne les 20 niveaux de cohérence spectrale (de 1 à 20), les régions situées entre deux lignes pleines diagonales.

Le graphique suivant montre un agrandissement de la zone D. Nous avons quadrillé cette zone et placé des lignes diagonales de façon à mettre en évidence les niveaux de cohérence.

Pour des raisons de lisibilité nous avons zoomé une seconde fois sur cette zone de cohérence de façon à ce que l'échelle des abscisses (les *rpbis SCT* de la réponse correcte) couvre une plage de 0,20 à 0,80 et celle des ordonnées (les moyennes des *rpbis SCT* des propositions incorrectes) une plage de 0 à -0,60. Le résultat apparaît sur le graphique en nuage de points ci-dessus. Les chiffres désignant les niveaux de qualité occupés par les points des QCM ont été placés en pourtour du graphique (de 4 à 13).

Zoom sur la zone de cohérence spectrale du graphe de la Gerbe Spectrale des questions (GSq) après turbo analyse (de T0 à T90 avec pas de 10)



Nous observons sur la gerbe spectrale du test de physique, une tendance au regroupement des points à la base des brins (en haut à gauche de la gerbe). Ces points correspondent aux valeurs *rpbis SCT* obtenues aux seuils exigés de réalisme les plus bas de la turbo analyse (de $Rs \geqslant 0$ à $Rs \geqslant 40$).

On observe aussi une autre tendance générale : plus les seuils exigés en réalisme de la turbo analyse sont élevés, plus les niveaux de qualité atteints par les *rpbis SCT* sont grands et atteignent les niveaux 9, 10, 11 et même 13 pour la question 5. Le Brin Spectral de la question 4 (*BSq4*) (signalé par la flèche en pointillés) semble avoir un comportement un peu différent des autres entre le seuil $Rs \geqslant 80$ et $Rs \geqslant 90$ de la turbo analyse : la valeur négative de la moyenne des *rpbis SCT* des réponses incorrectes remonte en passant de -0,30 à -0,28 au lieu de continuer à progresser vers le coin inférieur droit comme pour les autres *BSq*. Nous observons aussi un changement de trajectoire du *BSq* pour de la 8ème QCM (signalé par la flèche continue) : à T80, la valeur de la moyenne des *rpbis SCT* des réponses incorrectes vaut -0,31 et à T90 elle augmente d'un point (-0,30) au lieu de continuer à diminuer (à devenir de plus en plus négative) comme pour les 8 autres QCM.

Voici le tableau des valeurs exprimées dans ces graphiques :

		T0	*T10*	*T20*	*T30*	*T40*	*T50*	*T60*	*T70*	*T80*	*T90*
q1	$rpbis\ SCT^c$	0,31	0,31	0,32	0,33	0,34	0,35	0,39	0,46	0,54	0,60
	$\overline{rpbis\ SCT^i}$	-0,21	-0,21	-0,22	-0,22	-0,23	-0,24	-0,26	-0,30	-0,34	-0,36
q2	$rpbis\ SCT^c$	0,34	0,34	0,35	0,36	0,37	0,38	0,41	0,49	0,55	0,67
	$\overline{rpbis\ SCT^i}$	-0,17	-0,17	-0,17	-0,18	-0,18	-0,19	-0,21	-0,25	-0,27	-0,34
q3	$rpbis\ SCT^c$	0,26	0,26	0,27	0,28	0,29	0,30	0,33	0,39	0,51	0,67
	$\overline{rpbis\ SCT^i}$	-0,12	-0,12	-0,12	-0,13	-0,14	-0,14	-0,16	-0,19	-0,25	-0,34
q4	$rpbis\ SCT^c$	0,35	0,35	0,36	0,37	0,40	0,42	0,46	0,49	0,58	0,61
	$\overline{rpbis\ SCT^i}$	-0,18	-0,18	-0,19	-0,20	-0,21	-0,22	-0,24	-0,25	-0,30	-0,28
q5	$rpbis\ SCT^c$	0,31	0,31	0,32	0,32	0,33	0,35	0,39	0,46	0,58	0,76
	$\overline{rpbis\ SCT^i}$	-0,19	-0,19	-0,19	-0,20	-0,20	-0,22	-0,24	-0,29	-0,37	-0,44
q6	$rpbis\ SCT^c$	0,30	0,30	0,31	0,32	0,33	0,34	0,38	0,42	0,52	0,60
	$\overline{rpbis\ SCT^i}$	-0,18	-0,18	-0,19	-0,20	-0,20	-0,20	-0,23	-0,25	-0,30	-0,34
q7	$rpbis\ SCT^c$	0,27	0,27	0,27	0,28	0,28	0,28	0,32	0,36	0,45	0,60
	$\overline{rpbis\ SCT^i}$	-0,22	-0,22	-0,23	-0,23	-0,23	-0,24	-0,27	-0,30	-0,36	-0,40
q8	$rpbis\ SCT^c$	0,47	0,47	0,48	0,48	0,49	0,51	0,55	0,62	0,66	0,73
	$\overline{rpbis\ SCT^i}$	-0,23	-0,23	-0,23	-0,23	-0,24	-0,25	-0,26	-0,29	-0,31	-0,30
q9	$rpbis\ SCT^c$	0,29	0,29	0,30	0,31	0,32	0,35	0,38	0,42	0,47	0,57
	$\overline{rpbis\ SCT^i}$	-0,13	-0,13	-0,13	-0,14	-0,14	-0,15	-0,17	-0,19	-0,22	-0,30
q10	$rpbis\ SCT^c$	0,29	0,29	0,29	0,30	0,32	0,34	0,37	0,42	0,45	0,53
	$\overline{rpbis\ SCT^i}$	-0,18	-0,18	-0,19	-0,19	-0,20	-0,21	-0,23	-0,26	-0,29	-0,29

Nous avons mis en évidence dans ce tableau (fonds gris) les valeurs liées aux changements de trajectoire des *BSq* des 4ème et 8ème QCM du test. Nous verrons plus loin que ces deux questions ne posent pas de problèmes particuliers, ces deux très légères augmentations des $\overline{rpbis\ SCT^i}$ au palier T90 ne sont pas révélatrices d'anomalies dans ces deux QCM.

3. Calcul des Niveaux de Cohérence Spectrale des questions (*NCSq*) aux différents paliers de turbo analyse

Précédemment nous avons exposé une méthode de calcul des moyennes des *rpbis SCT* des propositions incorrectes (voir p.189) d'une QCM.

Les Niveaux de Cohérence Spectrale d'une question donnée (*NCSq*) peuvent être calculés de façon précise à chaque palier de turbo analyse en soustrayant la valeur de la moyenne pondérée des *rpbis SCT* des propositions incorrectes ($\overline{rpbis\ SCT^i}$ dans le tableau précédent) à la valeur du *rpbis SCT* de la réponse correcte ($rpbis\ SCT^c$ dans le tableau précédent).

La méthode de calcul du *N*iveau de *C*ohérence *S*pectrale d'une question *q* à un palier de *T*urbo analyse t ($NCSq\ T_t$) devient :

$$\boxed{NCSq\ T_t = rpbis\ SCT_t^c - \overline{rpbis\ SCT_t^i}} \quad (41)$$

avec

T_t = palier de Turbo analyse *t*

$rpbis\ SCT_t^c$ = valeur du *rpbis SCT* de la proposition correcte au palier t de *T*urbo analyse

$\overline{rpbis\ SCT_t^i}$ = moyenne des *rpbis SCT* des solutions *i*ncorrectes au palier t de *T*urbo analyse

et

$$\overline{rpbis\ SCT_t^i} = \frac{\sum_{j=1}^{nj} (rpbis\ SCT_j^i * nT_{t\,j}^i)}{NRT_t^i}$$

avec

j = indice des propositions incorrectes

nj = nombre de propositions incorrectes dans la question envisagée

$rpbis\ SCT_{t\,j}^i$ = la valeur du *rpbis SCT* d'une proposition incorrecte j au palier t de *T*urbo analyse

$nT_{t\,j}^i$ = nombre d'utilisations d'une proposition incorrecte j au palier t de *T*urbo analyse

NRT_t^i = *N*ombre total de *R*éponses incorrectes au palier t de *T*urbo analyse

La plage des valeurs possibles pour l'indice *NCSq* va de –2 à 2 :

- si $\overline{rpbis\ SCT^i} = 1$ et $rpbis\ SCT^c = -1$, alors $NCSq = (-1) - 1 = -2$;
- si $\overline{rpbis\ SCT^i} = -1$ et $rpbis\ SCT^c = 1$, alors $NCSq = 1 - (-1) = 2$.

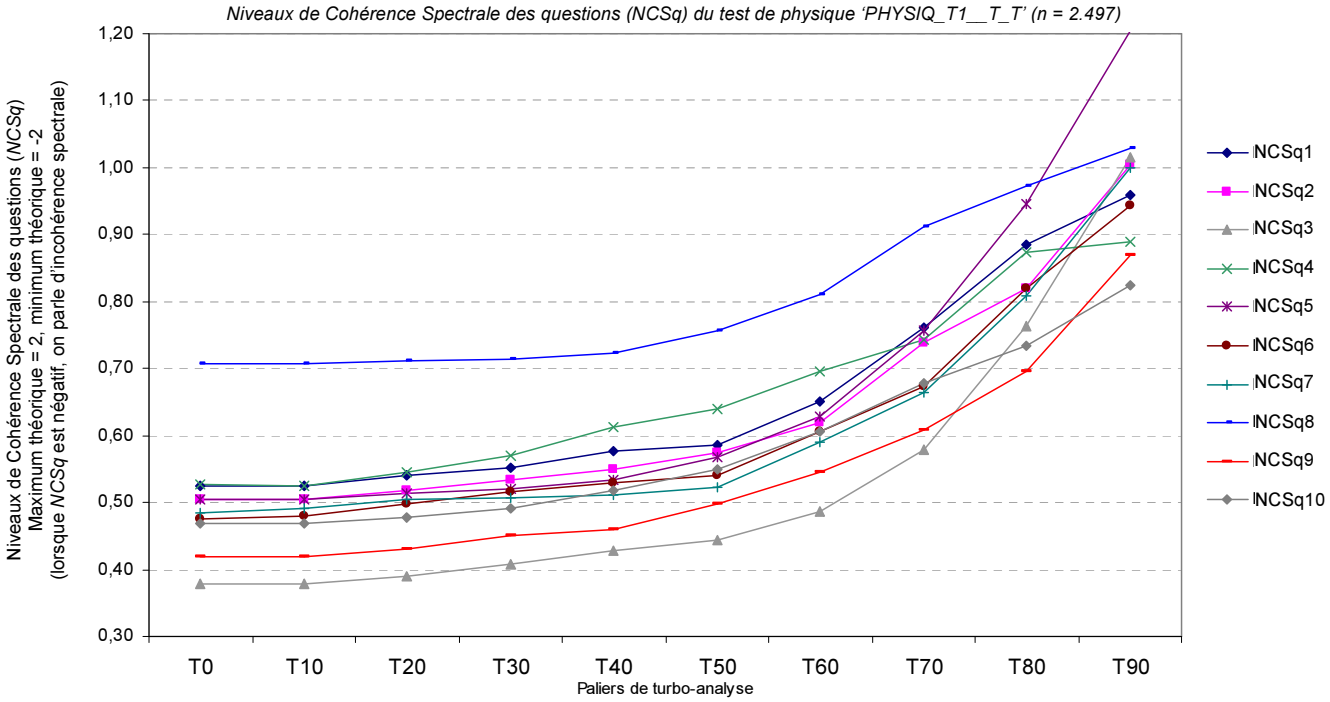
Dès lors, la cohérence spectrale maximale vaut 2 à l'indice *NCSq* et l'incohérence spectrale maximale vaut –2. Nous dirons qu'une question obtenant un *NCSq* entre 0 et 2 se situe dans une plage de cohérence spectrale. Lorsqu'une question récolte un *NCSq* entre –2 et 0 elle se situe dans une plage d'incohérence spectrale.

Donc, plus les valeurs de *NCSq* se rapprochent de 2 et plus elles indiquent un niveau de cohérence spectrale élevé. Plus elles tendent vers 0 et plus elles se rapprochent du niveau de cohérence spectrale minimal. Lorsque le *NCSq* est négatif, on parle alors d'incohérence spectrale.

Voici le tableau des valeurs *NCSq* récoltées par les 10 questions de l'épreuve de physique :

	T0	*T10*	*T20*	*T30*	*T40*	*T50*	*T60*	*T70*	*T80*	*T90*
NCSq1	0,52	0,53	0,54	0,55	0,58	0,59	0,65	0,76	0,89	0,96
NCSq2	0,51	0,51	0,52	0,53	0,55	0,57	0,62	0,74	0,82	1,00
NCSq3	0,38	0,38	0,39	0,41	0,43	0,44	0,49	0,58	0,76	1,02
NCSq4	0,53	0,53	0,55	0,57	0,61	0,64	0,70	0,74	0,87	0,89
NCSq5	0,51	0,51	0,51	0,52	0,53	0,57	0,63	0,76	0,95	1,20
NCSq6	0,48	0,48	0,50	0,51	0,53	0,54	0,61	0,67	0,82	0,94
NCSq7	0,49	0,49	0,51	0,51	0,51	0,52	0,59	0,67	0,81	1,00
NCSq8	0,71	0,71	0,71	0,72	0,72	0,76	0,81	0,91	0,97	1,03
NCSq9	0,42	0,42	0,43	0,45	0,46	0,50	0,54	0,61	0,70	0,87
NCSq10	0,47	0,47	0,48	0,49	0,52	0,55	0,61	0,68	0,73	0,82

Au palier de turbo analyse T90, la 5ème QCM (*NCSq5*) est celle qui obtient le niveau de cohérence spectrale le plus élevé (1,2). La 10ème (*NCSq10*) récolte le niveau le plus bas du test (0,82).



Les valeurs obtenues pour toutes les questions d'un test peuvent être représentées de façon graphique en plaçant en abscisse les paliers de la turbo analyse et en ordonnée les niveaux de qualité spectrale des questions atteints à chacun de ces paliers. Pour le test de physique on obtient le graphique ci-dessus (automatiquement généré par *SCANTEST 2.0*) :

Pour le test de physique, tous les *NCSq* sont positifs à tous les paliers de turbo analyse.

Les *NCSq* les plus élevés sont ceux de la 8ème QCM (*NCSq8*) sauf pour le dernier palier de turbo analyse T90 où la 5ème QCM (*NCSq5*) obtient un niveau de qualité spectrale de 1,20 contre 1,03 pour la 8ème QCM.

Remarquons que de T0 à T20 les valeurs obtenues sont relativement stables, ensuite, à partir de T30 et T40 les *NCSq* s'améliorent progressivement tout en gardant jusqu'à T60 *grosso modo* les tendances de départ. A partir de T70 les *NCSq* s'améliorent encore mais les classements sont bouleversés, par exemple la 3ème QCM qui jusqu'alors obtenait les moins bonnes valeurs se retrouve, au palier T90, propulsée de la 8ème à la 3ème place avec un *NCSq* de 1,02.

En ce qui concerne la 4ème QCM pour laquelle nous avions observé un *BSq* un peu différent des autres entre le seuil $R \geqslant 80$ et $R \geqslant 90$ de la turbo analyse (voir p. 230), nous observons ici aussi une courbe un peu différente dans la mesure où sa progression entre T80 et T90 est relativement plus faible.

4. Comparaison des Niveaux de Cohérence Spectrale des questions (*NCSq*) avec les Niveaux de Cohérence Interne des questions (*NCIq*)

La méthode de calcul des moyennes des rpbis des propositions incorrectes (voir p.189) peut aussi être appliquée aux *rpbis* classiques.

Pour chaque question nous pouvons alors soustraire au *rpbis* classique de la réponse correcte ($rpbis\ classique^{c}$) la moyenne pondérée des *rpbis* classiques des propositions incorrectes ($\overline{rpbis\ classique^{i}}$) pour obtenir un Niveau de Cohérence Interne de la question (*NCIq*).

$$NCIq = rpbis\ classique^{c} - \overline{rpbis\ classique^{i}} \quad (42)$$

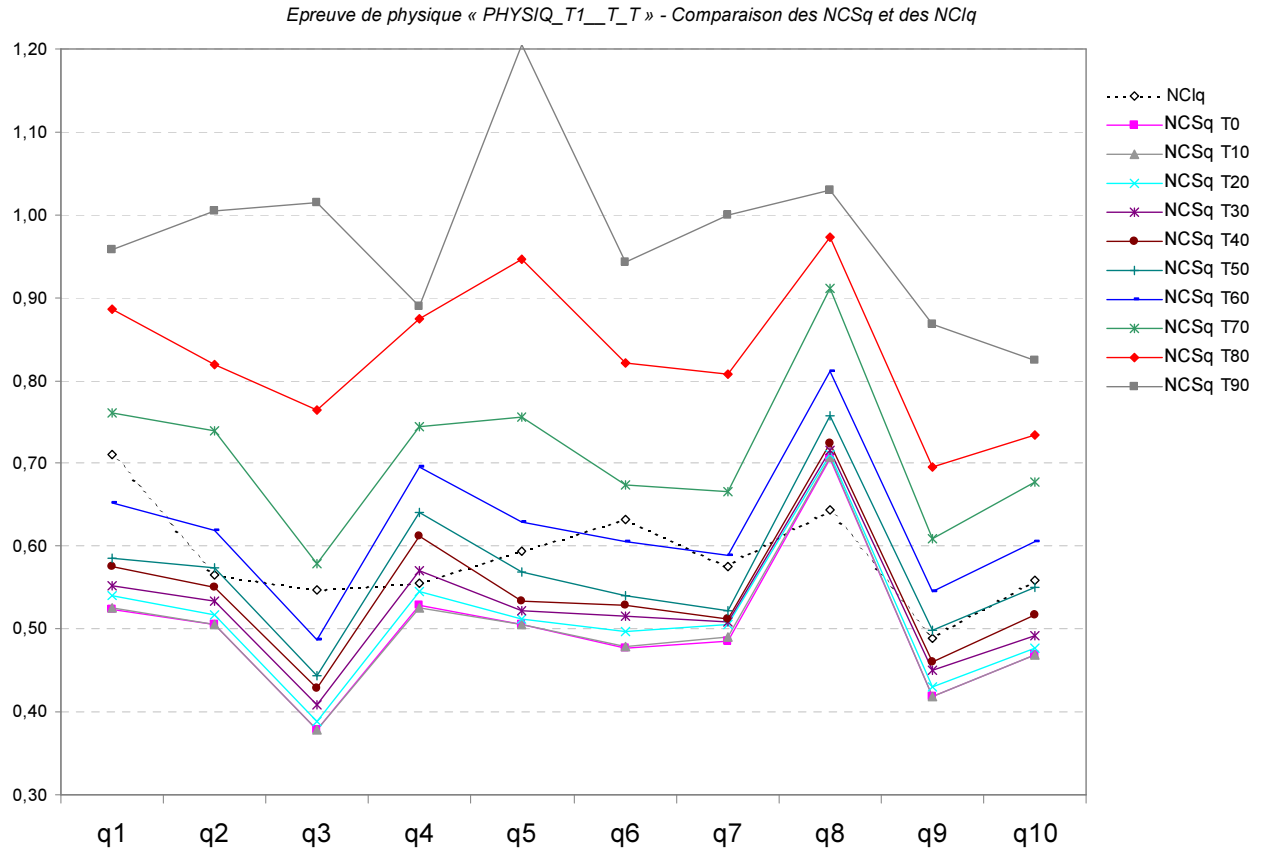
Comme pour le *NCSq*, la plage des valeurs possibles pour le *NCIq* va de –2 à 2 :

- si $\overline{rpbis\ classique^{i}} = 1$ et $rpbis\ classique^{c} = -1$, alors $NCIq = (-1) - 1 = -2$;
- si $\overline{rpbis\ classique^{i}} = -1$ et $rpbis\ classique^{c} = 1$, alors $NCIq = 1 - (-1) = 2$.

Nous obtenons les *NCIq* suivants pour les 10 QCM du test de physique :

	$rpbis\ classique^{c}$	$\overline{rpbis\ classique^{i}}$	NCIq
q1	0,515	-0,196	0,711
q2	0,395	-0,172	0,566
q3	0,392	-0,155	0,547
q4	0,371	-0,184	0,555
q5	0,448	-0,146	0,594
q6	0,455	-0,177	0,633
q7	0,428	-0,148	0,576
q8	0,466	-0,178	0,644
q9	0,342	-0,147	0,489
q10	0,344	-0,215	0,559

Les Niveaux de Cohérence Interne des questions (dernière colonne du tableau précédent) peuvent être comparés aux Niveaux de Cohérence Spectrale des questions (*NCSq*) à l'aide de ce graphique :



Dans le contexte des 10 questions du test de physique nous constatons qu'à partir du seuil $R \geqslant 70$ (*NCSq T70*) pour toutes les questions (q1 à q10) les *NCSq* calculés à partir des *rpbis SCT* sont supérieurs aux *NCIq* calculés à partir des *rpbis* classiques (courbe en pointillés). Rappelons que les traits reliant les valeurs ne sont là que pour guider l'œil du lecteur.

5. Constats et questions à propos des analyses des niveaux de qualité spectrale des 10 QCM de l'épreuve de physique

Comme nous l'avons fait dans la partie « Constats et questions à propos des analyses spectrales des propositions… » (voir p. 217), nous allons dresser une série de constats liés à l'analyse spectrale des niveaux de qualité des 10 QCM de l'épreuve de physique. Ces constats amènent des questions auxquelles nous ne pourrons répondre qu'après avoir analysé les protocoles de l'ensemble des QCM des épreuves MOHICAN.

Nous signalerons par un « ? » suivi d'un numéro d'ordre, le tout entre crochets, les premières questions que nous nous posons au départ des constats effectués à propos du test de physique. Ces questions seront reprises plus loin où, après analyse des données de l'ensemble des questions des 10 épreuves MOHICAN, nous tenterons de leur apporter des réponses.

Sur le graphique de la Gerbe Spectrale du test (*GSt*) de physique (voir p. 228), nous constatons que les 10 QCM représentés par les Brins Spectraux des questions (*BSq*) figurent dans la zone du 4ème quadrant (D).

[2.1] Les *BSq* des questions des autres épreuves MOHICAN se situent-ils aussi dans cette zone de qualité « D » (zone de cohérence spectrale) ?

Nous observons aussi sur l'agrandissement de la gerbe spectrale (voir p. 229) un regroupement des point à la base des brins.

[2.2] Tous les *BSq* des épreuves MOHICAN possèdent-ils cette caractéristique de regroupement des points aux paliers turbo T0 à T40 ? Comment peut-on expliquer de tels regroupements ?

6. En synthèse, ce que mesure *NCSq*, son intérêt

Rappelons que dans le cas d'une QCM qui fonctionne normalement du point de vue de l'utilisation des pourcentages de certitude, nous nous attendons à observer une tendance à fournir des certitudes plus élevées chez les étudiants qui choisissent la réponse correcte que chez les étudiants qui choisissent une solution incorrecte. Lorsque ce cas de figure se présente, nous parlons de « cohérence spectrale ».

Précédemment, nous avons vu que les *rpbis Spectraux Contrastés* (*rpbis SC*) permettaient de mesurer la cohérence spectrale au niveau des propositions d'une QCM (synthèse p. 191).

Pour chiffrer la cohérence spectrale au niveau d'une QCM, nous utilisons les valeurs récoltées par les *rpbis SC* des propositions de cette question, l'idée étant de soustraire au *rpbis SC* de la réponse correcte, la valeur de la moyenne pondérée des *rpbis SC* des propositions incorrectes. C'est le principe de calcul du *Niveau de Cohérence Spectral d'une question* (*NCSq*).

En nous référant aux trois conditions pour qu'un rpbis récolte la valeur « 1 » (voir p. 158), nous pouvons dire que la cohérence spectrale d'une question est maximale lorsque simultanément :

[1] les sujets qui ont choisi la solution correcte accompagnent tous cette réponse d'un même pourcentage de certitude ;

[2] les sujets qui n'ont pas choisi la solution correcte accompagnent tous leur choix d'une autre solution d'un même pourcentage de certitude ;

[3] le pourcentage de certitude choisi par les sujets qui répondent correctement est supérieur au pourcentage de certitude observé chez ceux qui répondent incorrectement.

Dans ce cas de figure, le *rpbis SC* de la réponse correcte est égal à 1 et la moyenne pondérée des *rpbis SC* des solutions incorrectes vaut –1. Dès lors, le *NCSq* de la question vaut 2.

A l'inverse, une situation d'incohérence spectrale maximale est théoriquement possible. Dans ce cas de figure, la condition [3] devient : « le pourcentage de certitude choisi par les sujets qui répondent correctement est inférieur au pourcentage de certitude observé chez ceux qui répondent incorrectement ». La question obtient alors pour la réponse correcte un *rpbis SC* dont la valeur est égale à –1 et pour la moyenne pondérée des *rpbis SC* des solutions incorrectes, une valeur égale à 1. Ce qui donne un *NCSq* qui vaut –2. Ces exemples extrêmes montrent que théoriquement, l'indice *NCSq* peut varier entre –2 et +2.

En pratique, pour les 173 questions des tests MOHICAN, aucune proposition récolte un *rpbis SC* égal à « 1 » ou à « -1 » (voir annexe F, p. 544) et les *NCSq* qui ont été calculés au départ des *rpbis SC* n'ont donc jamais atteint les valeurs extrêmes +2 et -2.

Lorsque nous avons abordé l'interprétation des *rpbis classiques* et présenté une série de graphiques montrant des simulations de données (p. 157) nous avons notamment montré, lorsque le rpbis est différent de 1 ou de –1, que l'écart entre Mx et Ma influence la valeur du *rpbis classique*. Il en va de même en ce qui concerne les *rpbis SC*, l'écart entre la moyenne des degrés de certitude des étudiants qui ont choisi la solution (Dx) et la moyenne des degrés de certitude des autres (Da) influence la valeur du *rpbis SC*. Lorsque la différence Dx – Da est positive, le *rpbis SC* est positif, plus l'écart est grand et plus le *rpbis SC* aura tendance à se rapprocher de 1. Dans le cas d'une différence Dx – Da dont le résultat est négatif, le *rpbis SC* est négatif, plus l'écart est grand, plus le *rpbis SC* se rapproche de -1.

Pour ce qui est des valeurs *NCSq* calculées dans le cadre des épreuves MOHICAN, nous parlerons de « cohérence spectrale dans les résultats des questions » lorsqu'elles varient entre 0 et 2 et « d'incohérence spectrale » lorsque les valeurs se situent entre 0 et –2.

Notons à propos de ce que mesure le *NCSq* qu'on pourrait aussi parler de l'évaluation de la « limpidité spectrale » dans les résultats d'une question, c'est-à-dire de la netteté de conscience de connaître pour ceux qui connaissent (en moyenne leurs certitudes devraient être plus élevées que celles de ceux qui se sont trompés) et de ne pas connaître pour ceux qui se trompent (en moyenne les certitudes de ces derniers devraient être moins élevées).

Enfin, soulignons qu'avec le nouvel indice *NCSq*, nous disposons d'un instrument d'évaluation de la cohérence spectrale qui nous permet d'épingler directement les QCM dont certaines propositions récoltent des valeurs anormales sans devoir passer en revue tous les *rpbis spectraux* de toutes les propositions de toutes les questions d'une épreuve.

B. Profils Spectraux des questions (PSq)

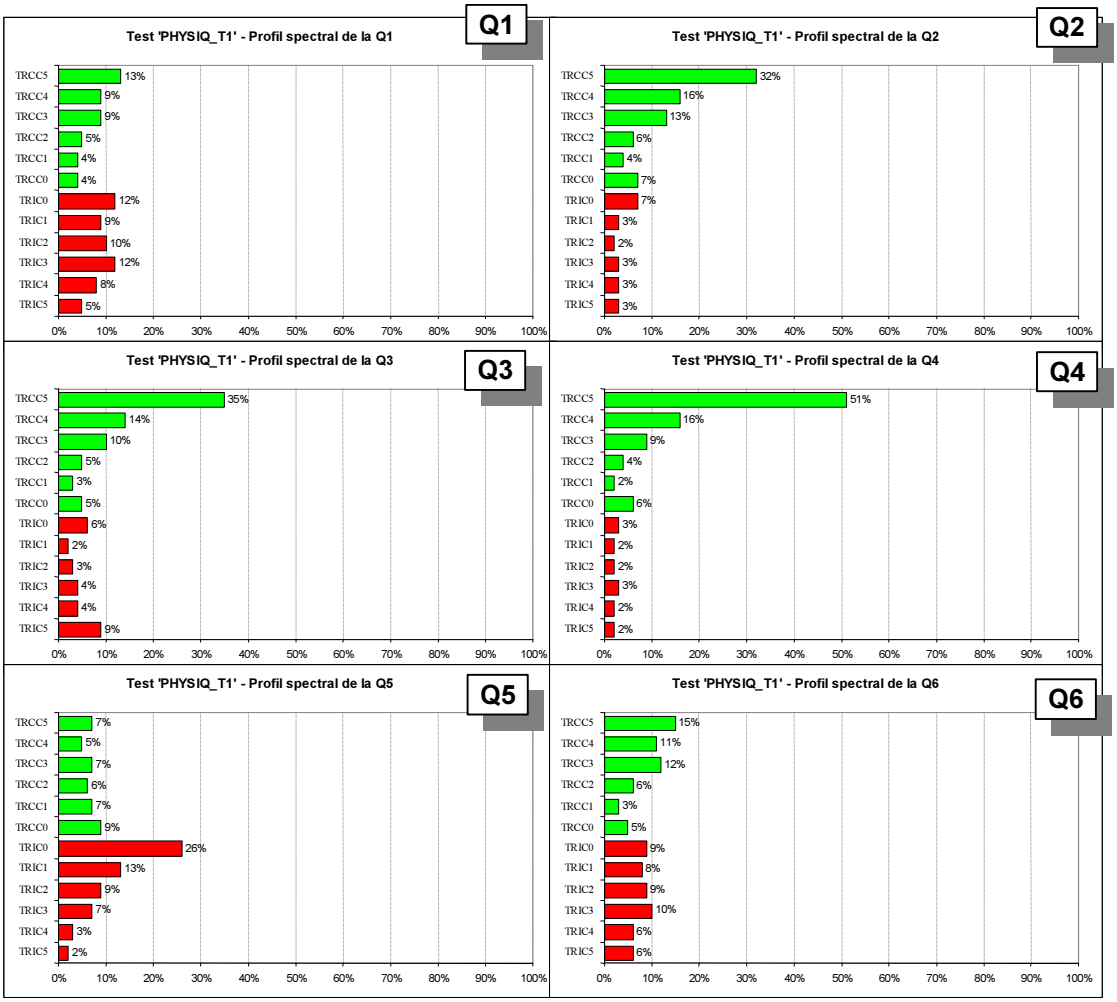
Nous allons maintenant présenter un deuxième type d'outil d'évaluation de la qualité spectrale des QCM, les Profils Spectraux par questions (*PSq*).

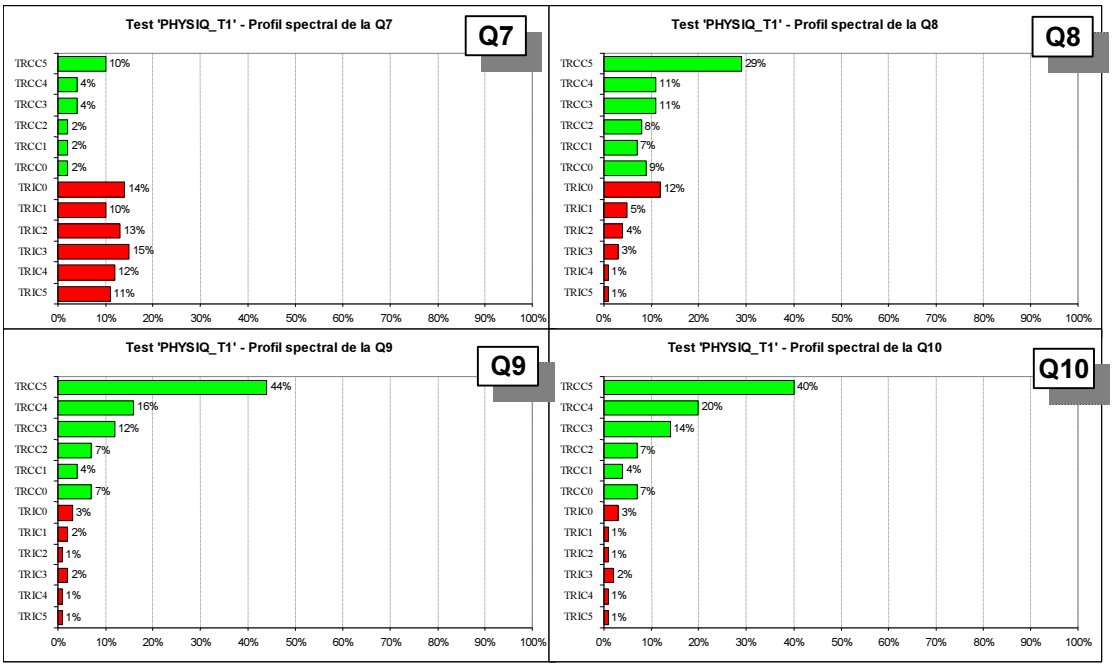
1. Principe d'élaboration des *PSq*

Le calcul des pourcentages de réponses correctes et incorrectes pour chaque degré de certitude permet de visualiser le Profil Spectral d'une question (*PSq*). Nous noterons les Taux de Réponses Correctes « *TRC* » et Incorrectes « *TRI* » et lui adjoindrons « *C0* » pour un pourcentage de certitude 0%, « *C1* » pour 20%, « *C2* » pour 40%, « *C3* » pour 60%, « *C4* » pour 80% et « *C5* » pour 100%.

La première version du logiciel *SCANTEST* permettait de visualiser l'hémi-spectre des pourcentages de réponses correctes (en vert) ainsi que l'hémi-spectre des pourcentages de réponses incorrectes (en rouge, pour toutes les réponses incorrectes confondues) accompagnant chaque degré de certitude, et ce, pour chaque question d'un test (Gilles, 1998a). Lors de cette recherche sur les apports des mesures métacognitives lors d'un test de compréhension d'un article scientifique, nous avons pu observer des profils spectraux remarquablement similaires après analyse spectrale des résultats de deux groupes d'étudiants (N_a=199 et N_b=125) soumis à un même test à une année académique d'intervalle. La corrélation moyenne pour les 16 paires de *PSq* du test était de 0,81.

Nous avons donc aussi introduit ces procédures dans la nouvelle version du logiciel *SCANTEST 2.0* pour les tests MOHICAN. Voici les profils obtenus pour les 10 questions du test de physique.





Les résultats ont été calculés au départ des données des 2.497 étudiants du test. Les représentations graphiques montrent que les profils spectraux par question peuvent être très différents d'un item à l'autre.

L'hémi-spectre des réponses correctes (en vert, sur le graphique « Q8 » ci-dessus) du *PSq* de la question 8 nous permet de visualiser un plus grand pourcentage (29%) de réponses correctes accompagnées du degré de certitude 5 (certitude 100%) suivi ensuite de pourcentages moins élevés pour les autres degrés de certitude. Sur l'hémi-spectre des réponses incorrectes de cette question (en rouge), nous observons un pourcentage plus élevé (12%) d'utilisation du degré de certitude 0 suivi en *decrescendo* par des utilisations de moins en moins fréquentes des autres degrés de certitude. Ce type de profil semble assez « sain » dans la mesure où les certitudes élevées ont été plus utilisés par les étudiants qui ont répondu correctement et pratiquement pas (seulement 1%) par des étudiants qui ont fourni une réponse incorrecte.

L'exemple du *PSq* de la question 7 montre une utilisation très différente des degrés de certitude dans l'hémi-spectre des réponses incorrectes : les pourcentages d'utilisation des six degrés de certitude se situent tous dans une fourchette comprise entre 10% et 15%, ce qui semble indiquer que la difficulté de cette question peu réussie (24% de réponses correctes) est mal perçue par certains étudiants, parmi ceux qui répondent incorrectement, on observe 14% de sujets qui accompagnent leur erreur d'une certitude 0 (ce qui témoigne d'une bonne auto-estimation) et 11% qui fournissent la certitude 5, c'est le pourcentage de prétentions de connaissances erronées associées à une question (réponses incorrectes et certitude maximum) le plus élevé du test.

Lors de travaux ultérieurs nous comptons introduire dans l'analyse de ces profils spectraux par question les coefficients d'asymétrie (*skewness*) et de voussure (*kurtosis*) calculés au départ de chaque hémi-spectre. Leclercq et Jans ont par ailleurs étudiés récemment ces coefficients dans le cadre d'analyses spectrales des performances des étudiants (Leclercq & Jans, 1999 ; Jans, 2000).

Voici le tableau des valeurs exprimées dans les graphiques *PSq* ci-avant.

	TRI C5	TRI C4	TRI C3	TRI C2	TRI C1	TRI C0	TRC C0	TRC C1	TRC C2	TRC C3	TRC C4	TRC C5
q1	0,05	0,08	0,12	0,1	0,09	0,12	0,04	0,04	0,05	0,09	0,09	0,13
q2	0,03	0,03	0,03	0,02	0,03	0,07	0,07	0,04	0,06	0,13	0,16	0,32
q3	0,09	0,04	0,04	0,03	0,02	0,06	0,05	0,03	0,05	0,1	0,14	0,35
q4	0,02	0,02	0,03	0,02	0,02	0,03	0,06	0,02	0,04	0,09	0,16	0,51
q5	0,02	0,03	0,07	0,09	0,13	0,26	0,09	0,07	0,06	0,07	0,05	0,07
q6	0,06	0,06	0,1	0,09	0,08	0,09	0,05	0,03	0,06	0,12	0,11	0,15
q7	0,11	0,12	0,15	0,13	0,1	0,14	0,02	0,02	0,02	0,04	0,04	0,1
q8	0,01	0,01	0,03	0,04	0,05	0,12	0,09	0,07	0,08	0,11	0,11	0,29
q9	0,01	0,01	0,02	0,01	0,02	0,03	0,07	0,04	0,07	0,12	0,16	0,44
q10	0,01	0,01	0,02	0,01	0,01	0,03	0,07	0,04	0,07	0,14	0,2	0,4

Dans le contexte d'une mise en évidence de la qualité spectrale des questions, il est important de pouvoir observer la façon dont les certitudes sont utilisées au sein des items. Ces observations sont à mettre en relation avec les valeurs obtenues par des indices spectraux tels que l'indice de Réalisation des prédictions par question (*Rq,* p. 242). D'un point de vue pratique, nous proposons ces informations en vue d'aider l'évaluateur à prendre des décisions concernant les questions à rectifier ou éventuellement à éliminer dans la version finale des résultats.

Signalons qu'il est possible de comparer les profils des questions à une typologie de Profils Spectraux par question. Voici les profils auxquels on peut rapprocher les 10 questions du test de physique :

Ceux qui réussissent (en vert) et ceux qui se trompent (en rouge) sont réalistes.
Profil observé pour la question 8.

Ceux qui réussissent ont tendance à être très sûrs, ceux qui se trompent n'en sont pas conscients.
Profil des questions 2 et 3.

Ceux qui réussissent n'en sont guère conscient et ceux qui se trompent ont tendance à être peu sûrs.
Profil de la question 5.

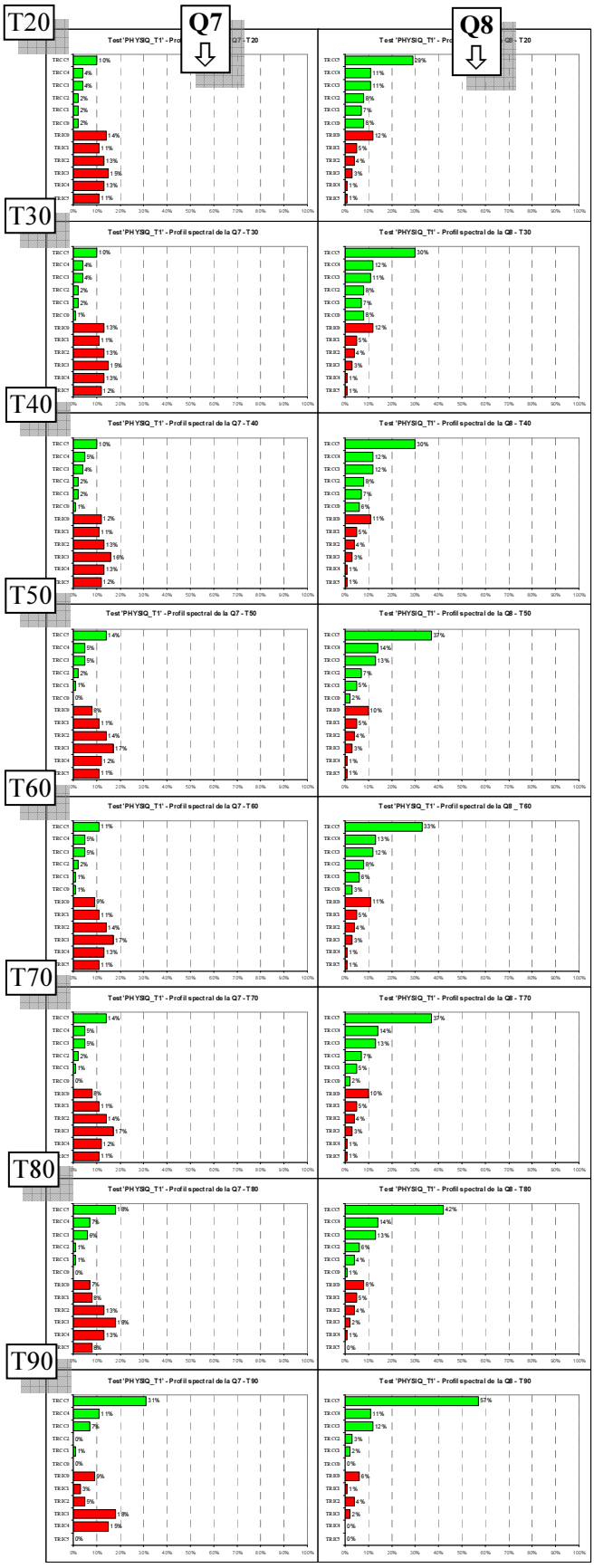
Question assez mal réussie et ceux qui se trompent n'en sont guère conscients.
Profil des questions 1 et 6.

Question très réussie et ceux qui se trompent n'en sont pas conscients.
Profil des questions 4, 9 et 10.

Question très mal réussie et ceux qui se trompent n'en sont pas conscients.
Profil de la question 7.

Le principe de la turbo analyse peut aussi être appliqué aux *PSq*, à des seuils de réalisme élevés ($Rs \geqslant 80$), elle renforce alors la validité des informations obtenues. C'est ce que nous allons envisager dans la section qui suit.

2. Turbo analyse appliquée aux Profils Spectraux par question



Le principe de la turbo analyse (voir p. 186) peut aussi être appliqué à la réalisation des Profils Spectraux par question (*PSq*).

Ci-contre figurent les *PSq* des 7ème et 8ème QCM du test de physique (n = 2497). Ces *PSq* ont été réalisés au départ des résultats filtrés en fonction du réalisme (*Rs*) des sujets calculé à partir des réponses fournies aux 10 questions du test. Rappelons que « T20 » signifie que ce sont les résultats des étudiants qui obtiennent un score de réalisme (*Rs*) plus grand ou égal à 20 qui ont été utilisés pour tracer les premiers profils spectraux des deux questions. Nous n'avons pas repris ici les graphiques obtenus aux paliers de turbo analyse « T0 » et « T10 » car ils sont quasi identiques à ceux qu'on observe ci-contre au palier « T20 ».

Tout comme les courbes des Niveaux de Cohérence Spectrale des questions (*NCSq*) montraient peu de différences aux paliers de turbo analyse T0 à T40 (voir p. 232), on observe pour les *PSq* des 7ème et 8ème QCM peu de changements à ces paliers.

A T50, les réponses correctes (en vert) accompagnées de la certitude la plus élevée (5) augmentent dans les deux profils tandis que les réponses incorrectes (en rouge) accompagnées du degré de certitude 0 diminuent. Rappelons qu'à ce niveau de turbo analyse les TRC et TRI sont calculés à partir des données de sujets dont $Rs \geqslant 50$, donc à l'aide des résultats des étudiants qui, par rapport aux paliers turbo précédents, utilisent en moyenne avec plus de pertinence les pourcentages de certitude.

A partir de T50 jusqu'à T80, pour la 7ème QCM, l'hémi-spectre des réponses incorrectes (en rouge) tend à prendre une forme en cloche. Remarquons que le pourcentage de réponses incorrectes accompagnées du degré de certitude 5 reste assez stable, or on s'attendrait à une diminution étant donné l'amélioration en *R*. Par contre les réponses correctes accompagnées de certitudes élevées augmentent. En ce qui concerne la 8ème QCM, on observe aussi une augmentation du pourcentage de réponses correctes avec la certitude 5. De plus, les réponses incorrectes avec la certitude 5 disparaissent.

A T90, les pourcentages de réponses correctes avec degrés de certitude élevés augmentent

encore. Pour la 7ème QCM, l'hémi-spectre des réponses incorrectes montre encore des pourcentages relativement élevés aux certitudes 3 (18%) et 4 (15%), mais une disparition totale des réponses incorrectes avec certitude 5. Pour la 8ème QCM, les réponses correctes avec la certitude 5 augmentent encore et il n'y a plus aucune réponse incorrecte avec la certitude 4.

Comparée à la 7ème QCM, la 8ème QCM montre à notre avis un meilleur profil dans la mesure où d'une part, pour les réponses correctes, plus les certitudes sont élevées plus les TRC sont élevés, et d'autre part pour les réponses incorrectes, plus les certitudes sont faibles et plus les TRI sont élevés.

L'intérêt des *PSq* avec turbo analyse est de nous permettre d'affiner notre diagnostic de la qualité spectrale des questions d'un test.

3. Les *PSq*, un nouveau champ de recherches docimologiques

Nous sommes conscient que l'étude des Profils Spectraux par question ouvre un nouvel espace de recherche en matière de qualité spectrale des questions.

Le but de cette section était d'entrouvrir cette porte sans pour autant poursuivre dans la direction d'une étude approfondie des *PSq* qui mériterait en soi qu'on y consacre une thèse entière.

Parmi les différentes pistes à suivre dans le cadre de nos prochains travaux, celle de l'élaboration d'une typologie des profils spectraux de questions nous paraît constituer une voie particulièrement prometteuse. Nous comptons notamment étudier les coefficients d'asymétrie (*skewness*) et de voussure (*kurtosis*) qui permettent de représenter les hémi-spectres et les comparer aux indices d'évaluation de la qualité spectrale des questions que nous proposons dans le cadre de notre thèse. Il serait par exemple intéressant de comparer les valeurs des *NCSq* aux valeurs observées aux indices de *skewness* et de *kurtosis*. Les *NCSq* constituant une sorte de résumé des *PSq*. En effet, comme nous l'avons déjà signalé (p. 235) le *NCSq* permet de chiffrer la limpidité spectrale des résultats d'une question, c'est-à-dire la netteté de conscience de connaître pour ceux qui connaissent (en moyenne leurs certitudes devraient être plus élevées que celles de ceux qui se sont trompés) et de ne pas connaître pour ceux qui se trompent (en moyenne les certitudes de ces derniers devraient être moins élevées), ce que les hémi-spectres des *PSq* permettent aussi de visualiser.

Terminons ce point en signalant à quel point nous sommes étonné par la fécondité de l'analyse spectrale. Il nous faut bien avouer que lorsque nous avons débuté cette thèse nous n'imaginions pas que l'utilisation des informations récoltées à l'aide des pourcentages de certitude pour fournir une série d'indications sur la qualité des questions ouvriraient autant de nouvelles perspectives de recherches… Non seulement une série d'indices spectraux sont possibles au niveau des propositions d'une question (le chapitre précédent avec les indices *rpbis S*, *rpbis SC* et *rpbis SCT*), mais également au niveau d'une QCM (ce chapitre avec les indices *NCSq*, *Psq*, *Rq*, *piq* et *Cq*) et au niveau des tests (le chapitre suivant avec les indices *NCSt*, *Rt*, *pit* et *Ct*). Ce foisonnement de possibilités d'instruments nouveaux pour l'évaluation de la qualité spectrale des épreuves nous a amené à prendre une série de décisions dans le cadre de cette thèse, dont celle de reporter à des travaux ultérieurs l'approfondissement des *PSq*.

C. Indice de Réalisation des prédictions par question (Rq)

Tout comme il est possible de calculer un indice de *R*éalisme des *s*ujets (*Rs*) (voir p. 184), il est possible d'évaluer comment les taux d'exactitude des réponses fournies à une question par l'ensemble du groupe s'ajustent aux probabilités des certitudes qui accompagnent ces réponses. Bien que le principe de calcul soit similaire, nous proposons d'éviter le terme « réalisme » parce que nous traitons ici d'une propriété qui concerne des QCM, objets qui ne peuvent pas faire preuve de réalisme au même titre que des personnes. Comme la formule de calcul est la même que celle de l'indice de *R*éalisme des *s*ujets (*Rs*) nous garderons la lettre « *R* » et lui ajouterons un « *q* » de manière à noter ce nouvel indice « *Rq* » et lui donner la signification « *Réalisation des prédictions par question »* (*Rq*).

1. Procédure de calcul

Nous l'avons signalé ci-dessus, la méthode de calcul correspond à celle employée pour l'indice de *R*éalisme des *s*ujets (*Rs*) utilisé pour l'évaluation des performances d'auto-estimation des étudiants. Elle s'en différencie cependant par la prise en compte des certitudes récoltées non plus pour chaque étudiant à toutes les questions mais par chaque question à tous les étudiants. Il s'agit donc ici de tirer parti des informations livrées par les sujets à propos de leur réalisme pour établir un indice qui concerne l'évaluation d'une qualité de la question : sa propension à récolter ou non des prédictions ou Taux d'Exactitude Annoncés (TEA) en concordance avec la réalité des Taux d'Exactitude Observés (TEO). La formule de l'indice de *R*éalisation des prédictions par *q*uestion (*Rq*) est similaire à celle de l'indice de réalisme des sujets :

$$\boxed{Rq = 100 - EMACq} \quad (43)$$

Avec

$EMACq = \frac{\sum_i (|\, C_i - TE_i \,|).NU_i}{NRq}$ = l'*E*rreur *M*oyenne *A*bsolue de *C*ertitude calculée pour la *q*uestion

i = *i*ndice des degrés de certitude (0, 1, 2, 3, 4 et 5)

C_i = Valeur de la *C*ertitude *i* (en pourcents) ou Taux d'Exactitude Annoncé (TEA)

NC_i = *N*ombre de réponses *C*orrectes pour la certitude *i*

NU_i = *N*ombre d'*U*tilisations de la certitude *i* (*si* $NU_i = 0$ alors l'indice *i* est ignoré)

TE_i = *T*aux d'*E*xactitude des réponses fournies avec la certitude *i* (en pourcents) ou Taux d'Exactitude Observé (TEO) = 100 x NC_i / NU_i

NRq = *N*ombre total de *R*éponses à la *q*uestion ($\Sigma_i NU_i$)

NC = *N*ombre total de réponses *C*orrectes ($\Sigma_i NC_i$)

Voici un exemple d'application de la procédure à la 1ère question du test de physique :

Degrés de Certitude	0	1	2	3	4	5	
C_i	0	20	40	60	80	100	Σ
NU_i	385	329	370	526	431	456	2497
NC_i	97	101	116	222	235	331	1102
TE_i	25	31	31	42	55	73	
$\|C_i - TE_i\|$	25	11	9	18	25	27	
$(\|\, C_i - TE_i \,|).\, NU_i$	9625	3619	3330	9468	10775	12312	49129

Nous commençons par compter le nombre d'utilisations (NU_i) et le nombre de réponses correctes pour chaque degré de certitude de la question considérée. Nous calculons ensuite les taux d'exactitude pour chaque certitude (TE_i) en divisant le nombre de réponses correctes par les utilisations (NU_i). Ensuite nous calculons les écarts ($|C_i - TE_i|$) pour chaque degré de certitude entre les taux d'exactitude et les

probabilités subjectives de réussite associées à chaque certitude (C_i). Après quoi nous pondérons les écarts par le nombre d'utilisations des degrés de certitude pour obtenir la pondération des écarts par certitude (($|C_i - TE_i|$).NU_i).

L'étape suivante de la procédure de calcul consiste à calculer la moyenne des écarts des certitudes par question :

$$EMACq = \frac{49129}{2497} = 19{,}7$$

Ce qui donne 49129 / 2497 = 19,7, ce qui signifie qu'en moyenne pour cette question du test les écarts entre les prédictions et la réalité des taux d'exactitude sont de 19,7%.

A partir de l'indice de moyenne des écarts de certitude par question (*EMACq*) il est possible de calculer un indice de concordance des prédictions par question en le soustrayant à la valeur 100 :

$$Rq = 100 - EMACq$$

Ce qui donne pour l'indice *Rq* de la 1ère question (nous dirons le *Rq1*) du test de physique (n = 2.497) : *Rq1* = 100 – 19,7 = 80,3.

Dans l'idéal, si les écarts entre les taux d'exactitude observés et les taux d'exactitude annoncés (les prédictions en %) étaient nuls pour chacun des degrés de certitude employés par les sujets pour répondre à la question, l'indice *Rq* vaudrait 100.

Dans la pire des situations, si les écarts entre les taux d'exactitude observés et les taux d'exactitude annoncés étaient égaux à 100 (écart maximum), ce qui signifierait que les sujets utilisent systématiquement soit la certitude 0 pour accompagner une réponse correcte soit la certitude 100 pour une réponse incorrecte (les autres pourcentages de certitude n'étant pas utilisés), alors l'indice *Rq* vaudrait 0.

2. Représentation graphique des *Rq*

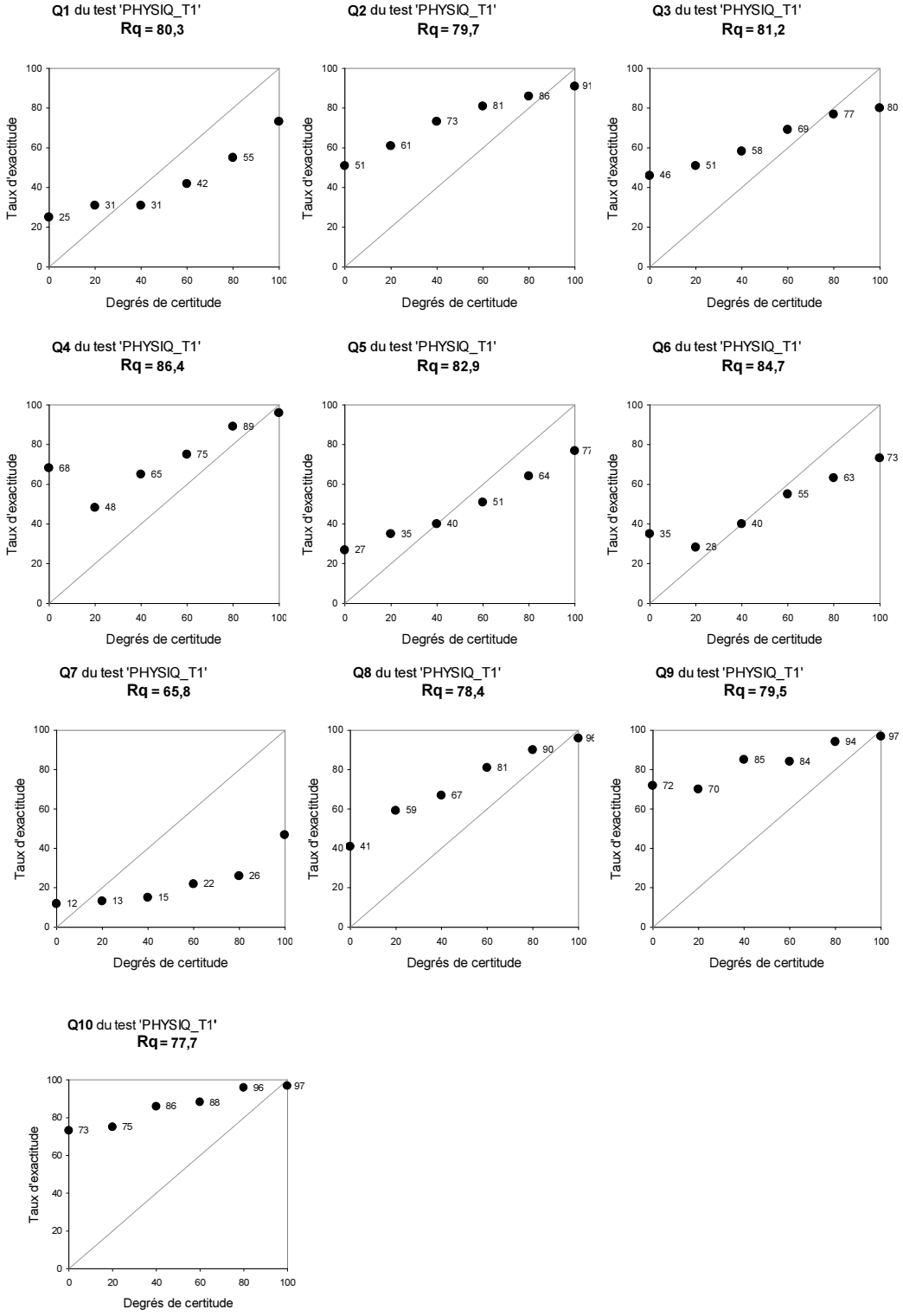
A l'aide de graphiques où nous portons d'une part sur l'axe des abscisses les probabilités subjectives de réussite associées à chaque degré de certitude 0, 20, 40, 60, 80 et 100 (ici exprimés par leurs pourcentages) et d'autre part sur l'axe des ordonnées les taux d'exactitude associés à chaque degré de certitude, nous pouvons pour chaque question visualiser à quel point les taux d'exactitude concordent avec les degrés de certitude.

La diagonale qui figure sur chacun des graphiques constitue le repère d'une situation idéale où les taux d'exactitude observés correspondent aux pourcentages de certitude annoncés par les sujets. Plus les points sont proches de cette diagonale, meilleures sont les concordances entre les prédictions (degrés de certitude) et les réalisations de celles-ci (les taux d'exactitude).

Nous avions introduit la création automatisée de ce type de graphique dans la 1ère version de *SCANTEST* (Gilles, 1998a) et nous l'avons reprogrammée dans le module « Profils Spectraux des questions (*PSq*) & Indices de Réalisation des prédictions par question (*Rq*) » du logiciel *SCANTEST 2.0* pour les tests MOHICAN (voir p. 205).

Les dix graphiques ci-dessous ont donc été générés automatiquement par *SCANTEST 2.0.* Ils reprennent pour chaque question, ici les 10 QCM du test de physique, les taux d'exactitude correspondant à chaque degré de certitude. Les résultats ont été calculés au départ des données des 2.497 étudiants ayant participés au test.

Ces graphiques nous permettent d'observer chez les 2.497 étudiants testés une tendance à la sous-estimation pour six questions : 2, 3, 4, 8, 9 et 10. Pratiquement tous les points sont alors alignés au-dessus du trait oblique (sauf pour le degré 100), les probabilités prédites étant inférieures aux taux d'exactitudes récoltés.



Les questions 5 et 6 se caractérisent par une sous-estimation lorsque les Degrés de Certitude (DC) faibles sont utilisés (0% et 20%) et une surestimation dans le cas des DC élevés (60%, 80% et 100%). Pour le DC 40 on remarque une concordance parfaite des prédictions par rapport aux taux d'exactitude observés.

Seule la question 7 dont l'indice *Rq* est le plus faible (65,8) est l'objet d'une surestimation pour 5 certitudes sur 6 (20, 40, 60, 80 et 100) et on peut se demander si il ne s'agit pas là de la manifestation d'un « effet surprise », le groupe témoignant par ses DC plus élevés que ses taux d'exactitude qu'il ne s'attendait pas à ce que la réponse correcte (P2) de cette QCM soit finalement celle-là (précédemment, p. 122, nous avons vu que la 7ème QCM est aussi la plus difficile du test).

Pour les 10 QCM de ce test de physique (n = 2.497) nous nous retrouvons donc face à trois types de graphiques :

- les tendances à la sous-estimation (6 QCM : 2, 3, 4, 8, 9 et 10) ;
- les tendances à la sous-estimation avec les DC faibles et à la surestimation avec les DC élevés (2 QCM : 5 et 6) ;
- une tendance à la surestimation quasi systématique (1 QCM : 7).

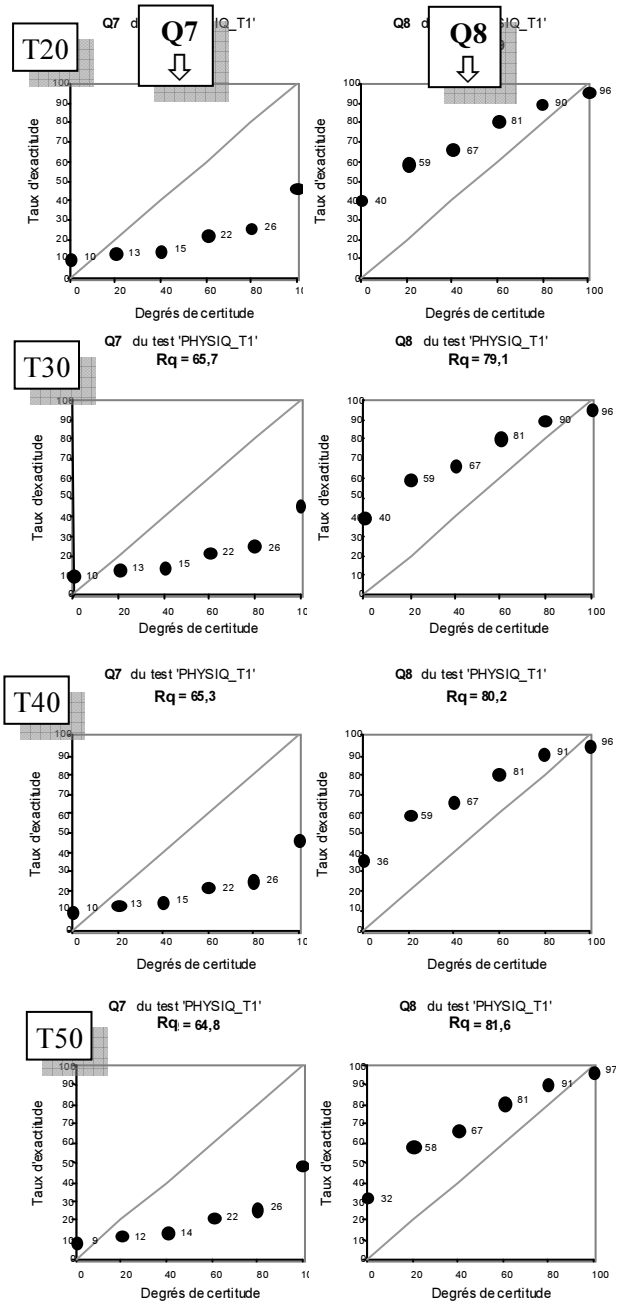
Remarquons que nous n'observons pas ici de question où plus les DC sont élevés, plus les taux d'exactitude sont faibles. En effet, la tendance générale est plutôt à une augmentation des taux d'exactitude lorsque les pourcentages de certitude sont plus élevés, ce qui montre une cohérence d'utilisation des degrés de certitude par les sujets, vérifiée question par question, les seules exceptions concernant le DC 0% des QCM 4, 6 et 9.

3. Application de la turbo analyse au calcul des *Rq*

a) Evolution de l'indice *Rq* aux paliers de turbo analyse : les cas des QCM 7 et 8 du test de physique (*n* = 2.497)

Le principe de la turbo analyse (p. 186) peut aussi être appliqué au calcul des indices de Réalisation des prédictions par question (*Rq*). Voici ce que donne la comparaison des graphiques *Rq* des questions 7 et 8 du test de physique (*n* = 2.497) à huit paliers de turbo analyse, de T20 à T90 avec un pas de 10. Nous n'avons pas repris ici les graphiques obtenus aux paliers de turbo analyse T0 et T10 car ils sont quasi identiques à ceux qu'on observe ci-contre au palier T20.

On constate peu de changements entre les paliers T20 et T80 tandis qu'au palier T90 les graphiques montrent de plus grosses différences.



Les valeurs des *Rq* aux huit paliers de turbo analyse sont reprises dans le tableau ci-dessous dans les colonnes « *Rq7* » et « *Rq8* ». La colonne « *n* » reprend les effectifs pour chaque palier. Le score de réalisme exigé par la turbo analyse étant de plus en plus élevé au fur et à mesure que l'on monte dans les paliers, ces effectifs diminuent systématiquement. La colonne « ≠ » contient les différences entre l'effectif d'un palier de turbo analyse et l'effectif du palier précédent, la dernière colonne « ≠ en % » nous donne l'équivalent en pourcentage du *n*.

Paliers Turbo	*Rq7*	*Rq8*	*n*	≠	≠en%
T20	65,8	78,9	2472		
T30	65,7	79,1	2457	15	-1%
T40	65,3	80,2	2392	65	-3%
T50	64,8	81,6	2309	83	-3%
T60	65,0	83,7	2097	212	-8%
T70	65,8	86,7	1669	428	-17%
T80	68,6	89,0	946	723	-29%
T90	80,6	92,8	235	711	-28%

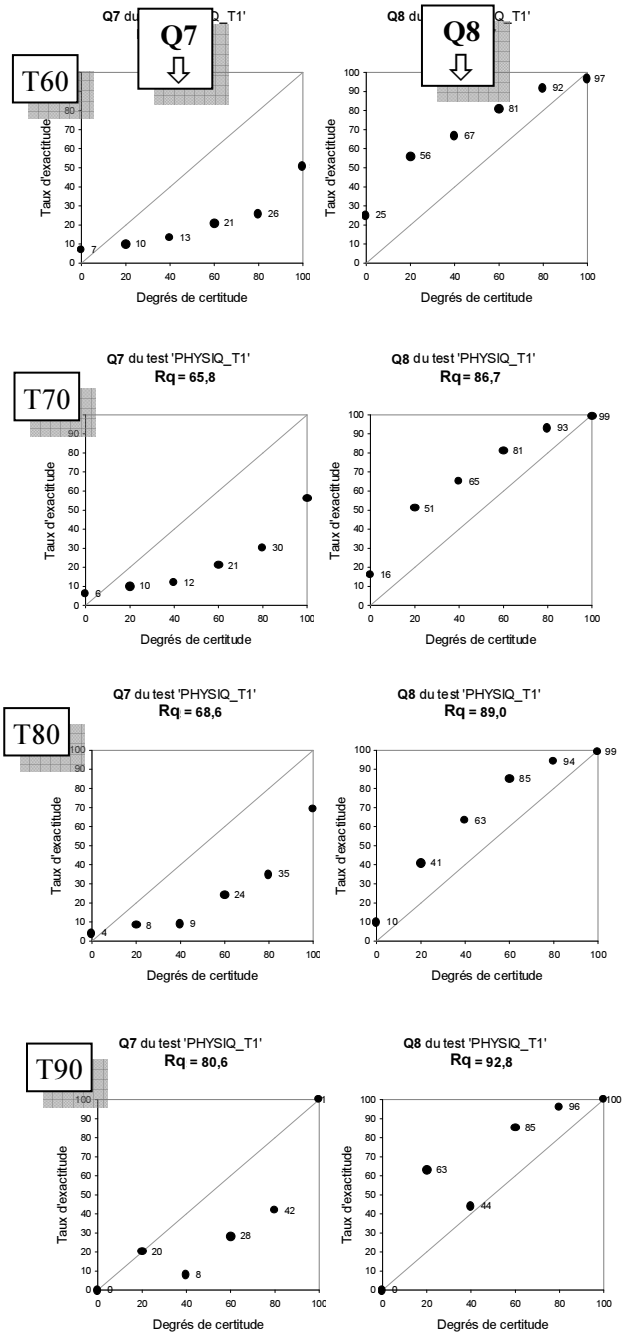
Plus on monte dans les paliers de turbo analyse, plus les *Rq* augmentent ce qui est logique dans la mesure où ces indices sont calculés à partir des données d'étudiants dont le score de réalisme est en moyenne de plus en plus élevé, donc dont les taux d'exactitude sont de mieux en mieux ajustés à leurs prédictions.

Pour la question 7, l'indice *Rq7* reste stable, voire diminue très légèrement du palier 20 au palier 70.

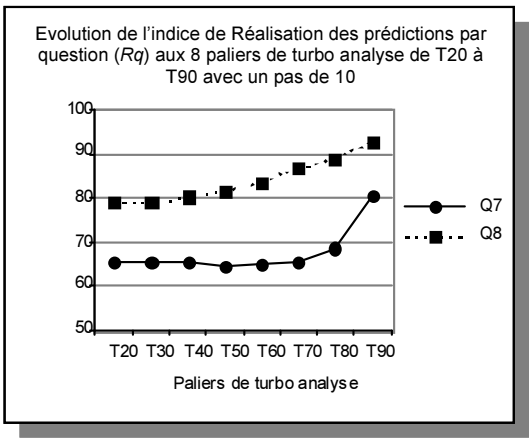
Pour la question 8, l'indice *Rq8* évolue peu du palier 20 à 30 et commence à augmenter sensiblement à partir du palier 40.

Il s'agit donc de deux évolutions assez différentes de l'indice *Rq*, l'une, pour la

7ème question, étant relativement continue (à partir de T40) et l'autre, pour la 8ème question, ne commençant qu'à partir du palier T80.



Le graphique ci-dessous montre ces deux évolutions différentes des *Rq* aux paliers de turbo analyse T20 à T90. On remarque un infléchissement de la courbe de la 8ème QCM à partir de T80.



Le tableau ci-dessous reprend les effectifs par degré de certitude :

		C0	C20	C40	C60	C80	C100
T20	**QCM 7**	376	300	368	476	420	532
	QCM 8	495	281	279	346	314	757
T30	**QCM 7**	362	300	368	476	420	531
	QCM 8	481	281	279	346	314	756
T40	**QCM 7**	307	298	367	475	417	528
	QCM 8	422	280	279	345	313	753
T50	**QCM 7**	261	289	362	471	411	515
	QCM 8	368	268	276	342	310	745
T60	**QCM 7**	208	262	334	445	378	470
	QCM 8	297	230	245	320	292	713
T70	**QCM 7**	136	200	264	369	294	406
	QCM 8	200	155	179	268	243	624
T80	**QCM 7**	68	86	137	220	191	244
	QCM 8	79	83	94	142	143	405
T90	**QCM 7**	21	10	12	58	60	74
	QCM 8	15	8	16	34	27	135

Jusqu'au palier de turbo analyse T80, le tableau montre pour les deux QCM des effectifs assez élevés pour toutes les certitudes. Mais à partir de T90, nous observons des effectifs relativement faibles pour les certitudes 0%, 20% et 40%.

Nous observons sur les graphiques que toutes les courbes sont relativement lisses sauf celles qui correspondent à T90 et nous remarquons qu'à ce palier de turbo analyse les effectifs sont faibles. Moins il y a de sujets plus l'erreur statistique est grande et nous devrons donc signaler les erreurs de mesure sur ce type de graphique. Dans l'idéal et en particulier pour les graphiques à T90 où il y a peu d'étudiants, il faudrait que chaque point soit affecté d'une barre d'erreur de part et d'autre du point.

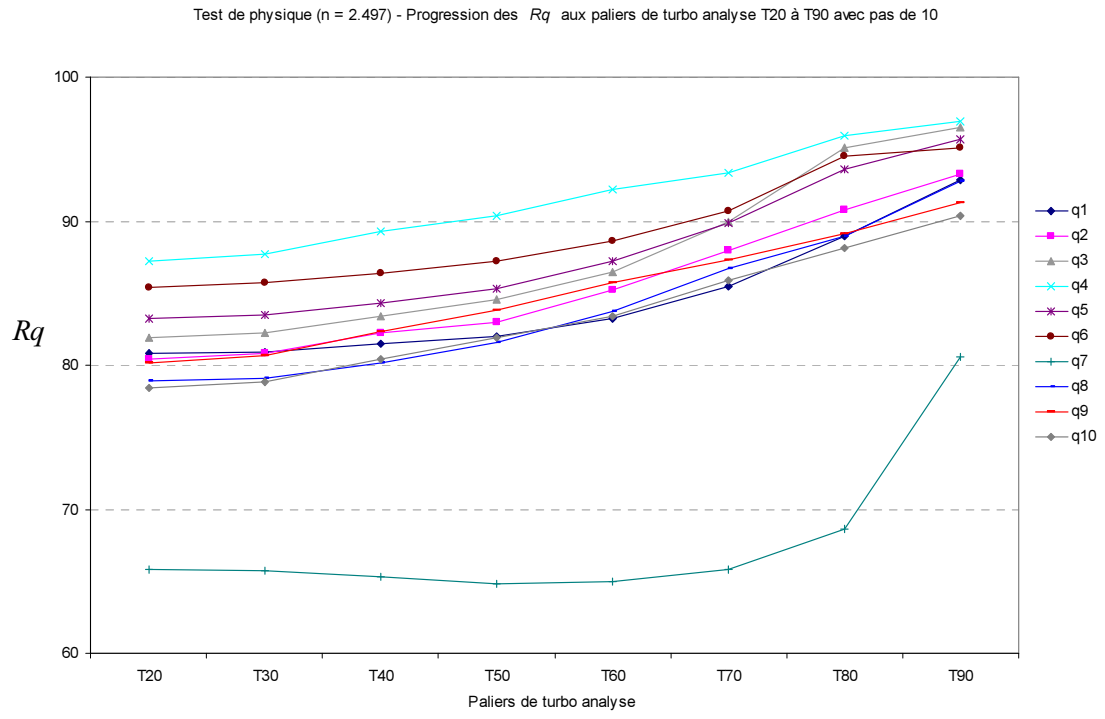
A T90 le taux d'exactitude du degré de certitude 20% de la 8ème question vaut 63% (5 réponses correctes sur 8). Bien qu'à ce palier de turbo analyse les sujets soient de plus en plus réalistes, l'erreur statistique liée à ce taux d'exactitude calculé sur la base de 8 réponses seulement est de 38%, ce qui est énorme. Dans le contexte de ces deux QCM, il nous paraît raisonnable de nous limiter au palier T80 vu le manque d'effectifs au palier T90 dans les cellules des pourcentages de certitude 0%, 20%, 40% de la 7ème question et 0%, 20%, 40%, 60%, 80% de la 8ème question.

b) Valeurs obtenues par les 10 QCM du test de physique (*n* = 2.497) à l'indice *Rq*

Voici le tableau des valeurs obtenues aux indices de Réalisation des prédictions par question (*Rq*) du test de physique (*n* = 2.497 étudiants) pour huit paliers de turbo analyse (de *T20* à *T90* avec un pas de 10).

	T20	*T30*	*T40*	*T50*	*T60*	*T70*	*T80*	*T90*
Rq1	80,8	80,9	81,5	82	83,2	85,5	89	92,9
Rq2	80,4	80,8	82,2	83	85,2	88	90,8	93,3
Rq3	81,9	82,2	83,4	84,6	86,5	90	95,1	96,5
Rq4	87,2	87,7	89,3	90,4	92,2	93,4	95,9	96,9
Rq5	83,2	83,5	84,3	85,3	87,2	89,9	93,6	95,7
Rq6	85,4	85,7	86,4	87,2	88,6	90,7	94,5	95,1
Rq7	65,8	65,7	65,3	64,8	65	65,8	68,6	80,6
Rq8	78,9	79,1	80,2	81,6	83,7	86,7	89	92,8
Rq9	80,2	80,7	82,3	83,8	85,7	87,3	89,1	91,3
Rq10	78,4	78,8	80,4	81,9	83,4	85,9	88,1	90,4

Le graphique ci-dessous montre la progression des valeurs des *rq* reprises dans le tableau ci-dessus. Pour des raisons liées à la lisibilité du graphique nous avons débuté l'échelle de l'axe des ordonnées à 60 bien que la valeur minimum possible pour l'indice *Rq* puisse être égale à 0 (rappelons que le maximum = 100).



La progression continue de 9 QCM sur 10 est logique dans la mesure où à chaque augmentation du palier de turbo analyse on calcule les *Rq* à partir de données d'étudiants en moyenne plus réalistes qu'au palier précédent, donc dont les prédictions se réalisent de mieux en mieux.

A la différence des autres questions, la courbe de la 7ème QCM ne progresse pas de façon continue à tous les paliers de la turbo analyse. Elle se détache des autres dès le palier T20 et reste assez stable jusqu'au palier T60. A T70 les valeurs de *Rq7* commencent seulement à s'élever. Cependant, même à T90, la valeur de *Rq7* (80,4) reste à environ 10 points en dessous de la moins bonne valeur *Rq* des 9 autres questions, celle de la 10ème QCM (90,4).

4. Constats et questions à propos des *Rq*

Lorsque nous comparons le *Rq* avec le *NCSq* (p. 231) de la 7ème QCM nous constatons que cette question obtient des valeurs assez opposées. Ces différences observées aux indices *Rq* et *NCSq* pourraient améliorer la fiabilité de détection des questions problématiques par l'analyse spectrale. En effet, si l'un des deux indices spectraux « laissait passer » un item défectueux, l'autre pourrait le signaler. Inversement, une « fausse alerte » déclenchée par un indice spectral pourrait être relativisée si l'autre indice spectral récolte une valeur normale.

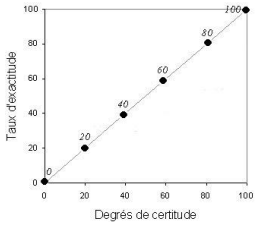
[2.3] Ces performances opposées des indices *Rq* et *NCSq* s'observent-elles pour d'autres QCM des épreuves MOHICAN ?

5. En synthèse, ce que mesure *Rq*, ses limites, son intérêt

Rappelons le principe de calcul de l'indice de Réalisation des prédictions par question (*Rq*). Il s'agit d'évaluer au niveau des résultats d'une question, dans quelle mesure les pourcentages de certitude sont proches des pourcentages de réponses correctes qui leurs sont associés. Ce qui nous amène à calculer une erreur en valeur absolue pour chaque pourcentage de certitude pondérée par le nombre d'utilisations de cette certitude. On effectue ensuite la moyenne de ces erreurs ce qui donne un indice d'Erreur Moyenne Absolue de Certitude calculée pour la question (*EMACq*). Enfin, le *Rq* de la question est obtenu en soustrayant 100 à la valeur d'*EMACq* (*Rq* = 100 – *EMACq*).

Pour que *Rq* soit maximal (égal à 100), l'*EMACq* doit être minimale (égale à 0). Ce qui implique qu'au niveau des résultats d'une question on observe :

- pour la certitude 0%, 0% d'étudiants qui fournissent la réponse correcte ;
- pour la certitude 20%, 20% d'étudiants qui fournissent la réponse correcte ;
- pour la certitude 40%, 40% d'étudiants qui fournissent la réponse correcte ;
- pour la certitude 60%, 60% d'étudiants qui fournissent la réponse correcte ;
- pour la certitude 80%, 80% d'étudiants qui fournissent la réponse correcte ;
- pour la certitude 100%, 100% d'étudiants qui fournissent la réponse correcte .

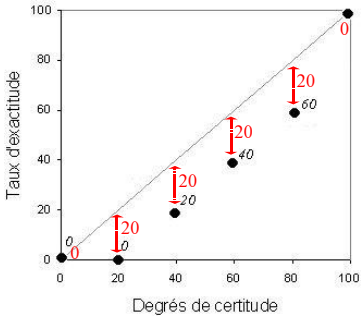


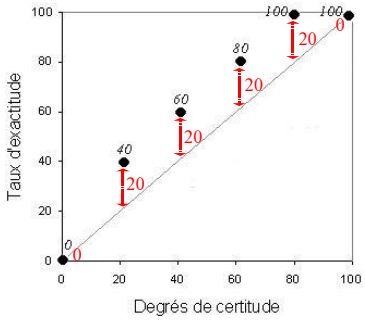
Dans ce cas de figure idéal où *Rq* vaut 100, il n'y a pas de « surprise » dans les résultats de la question, les pronostics de réussites livrés par les pourcentages de certitude « collent » aux taux d'exactitude observés pour chaque certitude.

L'indice *Rq* fournit donc au niveau des résultats d'une question, une information globale sur le niveau de réalisation des prédictions liées aux différents pourcentages de certitude.

Une limitation de *Rq* réside dans le fait que l'indice ne permet pas de nous informer sur la tendance à la sur ou sous-estimation contenue dans les résultats de la question (nous verrons plus loin qu'un autre indice, *Cq*, fournit cette information). Cette limitation est liée au fait que les erreurs de certitude sont calculées en valeurs absolues.

Sur les deux figures ci-dessous les « quantités » d'erreurs absolues de certitude sont identiques (elles sont indiquées en rouge). Par conséquent, dans les deux cas, *EMACq* vaut 13,3 et *Rq* vaut 86,7. Cependant les situations illustrées sont différentes.



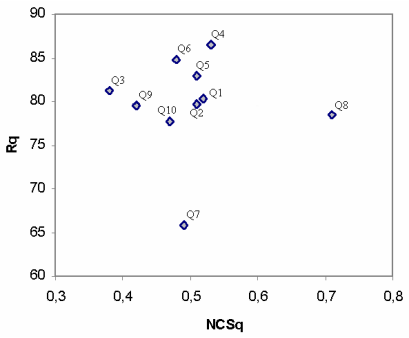


En ce qui concerne le graphique de gauche, les résultats de la question contiennent une tendance à la surestimation. Celle-ci est due aux taux d'exactitude systématiquement inférieurs pour les pourcentages de certitude 20%, 40%, 60% et 80%.

Le graphique de droite montre quant à lui une tendance à la sous-estimation dans les résulats, liée elle aux taux d'exactitude systématiquement supérieurs pour les pourcentages de certitude 20%, 40%, 60% et 80%.

Nous avons vu précédemment qu'un autre indice, le *NCSq* (calculé à partir des *rpbis spectraux*), permet lui d'évaluer dans quelle mesure on observe une tendance à fournir des certitudes plus élevées chez les étudiants qui ont choisi la réponse correcte, et, en parallèle, une tendance à fournir des certitudes plus faibles chez ceux qui ont opté pour un distracteur. Dans le cas du *NCSq* on peut aussi parler d'une mesure de la « limpidité spectrale » des résultats de la question, c'est-à-dire la netteté de conscience de connaître pour ceux qui connaissent (en moyenne leurs certitudes devraient être plus élevées que celles de ceux qui se sont trompés) et de ne pas connaître pour ceux qui se trompent (en moyenne les certitudes de ces derniers devraient être moins élevées).

Dans la mesure où il offre une autre information liée au niveau d'ajustement des prédictions aux taux d'exactitude pour les six pourcentages de certitude, l'indice *Rq* apporte une information complémentaire à celle offerte par le *NCSq*. Théoriquement, il devrait donc être possible de rencontrer des questions où les résultats seraient d'une bonne « limpidité spectrale » (*NCSq* élevé) sans pour autant obtenir une *EMACq* faible et donc un *Rq* élevé.



Le graphique en nuage de points des questions du test de physique montre que des performances relativement élevées à un des deux indices n'impliquent pas forcément des performances relativement élevées à l'autre. Remarquons la question 7, précédemment signalée, qui se détache par une valeur *Rq* relativement faible et un indice *NCSq* relativement moyen.

Plus loin, dans le cadre de l'exploration du niveau « QCM » des tests MOHICAN (chapitre X) nous analyserons les corrélations entre les valeurs de *NCSq* et de *Rq* pour les questions de trois autres épreuves (Vocabulaire, Connaissances artistiques et Mathématiques).

Le fait de disposer de plusieurs « filtres spectraux » nous paraît intéressant du point de la fiabilité de détection des questions défectueuses. En effet, si un des indices « laisse passer » un item suspect, l'autre pourrait bien le signaler.

D. Indice de facilité introspective des questions (piq)

Reprenons l'exemple des résultats liés aux propositions de la 1ère question du test de physique (voir p. 217) mis en forme à l'aide du module « Protocole d'analyse spectrale de l'épreuve et des items » de *SCANTEST 2.0* (p. 207). L'exemple montre les résultats de deux paliers de turbo analyse : T80 (1er tableau) et T90 (2ème tableau).

8. Palier de Turbo analyse : T80

	OM	**P1**	P2	P3	P4	P5	P6	P7	P8	P9
8.1 N Rép. T80	18	**540**	16	222	28	114	114	2	0	0
8.2 % Rép. T80	2%	**51%**	2%	21%	3%	11%	11%	0%	0%	0%
8.3 C. Moy. T80	3%	**78%**	46%	49%	48%	47%	45%	10%	xxxx	xxxx
8.4 rpbis SC T80	-0,31	**0,54**	-0,13	-0,40	-0,17	-0,32	-0,35	-0,10	xxxx	xxxx

9. Palier de Turbo analyse : T90

	OM	**P1**	P2	P3	P4	P5	P6	P7	P8	P9
9.1 N Rép. T90	5	**199**	4	46	7	20	31	0	0	0
9.2 % Rép. T90	2%	**64%**	1%	15%	2%	6%	10%	0%	0%	0%
9.3 C. Moy. T90	8%	**85%**	60%	51%	57%	49%	43%	xxxx	xxxx	xxxx
9.4 rpbis SC T90	-0,33	**0,60**	-0,09	-0,42	-0,11	-0,28	-0,41	xxxx	xxxx	xxxx

Le 1er tableau intitulé « 8. Palier de Turbo analyse : T80 » contient la ligne « 8.3 » intitulée « 8.3 C. Moy. T80 ». Cette ligne donne les certitudes moyennes obtenues avec une sélection d'étudiants dont le réalisme est supérieur ou égal à 80. Elle indique pour la réponse correcte (ici la proposition P1, en gras) 78% de certitude moyenne. En moyenne, les étudiants qui ont choisi la réponse correcte (P1) lui accordent 78% de chances d'être correcte. La ligne « 8.1 N Rép. T80 » indique combien de sujets ont choisi P1, à T80 ils étaient 540.

La ligne « 9.3 » du 2ème tableau (palier turbo T90) montre les certitudes moyennes calculées à l'aide des données d'une sélection d'étudiants encore plus performants dans leurs auto-estimations étant donné leur réalisme supérieur ou égal à 90. La certitude moyenne de P1 vaut cette fois 85% et les étudiants étaient 199 à avoir choisi P1.

A l'aide des données de ces tableaux nous pouvons calculer la facilité introspective au niveau de toute une question (*piq*), et ce, aux différents paliers de turbo analyse.

Nous définissons la facilité introspective d'une question par :

$$piq = \frac{\sum_{j=1}^{nj} (CMp_j * nup_j)}{NR} \quad (44)$$

avec :

j = indice des propositions pour une question donnée

nup_j = *n*ombre d'*u*tilisations de la *p*roposition j

CMp_j = la valeur de la *C*ertitude *M*oyenne de la *p*roposition j

NR = le *N*ombre total de *R*éponses à la question (= $\sum_j nup_j$)

et la *C*ertitude *M*oyenne d'une *p*roposition *j* par :

$$CMp_j = \frac{\sum_{s=1}^{nup_j} C_{j\,s}}{nup_j} \quad (45)$$

avec :

s = indice des *s*ujets ayant choisi la proposition *j*

$C_{j\,s}$ = *C*ertitude fournie pour la proposition *j* par un sujet *s*

nup_j = *n*ombre d'*u*tilisations de la *p*roposition *j*

Appliquée aux données sélectionnées lors d'une turbo analyse, la facilité introspective d'une question devient :

$$piq\ T_t = \frac{\sum_{j=1}^{nj} (CMp_j\ T_t * nup_j\ T_t)}{NR\ T_t} \quad (46)$$

avec :

j = indice des propositions pour une question donnée

nj = nombre de propositions pour une question donnée

$nup_j\ T_t$ = nombre d'utilisations de la proposition *j* au palier *t* de la *T*urbo analyse

$CMp_j\ T_t$ = la valeur de la *C*ertitude *M*oyenne de la *p*roposition *j* au palier *t* de la *T*urbo analyse

NR = le *N*ombre total de *R*éponses à la question au palier *t* de la *T*urbo analyse ($= \sum_j nup_j\ T_t$)

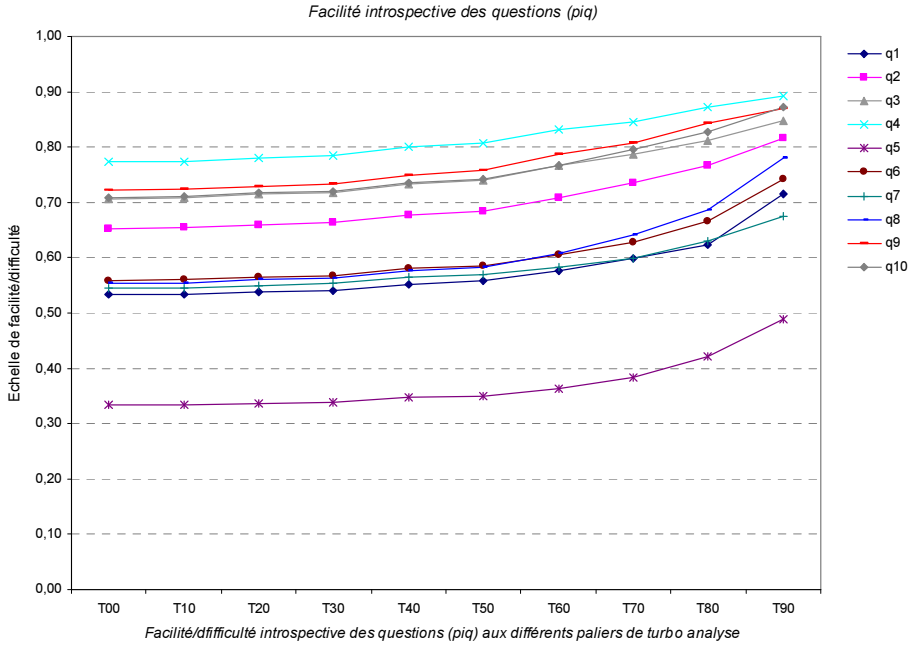
Plus l'indice *piq* est élevé, plus la question récolte des pourcentages de certitude élevés, plus les étudiants sont sûrs des réponses qu'ils fournissent et donc, plus la question leur paraît facile.

A l'inverse, lorsque la question récolte des pourcentages de certitudes faibles, l'indice *piq* est peu élevé. Les étudiants sont alors peu confiants en ce qui concerne leurs réponses, la question leur paraît plus difficile.

Voici le tableau des valeurs des indices de facilité introspective par question (*piq*) qui est généré par *SCANTEST 2.0* lors de la réalisation du protocole d'analyse de l'épreuve de physique (n = 2.497).

	T00	T10	T20	T30	T40	T50	T60	T70	T80	T90
q1	0,53	0,53	0,54	0,54	0,55	0,56	0,58	0,60	0,62	0,72
q2	0,65	0,65	0,66	0,66	0,68	0,68	0,71	0,73	0,77	0,82
q3	0,71	0,71	0,71	0,72	0,73	0,74	0,77	0,79	0,81	0,85
q4	0,77	0,77	0,78	0,78	0,80	0,81	0,83	0,85	0,87	0,89
q5	0,33	0,33	0,34	0,34	0,35	0,35	0,36	0,38	0,42	0,49
q6	0,56	0,56	0,56	0,57	0,58	0,59	0,61	0,63	0,67	0,74
q7	0,54	0,55	0,55	0,55	0,56	0,57	0,58	0,60	0,63	0,68
q8	0,55	0,55	0,56	0,56	0,58	0,58	0,61	0,64	0,69	0,78
q9	0,72	0,72	0,73	0,73	0,75	0,76	0,79	0,81	0,84	0,87
q10	0,71	0,71	0,72	0,72	0,74	0,74	0,77	0,80	0,83	0,87

Le graphique des *piq* ci-dessous (généré par *SCANTEST 2.0*) visualise les données de ce tableau.



La courbe de la question 5 se détache des autres à tous les paliers de la turbo analyse par son niveau de facilité introspective moins élevé que les autres questions. C'est la question du test qui pour les étudiants paraît la plus difficile.

E. Indice de Centration par question (Cq)

A l'aide de l'indice de facilité introspective d'une question (*piq*) et de l'indice de facilité objective d'une question (*poq*) nous pouvons envisager un indice de Centration par question (*Cq*).

1. Principe

Nous définissons la Centration par question par :

$$\boxed{Cq = piq - poq} \quad (47)$$

Avec :

piq = la facilité introspective d'une question (voir définition p. 251)
poq = la facilité objective d'une question

et la facilité objective d'une question (*poq*) par :

$$\boxed{poq = \frac{\sum_{s=1}^{ns} RC_{qs}}{ns}} \quad (48)$$

Avec :

s = indice des *s*ujets ayant choisi la réponse correcte à une question *q*
RC_{qs} = le choix de la *R*éponse *C*orrecte *RC* à une question *q* par le sujet *s*
ns = le *n*ombre total de *s*ujets ayant répondu à la question

Nous pouvons aussi calculer *Cq* pour chaque palier défini lors d'une turbo analyse :

$$\boxed{Cq\ T_t = piq\ T_t - poq\ T_t} \quad (49)$$

Avec :

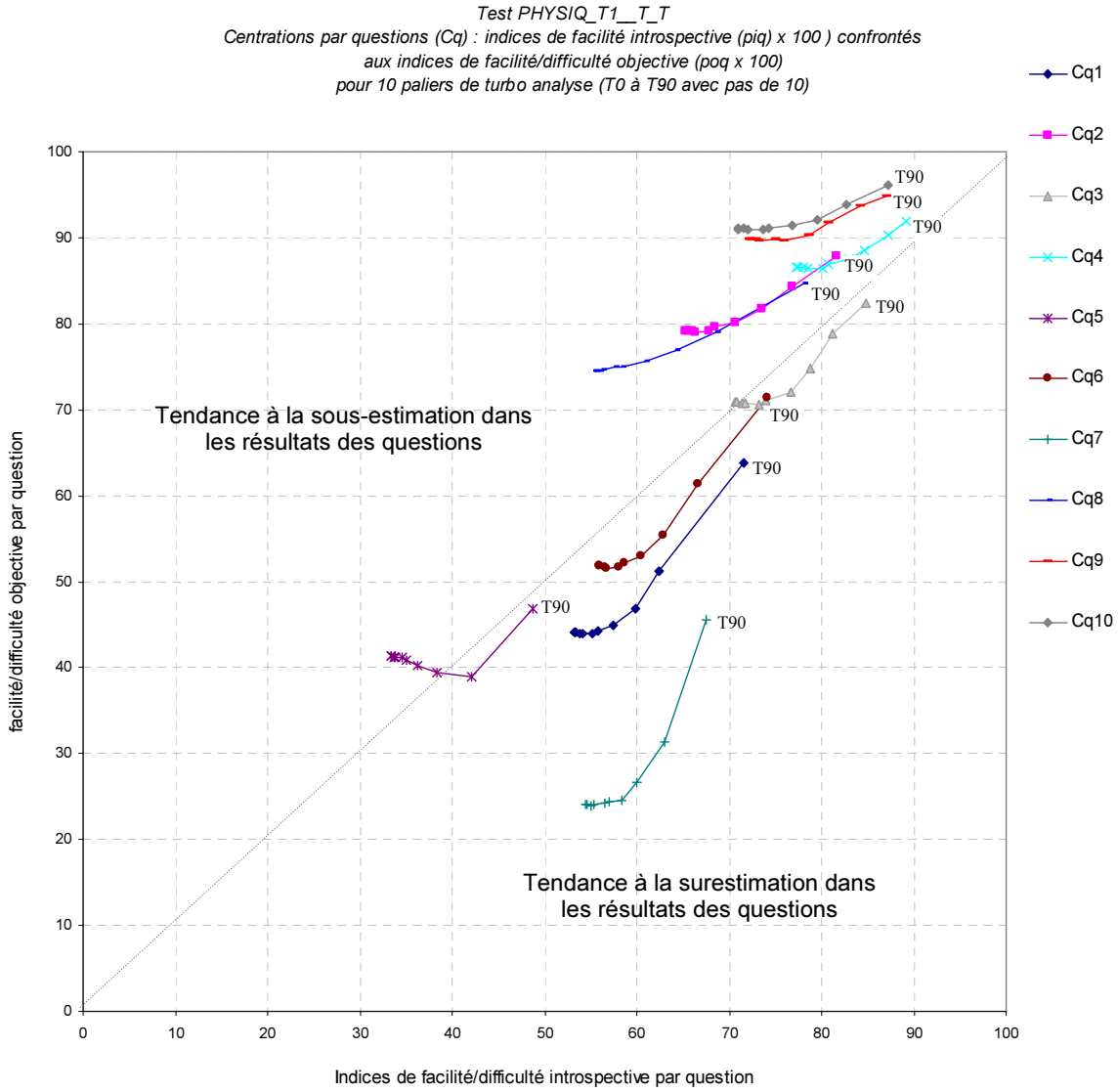
T_t = palier de *T*urbo analyse *t*
piq T_t = la *facilité i*ntrospective d'une question au palier de *T*urbo analyse *t*
poq T_t = la *facilité o*bjective d'une question au palier de *T*urbo analyse *t*

Voici le tableau des *piq*, *poq* et *Cq* aux 10 paliers de turbo analyse (T0 à T90 et pas de 10) :

		T00	T10	T20	T30	T40	T50	T60	T70	T80	T90
q1	*piq*	53,3	53,3	53,8	54,1	55,2	55,7	57,5	59,8	62,4	71,6
	poq	44,1	44,1	44,0	43,9	43,9	44,3	44,8	46,9	51,2	63,8
	Cq	**9,1**	**9,2**	**9,9**	**10,2**	**11,2**	**11,4**	**12,7**	**12,9**	**11,2**	**7,8**
q2	*piq*	65,3	65,4	66,0	66,3	67,8	68,4	70,7	73,5	76,8	81,7
	poq	79,2	79,2	79,1	79,0	79,2	79,6	80,2	81,8	84,3	87,8
	Cq	**-13,9**	**-13,8**	**-13,1**	**-12,7**	**-11,4**	**-11,2**	**-9,4**	**-8,3**	**-7,5**	**-6,1**
q3	*piq*	70,7	70,8	71,4	71,8	73,3	74,0	76,8	78,8	81,2	84,8
	poq	70,8	70,9	70,8	70,7	70,6	71,1	72,1	74,8	78,8	82,4
	Cq	**-0,2**	**-0,1**	**0,6**	**1,1**	**2,7**	**3,0**	**4,7**	**4,0**	**2,4**	**2,4**
q4	*piq*	77,3	77,4	78,1	78,5	80,1	80,8	83,1	84,6	87,2	89,2
	poq	86,5	86,6	86,5	86,5	86,5	86,9	87,5	88,5	90,2	92,0
	Cq	**-9,3**	**-9,2**	**-8,4**	**-8,0**	**-6,3**	**-6,2**	**-4,4**	**-3,9**	**-3,1**	**-2,8**
q5	*piq*	33,4	33,4	33,7	33,8	34,7	35,0	36,3	38,3	42,1	48,8
	poq	41,3	41,3	41,2	41,2	41,2	40,9	40,3	39,5	38,9	46,8
	Cq	**-7,9**	**-7,9**	**-7,5**	**-7,3**	**-6,5**	**-5,8**	**-4,0**	**-1,2**	**3,2**	**2,0**
q6	*piq*	55,9	56,0	56,5	56,7	58,0	58,6	60,5	62,8	66,6	74,1
	poq	51,9	51,9	51,7	51,5	51,7	52,2	52,9	55,5	61,4	71,5
	Cq	**4,0**	**4,1**	**4,8**	**5,2**	**6,3**	**6,4**	**7,6**	**7,4**	**5,2**	**2,6**
q7	*piq*	54,5	54,6	55,0	55,3	56,5	57,0	58,3	59,9	62,9	67,5
	poq	24,1	24,1	24,0	24,0	24,2	24,4	24,5	26,7	31,3	45,5
	Cq	**30,4**	**30,5**	**31,1**	**31,3**	**32,3**	**32,6**	**33,8**	**33,3**	**31,6**	**22,0**
q8	*piq*	55,4	55,5	56,0	56,2	57,6	58,3	60,9	64,2	68,6	78,0
	poq	74,4	74,4	74,5	74,7	75,0	75,0	75,7	76,8	78,9	84,6
	Cq	**-19,0**	**-18,9**	**-18,5**	**-18,4**	**-17,4**	**-16,7**	**-14,8**	**-12,7**	**-10,3**	**-6,6**
q9	*piq*	72,2	72,3	73,0	73,3	75,0	75,9	78,7	80,7	84,2	87,1
	poq	89,8	89,8	89,8	89,7	89,8	89,7	90,3	91,8	93,6	94,9
	Cq	**-17,6**	**-17,5**	**-16,8**	**-16,4**	**-14,8**	**-13,9**	**-11,7**	**-11,1**	**-9,4**	**-7,8**
q10	*piq*	70,9	71,0	71,6	72,0	73,6	74,3	76,8	79,6	82,8	87,2
	poq	91,0	91,1	91,0	91,0	90,9	91,1	91,5	92,1	93,8	96,2
	Cq	**-20,1**	**-20,0**	**-19,4**	**-19,0**	**-17,3**	**-16,8**	**-14,7**	**-12,5**	**-11,1**	**-9,0**

Pour chaque question reprise dans le tableau ci-dessus nous remarquons une tendance globale de l'indice *Cq* : plus les paliers de turbo analyse sont élevés, plus l'indice *Cq* se rapproche de zéro.

Les valeurs du tableau peuvent être visualisées à l'aide du graphique en nuage de points suivant (généré par *SCANTEST 2.0*). Pour chaque question on obtient 10 points qui représentent les valeurs de l'indice *Cq* aux 10 paliers de turbo analyse. Chaque ensemble de points reliés entre eux représente les *Cq* d'une question aux 10 paliers de turbo analyse.



La diagonale qui figure sur le graphique est un repère qui montre les positions que pourraient prendre les points si il y avait concordance parfaite entre la facilité introspective et la facilité objective d'une question.

Comme pour les *BSq* (voir p. 228) on remarque un regroupement des points à la base des lignes représentant les questions, traduisant des valeurs relativement proches aux paliers de turbo analyse les moins élevés (de T0 à T60).

Précisons ici que l'indice *Cq* peut varier dans une plage de –100 à 100. Lorsque *Cq* est négatif, il traduit pour la question une propension à la sous-estimation dans les résultats des sujets. A l'inverse, lorsque *Cq* est positif, il traduit une propension à la surestimation. Lorsque *Cq* est proche de 0, il n'y a ni sous-estimation, ni surestimation.

Dans le cadre de cette épreuve de physique, lorsque les points à l'extrémité droite d'une série sont proches de la diagonale, cela signifie que la facilité introspective (*piq*) ressentie par les étudiants les plus réalistes du groupe (T90) est proche de la facilité objective (*poq*), la valeur de *Cq* est alors proche de 0. Au palier de turbo analyse T90, c'est le cas pour : *Cq3* (2,4), *Cq4* (–2,8), *Cq5* (2) et *Cq6* (2,6).

D'autres séries de points situés plus au-dessus de la diagonale montrent des situations où les étudiants sous-estiment la facilité des questions, les *poq* sont plus élevés que les *piq* et l'indice *Cq* est alors négatif. c'est le cas pour *Cq2* (-6,1), *Cq8* (-6,6), *Cq9* (-7,8) et *Cq10* (-9).

A l'inverse, lorsque les séries de points sont sous la diagonale, on assiste alors à une tendance à surestimer la facilité de la question dans le groupe, c'est le cas des *Cq1* (7,8) et *Cq7* (22)

On remarque que la série de points de la question 7 se détache des autres. Même à un palier de turbo analyse élevé T90, cette question est relativement éloignée de la diagonale.

Bien qu'au palier de turbo analyse T90 cette question 7 présente un niveau de facilité objective (*poq7* = 45,5) très proche de celui de la question 5 (*poq5* = 46,8), le niveau de facilité introspective est très différent (*piq7* = 67,5 et *poq8* = 48,8), ce qui donne pour deux questions dont le niveau de difficulté/facilité objective est quasi identique, des centrations très différentes (*Cq7* = 22 et *Cq* = 2).

2. Constats et questions à propos des *piq et Cq*

Concernant la question 7, d'abord nous constatons sur le graphique de *Rq*7 (voir p. 244) un alignement des points sous la diagonale. Ensuite, le graphique des *piq* (p. 253) montre que cette question 7 n'est pas ressentie comme étant la plus difficile du test, c'est en fait la 5ème question qui est ressentie comme particulièrement difficile. Enfin, le graphique des *Cq* (p. 255) nous montre que les points liés à la question 5 sont plus proches de la diagonale idéale que les points de la question 7. L'indice *Cq7* chiffre la surestimation à 22.

[2.4] Rencontre-t-on dans les autres épreuves MOHICAN des questions qui, relativement aux autres, se caractérisent par une forte surestimation ?

[2.5] Existe-t-il dans les autres épreuves MOHICAN des questions qui, relativement aux autres, enferment une forte tendance à la sous-estimation dans leurs résultats ?

3. En synthèse, ce que mesurent *piq et Cq*, l'intérêt de ces indices

Lorsque les étudiants ont été invités à accompagner d'un pourcentage de certitude le choix d'une proposition dans le cadre d'une QCM, il devient possible de chiffrer le degré de conviction avec laquelle chaque solution a été choisie. C'est l'information qui est livrée par la Certitude Moyenne d'une proposition (*CMp*). Cette *CMp* constitue aussi ce qu'on pourrait appeler « le pourcentage moyen de chances attribué à une proposition d'être correcte ».

Nous avons vu que cette information peut être globalisée au niveau de la question entière. Il s'agit alors d'effectuer la moyenne des *CMp* pondérée par les fréquences d'utilisation de chacune des propositions. On obtient ainsi à l'aide de cette moyenne pondérée des Certitudes Moyennes des propositions, une mesure de la facilité introspective d'une question, l'indice *piq*. On pourrait dès lors dire que l'indice *piq* constitue « la moyenne pondérée des pourcentages moyens de chances attribuée à l'ensemble des propositions », en d'autres termes, la facilité subjective de la question.

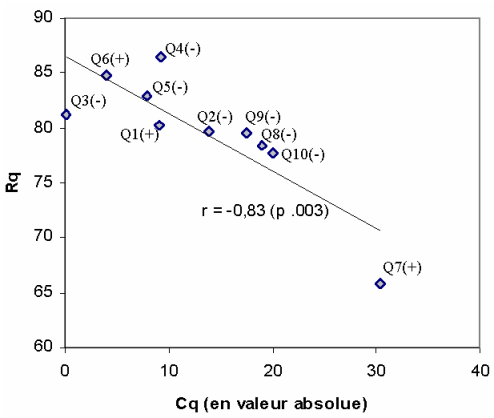
En effet, plus une question paraît facile aux yeux des étudiants, plus ceux-ci estiment les probabilités de voir leur réponse correcte élevées. Dans ce cas, ils accompagnent leurs choix de pourcentages de certitude élevés ce qui entraîne un indice *piq* élevé. Donc, plus la valeur de l'indice *piq* est élevée, plus les étudiants ont tendance à considérer la question comme étant facile.

Il est également possible de mesurer la facilité objective d'une question (*poq*). Nous avons vu que l'indice *poq* revient à calculer le pourcentage de réponses correctes. Plus ce pourcentage (le taux d'exactitude) sera élevé et plus la question pourra être considérée comme objectivement facile.

Dans le cadre de l'évaluation de la qualité spectrale d'une question, nous proposons de confronter la facilité subjective à la facilité objective. En soustrayant la valeur de l'indice *poq* à celle de l'indice *piq*, nous sommes en mesure de chiffrer une éventuelle tendance globale à la sur ou sous-estimation dans les résultats d'une question. C'est ce que permet l'indice de Centration par question (*Cq*) dont le signe nous informe à propos de l'orientation d'une erreur globale dans l'utilisation des pourcentages de certitude : vers de la surestimation lorsqu'il est positif, vers de la sous-estimation lorsqu'il est négatif.

L'intérêt de l'indice *Cq* réside dans une utilisation en complémentarité avec l'indice *Rq*. Précédemment, nous avons vu que la valeur du *Rq* fournit une bonne mesure du niveau de réalisation des prédictions liées aux différents pourcentages de certitude d'une question (elle prend en compte de façon précise les erreurs d'estimation liées à chacune des certitudes). Par contre, elle n'informe pas sur la tendance à la sur ou sous-estimation dans les résultats.

Le graphique en nuage de points des questions du test de physique reprend les valeurs de l'indice *Rq* et de l'indice *Cq* calculé en valeur absolue. Nous remarquons que les valeurs obtenues à l'indice *Cq* sont fortement corrélées avec celles de l'indice *Rq*. Plus loin dans la troisième partie « exploration du niveau QCM » (chapitre X) nous étudierons les corrélations entre les indices *NCSq*, *Rq*, *piq* et *Cq* de trois autres épreuves MOHICAN (Vocabulaire, Connaissances artistiques et Mathématiques).



Le signe (-) qui suit le numéro de la question signale qu'elle contient une tendance à la sous-estimation dans les résultats. Le signe (+) indique qu'elle contient une tendance à la surestimation.

Remarquons la question 7 qui se détache du lot par une forte tendance à la surestimation dans ses résultats.

Dans une perspective de détection des questions litigieuses, l'indice *Cq* pourrait s'avérer utile pour détecter des « effets de surprise » dans les résultats. On peut en effet suspecter un problème dans un item qui présente une forte surestimation (comme la question 7), cela pourrait signifier que la question était d'une difficulté à laquelle les étudiants ne pouvaient pas s'attendre. Peut-être que la question 7 contenait un piège qu'un bon nombre d'étudiants n'a pas pu déjouer (elle est réussie par seulement 24,1% des sujets).

Chapitre VIII :

Outils d'analyse de la qualité spectrale des tests

Sommaire

Nous venons de présenter des outils d'aide à l'analyse de la qualité spectrale des PROPOSITIONS (p. 211) et des QCM (p. 225). Dans la partie qui suit, nous proposerons une série d'outils dont le but est de nous aider à évaluer la qualité spectrale d'un TEST.

A. Niveaux de Cohérence Spectrale d'un test (NCSt) comparés au Niveau de Cohérence Interne (NCIt)

1. Définition du *NCSt*

Il est possible d'identifier le Niveau de Cohérence Spectrale d'un test (*NCSt*) à chaque palier de turbo analyse, soit à partir des valeurs des Niveaux de Cohérence Spectrale de chaque question (*NCSq*) positionnée dans les zones de qualité (voir p. 228), soit en calculant de façon plus précise les *NCSq* (voir p. 231).

Voici le tableau des Niveaux de Cohérence Spectrale des 10 questions (*NCSq*) du test de physique (n = 2.497 étudiants) calculés selon la méthode exposée en amont (p. 231) pour chaque palier de turbo analyse (de *T0* à *T90* avec un pas de 10).

	T0	*T10*	*T20*	*T30*	*T40*	*T50*	*T60*	*T70*	*T80*	*T90*
NCSq1	0,52	0,53	0,54	0,55	0,58	0,59	0,65	0,76	0,89	0,96
NCSq2	0,51	0,51	0,52	0,53	0,55	0,57	0,62	0,74	0,82	1,00
NCSq3	0,38	0,38	0,39	0,41	0,43	0,44	0,49	0,58	0,76	1,02
NCSq4	0,53	0,53	0,55	0,57	0,61	0,64	0,70	0,74	0,87	0,89
NCSq5	0,51	0,51	0,51	0,52	0,53	0,57	0,63	0,76	0,95	1,20
NCSq6	0,48	0,48	0,50	0,51	0,53	0,54	0,61	0,67	0,82	0,94
NCSq7	0,49	0,49	0,51	0,51	0,51	0,52	0,59	0,67	0,81	1,00
NCSq8	0,71	0,71	0,71	0,72	0,72	0,76	0,81	0,91	0,97	1,03
NCSq9	0,42	0,42	0,43	0,45	0,46	0,50	0,54	0,61	0,70	0,87
NCSq10	0,47	0,47	0,48	0,49	0,52	0,55	0,61	0,68	0,73	0,82
Moyenne = ***NCSt***	**0,47**	**0,47**	**0,48**	**0,49**	**0,50**	**0,53**	**0,59**	**0,68**	**0,80**	**0,95**

La dernière ligne du tableau reprend la moyenne des *NCSq* à chaque palier de turbo analyse. Rappelons que l'indice *NCSq* varie de –2 à 2 (p. 231) et :
- qu'entre 0 et 2 une QCM se situe dans une plage de qualité spectrale ;
- qu'entre –2 et 0 une QCM se situe dans une plage de non qualité spectrale.

Nous définissons Niveau de Cohérence Spectrale d'un test (*NCSt*) pour un palier de Turbo analyse t donné (T_t) :

$$NCSt\ T_t = \frac{\sum_{i=1}^{nq} NCSq_i\ T_t}{nq} \quad (50)$$

avec :

T_t = palier de *T*urbo analyse *t*
i = l'*i*ndice des questions du test
$NCSq_i\ T_t$ = les *N*iveaux de *C*ohérence *S*pectrale de chaque *q*uestion pour un palier de *T*urbo analyse *t*
nq = le *n*ombre de *q*uestions figurant dans le test

2. Définition du *NCIt*

Nous avons vu précédemment qu'il était possible de calculer les Niveaux de Cohérence Interne des questions (*NCIq*) à l'aide des *rpbis* classiques (voir p. 233) :

Nous obtenons les *NCIq* suivants pour les 10 QCM du test de physique (*n* = 2.497) :

	NCIq
q1	0,711
q2	0,566
q3	0,547
q4	0,555
q5	0,594
q6	0,633
q7	0,576
q8	0,644
q9	0,489
q10	0,559
Moyenne = ***NCIt***	**0,588**

La dernière ligne du tableau reprend la moyenne des *NCIq* et donne le Niveau de Cohérence Interne du test basé sur les *rpbis classiques* (*NCIt*) qui vaut ici 0,588. Nous définissons le *NCIt* :

$$NCIt = \frac{\sum_{i=1}^{nq} NCIq_i}{nq} \quad (51)$$

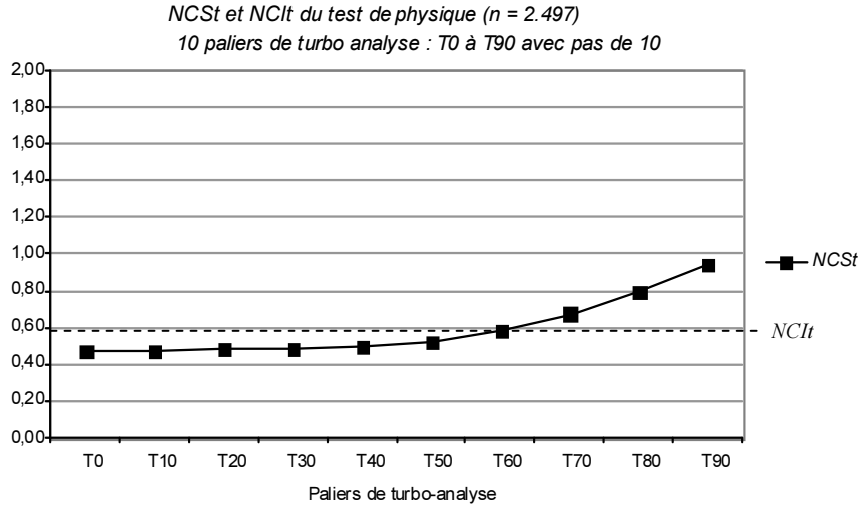
avec :

$NCIq_i$ = les *N*iveaux de *C*ohérence *I*nterne de chaque *q*uestion
i : l'*i*ndice des questions du test
nq : le *n*ombre de *q*uestions figurant dans le test

Le *NCSt* et le *NCIt*, varient entre –2 et 2. Lorsque les valeurs de ces indices se situent entre 0 et 2 nous disons du *NCSt* qu'il est dans une plage de qualité spectrale (de cohérence spectrale, voir p. 227, où le maximum vaut 2) et du *NCIt* qu'il est dans une plage de cohérence interne (où le maximum vaut 2 aussi). Lorsque ces indices valent entre –2 et 0, nous disons du *NCSt* qu'il se situe dans une plage de non qualité spectrale (où la valeur la plus basse = –2) et du *NCIt* qu'il se situe dans une plage d'incohérence interne (où la valeur la plus basse = –2).

3. Représentation graphique

Les valeurs des *NCSt* aux différents paliers de turbo analyse (voir tableau p. 261) peuvent être visualisées à l'aide d'un graphique tel que ci-dessous (automatiquement généré par *SCANTEST 2.0*).



Nous y avons tracé une ligne en pointillés qui reprend le Niveau de Cohérence Interne du test (*NCIt*) calculé à partir des *rpbis* classiques (*NCIt* = 0,588). Nous constatons qu'au palier T60 de la turbo analyse, le *NCSt* calculé à partir des *rpbis SCT* est pratiquement égal au Niveau de Cohérence Interne du test (*NCIt*) calculé à l'aide des *rpbis* classiques.

Dans le contexte de ce test de physique nous remarquons aussi qu'à partir du palier T70 le *NCSt* est plus élevé que le *NCIt*.

4. Constat et question à propos des *NCSt* et *NCIt*

Nous venons de constater qu'au palier de turbo analyse T60 le *NCSt* est égal au *NCIt* et qu'à partir de T70 les valeurs obtenues au *NCSt* sont supérieures à celles du *NCIt*.

[3.1] Lorsqu'on compare les valeurs obtenues par les autres épreuves MOHICAN aux *NCSt* et *NCIt*, observe-t-on aussi ce dépassement du *NCIt* par les valeurs *NCSt* à partir du palier T60 ?

B. Indice de Réalisation des prédictions par test (Rt)

1. Méthode de calcul

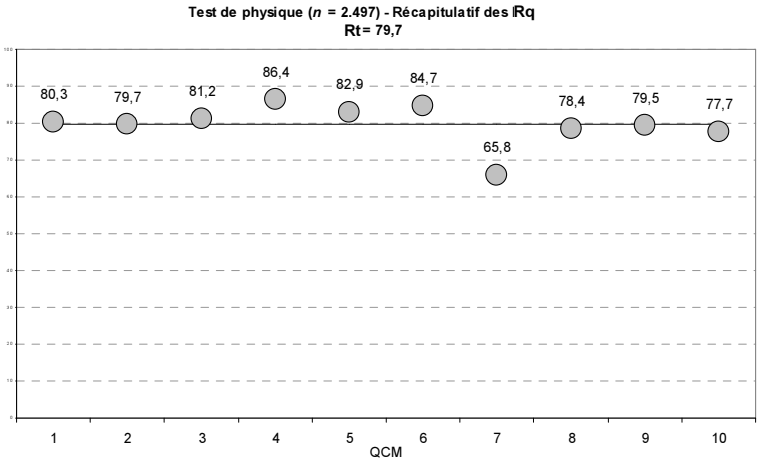
Nous proposons de calculer l'indice de *R*éalisation des prédictions par *t*est (*Rt*) (exprimé en pourcents) en effectuant la moyenne des indices de *R*éalisation des prédictions par *q*uestion (*Rq* - voir p. 242) :

$$Rt = \frac{\sum_{i=1}^{nq} Rq_i}{nq} \quad (52)$$

Avec :
i = l'*i*ndice des questions du test
Rq_i = l'indice de *R*éalisation des prédictions pour la *q*uestion *i* (en pourcents)
nq = le *n*ombre de *q*uestions dans le test

2. Représentation graphique

Voici un exemple de récapitulatif pour les 10 questions du test de physique. Les valeurs des *Rq* des questions (de 1 à 10) y sont représentées par des ronds grisés. La moyenne des *Rq*, c'est-à-dire l'indice *Rt*, vaut 79,7 est visualisée par un trait horizontal continu (l'écart type non repris dans le graphique figure dans les protocoles d'analyse et vaut 5,3).



On remarque sur le graphique (automatiquement généré par *SCANTEST 2.0*) que la plupart des QCM sont soit très proches de la moyenne soit au-dessus, sauf la question 7 dont le *Rq* à 65,8 se détache des autres.

Rappelons que le *Rq* nous informe sur la propension d'une question à récolter ou non des prédictions ou Taux d'Exactitude Annoncés (TEA) en concordance avec la réalité des Taux d'Exactitude Observés (TEO). Lorsque nous globalisons l'information au niveau d'un test à l'aide de l'indice *Rt*, nous obtenons un éclairage sur la propension de l'ensemble des questions d'une épreuve à récolter des TEA en concordance avec les TEO.

3. *Rt* après turbo analyse

Nous définissons l'indice de *R*éalisation des prédictions d'un *t*est (*Rt*) (exprimé en pourcents) pour un palier de Turbo analyse *t* donné (T_t) :

$$Rt\ T_t = \frac{\sum_{i=1}^{nq} Rq_i\ T_t}{nq} \quad (53)$$

avec :

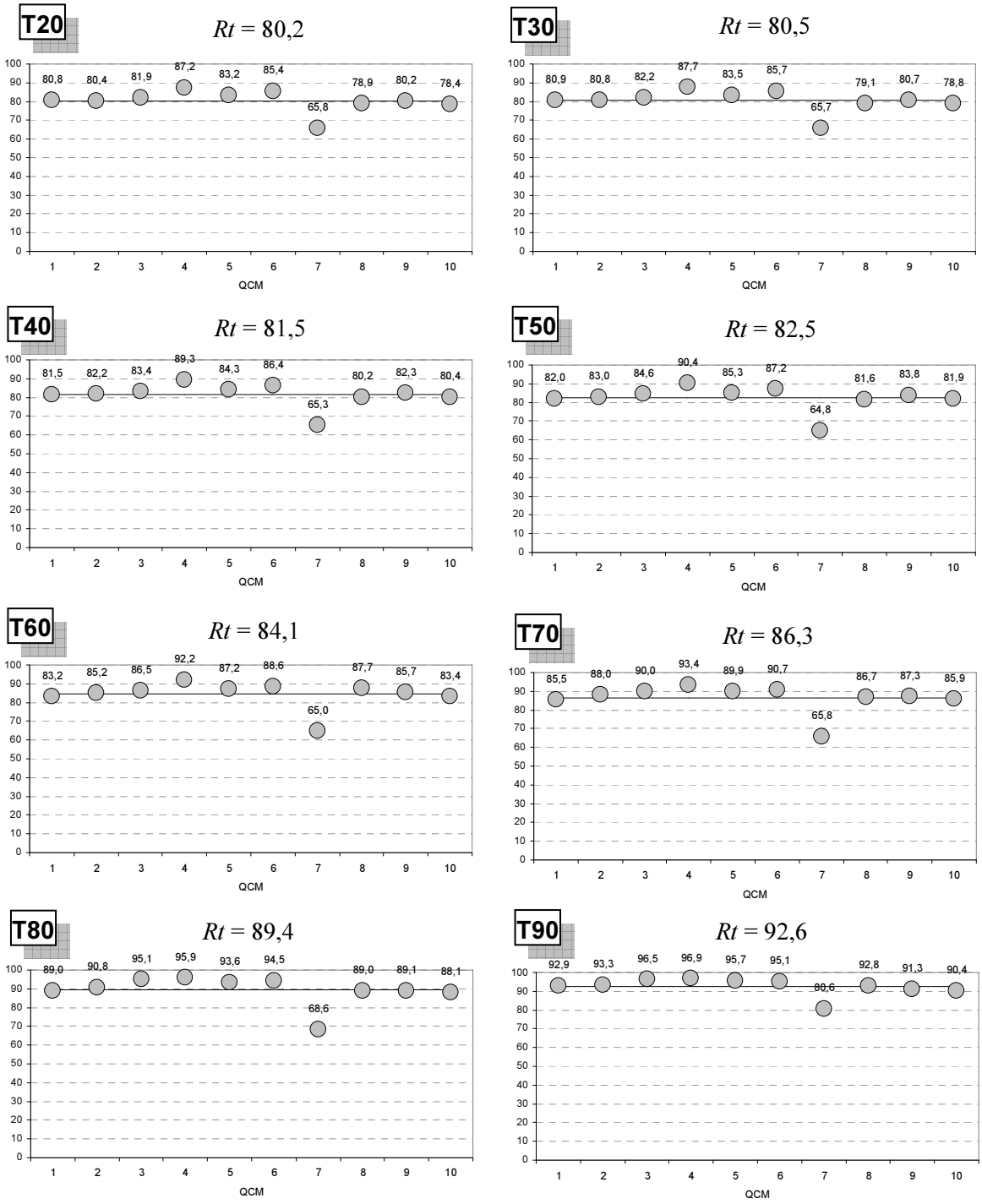
T_t = palier de *T*urbo analyse *t*
i = l'*i*ndice des questions du test
$Rq_i\ T_t$ = l'indice de *R*éalisation des prédictions de la *q*uestion *i* pour un palier de *T*urbo analyse *t*
nq = le *n*ombre de *q*uestions figurant dans le test

Voici le tableau des indices de Réalisation des prédictions par question (*Rq*) du test de physique (*n* = 2.497 étudiants) pour huit paliers de turbo analyse (de *T20* à *T90* avec un pas de 10).

	T20	*T30*	*T40*	*T50*	*T60*	*T70*	*T80*	*T90*
Rq1	80,8	80,9	81,5	82	83,2	85,5	89	92,9
Rq2	80,4	80,8	82,2	83	85,2	88	90,8	93,3
Rq3	81,9	82,2	83,4	84,6	86,5	90	95,1	96,5
Rq4	87,2	87,7	89,3	90,4	92,2	93,4	95,9	96,9
Rq5	83,2	83,5	84,3	85,3	87,2	89,9	93,6	95,7
Rq6	85,4	85,7	86,4	87,2	88,6	90,7	94,5	95,1
Rq7	65,8	65,7	65,3	64,8	65	65,8	68,6	80,6
Rq8	78,9	79,1	80,2	81,6	83,7	86,7	89	92,8
Rq9	80,2	80,7	82,3	83,8	85,7	87,3	89,1	91,3
Rq10	78,4	78,8	80,4	81,9	83,4	85,9	88,1	90,4
Moyenne = ***Rt***	**80,2**	**80,5**	**81,5**	**82,5**	**84,1**	**86,3**	**89,4**	**92,6**

La dernière ligne du tableau montre qu'à chaque augmentation du palier de la turbo analyse correspond une élévation de l'indice *Rt*.

Les huit graphiques récapitulatifs des indices de concordance des prédictions ci-dessous ont été réalisés à partir des données du test de physique (n = 2.497). Chaque graphique reprend à un niveau « t » de turbo analyse (de T20 à T90 avec un pas de 10 points en réalisme) les valeurs des indices de réalisation des prédictions par question (les ronds grisés) ainsi que l'indice *Rt* calculé en effectuant la moyenne des *Rq* (le trait noir horizontal).



L'élévation de l'indice *Rt* (la moyenne des *Rq*), à chaque palier de turbo analyse est logique. A chaque niveau de turbo analyse, les données à partir desquelles les *Rq* sont calculés proviennent en effet d'étudiants en moyenne de plus en plus réalistes.

La série de graphiques montre qu'au palier de turbo analyse T90 on assiste à un rapprochement de tous les *Rq* vers la moyenne, à ce niveau de turbo analyse réalisée à partir des données des étudiants les plus réalistes, les prédictions sont de plus en plus en concordance avec les taux d'exactitudes. Cependant, même à ce palier T90 on remarque que la QCM n°7 reste à l'écart de la moyenne tandis que les ronds grisés des autres questions « touchent » la ligne horizontale tracée à la valeur de la moyenne.

*C. Indice de facilité introspective du test (*pit*)*

Précédemment nous avons défini comment calculer l'indice *piq*, c'est-à-dire la facilité/difficulté introspective d'une question (voir p. 251). Nous proposons maintenant de calculer la facilité/difficulté d'un test, l'indice *pit*, en effectuant la moyenne des *piq* d'une épreuve.

1. Méthode de calcul

Nous définissons la facilité/difficulté introspective d'un test par :

$$pit = \frac{\sum_i piq_i}{nq} \quad (54)$$

avec :

i = l'indice des questions du test
piq_i = la facilité/difficulté introspective d'une question *i*
nq = le *n*ombre de *q*uestions dans le test

Aux paliers élevés de la turbo analyse les valeurs de *pit* reposent sur des données plus valides car provenant des sujets les plus réalistes. Nous allons donc aussi calculer la facilité/difficulté d'un test (*pit*) aux différents paliers *t* d'une Turbo analyse (T_t) :

$$pit\ T_t = \frac{\sum_i piq_i\ T_t}{nq} \quad (55)$$

avec :

i = l'indice des questions du test
$piq_i\ T_t$ = la facilité/difficulté d'une question *i* au palier *t* de la Turbo analyse
nq = le *n*ombre de *q*uestions dans le test

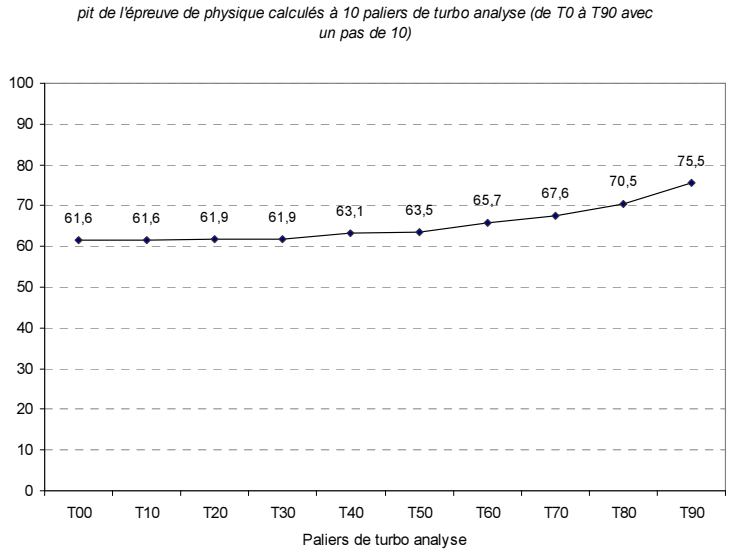
Voici le tableau des valeurs récoltées par le test de physique à l'indice *pit* calculé aux différents paliers de la turbo analyse (T0 à T90 avec un pas de 10) :

pit T00	61,6
pit T10	61,6
pit T20	61,9
pit T30	61,9
pit T40	63,1
pit T50	63,5
pit T60	65,7
pit T70	67,6
pit T80	70,5
pit T90	75,5

2. Représentation graphique

Lorsqu'on trace la courbe des indices pit aux 10 paliers de la turbo analyse on constate que valeurs n'évoluent pratiquement pas de T0 à T30 (ce qui est logique dans la mesure où à ces paliers les différences d'effectifs sont très faibles, voir p. 219), puis augmentent lentement jusqu'à T90.

Dans le cadre de l'épreuve de physique, plus les étudiants sont réalistes (les paliers de turbo analyse élevés), plus ils ont tendance à accompagner leurs réponses correctes de pourcentages de certitude élevés et donc plus l'épreuve leur paraît facile.



3. Constat et questions à propos de l'indice *pit*

Nous constatons que plus on monte dans les paliers de la turbo analyse de l'épreuve de physique, plus elle récolte un score *pit* élevé et donc plus les sujets la ressentent comme étant facile.

[3.2] Observe-t-on une augmentation des taux d'exactitude en parallèle avec cette augmentation de la facilité ressentie par les étudiants aux paliers turbo les plus élevés ?

D. Indice de Centration par test (Ct)

Précédemment nous avons présenté l'indice de centration par question (*Cq*). L'indice *Cq* nous permet de chiffrer la propension à la surestimation ou à la sous-estimation pour chaque question (voir p. 254). Nous proposons ici de calculer le même type d'indice mais au niveau d'une épreuve entière.

1. Méthode de calcul

Nous définissons la Centration d'un test par :

$$\boxed{Ct = pit - pot} \quad (56)$$

Avec :

pit = la facilité introspective d'un test (voir définition p. 268)
pot = la facilité objective d'un test

et la facilité objective d'un test (*pot*) par :

$$\boxed{pot = \frac{\sum_i poq_i}{nq}} \quad (57)$$

avec :

i = l'indice des questions du test
poq_i = la facilité objective d'une question *i* (voir définition p. 254)
nq = le *n*ombre de *q*uestions dans le test

Nous pouvons aussi calculer *Ct* pour chaque palier défini lors d'une turbo analyse :

$$\boxed{Ct\ T_t = pit\ T_t - pot\ T_t} \quad (58)$$

Avec :

T_t = palier de Turbo analyse *t*
pit T_t = la facilité/difficulté introspective d'un test au palier de Turbo analyse *t*
pot T_t = la facilité/difficulté objective d'un test au palier de Turbo analyse *t*

Appliqué aux données de l'épreuve de physique, cela donne :

	pit	*pot*	*Ct*
T00	61,6	65,5	-3,9
T10	61,6	65,5	-3,9
T20	61,9	65,4	-3,6
T30	61,9	65,4	-3,5
T40	63,1	65,4	-2,3
T50	63,5	65,5	-1,9
T60	65,7	65,8	-0,1
T70	67,6	67,0	0,5
T80	70,5	69,3	1,1
❍ *T90*	75,5	75,1	0,5

Jusqu'à T30, les valeurs de l'indice *pit* sont moins élevées que celles du *pot* et restent stables, mais à partir de T40 elles augmentent progressivement et rejoignent celles du *pot*. Les valeurs de l'indice *pot* restent stables plus longtemps jusqu'à T60 puis augmentent aussi jusqu'à T90.

SCANTEST 2.0 calcule automatiquement ce tableau et génère le graphique présenté dans la section qui suit.

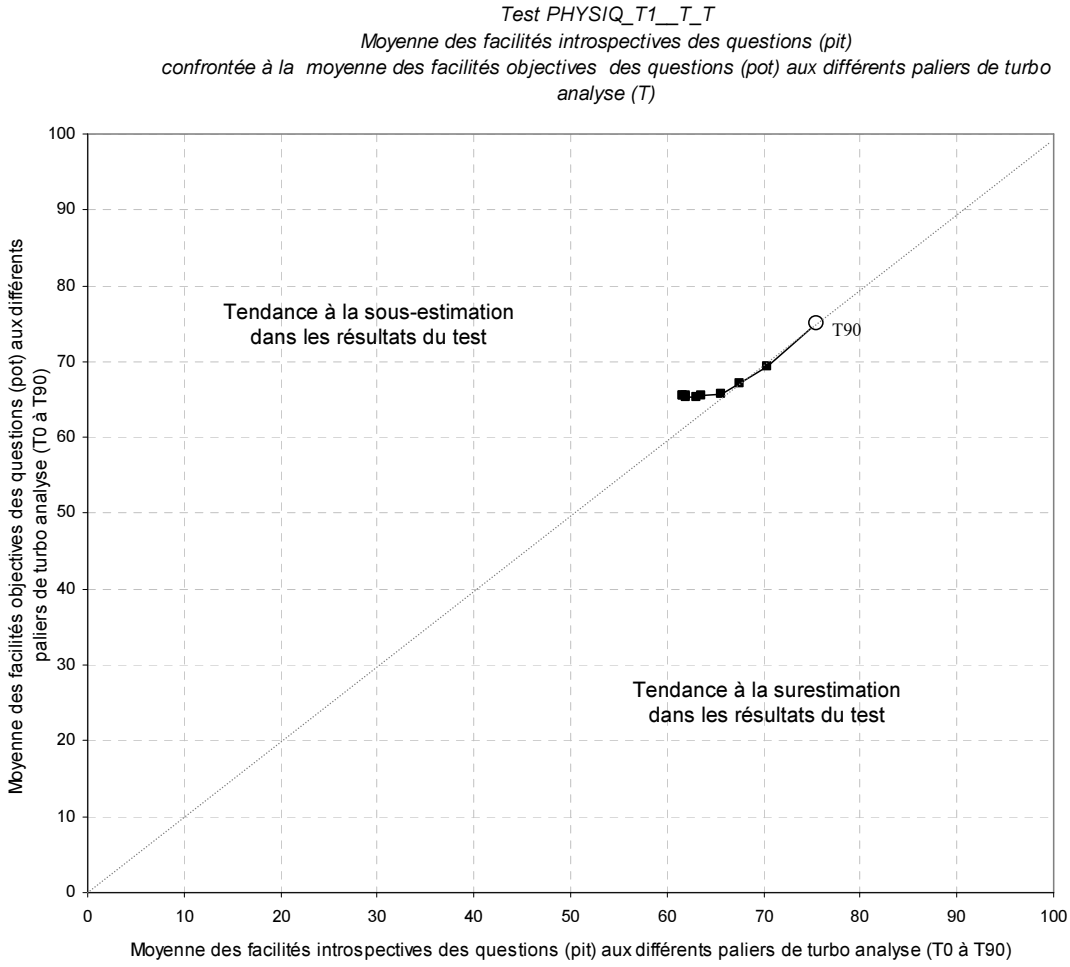
2. Représentation graphique des *Ct* aux paliers d'une turbo analyse

Le point correspondant à l'indice *Ct* calculé à T90 est signalé par le rond ombré « ❍ ».

La diagonale qui traverse le graphique est un repère qui montre les situations où la valeur de l'indice *pit* est systématiquement égale à la valeur de l'indice *pot*. Dans ce cas il n'y a ni sous-estimation ni surestimation.

Lorsqu'un point représentant la valeur de *Ct* se situe au-dessus de la diagonale, il y a une propension à la sous-estimation dans les réponses fournies aux questions du test et *Ct* est alors négatif.

Lorsque les points sont en dessous de la diagonale, il y a surestimation et *Ct* est positif.



Nous observons à partir de T60 une quasi égalité entre les *pit* et les *pot*, les points représentant Ct à T60, T70, T80 et T90 sont alors alignés sur la diagonale repère. Non seulement les questions paraissent plus faciles aux sujets sélectionnés dans les paliers de turbo analyse élevés, mais en plus, dans le cadre de cette épreuve de physique et pour ces sujets réalistes, les questions sont réellement plus faciles !

E. Fréquences et statistiques descriptives des performances en Réalisme des groupes (Rg)

Les informations liées aux scores de *R*éalisme obtenus par les *s*ujets (*Rs*) doivent être prises en compte lors de l'analyse de la qualité spectrale d'un test. Rappelons que c'est sur base de l'indice *Rs* que sont sélectionnées les données qui alimentent les turbo analyses pour le calcul des indices *rpbis SCT* (voir principe de turbo analyse, p. 186).

1. Implications de la consigne de recueil des degrés de certitude sur le calcul de l'indice de réalisme des sujets (*Rs*)

De nombreuses recherches ont porté sur l'élaboration d'indices de mesure de la qualité de l'auto-estimation des étudiants et notamment sur l'indice de réalisme (Brier, 1950; Adams & Adams, 1961; Oskamp, 1962; Murphy, 1974 ; Lichtenstein & al. 1977; Leclercq, 1982). Habituellement, on évalue le réalisme d'un sujet en se basant sur la moyenne de ses erreurs de certitude.

a) Consigne et formule « FAPSE »

Dans le cadre de nos recherches sur l'utilisation des degrés de certitude par les étudiants de la Faculté de Psychologie et des Sciences de l'Education (FAPSE) de l'Université de Liège (ULg) (Gilles, 1995,1996a, 1996b, 1997, 1998a, 1998b) nous avons utilisé l'échelle des Degrés de Certitude (DC) proposée par Leclercq (1982) qui entraîne l'utilisation de Valeurs Centrales (VC) dans le calcul du réalisme.

```
DC =       0         1          2      3     4   5
     |----------|----------|--------|------|----|--|
%  = 0         25          50       70     85   95  100
VC =    12,5        37,5       60     77,5   90  97,5
```

La formule de calcul du réalisme des sujets (*Rs*) que nous avons utilisée est celle de Leclercq (1982). Pour faciliter l'interprétation par les étudiants, nous l'avons adaptée de façon à ce que le minimum en réalisme soit égal à 0 et le maximum possible égal à 1 pour aboutir à la formule suivante (en fait, avec cette formule le maximum possible dépasse très légèrement 1) :

$$\boxed{Rs = [(1 - MEC) - \beta].\alpha} \quad (59)$$

avec :

$MEC = \dfrac{\sum_{i=1}^{nc} (|\ TE_i - VC_i\ |).NU_i}{NR}$ = la Moyenne des Erreurs de Certitude

i = *i*ndice des degrés de certitude

nc = *n*ombre de degrés de *c*ertitude (ici $nc = 6$)

NC_i = *n*ombre de réponses correctes pour la certitude i

NU_i = Nombre d'utilisations de la certitude i

TE_i = Taux d'Exactitude de la certitude $i = NC_i / NU_i$

VC_i = Valeur de la Centrale de la certitude i

NR = Nombre total de Réponses = $\sum_i NU_i$

β = correction pour l'erreur minimale, dans le cas des épreuves FAPSE (étant donné l'échelle ci-dessus), β = ERR MIN (voir ci-dessous) ici $\beta = 0{,}025$

α = correction d'étendue, dans le cas des épreuves FAPSE (étant donné l'échelle ci-dessus), pour étendre la plage de 0,95 à 1, $\alpha = 1$ / (ERR MAX – ERR MIN) ici $\alpha = 1{,}0526$ (voir ci-dessous pour ERR MAX et ERR MIN).

Nous avons introduit les coefficients α et β pour ramener les scores sur une plage de 0 à 1. En effet, étant donné l'échelle des degrés de certitude utilisée (voir ci-dessus), la procédure définie par Leclercq (1982) entraîne une ERR MAX (l'erreur maximum) égale à 0,975 pour un sujet qui se trompe systématiquement avec le degré de certitude 5, et, une ERR MIN (l'erreur minimum) égale à 0,025 pour un sujet qui répond correctement à toutes les questions avec un degré de certitude 5.

La valeur centrale 97,5 pour le degré de certitude 5 provoque par exemple pour un étudiant qui répond correctement à toutes les questions avec cette certitude maximum, non pas une Moyenne des Erreurs de Certitude (*MEC*) égale à zéro, mais une *MEC* égale à 0,025 (1-0,975) et, en conséquence, un *Rs* qui vaut 0,975 (1-*MEC*) plutôt que 1 (le score de réalisme que cet étudiant mérite si on considère que 1 est le maximum).

C'est la raison pour laquelle nous appliquons :

- une correction pour l'étendue (α = 1,0526) étant donné les deux valeurs extrêmes (ERR MIN et ERR MAX) qui déterminent une plage de 0,95 (0,975 – 0,025) ;
- une autre correction pour l'erreur minimale (β = 0,025).

Ces deux corrections permettent alors d'approcher la valeur maximum 1 sur une plage de 0 à 1.

Dans l'exemple de l'étudiant ayant répondu correctement à toutes les questions avec la certitude 5 cela donne :

$$Rs = [(1 - 0{,}025) - 0{,}025].1{,}0526 = 0{,}95.1{,}0526 \approx 1$$

Les coefficients de correction α et β proposés dans la formule FAPSE peuvent malgré tout déboucher sur des indices *Rs* légèrement supérieurs à 1. Par exemple dans le cadre d'un test comportant 10 questions, si un sujet accompagne systématiquement ses 10 réponses d'un degré de certitude 2 (valeur centrale = 60%) et si 6 réponses sur 10 sont correctes (taux d'exactitude = 60%) alors MEC vaut 0 et 1 – MEC vaut 1. Dès lors, si on soustrait le coefficient β (0,025) et qu'on multiplie ensuite par le coefficient α (1,0526), Rs vaut 1,0262… Cette valeur légèrement plus élevée que 1 provient du fait que nous avons opté dans le cadre de cette consigne FAPSE pour un coefficient β égal à 0,025 alors que nous venons de voir qu'il existe des situations où l'erreur minimale peut être égale à 0 (dans ce cas le coefficient β n'est pas nécessaire et le coefficient d'étendue α devrait aussi être modifié [1/(0,975-0)]). Si à l'époque nous avions opté pour un coefficient β égal à 0,025 c'est parce qu'il existe aussi le cas de figure du sujet dont le taux d'exactitude vaut 0% et qui utilise systématiquement le degré de certitude 5 (dont la valeur centrale est égale à 97,5%). Pour ce sujet dont la méconnaissance est totale, MEC vaut 0,975 et [1 – MEC] vaut 0,025. Ce qui entraînerait SANS les coefficients un *Rs* égal à 0,025, légèrement supérieur à la valeur 0 attendue, d'où le choix du coefficient β pour correction de l'erreur minimale égal à 0,025 et du coefficient α pour correction d'étendue égal à 1,0526 (1/(0,975-0,025)).

A partir du moment où la nouvelle échelle simplifiée des pourcentages de certitude adoptée dans le cadre des épreuves MOHICAN ne faisait plus appel aux valeurs centrales et reprenait les pourcentages de certitude 0% et 100%, une nouvelle formule plus élégante pouvait être envisagée. Il s'agit de la formule « MOHICAN » que nous allons détailler ci-après.

b) Consigne et formule « MOHICAN »

Etant donné le public visé par les épreuves MOHICAN (les étudiants entrant en 1ère candidature à l'université qui ne sont pas habitués à l'utilisation des degrés de certitude), les concepteurs du projet ont préféré simplifier l'échelle des certitudes (DC = Degré de Certitude, %C = pourcentage de Certitude) :

```
DC = 0        1        2        3        4        5
     |--------|--------|--------|--------|--------|
%C = 0%      20%      40%      60%      80%     100%
```

D'une part l'échelle n'est plus dissymétrique et d'autre part, la lecture qui en est demandée aux étudiants est du type : « *Voici une échelle comportant six probabilités subjectives de réussite (les certitudes) : 0%, 20%, 40%, 60%, 80% et 100% de chances que la réponse soit correcte. A vous de choisir l'un de ces six pourcentages en fonction des chances que vous accordez à votre réponse d'être correcte*. ». Il ne s'agit donc plus de dire dans quel intervalle de certitude on se situe, mais de quel pourcentage proposé dans la consigne on est le plus proche. Dès lors la formule du score de *R*éalisme d'un *s*ujet devient :

$$\boxed{Rs = 100 - EMAC}$$ (voir p. 184)

Avec :

$EMAC = \dfrac{\sum_{i=1}^{nc} (|C_i - TE_i|).NU_i}{NR}$ = l'*E*rreur *M*oyenne *A*bsolue de *C*ertitude

i = *i*ndice des degrés de certitude

nc = *n*ombre de degrés de *c*ertitude (ici $nc = 6$)

C_i = valeur de la *C*ertitude i (en pourcents)

NC_i = *N*ombre de réponses *C*orrectes pour la certitude i

NU_i = *N*ombre d'*U*tilisations de la certitude i *(si* $NU_i = 0$ alors l'indice i est ignoré)

TE_i = *T*aux d'*E*xactitude de la certitude i (en pourcents) = 100 x NC_i / NU_i

NR = *N*ombre total de *R*éponses pour test ($\Sigma_i NU_i$)

Cette formule a été exposée en détail lors de la présentation du principe de la turbo analyse (p. 186)

c) Réalisme moyen du groupe (*Rg*)

Calculé sur les données de l'ensemble des scores de *R*éalisme des *s*ujets (*Rs*) ayant participé à un test, le réalisme moyen pour le groupe (*Rg*) nous fournit une indication sur la façon dont le groupe s'est auto-évalué face aux questions proposées lors de l'épreuve.

$$\boxed{Rg = \frac{\sum_{i=1}^{ns} Rs_i}{ns}}$$ (60)

avec :

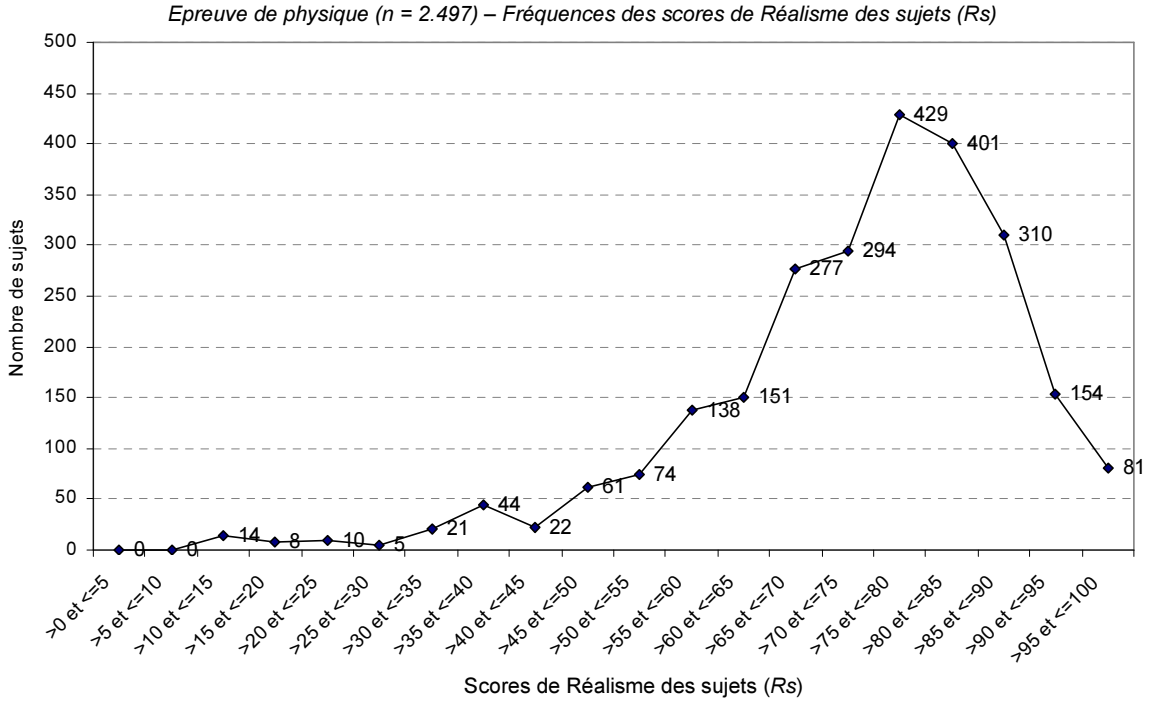
i = *i*ndice de chaque sujet

Rs_i = score de *R*éalisme obtenu par chaque *s*ujet

ns = le *n*ombre de *s*ujets ayant participé au test

2. Comparaison avec les normes de la FAPSE-ULg

Voici la distribution des fréquences des scores de *R*éalisme des *s*ujets (*Rs*) pour l'épreuve de physique soumise à 2.497 étudiants (graphique généré par *SCANTEST 2.0*).



Les performances en réalisme semblent bonnes pour ce test dans la mesure où on observe un regroupement des scores *Rs* vers la droite du graphique.

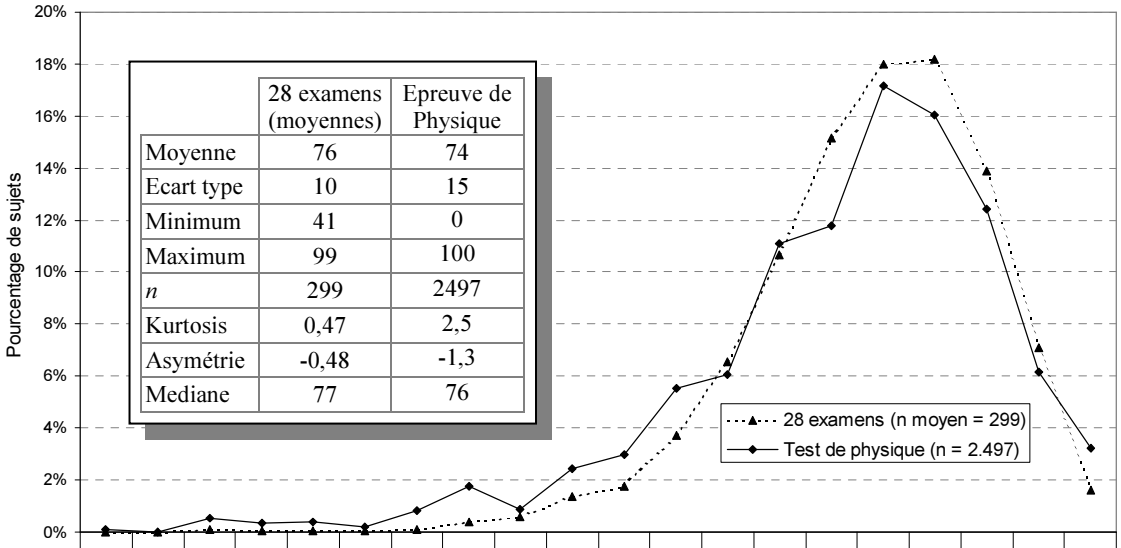
Lors d'une recherche portant sur les données de 28 examens ayant recours aux QCM et aux degrés de certitude entre 1994 et 1996 à la Faculté de Psychologie et des Sciences de l'Education (FAPSE) de l'Université de Liège (ULg), nous avons déjà pu observer cette tendance aux scores de réalisme élevés (Gilles, 1996a).

Dans le cadre de cette étude le réalisme avait alors été calculé à l'aide de la formule de la « Consigne FAPSE » exposée plus haut. Le nombre moyen de sujets pour ces 28 examens[65] qui eurent lieu dans le premier cycle d'étude était de 299.

[65] Le *n* total pour les 28 examens confondus s'élève à 8.385.

Le graphique ci-dessous permet de comparer les deux courbes des fréquences en pourcentages des scores de réalisme. La courbe en pointillés représente les moyennes des fréquences des scores *Rs* obtenus aux 28 examens FAPSE et la courbe en trait continu les fréquences des scores *Rs* pour l'épreuve de physique MOHICAN.

Fréquence des scores de Réalisme des sujets (Rs) en pourcentages



	28 examens (moyennes)	Epreuve de Physique
Moyenne	76	74
Ecart type	10	15
Minimum	41	0
Maximum	99	100
n	299	2497
Kurtosis	0,47	2,5
Asymétrie	-0,48	-1,3
Mediane	77	76

	>0 et <=5	>5 et <=10	>10 et <=15	>15 et <=20	>20 et <=25	>25 et <=30	>30 et <=35	>35 et <=40	>40 et <=45	>45 et <=50	>50 et <=55	>55 et <=60	>60 et <=65	>65 et <=70	>70 et <=75	>75 et <=80	>80 et <=85	>85 et <=90	>90 et <=95	>95 et <=100
28 examens (n moyen = 299)	0,0%	0,0%	0,1%	0,0%	0,0%	0,1%	0,1%	0,4%	0,6%	1,4%	1,8%	3,7%	6,5%	10,7%	15,2%	18,0%	18,2%	13,9%	7,1%	1,6%
Test de physique (n = 2.497)	0,1%	0,0%	0,6%	0,3%	0,4%	0,2%	0,8%	1,8%	0,9%	2,4%	3,0%	5,5%	6,0%	11,1%	11,8%	17,2%	16,1%	12,4%	6,2%	3,2%

Réalisme

3. Constat et question à propos des distributions des fréquences des scores *Rs*

Nous constatons que malgré des publics et des contextes d'évaluation différents, les deux distributions des fréquences en pourcentages sont très proches et très décentrées par leurs modes vers la droite, donc plutôt en J.

Les valeurs des moyennes et écarts type, sont assez proches : d'une part pour l'épreuve de physique *Rs* moyen = 74 et écart type = 15 et d'autre part pour les 28 examens FAPSE *Rs* moyen[66] = 76 et écart type = 10. [3.3] Cette similitude avec la distribution des fréquences des scores de réalisme des 28 examens FAPSE s'observe-t-elle dans les autres épreuves MOHICAN ?

Précédemment nous nous sommes posé la question d'une éventuelle relation entre les scores de réalisme des sujets et leurs performances en termes de pourcentages de réponses correctes. Autrement dit, plus les étudiants sont réalistes, plus leurs taux d'exactitude sont-ils élevés ? [3.4] Observe-t-on une corrélation entre les Taux d'Exactitude (*TE*) et les scores de Réalisme des sujets (*Rs*) lors des épreuves MOHICAN ?

[66] En fait 0,76 sur l'échelle allant de 0 à 1 dans la consigne FAPSE. Nous avons multiplié par 100 la moyenne, l'écart type, le minimum, le maximum et la médiane obtenus pour les 28 examens FAPSE afin de faciliter la comparaison avec les statistiques de l'épreuve de physique calculées à l'aide de la formule MOHICAN.

F. Fréquences et statistiques descriptives des scores de Centration *moyenne du* groupe *(*Cg*)*

1. L'indice de *C*entration d'un *s*ujet (*Cs*)

L'indice de Centration calculé pour chaque sujet (*Cs*) est basé sur la différence entre la Certitude Moyenne (*CMs*) et le Taux d'Exactitude (*TEs*) calculée selon la formule proposée par Leclercq (1982, 1993) « *…difference between Average central values of confidence degrees (or average confidence) and Rate of correct answers (percentage)* » pour un étudiant donné. Les *Cs* montrent en fonction du signe une tendance à la surestimation lorsque le signe est positif ou une tendance à la sous-estimation lorsque le signe est négatif.

$$\boxed{Cs = CMs - TEs} \quad (61)$$

avec pour chaque sujet *s* :

CMs = la *C*ertitude *M*oyenne du sujet pour l'ensemble des questions du test (exprimé en pourcents)
TEs = le *T*aux d'*E*xactitude de ce sujet (exprimé en pourcents)

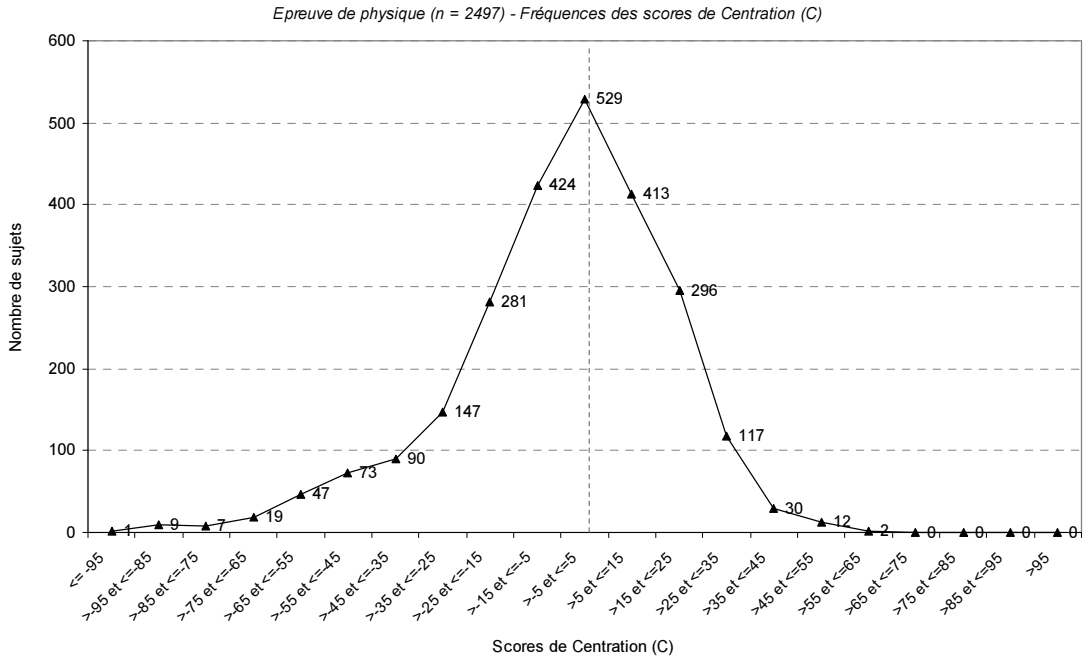
et la *C*ertitude *M*oyenne d'un *s*ujet *s* est donnée par :

$$\boxed{CMs = \frac{\sum_{q=1}^{nq} C_{sq}}{nq}} \quad (62)$$

avec :

q = l'indice des questions
C_{sq} = *C*ertitude fournie par le sujet *s* pour la question *q*
nq = nombre de questions dans le test

Voici la distribution des scores de centration pour le test de physique (n = 2.497).



Pour ce test de physique, on observe plus d'étudiants qui se sous-estiment (Cs⩽-5, *n* = 870, 35%) que d'étudiants qui se surestiment (Cs>5, *n* = 1098, 44%) et une minorité d'étudiants dont l'auto-estimation est excellente (-5<*Cs*⩽5, *n* = 524, = 21%). Le nombre 5% est ici pris de manière purement conventionnel, on aurait par exemple pu prendre 10% ou 15%.

2. L'indice de Centration moyenne du *g*roupe (*Cg*)

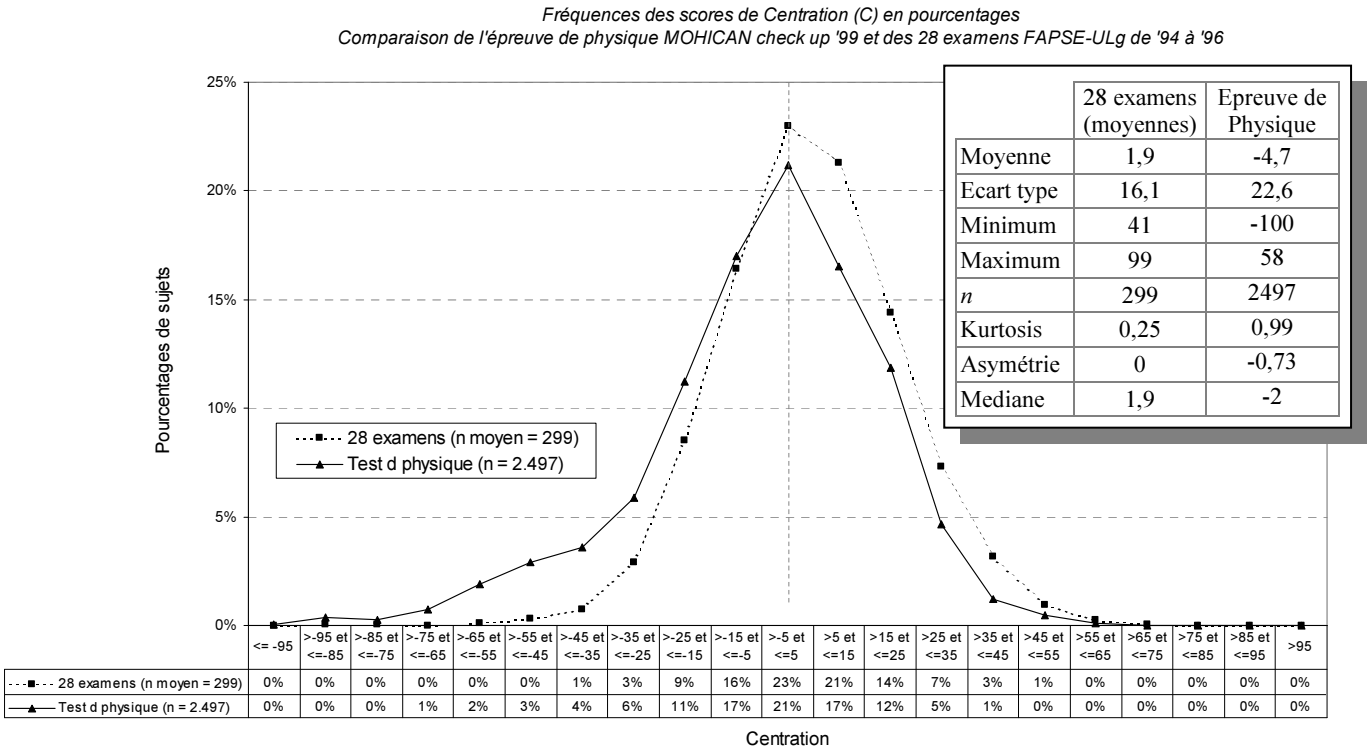
Comme pour l'indice de réalisme, nous proposons de calculer un score moyen de Centration pour le groupe (*Cg*) (exprimé en pourcents) en sommant toutes les *C*entrations des sujets (*Cs*) ayant participé à l'épreuve et en divisant cette somme par le nombre de participants :

$$Cg = \frac{\sum_{s=1}^{ns} Cs_s}{ns} \quad (63)$$

avec :

s = l'indice des *s*ujets
Cs_s = score de *C*entration obtenu par le *s*ujet *s* pour l'ensemble des questions du test (en pourcents)
ns = le *n*ombre de *s*ujets ayant participé au test.

3. Comparaison avec les normes de la FAPSE-Ulg



	28 examens (moyennes)	Epreuve de Physique
Moyenne	1,9	-4,7
Ecart type	16,1	22,6
Minimum	41	-100
Maximum	99	58
n	299	2497
Kurtosis	0,25	0,99
Asymétrie	0	-0,73
Mediane	1,9	-2

	<= -95	>-95 et <=-85	>-85 et <=-75	>-75 et <=-65	>-65 et <=-55	>-55 et <=-45	>-45 et <=-35	>-35 et <=-25	>-25 et <=-15	>-15 et <=-5	>-5 et <=5	>5 et <=15	>15 et <=25	>25 et <=35	>35 et <=45	>45 et <=55	>55 et <=65	>65 et <=75	>75 et <=85	>85 et <=95	>95
28 examens (n moyen = 299)	0%	0%	0%	0%	0%	0%	1%	3%	9%	16%	23%	21%	14%	7%	3%	1%	0%	0%	0%	0%	0%
Test d physique (n = 2.497)	0%	0%	0%	1%	2%	3%	4%	6%	11%	17%	21%	17%	12%	5%	1%	0%	0%	0%	0%	0%	0%

Comme pour l'indice de réalisme, malgré des publics et des contextes d'évaluation différents, nous observons des courbes de distribution forts proches. Les pourcentages d'auto-estimations correctes des sujets (*Cs* de -5 à +5), sans trop de surestimations ni trop de sous-estimations, sont forts proches : 21% pour le test de physique et 23% pour les 28 examens de 1994 à 1996. Lors des 28 examens, la tendance à la

surestimation était présente ($5<Cs\leqslant 100$ = 46%) par rapport aux sous-estimations moins fréquentes chez les étudiants ($-100<Cs\leqslant -5$ = 29%).

4. Constat et question à propos des distributions des fréquences des scores *Cs*

Nous constatons d'une part une tendance à la sous-estimation dans le cadre du test de physique alors que la tendance était à la surestimation dans le contexte des 28 examens de la FAPSE et d'autre part des pourcentages quasi similaires d'auto-estimations excellentes ($-5<Cs\leqslant 5$).

[3.4] Observe-t-on chez les sujets des autres épreuves MOHICAN une tendance à la sous-estimation comme dans l'épreuve de physique ?

PARTIE III

Exploration spectrale et classique des "check up '99" MOHICAN

Chapitre IX :

Exploration du niveau « TEST »

Sommaire

A. Introduction

Nous venons d'exposer différents indices et outils d'analyse spectrale de la qualité des épreuves (voir p. 211). Précédemment, nous avons aussi présenté une série d'indices classiques habituellement utilisés lors des analyses de tests (p. 118). Pour introduire ces instruments, nous avons jusqu'à présent exploité les données d'une seule épreuve : celles du test de physique soumis à 2.497 étudiants.

Nous allons maintenant examiner les données des 10 épreuves MOHICAN (p. 93) à l'aide de ces indices spectraux et classiques. Dans un premier temps, dans un souci de réflexion structurée, nous proposerons une classification des différents indices proposés. Ensuite nous explorerons les résultats des épreuves en segmentant le travail en fonction des niveaux d'analyse définis dans notre classification. Enfin, après exploration de chaque niveau d'analyse nous reprendrons les questions que nous nous sommes posées lors des présentations des indices spectraux (elles sont balisées par « [ ] » accompagné d'un numéro d'ordre et de la page de référence de la question) et y apporterons les réponses à la lumière des observations effectuées

Dans cette partie nous utiliserons les appellations suivantes pour désigner les 10 tests MOHICAN qui seront analysés :

11) **VOCABU** = épreuve de vocabulaire (45 questions, 3.846 étudiants, voir annexe p. 482) ;
12) **SYNTAX** = épreuve de syntaxe et articulation logique (12 questions, 3.739 étudiants, voir annexe p. 487) ;
13) **COMPRE** = épreuve de compréhension (6 questions, 3.420 étudiants, voir annexe p. 489) ;
14) **GEOGRA** = épreuve de lecture de documents et géographie (10 questions, 3.688 étudiants, voir annexe p. 490) ;
15) **HISTOI** = épreuve de connaissances en histoire et socio économie (25 questions, 1.410 étudiants, annexe p. 494) ;
16) **ARTACT** = épreuve de connaissances artistiques (25 questions, 1.392 étudiants, voir annexe p. 497) ;
17) **MATHEM** = épreuve de mathématiques (22 questions, 2.516 étudiants, voir annexe p. 500) ;
18) **BIOLOG** = épreuve de biologie (10 questions, 2.507 étudiants, voir annexe p. 505) ;
19) **CHIMIE** = épreuve de chimie (8 questions, 2.501 étudiants, voir annexe p. 507) ;
20) **PHYSIQ** = épreuve de physique (10 questions, 2.497 étudiants, voir annexe p. 509).

B. Classification des indices d'analyse de la qualité des épreuves MOHICAN

Nous proposons de ranger les indices selon trois niveaux de profondeur d'exploration des épreuves. Un premier niveau « TEST » où les indices doivent nous aider à obtenir une vue d'ensemble de la qualité globale de l'épreuve. Un second niveau « QCM » où le but est de nous fournir un aperçu qualitatif du fonctionnement global de chaque question et un troisième niveau « PROPOSITION » où l'idée est de livrer des informations qualitatives spécifiques à chaque proposition au sein des QCM.

Nous envisageons aussi pour chacun de ces trois niveaux d'analyse, trois catégories d'indices.

- Les indices spectraux qui livrent une série d'informations : (1) sur la cohérence d'utilisation des pourcentages de certitude au niveau des propositions (*rpbis SC*, p. 178, *rpbis SCT*, p. 184), des questions (*NCSq*, p. 225) et des épreuves (*NCSt*, p. 259), (2) sur la qualité de réalisation des prédictions par question (*Rq*, p. 242) et par test (*Rt*, p. 264) ainsi que les tendances à la surestimation ou à la sous-estimation au sein des questions (*Cq*, p. 254) ou au sein des tests (*Ct*, p. 270) et (3) sur la qualité des auto-estimations des groupes de sujets (*Rg*, p. 272 et *Cg*, p. 277). Nous intitulons ce premier ensemble « *Indices spectraux* ».

- Une seconde catégorie est constituée des « *Indices classiques sur matrices binaires* ». Nous utilisons l'appellation « …*classiques*… » car il s'agit de statistiques habituellement présentées lors des analyses de test (voir p. 118). Cette appellation est suive de « …*sur matrices binaires* » car nous calculons ici ces indices à l'aide de matrices où les réponses fournies par les étudiants sont codées de façon binaire (soit « 1 », soit « 0 ») en fonction de l'exactitude (voir p. 132).

- Le troisième ensemble d'indices appelé « *Indices classiques sur matrices spectrales* » concerne aussi les statistiques classiques mais ici elles sont calculées à l'aide d'un deuxième type de matrice où les réponses sont notées en utilisant les pourcentages de certitude que l'on fait précéder du signe « + » ou « - » selon l'aspect correct ou incorrect de la réponse (voir p. 132).

Le tableau ci-après reprend cette classification des indices. La signification des abréviations des statistiques qui figurent dans les différentes cases sont rappelées sous le tableau (nous indiquons entre parenthèses les pages où ces indices sont expliqués).

ANALYSE DES EPREUVES – INDICES :

NIVEAUX D'ANALYSE		SPECTRAUX	CLASSIQUES SUR MATRICE BINAIRE	CLASSIQUES SUR MATRICE SPECTRALE
	TEST	*a) NCSt* *b) Rt* *c) pit* *d) Ct* *e) Rg* *f) Cg*	*a)* pot_{mb} *b)* $r_{xx'mb}$ *c)* $rS_{xx'mb}$ *d)* rG_{mb} *e)* α_{mb} *f)* $kq[\alpha=0,8]_{mb}$ *g) NCIt*	*a)* $r_{xx'ms}$ *b)* $rS_{xx'ms}$ *c)* rG_{ms} *d)* α_{ms} *e)* $kq[\alpha=0,8]_{ms}$
	QCM	*g) NCSq* *h) Rq* *i) piq* *j) Cq*	*h)* poq_{mb} *i) NCIq* *j)* $\alpha\text{-}q_{mb\cdot}$ *k)* $r_{qt\,mb\cdot}$ *l)* $cov_{qxqy\,mb\cdot}$ *m)* $r_{qxqy\,mb}$	*f)* $\alpha\text{-}q_{ms\cdot}$ *g)* $r_{qt\,ms\cdot}$ *h)* $cov_{qxqy\,ms\cdot}$ *i)* $r_{qxqy\,ms}$
	PROP	*k) rpbis SC* *l) CMp* *m) N Rép.* *n) % Rép.*	*n) N Rép.* *o) % Rép.* *p) rpbis classique*	

1. Rappel des significations des abréviations des indices spectraux

*a) NCSt = N*iveau de *C*ohérence *S*pectrale d'un *t*est (p. 261) ;
*b) Rt = R*éalisation des prédictions au niveau d'un *t*est (p. 264) ;
*c) pit = facilité/difficulté i*ntrospective d'un *t*est (p. 268) ;
*d) Ct = C*entration moyenne d'un *t*est (p. 270) ;
*e) Rg = R*éalisme d'un *g*roupe de sujets (p. 272) ;
*f) Cg = C*entration moyenne d'un *g*roupe de sujets (p. 277) ;
*g) NCSq = N*iveau de *C*ohérence *S*pectrale d'une *q*uestion (p. 231) ;
*h) Rq = R*éalisation des prédictions au niveau d'une *q*uestion (p. 242) ;
*i) piq = facilité/difficulté i*ntrospective d'une *q*uestion (p. 251) ;
*j) Cq = C*entration moyenne d'une *q*uestion (p. 270) ;
*k) rpbis SC = C*oe*fficient de corrélation p*oint *bis*érial *S*pectral *C*ontrasté (p. 178) ;
*l) CMp = C*ertitude *M*oyenne par *p*roposition (pp. 217 et 252) ;
*m) N Rép. = n*ombre de *Rép*onses calculé après *T*urbo analyse pour une proposition (p. 217) ;
n) % Rép. = pourcentage de *Rép*onses calculé après *T*urbo analyse pour une proposition (p. 217).

Rappelons que le principe de la turbo analyse (voir p. 186) peut être appliqué au calcul de ces indices spectraux.

2. Rappel des significations des abréviations des indices classiques

a) Indices classiques calculés à l'aide de matrices binaires

a) pot_{mb} = *facilité/difficuté o*bjective d'un *t*est calculée à partir d'une *m*atrice *b*inaire (pp. 270 et 120) ;
b) $r_{xx'mb}$ = coefficient de bipartition d'un *t*est calculé à partir d'une *m*atrice *b*inaire (p. 130) ;
c) $rS_{xx'mb}$ = correction de Spearman-Brown appliquée au coefficient de bipartition calculé à partir d'une *m*atrice *b*inaire (p. 131) ;
d) rG_{mb} = coefficient de bipartition de Guttman calculé à partir d'une *m*atrice *b*inaire (p. 131) ;
e) α_{mb} = coefficient alpha de Cronbach calculé à partir d'une *m*atrice *b*inaire (p. 137) ;
f) $kq[\alpha 0{,}8]_{mb}$ = nombre de questions à ajouter ou retrancher au test pour obtenir un alpha égal à 0,8 dans le cas d'une matrice binaire (p. 139) ;
g) *NCIt* = *N*iveau de *C*ohérence *I*nterne du *t*est (p. 261) ;
h) poq_{mb} = indice de *facilité/difficulté o*bjective de la *q*uestion calculé à partir de la matrice binaire (p. 254) ;
i) *NCIq* = Niveau de Cohérence Interne de la question (p. 233) ;
j) $\alpha\text{-}q_{mb}$ = alpha obtenu par le test lorsqu'on retire la question *q* calculé à partir d'une *m*atrice *b*inaire (p. 141) ;
k) $r_{qt\,mb}$ = corrélation question-total calculé à partir d'une *m*atrice *b*inaire (p. 141) ;
l) $r_{qxqy\,mb}$ = corrélation des scores d'une question *x* avec les scores d'une question *y* à partir des résultats d'une *m*atrice *b*inaire (p. 143) ;
m) $cov_{qxqy\,mb}$ = covariance des scores d'une question *x* avec les scores d'une question *y* à partir des résultats d'une *m*atrice *b*inaire (p. 142) ;
n) *N Rép.* = *n*ombre de *Rép*onses pour une proposition (p. 216) ;
o) *% Rép.* = *pourcentage* de *Rép*onses pour une proposition (p. 216) ;
p) *rpbis* classique = *Coefficient de corrélation p*oint *bis*érial classique (pp. 171 et 216).

b) Indices classiques calculés à l'aide de matrices spectrales

a) $r_{xx'ms}$ = coefficient de bipartition d'un *t*est calculé à partir d'une *m*atrice *s*pectrale (p. 130) ;
b) $rS_{xx'ms}$ = correction de Spearman-Brown appliquée au coefficient de bipartition calculé à partir d'une *m*atrice *s*pectrale (p. 131) ;
c) rG_{ms} = coefficient de bipartition de Guttman calculé à partir d'une *m*atrice *s*pectrale (p. 131) ;
d) α_{ms} = coefficient alpha de Cronbach calculé à partir d'une *m*atrice *s*pectrale (p. 137) ;
e) $kq[\alpha 0{,}8]_{ms}$ =nombre de questions à ajouter ou retrancher au test pour obtenir un alpha égal à 0,8 dans le cas d'une *m*atrice *s*pectrale (p. 139) ;
f) $\alpha\text{-}q_{ms}$ = alpha obtenu par le test lorsqu'on retire la question *q* calculé à partir d'une *m*atrice *s*pectrale (p. 141) ;
g) $r_{qt\,ms}$ = corrélation question-total calculé à partir d'une *m*atrice *s*pectrale (p. 141) ;
h) $r_{qxqy\,ms}$ = corrélation des scores d'une question *x* avec les scores d'une question *y* à partir des résultats d'une *m*atrice *s*pectrale (p. 143) ;
i) $cov_{qxqy\,ms}$ = covariance des scores d'une question *x* avec les scores d'une question *y* à partir des résultats d'une *m*atrice *s*pectrale (p. 142).

C. Analyse de la qualité des tests à l'aide des indices spectraux

Dans la partie précédente nous avons défini une série d'instruments d'analyse de la qualité spectrale des tests (voir p. 259). Quelles sont les performances des 10 épreuves MOHICAN à ces nouveaux indices spectraux ? Quelles interprétations peut-on donner aux valeurs obtenues ? Des différences de qualité spectrale apparaissent-elles lorsqu'on compare les 10 épreuves ? Quelles est l'ampleur des liaisons entre indices spectraux ?

1. Comparaison des effectifs des épreuves aux paliers turbo

Les indices de qualité spectrale peuvent être calculés à différents niveaux de turbo analyse ce qui permet d'accroître la confiance dans les informations obtenues étant donné qu'aux niveaux les plus élevés nous ne prenons en compte que les données des sujets qui commenttent moins d'erreurs dans leurs auto-estimations. Cependant, plus nous « montons » dans les paliers de turbo analyse, moins nous avons de sujets. Dès lors, il est important de connaître les effectifs à ces différents paliers. En particulier, les effectifs sont-ils suffisants aux paliers T80 et T90 ?

Rappelons brièvement le paramétrage de la turbo analyse (voir principe, p. 186) qui a été appliquée au calcul des indices exposés ci-après pour chaque épreuve MOHICAN. Nous avons choisi de fixer le seuil de *R*éalisme des *s*ujets (*Rs*) (voir p. 184) inférieur initial à 0, le seuil *Rs* supérieur à 100 et le pas à 10. Ce paramétrage de la turbo analyse entraîne 10 traitements successifs. Le premier est fait en tenant compte des sujets dont le réalisme (*Rs*) est compris entre 0 (seuil inférieur initial) et 100 (seuil supérieur), c'est-à-dire tous les étudiants. Ensuite, un second traitement est fait en tenant compte des données des étudiants dont le réalisme est compris entre 10 (seuil inférieur initial augmenté du pas égal à 10) et 100 (niveau supérieur). Ce second traitement est suivi d'un troisième où les données sont celles des étudiants dont le réalisme est compris entre 20 (seuil inférieur de l'analyse précédente augmenté du pas) et 100 (niveau supérieur). Les traitements se succèdent ainsi franchissant à chaque fois un nouveau palier de turbo analyse jusqu'à ce que le seuil inférieur soit égal au seuil supérieur. En tout, ce paramétrage aura permis de calculer les indices spectraux à 10 paliers de Turbo analyse : T0, T10, T20, T30, T40, T50, T60, T70, T80 et T90. Voici les effectifs des 10 épreuves aux 10 paliers de turbo analyse de T0 à T90 avec un pas de 10 :

	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ
T0	3846	3739	3420	3685	1410	1392	2516	2507	2501	2497
T10	3846	3734	3413	3683	1409	1392	2514	2507	2500	2494
T20	3843	3720	3394	3678	1407	1391	2499	2497	2482	2472
T30	3838	3699	3348	3649	1405	1387	2465	2485	2461	2458
T40	3815	3658	3208	3600	1399	1380	2434	2442	2419	2394
T50	3773	3553	2860	3474	1391	1363	2398	2340	2341	2342
T60	3602	3318	2279	3116	1340	1323	2278	2036	2096	2129
T70	3089	2691	1481	2303	1101	1137	1956	1511	1661	1724
T80	1806	1611	756	1162	584	681	1276	757	909	1007
T90	350	439	251	253	73	124	339	151	232	275

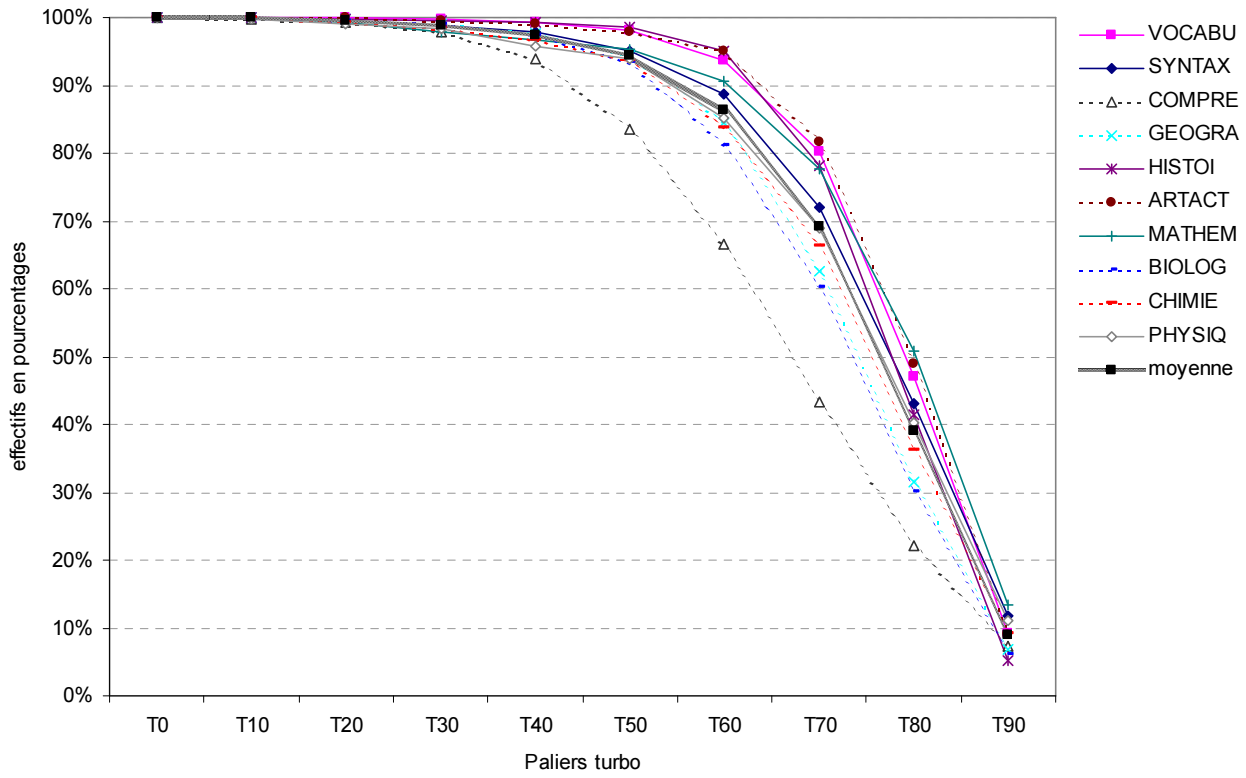
Pour permettre la comparaison nous avons transformé ces nombres en pourcentages :

	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ	% moyen
T0	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
T10	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
T20	100%	99%	99%	100%	100%	100%	99%	100%	99%	99%	100%
T30	100%	99%	98%	99%	100%	100%	98%	99%	98%	98%	99%
T40	99%	98%	94%	98%	99%	99%	97%	97%	97%	96%	97%
T50	98%	95%	84%	94%	99%	98%	95%	93%	94%	94%	94%
T60	94%	89%	67%	85%	95%	95%	91%	81%	84%	85%	86%
T70	80%	72%	43%	62%	78%	82%	78%	60%	66%	69%	69%
T80	47%	43%	22%	32%	41%	49%	51%	30%	36%	40%	39%
T90	9%	12%	7%	7%	5%	9%	13%	6%	9%	11%	9%

Nous avons ajouté une colonne « % moyen » au tableau précédent, cette colonne reprend la moyenne des pourcentages des 10 épreuves à chaque palier de turbo analyse.

Les courbes du graphique ci-dessous reprennent les pourcentages des 10 épreuves ainsi que la moyenne de ces pourcentages.

Effectifs des épreuves MOHICAN aux 10 paliers de turbo analyse (de T0 à T90 avec un pas de 10)



Nous constatons à l'aide de ce graphique et du tableau qui le précède que les pourcentages d'effectifs aux différents paliers de turbo analyse sont très proches pour la plupart des épreuves MOHICAN, sauf pour l'épreuve de compréhension (COMPRE) qui se démarque des autres par des pourcentages plus faibles. Au palier de turbo analyse T80, l'effectif de COMPRE est à 22% de son effectif total alors que les effectifs des 9 autres épreuves se situent dans une fourchette approximative de 30 à 50% des effectifs totaux.

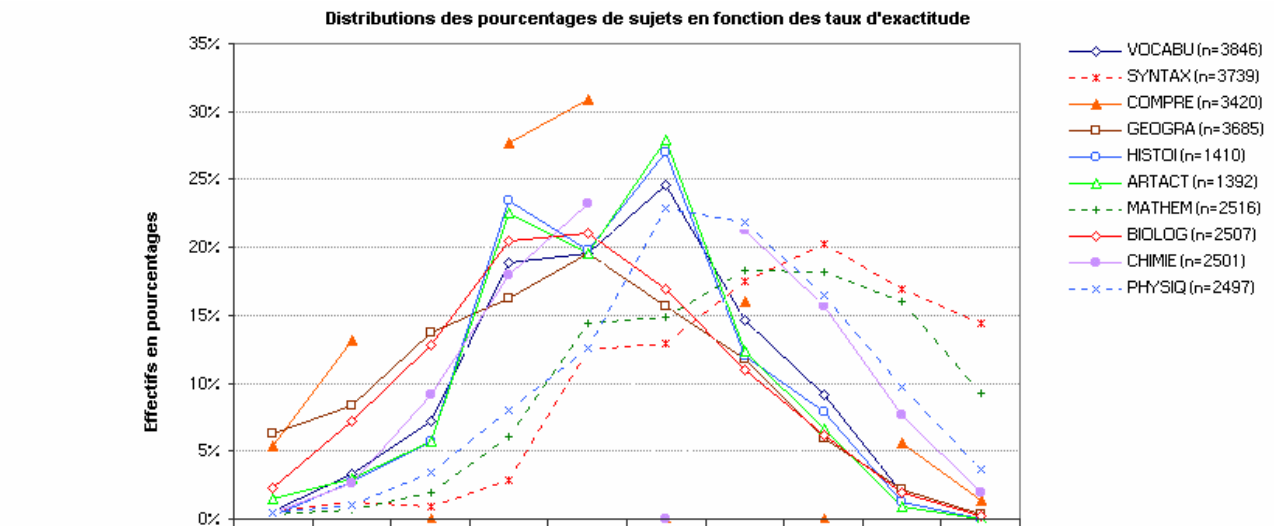
Sur le premier tableau des effectifs en nombres nous remarquons qu'au palier T90 l'épreuve de connaissances en histoire et socio-économie (HISTOI) ne comporte plus que 73 sujets. Les autres épreuves à ce palier turbo ont toutes des effectifs supérieurs à 150 sujets. Au palier de turbo analyse T80 les effectifs sont tous élevés avec un *n* égal à 583 dans le cas de l'épreuve « HISTOI ». Parmi les 10 tests, c'est celle qui compte le moins de sujets à ce palier de turbo analyse T80.

En ce qui concerne la moyenne des effectifs en pourcentage (dernière colonne du tableau précédent) pour le palier T80 nous voyons qu'elle se situe à 39 % des effectifs totaux. A T90, cette moyenne globale chute à 9%.

Nous observons donc pour les 10 épreuves MOHICAN d'importants effectifs de sujets dont le réalisme est supérieur ou égal à 80 (en moyenne 1.007), donc qui commettent moins de 20% d'erreurs de certitude dans leurs auto-estimations. Dans le contexte MOHICAN, les effectifs disponibles permettent donc d'envisager le calcul d'indices de qualité spectrale des tests à un palier élevé de turbo analyse tel que T80.

Nous venons de voir qu'aux paliers de turbo analyse T80 et T90 nous conservons d'importants effectifs de sujets. Voyons maintenant comment se profilent les distributions des pourcentages de sujets en fonction des taux d'exactitude à ces paliers turbo élevés. A T80 et T90 les courbes sont-elles différentes de celles observées lorsque tous les sujets sont pris en compte ?

Voici d'abord les courbes liées aux effectifs en pourcentages pour dix catégories de taux de réussites. Dans le graphique qui suit ces pourcentages ont été calculés au départ des données de tous les sujets ayant participé à chacune des 10 épreuves MOHICAN.



Taux d'exactitude -->	>=0% et <=10%	>10% et <=20%	>20% et <=30%	>30% et <=40%	>40% et <=50%	>50% et <=60%	>60% et <=70%	>70% et <=80%	>80% et <=90%	>90% et <=100%
VOCABU (n=3846)	0,6%	3,3%	7,2%	18,9%	19,6%	24,5%	14,6%	9,2%	1,9%	0,3%
SYNTAX (n=3739)	0,6%	1,2%	0,9%	2,9%	12,5%	12,9%	17,5%	20,2%	16,9%	14,4%
COMPRE (n=3420)	5,3%	13,2%	0,0%	27,6%	30,9%	0,0%	16,0%	0,0%	5,6%	1,4%
GEOGRA (n=3685)	6,2%	8,3%	13,8%	16,2%	19,5%	15,7%	11,8%	6,0%	2,2%	0,3%
HISTOI (n=1410)	0,3%	2,7%	5,7%	23,4%	19,8%	27,0%	12,0%	7,9%	1,3%	0,0%
ARTACT (n=1392)	1,5%	2,9%	5,7%	22,6%	19,5%	27,9%	12,4%	6,6%	0,9%	0,0%
MATHEM (n=2516)	0,3%	0,7%	1,9%	6,1%	14,4%	14,9%	18,3%	18,2%	16,0%	9,2%
BIOLOG (n=2507)	2,3%	7,3%	12,8%	20,4%	21,1%	17,0%	10,9%	6,1%	1,9%	0,2%
CHIMIE (n=2501)	0,5%	2,6%	9,2%	17,9%	23,2%	0,0%	21,3%	15,7%	7,6%	2,0%
PHYSIQ (n=2497)	0,4%	1,0%	3,4%	8,0%	12,6%	22,9%	21,8%	16,5%	9,7%	3,6%

Les courbes des 10 épreuves sont assez gaussiennes. Nous remarquons que 3 épreuves se dégagent des autres par un décalage vers la droite : il s'agit des tests SYNTAX, MATHEM et PHYSIQ (courbes en pointillés). Ces 3 épreuves se caractérisent donc par de plus gros pourcentages d'étudiants dans les catégories de taux d'exactitude les plus élevées. Remarquons que les courbes de COMPRE (test de 6 questions) et de CHIMIE (8 questions) sont incomplètes étant donné les 10 intervalles des taux d'exactitude et leur nombre de questions inférieur à 10, (il est donc impossible d'obtenir des données pour certaines catégories, par exemple si le test ne comporte que 6 questions alors la catégorie de taux d'exactitude « >20% et <=30% » est systématiquement vide).

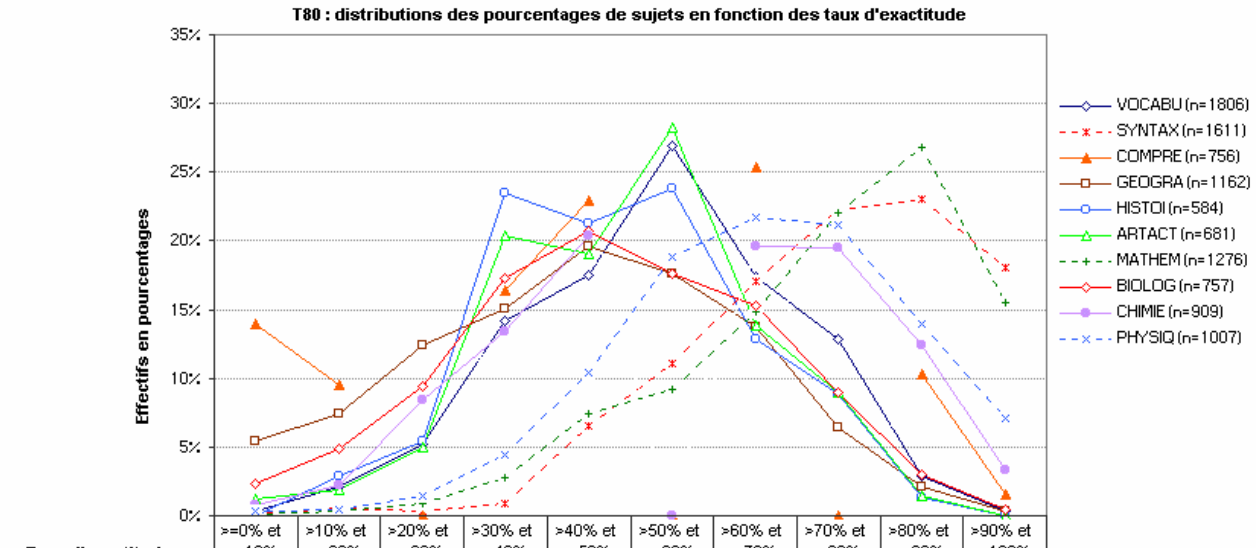
Voici maintenant le tableau des effectifs à T80 (lorsqu'on sélectionne les données des sujets dont le réalisme est supérieur ou égal à 80) pour les 10 catégories de taux d'exactitude que nous avons définies.

Taux d'exactitude	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ
>=0% et <=10%	6	2	53	63	0	8	2	18	7	3
>10% et <=20%	38	9	36	86	17	13	4	37	20	5
>20% et <=30%	94	5	0	144	32	34	12	71	77	15
>30% et <=40%	256	15	62	175	137	139	35	131	122	45
>40% et <=50%	317	106	87	228	124	130	94	157	185	105
>50% et <=60%	487	179	0	204	139	192	118	133	0	190
>60% et <=70%	315	274	96	159	75	94	190	116	178	219
>70% et <=80%	233	359	0	75	52	61	281	68	177	213
>80% et <=90%	53	372	39	24	8	10	342	23	113	141
>90% et <=100%	7	290	6	4	0	0	198	3	30	71

Niveau de turbo analyse T80 : tableau des nombres de sujets en fonction des taux d'exactitude

Remarquons que pour la dernière catégorie des taux d'exactitudes supérieur à 90% les effectifs des tests SYNTAX, MATHEM et PHYSIQ sont sensiblement plus élevés que pour les autres épreuves.

Comment se profilent les courbes des pourcentages d'effectifs lorsqu'on sélectionne les données des seuls sujets dont le réalisme est supérieur ou égal à 80 (T80) ?



Taux d'exactitude -->	>=0% et <=10%	>10% et <=20%	>20% et <=30%	>30% et <=40%	>40% et <=50%	>50% et <=60%	>60% et <=70%	>70% et <=80%	>80% et <=90%	>90% et <=100%
VOCABU (n=1806)	0,3%	2,1%	5,2%	14,2%	17,6%	27,0%	17,4%	12,9%	2,9%	0,4%
SYNTAX (n=1611)	0,1%	0,6%	0,3%	0,9%	6,6%	11,1%	17,0%	22,3%	23,1%	18,0%
COMPRE (n=756)	14,0%	9,5%	0,0%	16,4%	23,0%	0,0%	25,3%	0,0%	10,3%	1,6%
GEOGRA (n=1162)	5,4%	7,4%	12,4%	15,1%	19,6%	17,6%	13,7%	6,5%	2,1%	0,3%
HISTOI (n=584)	0,0%	2,9%	5,5%	23,5%	21,2%	23,8%	12,8%	8,9%	1,4%	0,0%
ARTACT (n=681)	1,2%	1,9%	5,0%	20,4%	19,1%	28,2%	13,8%	9,0%	1,5%	0,0%
MATHEM (n=1276)	0,2%	0,3%	0,9%	2,7%	7,4%	9,2%	14,9%	22,0%	26,8%	15,5%
BIOLOG (n=757)	2,4%	4,9%	9,4%	17,3%	20,7%	17,6%	15,3%	9,0%	3,0%	0,4%
CHIMIE (n=909)	0,8%	2,2%	8,5%	13,4%	20,4%	0,0%	19,6%	19,5%	12,4%	3,3%
PHYSIQ (n=1007)	0,3%	0,5%	1,5%	4,5%	10,4%	18,9%	21,7%	21,2%	14,0%	7,1%

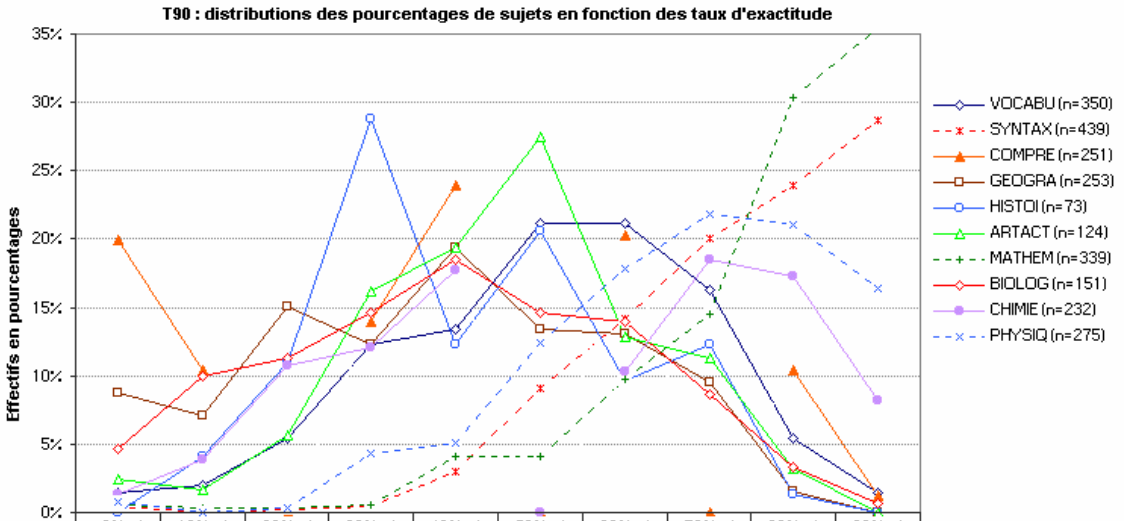
Grosso modo l'allure d'ensemble des 10 courbes est assez similaire à celle des 10 courbes du premier graphique où tous les sujets étaient pris en compte quelles que soient leurs performances en réalisme. Nous constatons que les trois épreuves qui se dégagent des autres par leur plus grande facilité (SYNTAX, MATHEM et PHYSIQ) sont ici un peu plus décalées vers la droite, elles sont encore mieux réussies par les sujets dont le réalisme est égal ou supérieur à 80.

Qu'en est-il lorsqu'on sélectionne les données des sujets dont le réalisme est supérieur ou égal à 90 (T90) ?

Taux d'exactitude	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ
>=0% et <=10%	5	2	50	22	0	3	2	7	3	2
>10% et <=20%	7	0	26	18	3	2	1	15	9	0
>20% et <=30%	19	1	0	38	8	7	1	17	25	1
>30% et <=40%	43	2	35	31	21	20	2	22	28	12
>40% et <=50%	47	13	60	49	9	24	14	28	41	14
>50% et <=60%	74	40	0	34	15	34	14	22	0	34
>60% et <=70%	74	62	51	33	7	16	33	21	24	49
>70% et <=80%	57	88	0	24	9	14	49	13	43	60
>80% et <=90%	19	105	26	4	1	4	103	5	40	58
>90% et <=100%	5	126	3	0	0	0	120	1	19	45

Niveau de turbo analyse T90 : tableau des nombres de sujets en fonction des taux d'exactitude

Le graphique suivant montre les courbes des 10 épreuves à T90.



Taux d'exactitude -->	>=0% et <=10%	>10% et <=20%	>20% et <=30%	>30% et <=40%	>40% et <=50%	>50% et <=60%	>60% et <=70%	>70% et <=80%	>80% et <=90%	>90% et <=100%
VOCABU (n=350)	1,4%	2,0%	5,4%	12,3%	13,4%	21,1%	21,1%	16,3%	5,4%	1,4%
SYNTAX (n=439)	0,5%	0,0%	0,2%	0,5%	3,0%	9,1%	14,1%	20,0%	23,9%	28,7%
COMPRE (n=251)	19,9%	10,4%	0,0%	13,9%	23,9%	0,0%	20,3%	0,0%	10,4%	1,2%
GEOGRA (n=253)	8,7%	7,1%	15,0%	12,3%	19,4%	13,4%	13,0%	9,5%	1,6%	0,0%
HISTOI (n=73)	0,0%	4,1%	11,0%	28,8%	12,3%	20,5%	9,6%	12,3%	1,4%	0,0%
ARTACT (n=124)	2,4%	1,6%	5,6%	16,1%	19,4%	27,4%	12,9%	11,3%	3,2%	0,0%
MATHEM (n=339)	0,6%	0,3%	0,3%	0,6%	4,1%	4,1%	9,7%	14,5%	30,4%	35,4%
BIOLOG (n=151)	4,6%	9,9%	11,3%	14,6%	18,5%	14,6%	13,9%	8,6%	3,3%	0,7%
CHIMIE (n=232)	1,3%	3,9%	10,8%	12,1%	17,7%	0,0%	10,3%	18,5%	17,2%	8,2%
PHYSIQ (n=275)	0,7%	0,0%	0,4%	4,4%	5,1%	12,4%	17,8%	21,8%	21,1%	16,4%

Les courbes des trois tests SYNTAX, MATHEM et PHYSIQ sont encore plus décalées vers la droite. Les effectifs en pourcentages des tests SYNTAX, MATHEM et PHYSIQ sont encore plus élevés que ceux des autres tests lorsque nous prenons en considération les deux dernières catégories de taux d'exactitudes entre 80% et 90% et entre 90% et 100%.

Trois tests se dégagent donc des autres par leur plus grande facilité objective, il s'agit de SYNTAX, MATHEM et PHYSIQ. En principe ces trois épreuves objectivement plus faciles devraient aussi être ressenties subjectivement comme telles par les sujets les plus réalistes. Nous remarquons qu'à T80 et T90 le décalage vers la droite des courbes de ces trois tests s'accentue. A cette augmentation de la facilité observée (les taux d'exactitude) devrait aussi correspondre une augmentation de la facilité prédite (les certitudes moyennes) étant donné qu'à ces deux paliers de turbo analyse les plus élevés les sujets commettent moins d'erreurs dans leurs auto-estimations. C'est ce que nous vérifierons plus loin à l'aide de l'indice de Centration par test (Ct).

2. L'indice du *N*iveau de *Cohérence* Spectrale d'un *t*est (*NCSt*)

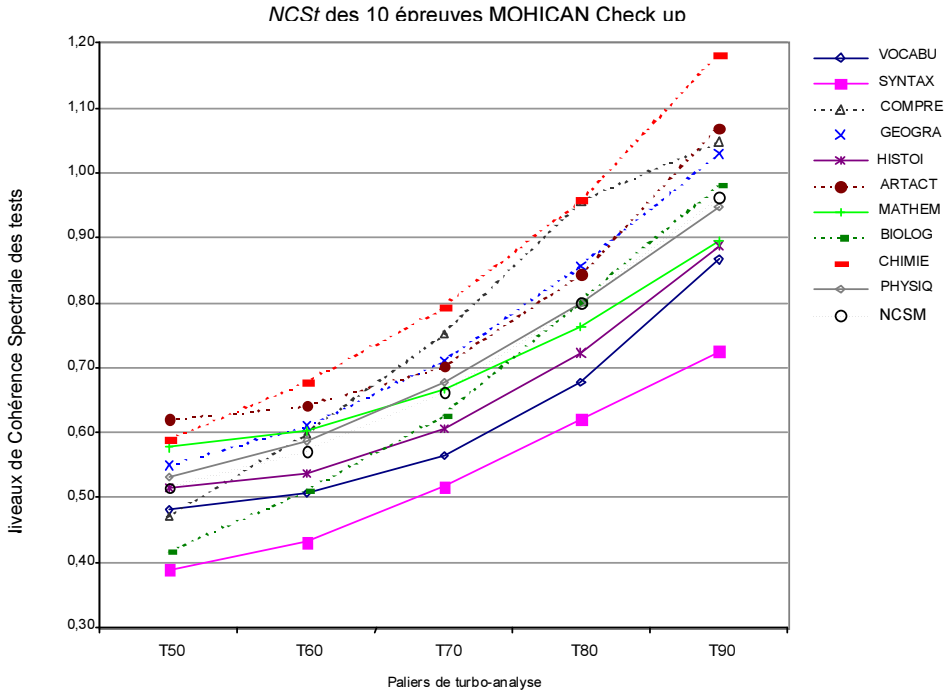
Globalement pour l'ensemble des réponses fournies, nous nous attendons à observer dans chaque test une propension des sujets à utiliser des certitudes plus élevées lorsqu'ils répondent correctement que lorsqu'ils répondent incorrectement (ce que nous appelons la « cohérence spectrale »). Mais cette propension mesurée à l'aide du NCSt sera-t-elle la même dans les 10 tests ? En ce qui concerne les valeurs récoltées après turbo analyse, nous nous attendons à une augmentation du NCSt au fur et à mesure de la progression dans les différents paliers turbo. Observera-t-on des différences de progression pour les 10 épreuves MOHICAN ?

L'indice du Niveau de Cohérence Spectrale d'un test (*NCSt*, p. 261) est calculé à partir des *rpbis SCT*. L'indice *NCSt* est en fait la moyenne des Niveaux de Cohérence Spectrale des questions (*NCSq*) de l'épreuve et les *NCSq* sont obtenus en soustrayant la valeur de la moyenne pondérée des *rpbis SCT* des propositions incorrectes à la valeur du *rpbis SCT* de la réponse correcte. La fiabilité de ces indices est donc renforcée par le fait qu'on leur applique une turbo analyse : plus le seuil exigé en réalisme est élevé, plus les sujets sélectionnés sont réalistes et plus leurs données sont cohérentes en ce qui concerne l'utilisation des pourcentages de certitude. Le *NCSt* peut être considéré comme un indice de la qualité de cohérence d'utilisation des pourcentages de certitude dans le cadre d'une épreuve. Rappelons que la valeur de l'indice *NCSt* varie entre –2 et +2. Quelle est l'évolution de l'indice *NCSt* aux différents paliers de turbo analyse pour les 10 épreuves MOHICAN ? Grosso modo, de 0,5 (T0) à 1 (T90).

	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ	*NCSM*
T0	0,46	0,33	0,32	0,50	0,50	0,60	0,51	0,35	0,52	0,47	0,46
T10	0,46	0,33	0,32	0,51	0,50	0,60	0,51	0,35	0,52	0,47	0,46
T20	0,46	0,34	0,33	0,51	0,50	0,60	0,52	0,35	0,53	0,48	0,46
T30	0,47	0,35	0,34	0,51	0,50	0,60	0,54	0,36	0,54	0,49	0,47
T40	0,47	0,37	0,39	0,52	0,51	0,61	0,56	0,37	0,56	0,50	0,49
T50	0,48	0,39	0,47	0,55	0,52	0,62	0,58	0,42	0,59	0,53	0,51
T60	0,51	0,43	0,60	0,61	0,54	0,64	0,60	0,51	0,68	0,59	0,57
T70	0,57	0,52	0,75	0,71	0,61	0,70	0,67	0,63	0,79	0,68	0,66
T80	0,68	0,62	0,96	0,86	0,72	0,84	0,76	0,80	0,96	0,80	0,80
T90	0,87	0,72	1,05	1,03	0,89	1,07	0,89	0,98	1,18	0,95	0,96

Nous avons ajouté une colonne *NCSM* (Niveaux de Cohérence Spectrale des épreuves Mohican) dans le tableau ci-dessus. Cette colonne reprend la moyenne des valeurs des *NCSt* des 10 tests aux 10 paliers de turbo analyse. Nous remarquons que les valeurs *NCSt* des 10 épreuves aux paliers de turbo analyse T0 à T50 sont très proches.

Afin d'alléger le graphique ci-contre nous n'avons repris que les valeurs des paliers T50 à T90.



Nous observons sur ce graphique qu'au palier T80, là où les étudiants sont performants dans leurs auto-estimations et où ils sont aussi encore très nombreux, six épreuves obtiennent un *NCSt* supérieur ou égal à la moyenne des *NCSt* de l'ensemble des épreuves *M*OHICAN (*NCSM* = 0,80). Il s'agit, dans l'ordre de classement (du meilleur au moins bon *NCSt*), des épreuves : (1) CHIMIE [0,96] et COMPRE [0,96], (2) GEOGRA [0,86], (3) ARTACT [0,84] et (4) PHYSIQ [0,80] et BIOLOG [0,80].

Quatre autres épreuves se situent sous cette moyenne, il s'agit dans le même ordre de classement des épreuves : (5) MATHEM [0,76], (6) HISTOI [0,72], (7) VOCABU [0,68] et (8) SYNTAX [0,62].

Les étudiants soumis à l'épreuve SYNTAX (la plus facile de toutes) ont donc utilisé les pourcentages de certitude de façon moins cohérente que dans les 9 autres épreuves MOHICAN. Nous reviendrons plus loin sur cette épreuve et sur les 3 autres dont le *NCSt* se situe sous la moyenne lors de l'exploration du niveau « QCM ».

Bien que la plage théorique des valeurs du NCSt varie de –2 à +2, la moyenne des 10 épreuves fluctue entre 0,46 et 0,96. Le fait qu'on observe uniquement des valeurs positives montre que globalement au niveau des tests il existe une propension à utiliser des pourcentages de certitude plus élevés lorsque les réponses sont correctes que lorsqu'elles sont incorrectes. Des différences entre tests apparaissent, ainsi à T80 le NCSt le plus élevé est celui de CHIMIE (0,96) et le moins élevé est celui de SYNTAX (0,62). Lorsque nous comparons les NCSt calculés aux différents paliers de la turbo analyse nous remarquons que les données des sujets qui commettent moins d'erreurs dans leurs auto-estimations sont plus cohérentes d'un point de vue spectral. Enfin, nous remarquons que les classements des épreuves aux différents paliers de turbo analyse restent relativement stables sauf pour COMPRE et GEOGRA.

3. L'indice de *R*éalisation des prédictions au niveau d'un *t*est (*Rt*)

Lorsque nous globalisons l'information livrée par les indices de Réalisation des prédictions par question (Rq) au niveau d'un test à l'aide de l'indice Rt, nous obtenons un éclairage sur la tendance à récolter en moyenne pour l'ensemble des questions de l'épreuve des Taux d'Exactitude Annoncés (TEA) en concordance avec les Taux d'Exactitude Observés (TEO). Nous nous attendons à des variations d'une épreuve à l'autre mais sans pour autant connaître l'ampleur des différences. Par ailleurs nous nous demandons si les valeurs de Rt sont corrélées avec celles de l'indice NCSt. Une liaison faible ou nulle renforcerait l'idée d'une complémentarité des deux indices qui, bien que voisins, ne mesurent pas selon nous les mêmes propriétés.

Rappelons que l'indice *Rt* (voir p. 264) d'une épreuve est calculé en effectuant la moyenne des indices *Rq* (voir méthode de calcul, p. 242) et que le *Rq* nous informe sur la concordance des *TEA* par rapport aux *TEO* dans les résultats de chaque QCM. Rappelons enfin que la plage des valeurs pour les indices *Rq* et *Rt* varie de 0 à 100.

Voici le tableau des valeurs obtenues par les indices *Rt* des 10 épreuves calculés au palier de turbo analyse T0 (les données de tous les étudiants sont prises en compte) et T80 (seules les données des sujets dont *Rs* ⩾ 80 sont utilisées). Nous avons ajouté une colonne « *RtM* » qui reprend la moyenne des *Rt* des 10 tests calculés à un palier donné.

	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ	*RtM*
Rt T0	80,9	83,3	71,7	82,4	80,2	82,7	79,7	80	87	79,7	80,8
Rt T80	86,1	87,5	83	85,1	84,8	87,8	89,6	90,2	92,7	89,4	87,6

Nous voyons que l'épreuve COMPRE récolte le moins bon score *Rt* à T0 (71,7) et à T80 (83).

Au palier de turbo analyse T80, cinq tests récoltent des *Rt* supérieurs à la moyenne (87,6) : CHIMIE (92,7), BIOLOG (90,2), MATHEM (89,6), PHYSIQ (89,4) et ARTACT (87,8). Les cinq autres

tests dont l'indice *Rt T80* se situe sous la moyenne sont : SYNTAX (87,5), VOCABU (86,1), GEOGRA (85,1), HISTOI (84,8) et COMPRE (83).

Lorsque nous corrélons les valeurs de *Rt T80* et de *NCSt T80* (voir point précédent, p. 295) calculées pour les dix tests nous n'observons pas de liaison (r = 0,09).

Lorsque nous comparons les valeurs de Rt calculé à T0 et à T80 au sein de chaque épreuve, nous remarquons que les différences sont peu élevées. Les variations entre les épreuves sont également peu marquées (surtout à T80). Les différences observées entre les épreuves nous paraissent donc relativement faibles.

L'absence de liaison entre les valeurs des Rt et des NCSt renforce l'idée que les deux indices, bien qu'utilisant les données liées aux pourcentages de certitude, ne mesurent pas les mêmes propriétés. Le NCSt permet d'évaluer au niveau d'une épreuve dans quelle mesure les résultats des questions sont cohérents du point de vue de l'utilisation des certitudes (la propension à utiliser des certitudes plus élevées lorsque la réponse est correcte que lorsqu'elle est incorrecte) tandis que le Rt nous donne une indication sur la « quantité » d'erreurs d'auto-estimations contenue dans les résultats du test (la concordance moyenne entre les TEA et les TEO pour l'ensemble des questions de l'épreuve).

4. Indice de *facilité i*ntrospective du *t*est (*pit*)

Rappelons que l'utilisation des pourcentages de certitude permet de calculer pour chaque question la facilité introspective (ou la facilité subjective ressentie par les sujets) en effectuant la moyenne des pourcentages de certitude qui ont accompagnés les réponses (voir p. 251). La facilité introspective par question (piq) peut être globalisée au niveau d'une épreuve en effectuant la somme des piq et en divisant cette somme par le nombre de questions (voir p. 268). On obtient alors l'indice de facilité introspective par test (pit). Le principe de turbo analyse peut aussi être appliqué à l'indice pit. Aux paliers élevés de la turbo analyse les valeurs de pit reposent alors sur les données des répondants qui s'auto-estiment en commettant en moyenne moins d'erreurs de certitude. Nous nous attendons à des différences entre les pit récoltés par les 10 épreuves. Les moyennes des pourcentages de certitude utilisés dans les données des questions seront-elles très différentes d'une épreuve à l'autre ? La facilité introspective calculée au niveau des tests augmentera-t-elle ou diminuera-t-elle au fur et à mesure des paliers de turbo analyse ?

Voici le tableau des valeurs obtenues par les 10 épreuves MOHICAN à l'indice *pit* calculé aux 10 paliers de turbo analyse. Nous y avons ajouté une colonne intitulée « *piM* » qui reprend la facilité/difficulté introspective moyenne pour l'ensemble des 10 épreuves Mohican.

	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ	*piM*
T0	0,52	0,69	0,60	0,48	0,50	0,50	0,70	0,55	0,54	0,62	0,57
T10	0,52	0,69	0,60	0,48	0,50	0,50	0,70	0,55	0,54	0,62	0,57
T20	0,52	0,70	0,60	0,48	0,50	0,50	0,70	0,55	0,55	0,62	0,57
T30	0,52	0,70	0,59	0,49	0,50	0,50	0,71	0,55	0,55	0,62	0,57
T40	0,52	0,70	0,59	0,49	0,50	0,50	0,71	0,55	0,55	0,63	0,57
T50	0,52	0,71	0,58	0,49	0,50	0,50	0,72	0,55	0,55	0,64	0,57
T60	0,52	0,72	0,55	0,49	0,50	0,50	0,73	0,56	0,56	0,66	0,58
T70	0,52	0,73	0,53	0,50	0,50	0,50	0,74	0,55	0,57	0,68	0,58
T80	0,53	0,75	0,48	0,52	0,50	0,50	0,76	0,55	0,58	0,71	0,59
T90	0,53	0,79	0,41	0,52	0,45	0,49	0,81	0,50	0,58	0,76	0,58

Nous constatons que les *pit* n'évoluent pratiquement pas entre le palier T0 et le palier T50, ce qui s'explique par les faibles différences d'effectifs observées pour ces paliers, ce sont pratiquement les mêmes sujets de T0 à T50 (voir p. 290).

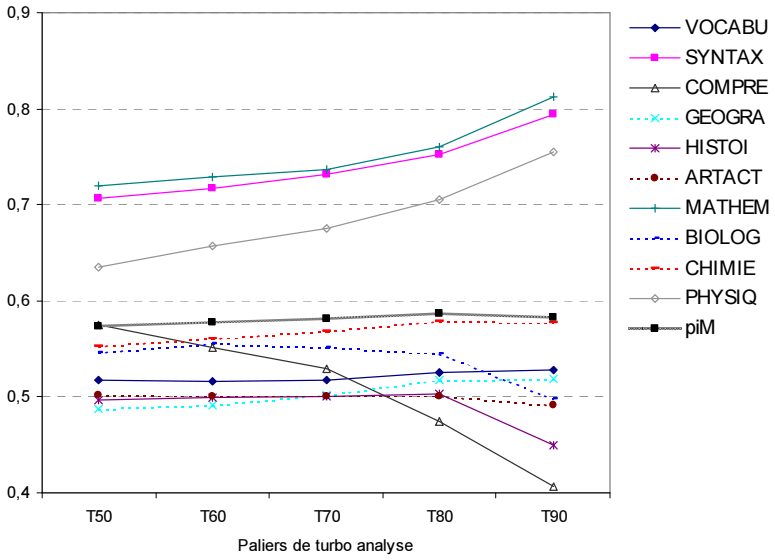
Etant donné les faibles différences observées entre T0 et T50 et dans le but d'alléger le graphique, nous reprenons dans les courbes qui suivent les valeurs des indices *pit* de T50 à T90. Bien que l'indice *pit* puisse varier entre 0 et 1, nous avons choisi 0,4 comme valeur de départ et 0,9 comme valeur finale pour l'échelle des ordonnées de manière à rendre le graphique encore plus lisible.

Nous rencontrons trois types de situations dans les épreuves MOHICAN :

- pour les tests « MATHEM », « SYNTAX » et « PHYSIQ » la facilité introspective augmente sensiblement avec l'élévation des paliers de turbo analyse, les épreuves paraissent plus faciles aux étudiants qui commettent en moyenne moins d'erreurs d'auto-estimations ;

- pour les épreuves « CHIMIE », « VOCABU », « GEOGRA » et « ARTACT » » la facilité introspective reste relativement stable ;

Comparaison des facilité/difficulté introspectives par test (pit)



- Enfin, pour les tests « COMPRE », « HISTOI » et « BIOLOG » la facilité introspective diminue, donc, pour les étudiants les plus réalistes ces épreuves paraissent plus difficiles que pour les étudiants qui commettent en moyenne plus d'erreurs dans leurs auto-estimations, c'est particulièrement le cas pour l'épreuve « COMPRE ».

Remarquons que pour les épreuves relativement stables et les épreuves qui diminuent du point de vue de la facilité introspective, les valeurs récoltées par les indices *pit* se situent en-dessous de la moyenne *piM* calculée pour les 10 épreuves.

Confrontons maintenant pour les 10 épreuves ces indices de facilité introspective avec les indices de facilité objective.

Pour quatre épreuves nous remarquons que les valeurs à l'indice de facilité subjective du test restent relativement stables (de T0 à T90) et se situent entre 0,48 et 0,58. Pour trois autres épreuves nous remarquons que plus nous « montons » dans les paliers de turbo analyse, à partir de T50, plus la facilité subjective a tendance à augmenter. Pour ces trois épreuves, moins les sujets commettent d'erreurs dans leurs auto-estimations et plus ils ressentent les questions comme étant faciles. Enfin, pour trois autres épreuves, nous constatons le phénomène inverse : moins les sujets commettent d'erreurs dans leurs auto-estimations et moins ils ressentent les questions comme étant faciles (à partir de T50).

Les prédictions (facilité subjective ressentie par les sujets) se vérifieront-elles dans les faits (facilité objective ou taux d'exactitude récoltés par les sujets) ? Logiquement oui car plus les paliers de turbo analyse seront élevés et plus les prédictions coïncideront avec la réalité, les sujets commettant de moins en moins d'erreurs dans leurs auto-estimations. C'est ce que nous allons vérifier dans la sous-section suivante.

5. Indice de Centration moyenne d'un test (*Ct*)

Nous venons de voir qu'à partir du palier T50 de turbo analyse, la facilité subjective des épreuves était soit stable (4 tests), soit augmentait (3 tests), soit diminuait (3 tests). Qu'en est-il lorsqu'on confronte la facilité subjective à la facilité objective (les taux d'exactitude) aux différents paliers de turbo analyse ? Etant donné la prise en compte des données des étudiants qui commettent de moins en moins d'erreurs dans leurs auto-estimations, plus les paliers turbo seront élevés et plus les valeurs des indices pit et pot devraient avoir tendance à se rapprocher, ce que nous allons vérifier dans cette sous-section à l'aide de l'indice de Centration par test (Ct). Par ailleurs étant donné que l'indice Ct mesure les sur (valeur positive) ou sous-estimations (valeur négative), nous nous attendons à des corrélations assez élevées entre l'indice Ct calculé en valeur absolue et l'indice Rt qui permet aussi d'évaluer la quantité d'erreurs d'auto-estimations contenue dans les résultats. Par contre la liaison entre les indices Ct et NCSt devrait, elle, être nulle ou très faible (comme dans le cas de Rt).

Rappelons qu'il est possible de calculer la *facilité o*bjective d'une *q*uestion (*poq*) en calculant la proportion de réponses correctes récoltées par la question (voir p. 254). En effectuant la moyenne des indices *poq* pour une épreuve on aboutit à un indice de *facilité o*bjective du test (*pot*, voir p. 270). En soustrayant la valeur de l'indice *pot* à la valeur de l'indice de *facilité i*ntrospective (ou subjective) du test (*pit*) on obtient l'indice de *C*entration du *t*est (*Ct*, voir p. 270).

Voici le tableau des valeurs récoltées par les indices *facilité o*bjective et de *facilité i*ntrospective des *t*ests MOHICAN. Nous les avons calculées aux 10 paliers de la turbo analyse :

		T00	T10	T20	T30	T40	T50	T60	T70	T80	T90
VOCABU	*pit*	0,52	0,52	0,52	0,52	0,52	0,52	0,52	0,52	0,53	0,53
	pot	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,51	0,52	0,53
SYNTAX	*pit*	0,69	0,69	0,70	0,70	0,70	0,71	0,72	0,73	0,75	0,79
	pot	0,69	0,69	0,69	0,69	0,69	0,70	0,71	0,72	0,74	0,78
COMPRE	*pit*	0,60	0,60	0,59	0,59	0,59	0,57	0,55	0,53	0,47	0,41
	pot	0,44	0,44	0,44	0,44	0,44	0,45	0,46	0,46	0,44	0,39
GEOGRA	*pit*	0,48	0,48	0,48	0,48	0,49	0,49	0,49	0,50	0,52	0,52
	pot	0,47	0,47	0,47	0,47	0,47	0,47	0,47	0,48	0,48	0,48
HISTOI	*pit*	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,45
	pot	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,49	0,45
ARTACT	*pit*	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,49
	pot	0,49	0,49	0,49	0,49	0,49	0,49	0,49	0,49	0,49	0,48
MATHEM	*pit*	0,70	0,70	0,70	0,71	0,71	0,72	0,73	0,74	0,76	0,81
	pot	0,66	0,66	0,66	0,65	0,66	0,66	0,67	0,69	0,73	0,80
BIOLOG	*pit*	0,55	0,55	0,55	0,55	0,55	0,55	0,56	0,55	0,55	0,50
	pot	0,49	0,49	0,49	0,49	0,49	0,49	0,50	0,51	0,52	0,48
CHIMIE	*pit*	0,54	0,54	0,55	0,55	0,55	0,55	0,56	0,57	0,58	0,58
	pot	0,55	0,55	0,55	0,54	0,54	0,55	0,55	0,56	0,57	0,57
PHYSIQ	*pit*	0,62	0,62	0,62	0,62	0,63	0,64	0,66	0,68	0,70	0,76
	pot	0,66	0,66	0,65	0,65	0,65	0,65	0,66	0,67	0,69	0,75

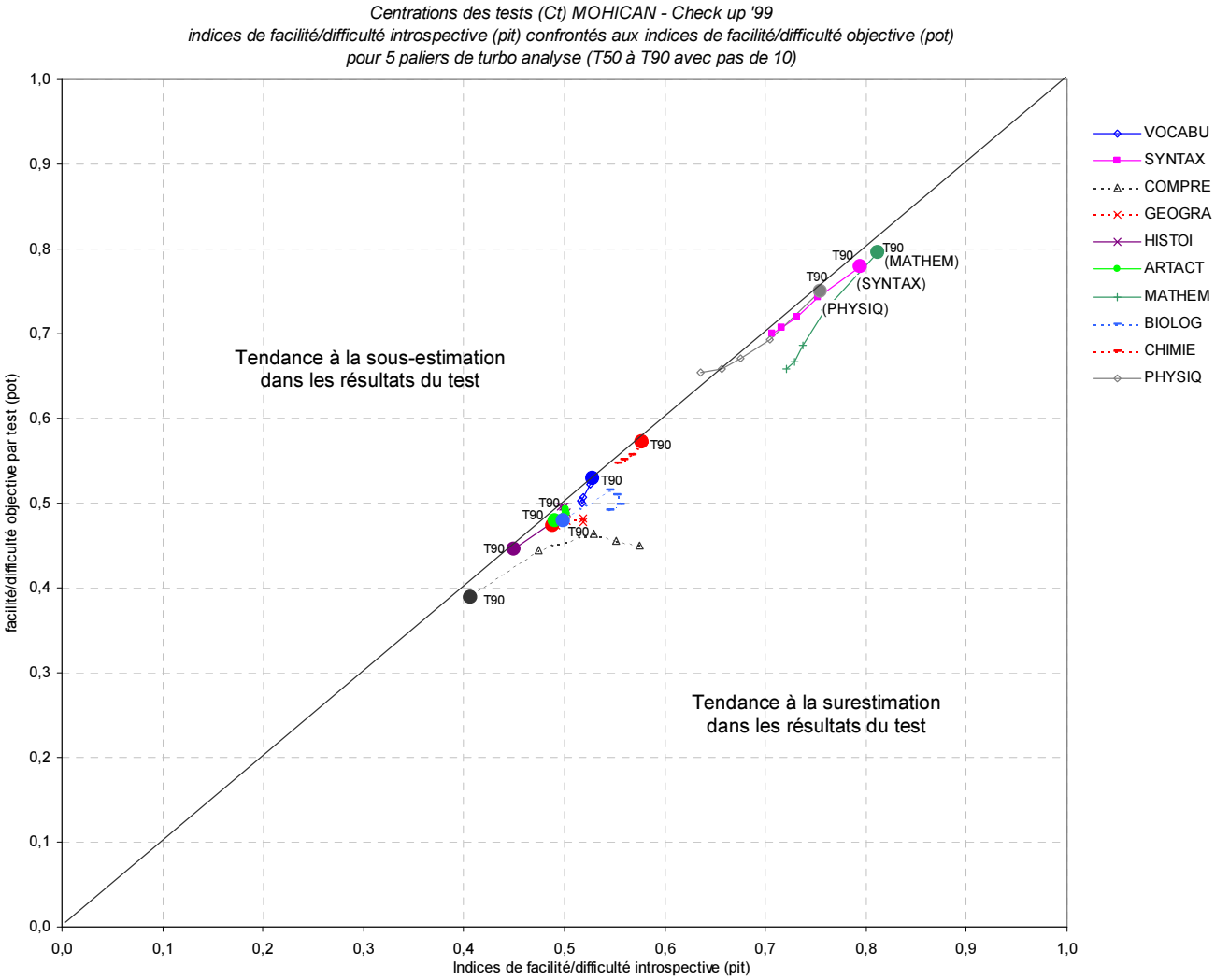
Lorsqu'on compare les données contenues dans ce tableau, on observe des valeurs pour l'indice *pot* en général légèrement inférieures à celles de l'indice *pit* (ce qui indique de la surestimation que nous mesurerons plus loin à l'aide de l'indice *Ct*). Plus on monte dans les paliers de la turbo analyse, plus ces différences s'atténuent, ce qui est logique dans la mesure où les indices sont calculés à partir des données des étudiants qui commettent de moins en moins d'erreurs dans leurs auto-estimations.

Afin de faciliter l'interprétation des données du tableau qui précède, nous proposons de visualiser les informations dans un graphique à deux dimensions.

Nous pouvons représenter par un point la valeur de l'indice *pit* (axe des abscisses) confrontée à celle de l'indice *pot* (axe des ordonnées) à l'aide d'un graphique en nuage de points tel que celui ci-

dessous. En reliant les différents points ainsi obtenus par une épreuve aux différents paliers de turbo analyse, nous traçons une courbe par test. Comme nous constatons que les valeurs des indices *pit* et *pot* ne varient guère de T0 à T50, nous n'avons repris sur le graphique que les valeurs correspondant aux paliers turbo T50 à T90. La valeur correspondant au palier turbo le plus élevé (T90) est signalée par un rond situé à un bout de la courbe de l'épreuve. Pour améliorer la lisibilité du graphique nous avons défini les valeurs minimum et maximum des échelles des axes respectivement à 0,4 et 0,9 (en fait les indices *pit* et *pot* varient de 0 à 1). La diagonale qui traverse le graphique est un repère qui montre les positions que pourraient prendre les points si il y avait concordance parfaite entre la facilité introspective et la facilité objective d'un test.

Rappelons que lorsque les points se situent sous la diagonale les valeurs des *pit* sont plus élevées que celles des *pot* ce qui indique une propension à la surestimation. Le graphique montre que c'est le cas pour les 10 épreuves MOHICAN, mais de façon peu marquée, les points étant en même temps proches de la diagonale.

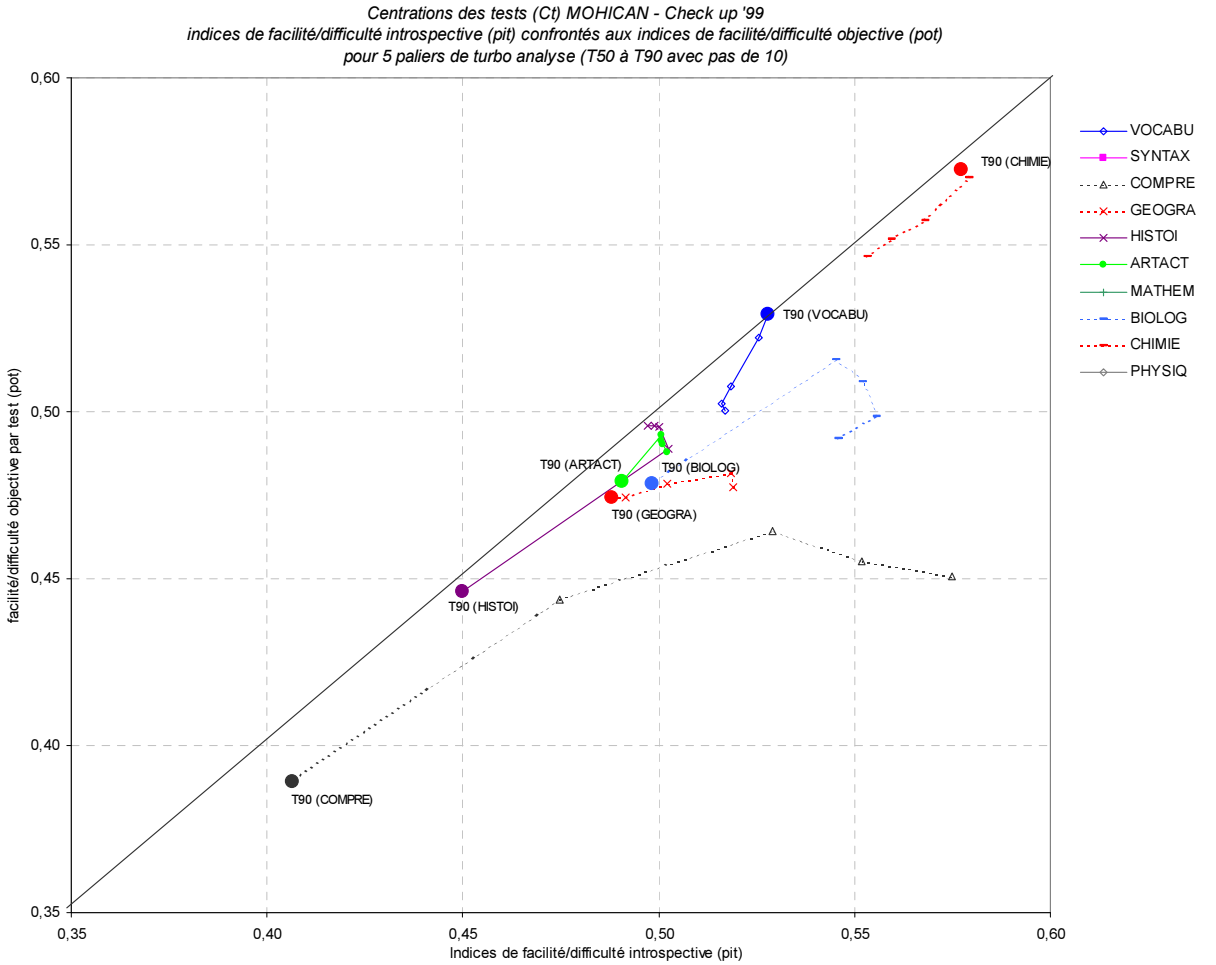


Nous constatons que tous les ronds correspondant aux points tracés à T90 sont tous proches de la diagonale. Nous remarquons trois3 tests qui se démarquent des autres par des courbes situées plus près du coin supérieur droit : MATHEM, SYNTAX et PHYSIQ. Ces épreuves déjà signalées lors de l'analyse des valeurs obtenues à l'indice *pit* (p. 298) sont ressenties comme étant faciles par les étudiants. Cette facilité introspective se confirme donc aussi d'un point de vue objectif. Pour ces trois tests, les traits réunissant les points obtenus aux paliers T50 à T80 forment des courbes relativement longues et parallèles

à la diagonale. Ces courbes montantes indiquent que les indices *pit* et *pot* sont assez proches et augmentent sensiblement et de concert aux paliers turbo élevés.

Agrandissons la zone du graphique qui contient les sept autres épreuves (voir ci-dessous). Nous remarquons deux types de situations. Premièrement, celle de deux tests dont les courbes sont montantes : VOCABU et CHIMIE. Donc, pour ces deux épreuves plus les étudiants sont réalistes, plus elles sont ressenties comme étant faciles et plus elles le sont objectivement.

Deuxièmement, pour les cinq épreuves qui restent et dont les *pot T90* se situent sous 0,5, les courbes sont descendantes : BIOLOG, ARTACT, GEOGRA, HISTOI et COMPRE. Pour ces cinq épreuves plus les étudiants sont réalistes, plus ils les considèrent comme étant difficiles et plus elles le sont dans les faits.



La distance des points par rapport la diagonale peut se chiffrer en soustrayant la valeur de l'indice *pot* à celle de l'indice *pit*, on obtient alors la valeur de l'indice de Centration du test (*Ct*) dont le signe indique la surestimation (+) ou la sous-estimation (-).

Le tableau ci-après reprend les valeurs obtenues par les épreuves MOHICAN à l'indice *Ct* calculé aux 10 paliers de la turbo analyse. Afin d'améliorer la lisibilité du tableau nous avons multiplié les valeurs obtenues à l'indice *Ct* par 100 (en théorie, l'indice *Ct* varie de –1 à +1 ; dans le tableau ci-dessous où les valeurs sont multipliées par 100, il pourrait donc varier entre –100 et +100). Nous avons ajouté une colonne *CtM* qui reprend la moyenne des *Ct* des 10 épreuves aux différents paliers de turbo analyse.

	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ	*CtM*
T0	1,6	0,1	16,0	1,0	0,2	1,6	4,1	5,7	-0,3	-3,9	2,6
T10	1,6	0,2	16,0	1,1	0,1	1,6	4,2	5,7	-0,2	-3,9	2,6
T20	1,6	0,3	15,8	1,1	0,1	1,6	4,4	5,7	0,0	-3,6	2,7
T30	1,6	0,4	15,4	1,2	0,1	1,7	5,2	5,6	0,4	-3,5	2,8
T40	1,6	0,6	14,3	1,2	0,2	1,6	5,9	6,0	0,9	-2,3	3,0
T50	1,7	0,7	12,4	1,4	0,1	1,4	6,2	5,4	0,7	-1,9	2,8
T60	1,4	0,9	9,7	1,7	0,3	1,1	6,3	5,7	0,8	-0,1	2,8
T70	1,1	1,2	6,5	2,3	0,5	0,9	5,1	4,3	1,1	0,5	2,3
T80	0,3	1	3,1	3,7	1,3	0,8	3,2	3,0	0,9	1,1	1,8
T90	-0,1	1,5	1,7	4,1	0,4	1,1	1,5	2,0	0,5	0,5	1,3

Globalement, nous constatons que lorsqu'on monte dans les paliers de la turbo analyse, les valeurs de l'indice *Ct* sont de moins en moins élevées et ont tendance à se rapprocher de 0.

Nous remarquons aux paliers turbo les moins élevés de l'épreuve COMPRE une tendance assez prononcée à la surestimation (*Ct* T0 = 16). Ce niveau de surestimation se réduit considérablement lorsqu'on monte dans les paliers turbo plus élevés (*Ct* T90 = 1,7), donc plus les étudiants sont réalistes et moins ils commettent d'erreurs de surestimations dans l'utilisation de leurs pourcentages de certitude, ce qui est logique. Pour l'épreuve PHYSIQ la tendance de départ était à la sous-estimation (*Cq* T0 = -3,9) et plus on monte dans les paliers turbo plus on se rapproche du zéro idéal avec un léger dépassement à T80 (*Cq* T80 = 1,1) qui se réduit à T90 (*Cq* T90 = 0,5).

Les valeurs récoltées à l'indice Ct montrent qu'il existe une relation entre certitude et compétence. Au palier de turbo analyse T0 (où les données de tous les sujets sont prises en compte), Ct est très proche de zéro (la centration idéale) dans 9 tests sur 10, seul COMPRE obtient un indice Ct égal à 16 (dans une plage qui varie de -100 à 100) mais c'est aussi le test qui contient le moins de questions (6). Ces valeurs montrent une faible quantité globale de sur ou sous-estimations dans les résultats des épreuves, ce qui témoigne d'une bonne fidélité de la certitude (globalement, à l'échelle d'un test, les prédictions des sujets correspondent assez bien aux taux d'exactitude observés). Logiquement, plus les sujets à partir desquels on calcule Ct sont réalistes (aux paliers de turbo analyse les plus élevés) et plus cette fidélité des certitudes augmente.

Qu'en est-il de la liaison entre les indices Rt T80, Ct Abs T80 et NCSt T80 ?

L'indice *Ct* lorsqu'il est calculé au départ des valeurs absolues des indices de centration par question ($|Cq|$) donne une idée globale de la « quantité » d'erreurs d'auto-estimations présente dans les résultats; nous appellerons cet indice *Ct Abs*. L'information fournie par *CT Abs* est fort proche de celle de l'indice *Rt*. Le tableau ci-contre montre les valeurs des corrélations obtenues entre les trois indices spectraux *NCSt T80*, *Rt T80* et *Ct Abs T80* dans le cadre des dix tests MOHICAN. La corrélation significative négative très élevée entre *Rt T80* et *Ct Abs T80* montre une relation logique : plus le taux de réalisation des prédictions est élevé, moins on observe de tendance à la sous ou surestimation dans les résultats (et inversement).

	NCSt T80	*Rt T80*
Rt T80	0,09 (ns)	
Ct Abs T80	-0,02 (ns)	-0,93 (*p* ,001)

L'absence de corrélations entre NCSt T80 et les deux autres indices spectraux montre que bien qu'étant aussi calculé à partir des données liées à l'utilisation des pourcentages de certitude, NCSt T80 livre une information différente liée à la cohérence d'utilisation des certitudes dans les résultats des questions, c'est-à-dire la tendance à fournir des certitudes plus élevées pour les réponses correctes que pour les incorrectes. La corrélation négative proche de –1 entre Rt T80 et Ct Abs T80 montre bien que plus le taux de réalisation des prédictions est élevé et moins on observe de tendance à la sur ou sous-estimation dans les résultats.

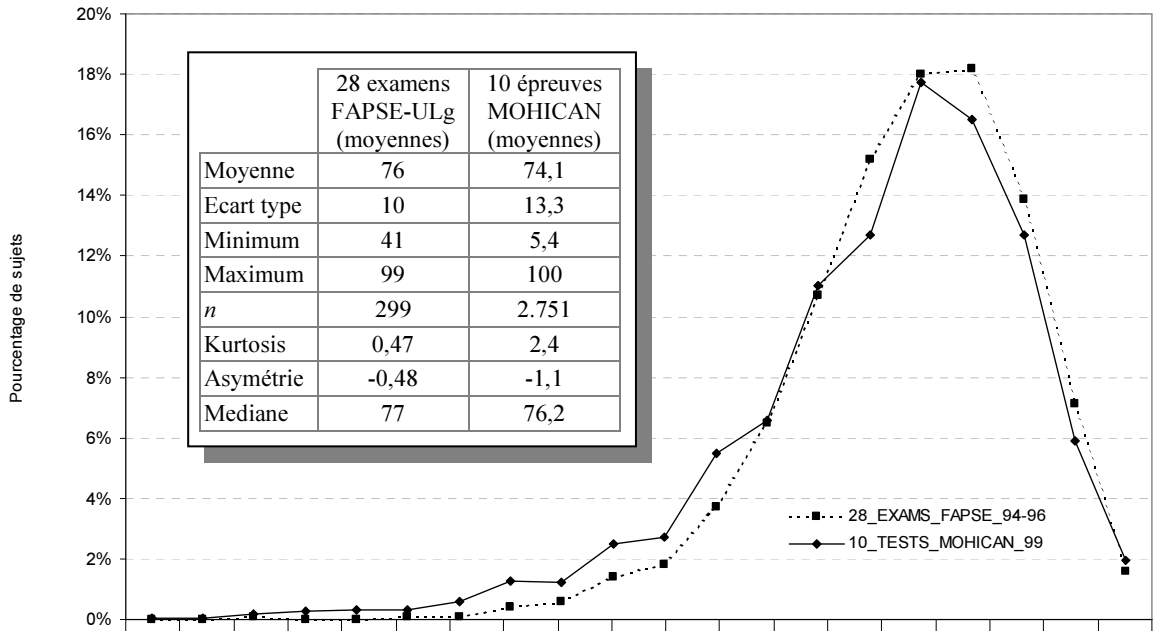
6. Performances en Réalisme des sujets (*Rs*) et des groupes (*Rg*)

Lors d'une recherche portant sur les données de 28 examens ayant recours aux degrés de certitude entre 1994 et 1996 (Gilles, 1996a), nous avons pu établir une distribution des fréquences de scores de Réalisme des sujets (Rs). A l'époque nous avions constaté une courbe en « J » très décentrée vers la droite. Qu'en est-il lorsqu'on compare avec les 10 épreuves MOHICAN ?

La procédure de calcul de l'indice de Réalisme d'un sujet (*Rs*) a été exposée précédemment (voir p.184). La moyenne des scores *Rs* de l'ensemble des sujets ayant participé au test donne l'indice de Réalisme moyen du groupe (*Rg*, voir p. 272). L'indice *Rg* est un indice global de la qualité des auto-estimations du groupe des répondants. Rappelons que *Rs* et *Rg* varient entre 0 et 100.

La distribution des fréquences des scores de réalisme nous permet de visualiser les niveaux de performance d'auto-estimation du groupe. Cette distribution peut être comparée à celle des moyennes des fréquences des scores de réalisme qui ont été calculées sur la base des résultats à 28 examens ayant eu lieu à la FAPSE-ULg entre 1994 et 1996. La distribution des fréquences des scores *Rs* est accompagnée des statistiques suivantes : moyenne (ou *Rg*), écart type, minimum, maximum, *n*, coefficient d'aplatissement, coefficient d'asymétrie et médiane (voir p. 275).

Fréquences en pourcentages des scores de réalisme des sujets
Scores moyens des 10 épreuves MOHICAN comparés aux scores moyens des 28 examens FAPSE-ULg de 1994 à 1996



	28 examens FAPSE-ULg (moyennes)	10 épreuves MOHICAN (moyennes)
Moyenne	76	74,1
Ecart type	10	13,3
Minimum	41	5,4
Maximum	99	100
n	299	2.751
Kurtosis	0,47	2,4
Asymétrie	-0,48	-1,1
Mediane	77	76,2

	>0 et <=5	>5 et <=10	>10 et <=15	>15 et <=20	>20 et <=25	>25 et <=30	>30 et <=35	>35 et <=40	>40 et <=45	>45 et <=50	>55 et <=60	>60 et <=65	>65 et <=70	>70 et <=75	>75 et <=80	>80 et <=85	>85 et <=90	>90 et <=95	>95
28_EXAMS_FAPSE_94-96	0,0%	0,0%	0,1%	0,0%	0,0%	0,1%	0,1%	0,4%	0,6%	1,4%	3,7%	6,5%	10,7%	15,2%	18,0%	18,2%	13,9%	7,1%	1,6%
10_TESTS_MOHICAN_99	0,0%	0,0%	0,2%	0,3%	0,3%	0,3%	0,6%	1,3%	1,2%	2,5%	5,5%	6,6%	11,0%	12,7%	17,7%	16,5%	12,7%	5,9%	1,9%

scores de réalisme des sujets

Précédemment lorsque nous avons comparé la courbe des 28 examens avec celle du test de physique nous avons constaté qu'elles étaient très proches (voir p. 275).

Lorsque nous prenons en compte les 10 épreuves MOHICAN, nous observons une similitude des distributions des fréquences. Les distributions des fréquences moyennes des scores de réalisme des 28 examens et des 10 épreuves MOHICAN sont toutes deux très décentrées vers la droite, « en J ».

Voici le tableau des fréquences en pourcentages des scores de réalisme des sujets :

Classes	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ	Moyennes
>0 et <=5	0,0%	0,1%	0,2%	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,0%
>5 et <=10	0,0%	0,1%	0,1%	0,0%	0,1%	0,0%	0,1%	0,0%	0,0%	0,0%	0,0%
>10 et <=15	0,0%	0,0%	0,1%	0,1%	0,0%	0,1%	0,1%	0,2%	0,5%	0,6%	0,2%
>15 et <=20	0,1%	0,3%	0,8%	0,2%	0,1%	0,0%	0,5%	0,2%	0,3%	0,3%	0,3%
>20 et <=25	0,1%	0,4%	0,4%	0,3%	0,0%	0,1%	0,9%	0,2%	0,6%	0,4%	0,3%
>25 et <=30	0,0%	0,2%	1,0%	0,4%	0,1%	0,1%	0,4%	0,4%	0,3%	0,2%	0,3%
>30 et <=35	0,3%	0,5%	1,8%	0,4%	0,1%	0,4%	0,5%	0,5%	0,6%	0,8%	0,6%
>35 et <=40	0,4%	0,6%	4,9%	1,2%	0,4%	0,1%	0,8%	1,4%	1,3%	1,8%	1,3%
>40 et <=45	0,3%	1,2%	3,2%	1,1%	0,1%	0,7%	0,8%	2,1%	1,7%	0,9%	1,2%
>45 et <=50	0,8%	2,0%	8,5%	2,8%	0,4%	0,5%	1,0%	3,2%	3,2%	2,4%	2,5%
>50 et <=55	1,7%	2,2%	5,4%	4,0%	0,8%	0,9%	1,6%	3,6%	3,9%	3,0%	2,7%
>55 et <=60	2,7%	4,1%	13,2%	6,8%	3,0%	2,1%	2,8%	8,3%	6,5%	5,5%	5,5%
>60 et <=65	5,1%	6,3%	7,4%	9,3%	5,6%	4,5%	5,3%	8,6%	7,9%	6,0%	6,6%
>65 et <=70	8,2%	9,8%	15,8%	12,6%	11,3%	8,9%	7,6%	13,3%	11,5%	11,1%	11,0%
>70 et <=75	13,3%	12,2%	7,0%	14,7%	14,8%	12,1%	11,5%	13,6%	16,1%	11,8%	12,7%
>75 et <=80	20,2%	16,7%	13,9%	17,2%	22,3%	21,5%	16,0%	16,6%	15,7%	17,2%	17,7%
>80 et <=85	20,6%	17,2%	5,4%	13,5%	22,6%	23,3%	19,1%	13,0%	14,0%	16,1%	16,5%
>85 et <=90	17,1%	14,1%	6,7%	9,4%	13,3%	15,9%	18,3%	10,1%	9,6%	12,4%	12,7%
>90 et <=95	7,6%	8,8%	1,9%	4,5%	4,3%	7,5%	10,0%	3,7%	4,5%	6,2%	5,9%
>95	1,5%	2,9%	2,3%	1,4%	0,8%	1,4%	2,8%	1,1%	1,9%	3,2%	1,9%

La dernière colonne du tableau précédent reprend les valeurs qui ont été exprimées dans le graphique des fréquences en pourcentages.

Voici le tableau des statistiques descriptives liées aux scores de réalisme des épreuves MOHICAN.

	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ	Moyennes
Rg	77,7	75,8	65,2	72,1	76,8	78,1	77,0	71,3	72,6	73,9	74,1
Ecart type	10,8	13,9	16,9	13,5	9,9	10,5	14,2	13,8	14,8	15,2	13,3
Minimum	14,3	0	0	0	8	12,5	9,1	10	0	0	5,4
Maximum	100	100	100	100	100	100	100	100	100	100	100
N	3846	3739	3420	3682	1410	1392	2516	2507	2501	2497	2751
Kurtosis	2,50	2,52	0,13	1,44	4,47	4,03	3,75	0,97	1,73	2,49	2,4
Asymétrie	-1,15	-1,24	-0,40	-0,88	-1,29	-1,40	-1,65	-0,81	-1,05	-1,26	-1,1
Médiane	79,3	78,3	66,7	74,0	78,4	80,0	80,0	74,0	75,0	76,0	76,2
Mode	87,1	85,0	70,0	80,0	79,2	80,0	90,0	80,0	80,0	82,0	81,3

Nous remarquons que tous les scores Rg sont plus élevés que 70 sauf pour le groupe d'étudiants soumis à l'épreuve COMPRE qui récolte un Rg égal à 65,2. Dans 9 épreuves sur 10, les sujets ont donc tendance à commettre en moyenne moins de 30% d'erreurs dans leurs auto-estimations.

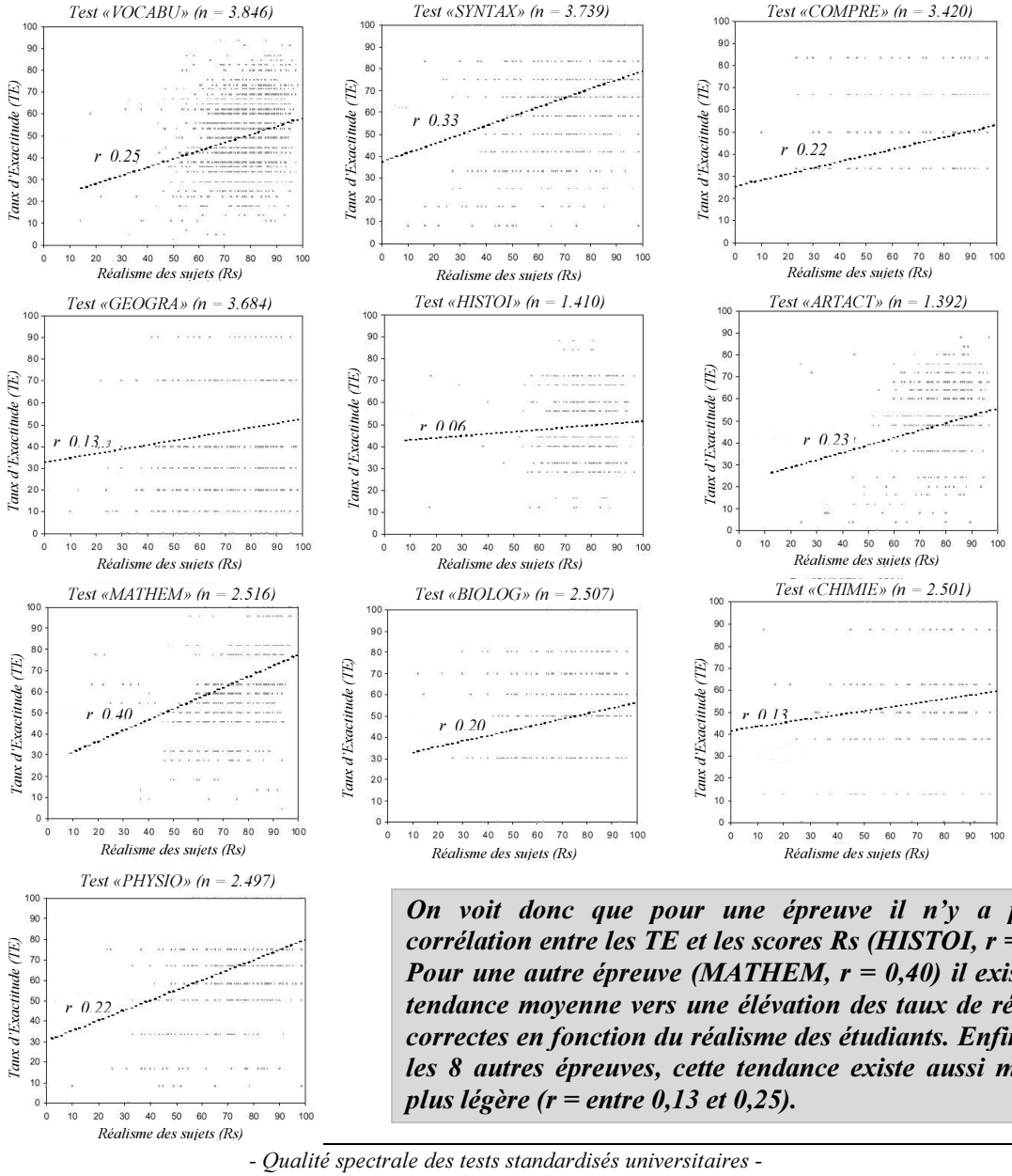
7. Les étudiants les plus réalistes obtiennent-ils de meilleurs taux d'exactitude ?

Précédemment, nous nous étions posé la question d'une éventuelle liaison entre les Taux d'Exactitude (TE) et les scores de Réalisme des sujets (Rs) ([1.3], p. 219).

Voici le tableau récapitulatif des corrélations entre scores *Rs* et *TE* pour les 10 épreuves MOHICAN :

	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ
r	0,25	0,33	0,22	0,13	0,06	0,23	0,40	0,20	0,13	0,22
n	3846	3739	3420	3682	1410	1392	2516	2507	2501	2497

Nous constatons que des corrélations élevées entre les scores *Rs* et les *TE* des étudiants existent mais sont très peu élevées pour les épreuves VOCABU, COMPRE, GEOGRA, ARTACT, BIOLOG, CHIMIE et PHYSIQ (toutes ces épreuves se situent entre 0,13 et 0,25). Pour l'épreuve HISTOI la corrélation est pratiquement nulle. Le test MATHEM possède la corrélation la plus élevée des 10 épreuves (r = 0,40) suivi de SYNTAX (0,33). Voici les nuages de points des valeurs des taux d'exactitude (axe des ordonnées) confrontées aux valeurs des scores de réalisme (axe des abscisses) des sujets pour chacune des épreuves MOHICAN.



On voit donc que pour une épreuve il n'y a pas de corrélation entre les TE et les scores Rs (HISTOI, r = 0,06). Pour une autre épreuve (MATHEM, r = 0,40) il existe une tendance moyenne vers une élévation des taux de réponses correctes en fonction du réalisme des étudiants. Enfin, pour les 8 autres épreuves, cette tendance existe aussi mais est plus légère (r = entre 0,13 et 0,25).

8. Performances en Centration des sujets (*Cs*) et des *g*roupes (*Cg*)

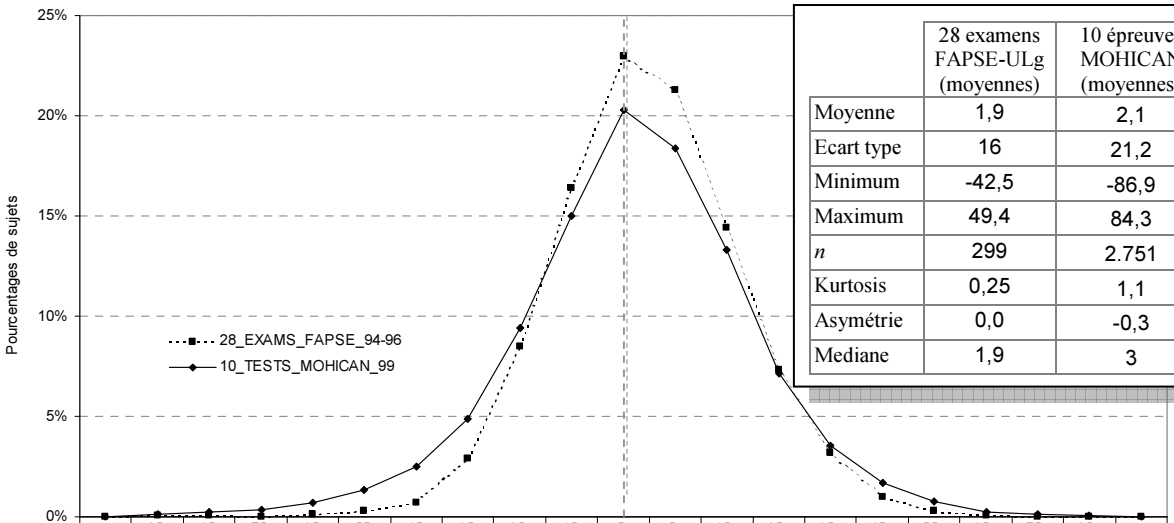
Nous allons comparer la distribution des fréquences des scores de Centration des sujets (Cs) des 10 épreuve MOHICAN à la distribution des scores de centration des sujets établie lors d'une recherche antérieure portant sur les données de 28 examens avec degrés de certitude (Gilles, 1996a). Nous nous attendons à des scores de centration forts proches pour les deux types d'épreuves étant donnés les performances en réalisme (Rs) similaires observées précédemment.

L'indice de centration moyen du groupe est calculé à partir des scores de Centration de chaque sujet (*Cs*, voir p. 277).

L'indice de centration d'un sujet se calcule en soustrayant le taux d'exactitude à la certitude moyenne et nous indique par son signe si il y a eu surestimation (+) ou sous-estimation (-). L'indice de Centration moyenne d'un groupe (*Cg*, voir p. 278) nous donne une information globale sur la tendance à se surestimer ou à se sous-estimer au niveau d'un groupe d'étudiants.

Voici la distribution des fréquences des scores de Centration des sujets *Cs* pour l'ensemble des 10 épreuves MOHICAN (en trait continu). Cette distribution est comparée à celle que nous obtenons pour 28 examens (en pointillés) qui ont eu lieu à la FAPSE-ULg entre 1994 et 1996 (voir p. 278). Nous accompagnons ces distributions des statistiques descriptives suivantes : moyenne (ou *Cg*), écart type, minimum, maximum, *n*, coefficient d'aplatissement, coefficient d'asymétrie et médiane.

Fréquences en pourcentages des scores de centration des sujets
Scores moyens des 10 épreuves MOHICAN comparés aux scores moyens des 28 examens FAPSE-ULg de 1994 à 1996



	28 examens FAPSE-ULg (moyennes)	10 épreuves MOHICAN (moyennes)
Moyenne	1,9	2,1
Ecart type	16	21,2
Minimum	-42,5	-86,9
Maximum	49,4	84,3
n	299	2.751
Kurtosis	0,25	1,1
Asymétrie	0,0	-0,3
Mediane	1,9	3

	<= -95	>-95 et <=-85	>-85 et <=-75	>-75 et <=-65	>-65 et <=-55	>-55 et <=-45	>-45 et <=-35	>-35 et <=-25	>-25 et <=-15	>-15 et <=-5	>-5 et <=5	>5 et <=15	>15 et <=25	>25 et <=35	>35 et <=45	>45 et <=55	>55 et <=65	>65 et <=75	>75 et <=85	>85 et <=95	>95
28_EXAMS_FAPSE_94-96	0%	0%	0%	0%	0%	0%	1%	3%	9%	16%	23%	21%	14%	7%	3%	1%	0%	0%	0%	0%	0%
10_TESTS_MOHICAN_99	0%	0%	0%	0%	1%	1%	3%	5%	9%	15%	20%	18%	13%	7%	4%	2%	1%	0%	0%	0%	0%

Scores de Centration des sujets (Cs)

Nous remarquons sur le graphique que les scores de centration excellents, c'est-à-dire proches de zéro, sans trop de surestimations ni trop de sous-estimations, sont un peu moins fréquents dans les épreuves MOHICAN (fréquence $-5 < Cs \leqslant 5 = 20\%$ pour les épreuves 10 MOHICAN contre 23 % pour les 28 examens FAPSE).

Nous constatons à l'aide des statistiques descriptives qui accompagnent le graphique que les moyennes sont forts proches (Cg = 1,9 pour les 28 examens et Cg = 2,1 pour les 10 épreuves MOHICAN) avec un écart type plus resserré autour de la moyenne pour les 28 examens (Et = 16) par rapport aux 10 épreuves MOHICAN (Et = 21,2).

Voici le tableau des fréquences en pourcentages des scores de centration des sujets :

Classes	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ	Moyennes
<= -95	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
>-95 et <=-85	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
>-85 et <=-75	0%	0%	0%	0%	0%	0%	1%	0%	1%	0%	0%
>-75 et <=-65	0%	0%	0%	0%	0%	0%	1%	0%	0%	1%	0%
>-65 et <=-55	0%	1%	0%	1%	0%	0%	1%	1%	1%	2%	1%
>-55 et <=-45	1%	2%	1%	1%	0%	0%	2%	2%	2%	3%	1%
>-45 et <=-35	2%	4%	1%	3%	2%	1%	1%	2%	4%	4%	3%
>-35 et <=-25	4%	6%	4%	7%	4%	2%	2%	4%	6%	6%	5%
>-25 et <=-15	10%	11%	6%	11%	11%	9%	5%	8%	11%	11%	9%
>-15 et <=-5	18%	16%	8%	16%	18%	21%	12%	12%	16%	17%	15%
>-5 et <=5	24%	19%	13%	20%	27%	29%	22%	17%	20%	21%	20%
>5 et <=15	19%	19%	14%	18%	20%	21%	24%	18%	17%	17%	18%
>15 et <=25	13%	12%	15%	12%	13%	11%	17%	16%	12%	12%	13%
>25 et <=35	6%	5%	14%	6%	3%	3%	8%	11%	6%	5%	7%
>35 et <=45	2%	2%	10%	3%	2%	2%	3%	6%	3%	1%	4%
>45 et <=55	1%	1%	7%	1%	0%	1%	1%	2%	1%	0%	2%
>55 et <=65	0%	1%	4%	0%	0%	0%	0%	1%	0%	0%	1%
>65 et <=75	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%
>75 et <=85	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%
>85 et <=95	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
>95	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

La dernière colonne du tableau précédent reprend les valeurs qui ont été exprimées dans le graphique des fréquences en pourcentages.

Voici le tableau des statistiques descriptives liées aux scores Centration des groupes (*Cg*) des épreuves MOHICAN.

	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ	Moyennes
Cg	1,3	-0,4	14,9	0,0	0,0	1,3	3,8	5,4	-0,5	-4,7	2,1
Ecart type	18,2	22,8	26,2	21,2	16,8	15,8	21,7	23,7	23,1	22,6	21,2
Minimum	-77	-100	-83	-80	-72	-76	-91	-90	-100	-100	-86,9
Maximum	80	90	100	90	92	73	83	90	88	58	84,3
N	3846	3739	3420	3685	1410	1392	2516	2507	2501	2497	2751
Kurtosis	0,66	1,05	0,05	0,51	1,56	1,93	2,93	0,41	0,92	0,99	1,1
Asymétrie	-0,10	-0,29	-0,25	-0,01	0,01	0,16	-1,19	-0,34	-0,45	-0,73	-0,3
Médiane	1,8	0,0	16,7	0,0	0,8	0,8	6,4	6,0	0,0	-2,0	3,0
Mode	0	0	10	-4	4,8	-4	10	14	0	0	3,1

Le tableau qui précède montre des scores de Centration des groupes (*Cg*) en général positifs et assez proches de zéro (rappelons que le minimum vaut –100 et le maximum 100) pour les épreuves VOCABU, SYNTAX, ARTACT, MATHEM, BIOLOG et CHIMIE. Deux épreuves sont même parfaitement « calibrées » du point de vue des scores *Cg* qui valent 0, il s'agit de GEOGRA et HISTOI.

Cependant, deux autres épreuves présentent des situations assez différentes des précédentes : le test COMPRE qui se caractérise par une plus forte tendance à la surestimation dans le groupe (*Cg* = 14,9) et le test PHYSIQ où nous remarquons une sous-estimation plus marquée (-4,7) que dans CHIMIE et SYNTAX.

D. Analyse de la qualité des tests à l'aide des indices classiques

Dans ce chapitre consacré au niveau « TEST » de l'analyse de la qualité des épreuves MOHICAN nous venons d'examiner les indices spectraux. Nous allons maintenant étudier la qualité des tests à l'aide des indices classiques. Quelles sont les valeurs obtenues aux indices classiques de fidélité des tests ? Observe-t-on des différences lorsque ces indices classiques sont calculés au départ des matrices binaires ou des matrices spectrales ? A quels niveaux de difficulté les tests MOHICAN permettent-ils de discriminer les étudiants ? Précédemment nous avons avons proposé un indice de cohérence interne qui repose sur l'exploitation des informations issues des rpbis classiques, le NCIt (p. 262), et, dans la section précédente nous venons d'examiner les valeurs obtenues par un indice de cohérence spectrale, le NCSt (p. 261 et p. 295), qui lui exploite les informations provenant des rpbis spectraux. Que donne la comparaison des valeurs obtenues à ces deux catégories d'indices ? Voilà, dans les grandes lignes, les questions que nous nous posons dans le contexte de cette section et auxquelles nous allons maintenant tenter d'apporter des réponses.

1. Indices de fidélité pour matrices binaires et spectrales

Les valeurs récoltées par les 10 épreuves MOHICAN à plusieurs indices classiques de fidélité seront analysées dans cette sous-section. Nous nous attendons à des niveaux de fidélité forts différents d'un test à l'autre car d'une part la longueur des épreuves MOHICAN est variable : la plus courte (COMPRE) comporte 6 questions et la plus longue, VOCABU, en comporte 45. D'autre part, un test comme GEOGRA est en fait composé de plusieurs sous-tests. Nous tenterons également de répondre à cette question : un indice de fidélité tel que l'alpha augmente-t-il ou diminue-t-il lorsqu'il est calculé à partir des données des sujets les plus réalistes ? Nous proposons de calculer ces indices de fidélité en utilisant des matrices binaires ainsi que des matrices spectrales (voir p. 132). Les indices de fidélité seront-ils meilleurs dans le cas des matrices binaires ou dans le cas des matrices spectrales ? Si les matrices spectrales procuraient une meilleure fidélité des résultats cela constituerait un argument de poids en faveur de l'utilisation des pourcentages de certitude.

Afin de faciliter les comparaisons nous avons regroupé les résultats dans un tableau récapitulatif. Lorsque les calculs ont été réalisés à l'aide de matrices binaires nous ajoutons « *mb* » en indice aux notations et lorsqu'ils ont été effectués à l'aide de matrice spectrales nous complétons par « *ms* ». Avant de commenter ce tableau, nous rappelons brièvement les différents indices qui sont utilisés.

Le coefficient de bipartition ($r_{xx'}$) est calculé en corrélant (coefficient *r* de Bravais-Pearson) les scores de deux moitiés d'un test administré en une fois aux mêmes sujets (voir p. 130). Il nous informe sur le niveau de fidélité de l'épreuve en nous donnant le niveau de variance des scores observés qui est attribuable à la variance des scores vrais. Le coefficient de bipartition est calculé à partir des sous-tests générés après répartition des items en fonction des niveaux de facilité objective (*pot*) (voir p. 134).

La valeur du coefficient de bipartition est systématiquement sous évaluée étant donné qu'il est calculé à partir de la moitié des questions, dès lors nous lui appliquons la correction de Spearman-Brown (*rS,* voir p. 131).

La fiabilité par moitié de Guttman (*rG*) est une méthode de calcul de la fidélité où on considère qu'il existe des différences dans les variances des résultats des deux moitiés (voir p. 131). Le coefficient de fiabilité par moitiés de Guttman est calculé à partir des sous-tests générés par la méthode de répartition des questions en fonction de leur facilité objective.

Le coefficient alpha de Cronbach (α) est très souvent utilisé pour calculer la cohérence interne d'un test. La formule de calcul de l'alpha est présentée page 137.

Il est possible de calculer le nombre de questions parallèles à ajouter à un test (ou éventuellement à retrancher lorsqu'on souhaite procéder à des comparaisons) pour atteindre une valeur alpha désirée. Le calcul s'effectue à l'aide de la formule de Spearman-Brown (voir p. 139). Nous utiliserons la notation suivante dans le cas de l'utilisation du coefficient d'allongement pour une fidélité désirée : « *kq[α=0,8]* ».

a) Comparaison des valeurs obtenues aux indices de fidélité

Voici le tableau récapitulatif des valeurs obtenues à ces indices classiques de fidélité calculés à partir de matrices binaires (…*mb*) et spectrales (…*ms*) :

	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ
$r_{xx'mb}$	0,714	0,410	0,279	0,373	0,484	0,510	0,637	0,250	0,270	0,341
$r_{xx'ms}$	0,714	**0,450**	0,192	0,310	**0,505**	0,503	**0,677**	0,248	**0,300**	**0,431**
rS_{mb}	0,833	0,581	0,437	0,543	0,652	0,676	0,779	0,400	0,425	0,509
rS_{ms}	0,833	**0,620**	0,322	0,473	**0,671**	0,662	**0,807**	0,397	**0,462**	**0,603**
rG_{mb}	0,833	0,581	0,436	0,543	0,651	0,673	0,778	0,398	0,423	0,504
rG_{ms}	0,833	**0,620**	0,322	0,473	**0,671**	0,669	**0,806**	0,393	**0,460**	**0,602**
α_{mb}	0,833	0,572	0,393	0,530	0,668	0,707	0,769	0,410	0,414	0,473
α_{ms}	0,832	**0,615**	0,221	0,437	0,652	0,686	**0,788**	0,374	**0,430**	**0,565**
$kq[\alpha=0,8]_{mb}$	-9	24	31	25	25	16	4	48	37	35
$kq[\alpha=0,8]_{ms}$	-9	18	79	42	28	21	2	57	34	21

Nous avons signalé par un fond grisé et des caractères gras les situations où les indices $r_{xx'}$, *rS*, *rG* et α calculés à partir des matrices spectrales sont plus élevés que les indices calculés à partir de matrices binaires. La présence d'un fond grisé sans caractères gras signale que les différences sont minimes, c'est-à-dire inférieures à 0,015.

Pour l'épreuve VOCABU les indices calculés à l'aide des matrices spectrales obtiennent les mêmes valeurs que les indices calculés à partir des matrices binaires (l'alpha est quasi identique : $\alpha_{mb} = 0,833$ et $\alpha_{ms} = 0,832$).

Pour 4 épreuves (SYNTAX, MATHEM, CHIMIE et PHYSIQ), les indices de fidélité calculés à partir des matrices spectrales sont systématiquement meilleurs.

Pour 2 épreuves, COMPRE et GEOGRA tous les indices de fidélité calculés sur matrices binaires sont plus élevés.

Dans le cas de l'épreuve HISTOI les coefficients de bipartition, de Spearman-Brown et de Guttman sont plus élevés lorsqu'ils sont calculés à partir des matrices spectrales, par contre l'alpha est plus élevé avec une matrice binaire.

Les valeurs obtenues aux coefficients $r_{xx'}$, *rS* et *rG* calculés sur matrices binaire et spectrale sont très proches dans le cas de BIOLOG et ARTACT par contre l'alpha est plus élevé lorsqu'il est calculé à partir d'une matrice binaire.

Nous constatons donc que dans une majorité de cas les indices de fidélité calculés à partir des matrices spectrales sont meilleurs. Par ailleurs, nous pourrions éliminer par raisonnement l'épreuve COMPRE qui ne comporte que 6 questions ainsi que l'épreuve GEOGRA qui est composée de deux sous-tests, l'un comportant 6 questions et l'autre 4. Nous reviendrons sur les valeurs obtenues aux 10 épreuves MOHICAN par les coefficients de cohérence interne dans les conclusions liées à cette partie (p. 320).

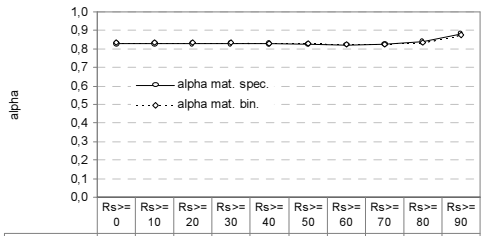
b) Stabilité de l'alpha lorsque les données sont sélectionnés sur base des performances en réalisme

Dans quelle mesure la qualité des auto-estimations des sujets peut-elle influencer l'alpha ? Autrement dit, l'alpha s'améliore-t-il lorsqu'il est calculé sur la base des résultats des étudiants les plus réalistes ?

Nous avons comparé les valeurs de l'alpha de Cronbach récoltées par chaque épreuve MOHICAN aux 10 paliers de la turbo analyse (de T0 à T90 avec un pas de 10). Les valeurs récoltées au palier T0 sont celles que l'ont retrouve dans le tableau récapitulatif des indices de fidélité (p. 309). Elles sont calculées sur l'ensemble des étudiants soumis au test étant donné qu'à T0 tous les sujets dont le réalisme (*Rs*) est supérieur ou égal à 0 sont sélectionnés.

A T90, le palier le plus élevé de cette turbo analyse, les alpha sont calculés à partir des données d'un plus petit nombre d'étudiants (en moyenne 9% des populations, voir tableau p. 290). Ces sujets sont particulièrement performants dans l'utilisation des pourcentages de certitude, puisqu'ils commettent en moyenne moins de 10% d'erreurs d'auto-estimations.

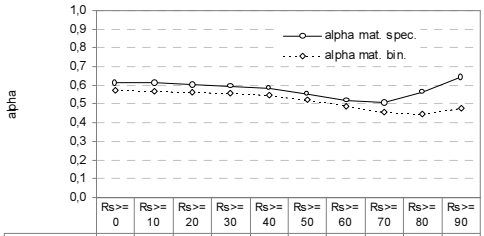
*Epreuve de **VOCABULAIRE***
Comparaison des coefficients alpha de Cronbach calculés sur matrice binaire et sur matrice spectrale . Données sélectionnées en fonction du Réalisme des sujets (Rs)



	Rs>= 0	Rs>= 10	Rs>= 20	Rs>= 30	Rs>= 40	Rs>= 50	Rs>= 60	Rs>= 70	Rs>= 80	Rs>= 90
alpha mat. spec.	0,832	0,832	0,831	0,83	0,828	0,827	0,821	0,825	0,84	0,884
alpha mat. bin.	0,833	0,833	0,833	0,833	0,831	0,83	0,825	0,826	0,834	0,878

Seuil de réalisme Rs pour la sélection des données

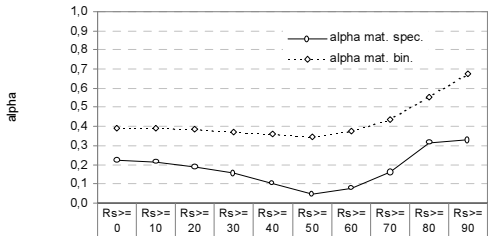
*Epreuve de **SYNTAXE ET ARTICULATION LOGIQUE***
Comparaison des coefficients alpha de Cronbach calculés sur matrice binaire et sur matrice spectrale . Données sélectionnées en fonction du Réalisme des sujets (Rs)



	Rs>= 0	Rs>= 10	Rs>= 20	Rs>= 30	Rs>= 40	Rs>= 50	Rs>= 60	Rs>= 70	Rs>= 80	Rs>= 90
alpha mat. spec.	0,615	0,613	0,604	0,595	0,582	0,551	0,52	0,510	0,561	0,645
alpha mat. bin.	0,572	0,57	0,565	0,56	0,548	0,521	0,485	0,455	0,445	0,477

Seuil de réalisme Rs pour la sélection des données

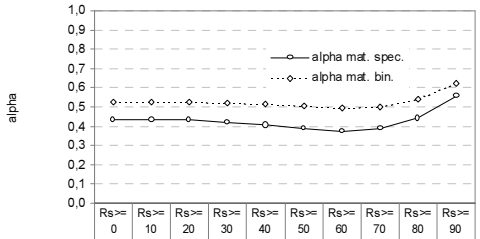
*Epreuve de **COMPREHENSION***
Comparaison des coefficients alpha de Cronbach calculés sur matrice binaire et sur matrice spectrale . Données sélectionnées en fonction du Réalisme des sujets (Rs)



	Rs>= 0	Rs>= 10	Rs>= 20	Rs>= 30	Rs>= 40	Rs>= 50	Rs>= 60	Rs>= 70	Rs>= 80	Rs>= 90
alpha mat. spec.	0,221	0,212	0,189	0,155	0,104	0,044	0,078	0,162	0,317	0,331
alpha mat. bin.	0,393	0,39	0,385	0,37	0,358	0,344	0,376	0,435	0,554	0,674

Seuil de réalisme Rs pour la sélection des données

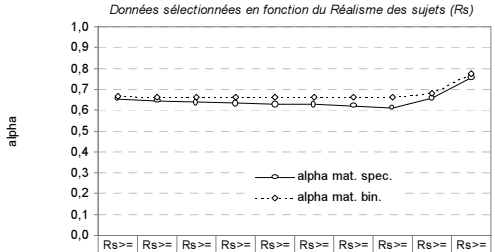
*Epreuve **GEOGRAPHIE ET LECTURE DE DOCUMENTS***
Comparaison des coefficients alpha de Cronbach calculés sur matrice binaire et sur matrice spectrale . Données sélectionnées en fonction du Réalisme des sujets (Rs)



	Rs>= 0	Rs>= 10	Rs>= 20	Rs>= 30	Rs>= 40	Rs>= 50	Rs>= 60	Rs>= 70	Rs>= 80	Rs>= 90
alpha mat. spec.	0,436	0,436	0,434	0,418	0,407	0,389	0,371	0,388	0,442	0,554
alpha mat. bin.	0,528	0,528	0,527	0,519	0,515	0,504	0,494	0,502	0,543	0,624

Seuil de réalisme Rs pour la sélection des données

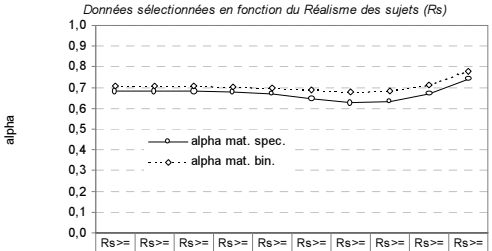
*Epreuve **CONNAISSANCES HIST. ET SOCIO-ECON.***
Comparaison des coefficients alpha de Cronbach calculés sur matrice binaire et sur matrice spectrale . Données sélectionnées en fonction du Réalisme des sujets (Rs)



	Rs>= 0	Rs>= 10	Rs>= 20	Rs>= 30	Rs>= 40	Rs>= 50	Rs>= 60	Rs>= 70	Rs>= 80	Rs>= 90
alpha mat. spec.	0,652	0,644	0,638	0,634	0,628	0,627	0,62	0,611	0,658	0,756
alpha mat. bin.	0,667	0,661	0,664	0,664	0,662	0,663	0,661	0,661	0,68	0,775

Seuil de réalisme Rs pour la sélection des données

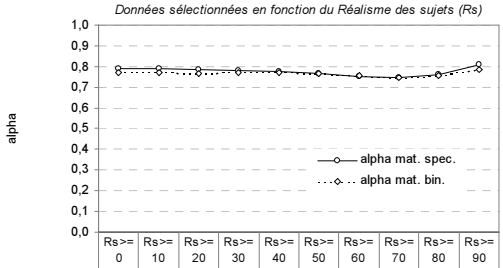
*Epreuve de **CONNAISSANCES ARTISTIQUES***
Comparaison des coefficients alpha de Cronbach calculés sur matrice binaire et sur matrice spectrale . Données sélectionnées en fonction du Réalisme des sujets (Rs)



	Rs>= 0	Rs>= 10	Rs>= 20	Rs>= 30	Rs>= 40	Rs>= 50	Rs>= 60	Rs>= 70	Rs>= 80	Rs>= 90
alpha mat. spec.	0,686	0,686	0,686	0,679	0,668	0,648	0,629	0,631	0,67	0,742
alpha mat. bin.	0,707	0,707	0,708	0,704	0,699	0,691	0,679	0,683	0,713	0,779

Seuil de réalisme Rs pour la sélection des données

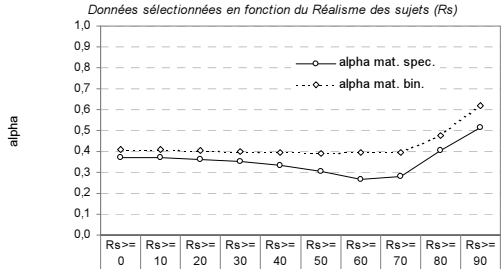
Epreuve de ***MATHEMATIQUES***
Comparaison des coefficients alpha de Cronbach calculés sur matrice binaire et sur matrice spectrale . Données sélectionnées en fonction du Réalisme des sujets (Rs)



	Rs>= 0	Rs>= 10	Rs>= 20	Rs>= 30	Rs>= 40	Rs>= 50	Rs>= 60	Rs>= 70	Rs>= 80	Rs>= 90
alpha mat. spec.	0,788	0,788	0,784	0,78	0,773	0,764	0,749	0,746	0,76	0,81
alpha mat. bin.	0,769	0,769	0,767	0,77	0,77	0,765	0,756	0,748	0,755	0,784

Seuil de réalisme Rs pour la sélection des données

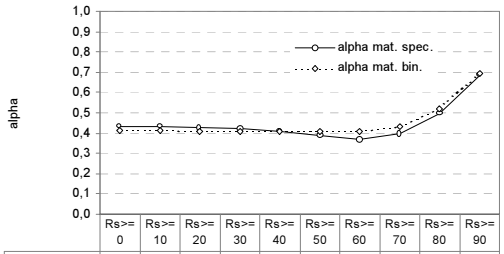
Epreuve de ***BIOLOGIE***
Comparaison des coefficients alpha de Cronbach calculés sur matrice binaire et sur matrice spectrale . Données sélectionnées en fonction du Réalisme des sujets (Rs)



	Rs>= 0	Rs>= 10	Rs>= 20	Rs>= 30	Rs>= 40	Rs>= 50	Rs>= 60	Rs>= 70	Rs>= 80	Rs>= 90
alpha mat. spec.	0,373	0,373	0,364	0,352	0,335	0,306	0,269	0,282	0,403	0,516
alpha mat. bin.	0,41	0,41	0,407	0,402	0,397	0,392	0,393	0,396	0,477	0,617

Seuil de réalisme Rs pour la sélection des données

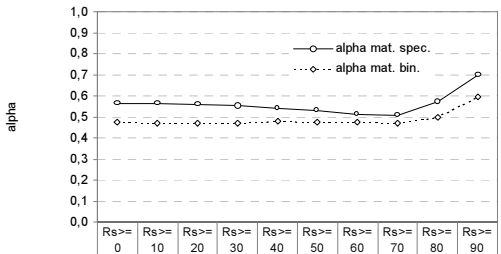
Epreuve de ***CHIMIE***
Comparaison des coefficients alpha de Cronbach calculés sur matrice binaire et sur matrice spectrale . Données sélectionnées en fonction du Réalisme des sujets (Rs)



	Rs>= 0	Rs>= 10	Rs>= 20	Rs>= 30	Rs>= 40	Rs>= 50	Rs>= 60	Rs>= 70	Rs>= 80	Rs>= 90
alpha mat. spec.	0,43	0,43	0,426	0,42	0,409	0,39	0,368	0,396	0,495	0,685
alpha mat. bin.	0,414	0,413	0,41	0,408	0,41	0,41	0,41	0,430	0,518	0,692

Seuil de réalisme Rs pour la sélection des données

Epreuve de ***PHYSIQUE***
Comparaison des coefficients alpha de Cronbach calculés sur matrice binaire et sur matrice spectrale . Données sélectionnées en fonction du Réalisme des sujets (Rs)



	Rs>= 0	Rs>= 10	Rs>= 20	Rs>= 30	Rs>= 40	Rs>= 50	Rs>= 60	Rs>= 70	Rs>= 80	Rs>= 90
alpha mat. spec.	0,565	0,565	0,559	0,555	0,539	0,532	0,512	0,506	0,572	0,696
alpha mat. bin.	0,473	0,47	0,47	0,47	0,48	0,475	0,475	0,472	0,498	0,594

Seuil de réalisme Rs pour la sélection des données

Chaque graphique ci-dessus concerne une épreuve MOHICAN et reprend deux courbes : en pointillés les coefficients alpha de Cronbach (α) calculés à partir des scores binaires et en trait continu les α calculés à partir des scores spectraux. Dans les deux cas les α sont calculés aux 10 paliers de la turbo analyse (seuils *Rs*>=…). Les effectifs de chaque palier turbo ont été présentés précédemment (voir p. 290).

Nous constatons que pour les matrices spectrales (traits continus), les coefficients alpha calculés à partir des données des sujets les plus réalistes sont plus élevés dans chaque épreuve. Lorsque nous observons les valeurs des alpha calculés à l'aide des matrices binaires, nous constatons une augmentation pour toutes épreuves sauf une : SYNTAX (alpha mat. Bin. à T0 = 0,572 et à T90 = 0,477).

Pour les deux types de matrices, nous remarquons une légère diminution des valeurs aux paliers turbo intermédiaires puis une remontée aux paliers turbo plus élevés. L'infléchissement des courbes (lorsque la valeur de l'alpha est la plus basse) se situe aux seuils de réalisme exigé *Rs*⩾50 (COMPRE), ⩾60 (ARTACT, CHIMIE, BIOLOGIE, GEOGRA), ⩾70 (MATHEM, PHYSIQ, SYNTAX dans le cas de la matrice spectrale) et à *Rs*⩾80 (SYNTAX dans le cas de la matrice binaire). Pour les épreuves VOCABU et HISTOI, cet infléchissement est pratiquement inexistant.

On voit donc que dans le cadre des épreuves MOHICAN, l'alpha s'améliore lorsqu'il est calculé sur la base des résultats des étudiants les plus réalistes : dans 9 cas sur 10 pour les matrices binaires et dans tous les cas pour les matrices spectrales. Pour certaines épreuves comme COMPRE, BIOLOG et CHIMIE ces gains en fidélité peuvent être relativement important.

(1) Comparaison de la fidélité des matrices binaires et spectrales à Rs⩾0 et Rs⩾90

Lorsqu'on compare les valeurs de l'alpha (α_{ms} et α_{mb}) en tenant compte d'une part des données de tous les sujets (Rs>0) et d'autre part des données des seuls sujets les plus réalistes (Rs>90), observe-t-on un gain plus élevé quand l'alpha est calculé à partir de matrices spectrales ou quand il est calculé à partir de matrices binaires ?

Observons maintenant les différences de fidélité entre les matrices spectrales et binaires des 10 épreuves à deux paliers de turbo analyse T0 et T90, donc d'une part lorsque l'alpha est calculé avec tous les étudiants quel que soit leur réalisme (palier turbo T0, $Rs \geqslant 0$) et d'autre part lorsqu'il est calculé à l'aide des données des sujets les plus réalistes (palier T90, $Rs \geqslant 90$). Remarquons que nous pourrions procéder à une étude spectrale de la fidélité des tests en envisageant les données d'étudiants dont le réalisme serait compris entre deux seuils A et B : $A \leqslant Rs < B$, ce que nous comptons explorer lors de travaux ultérieurs.

Si nous comparons la fidélité des matrices binaires et spectrales des 10 épreuves en prenant en compte pour le calcul de l'alpha de Cronbach d'une part les données de tous les sujets ($Rs \geqslant 0$) et d'autre part les données des sujets les plus réalistes ($Rs \geqslant 90$), nous obtenons le tableau ci-après.

Dans ce tableau nous notons « ↑ » lorsque la valeur α calculée à l'aide de la matrice spectrale (α_{ms}) est supérieure à celle calculée sur base de la matrice binaire (α_{mb}) et lorsque la différence est comprise entre 0,025 et 0,1. Nous notons « ≈ » lorsque les α_{ms} et α_{mb} sont égaux ou lorsque la différence est minime (moins de 0,025). Lorsque α_{ms} est légèrement inférieur (différence comprise entre 0,025 et 0,1) à α_{mb} nous notons « ↓ ».

Lorsque les différences sont plus marquées (supérieures à 0,1) et favorables à α_{ms} alors nous notons « ↑↑ ». Lorsqu'une différence plus marquée est en défaveur de la matrice spectrale (α_{ms} inférieur de plus de 0,1 à α_{mb}), nous notons « ↓↓ ».

	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ
Rs⩾0	$\alpha_{ms} \approx \alpha_{mb}$	$\alpha_{ms} \uparrow \alpha_{mb}$	$\alpha_{ms} \downarrow\downarrow \alpha_{mb}$	$\alpha_{ms} \downarrow \alpha_{mb}$	$\alpha_{ms} \approx \alpha_{mb}$	$\alpha_{ms} \approx \alpha_{mb}$	$\alpha_{ms} \approx \alpha_{mb}$	$\alpha_{ms} \downarrow \alpha_{mb}$	$\alpha_{ms} \approx \alpha_{mb}$	$\alpha_{ms} \uparrow \alpha_{mb}$
Rs⩾90	$\alpha_{ms} \approx \alpha_{mb}$	$\alpha_{ms} \uparrow\uparrow \alpha_{mb}$	$\alpha_{ms} \downarrow\downarrow \alpha_{mb}$	$\alpha_{ms} \downarrow \alpha_{mb}$	$\alpha_{ms} \approx \alpha_{mb}$	$\alpha_{ms} \approx \alpha_{mb}$	$\alpha_{ms} \approx \alpha_{mb}$	$\alpha_{ms} \downarrow\downarrow \alpha_{mb}$	$\alpha_{ms} \approx \alpha_{mb}$	$\alpha_{ms} \uparrow\uparrow \alpha_{mb}$

Nous constatons qu'en termes de différences de fidélité entre matrices spectrales et matrices binaires, il y a statu quo pour la majorité des épreuves lorsqu'on compare les α calculés sur les données de tous les étudiants (Rs⩾0) ou sur celles d'un sous-ensemble de sujets plus réalistes (Rs⩾90). Les différences observées entre α_{ms} et α_{mb} sont quasi identiques à Rs⩾0 ou Rs⩾90 pour 7 épreuves : VOCABU, COMPRE, GEOGRA, HISTOI, ARTACT, MATHEM et CHIMIE. Par contre les différences entre α_{ms} et α_{mb} augmentent pour 3 épreuves lorsqu'on considère les sujets les plus réalistes (Rs⩾90). Pour deux de ces épreuves (SYNTAX et PHYSIQ) la fidélité des matrices spectrales augmente et pour le troisième test (BIOLOG) c'est la fidélité de la matrice binaire qui augmente. Pour ces trois tests lorsqu'on compare les valeurs obtenues à Rs⩾0 et Rs⩾90, l'écart se creuse entre l'alpha calculé d'une part à l'aide d'une matrice spectrale et d'autre part à l'aide d'une matrice binaire.

(2) Evolution des α_{ms} et α_{mb}

Nous venons d'observer une amélioration de la fidélité dans les épreuves MOHICAN lorsque l'alpha est calculé à partir des données des étudiants les plus réalistes (Rs>90, voir graphiques p. 310). Comment expliquer d'un point de vue mathématique ces meilleures valeurs de l'alpha lorsque nous prenons en compte les données des sujets qui commettent moins d'erreurs dans leurs auto-estimations ?

Nous remarquons sur les 10 graphiques des épreuves MOHICAN (p. 310) que les coefficients alpha de Cronbach calculés à partir des données des étudiants les plus réalistes ($Rs \geqslant 90$) comparés à ceux de tous les sujets quel que soit leur niveau de réalisme ($Rs \geqslant 0$) sont plus élevés dans 9 tests sur 10 lorsqu'on utilise une matrice binaire et dans les 10 tests avec une matrice spectrale.

En nous référant à la formule de l'alpha de Cronbach (voir p. 137), cela signifie qu'après sélection des données en fonction du Réalisme des sujets (*Rs*) lorsque nous calculons l'alpha sur ces données des plus réalistes, la variance des scores du total (σ_t^2) du test tend à augmenter tandis que la somme des variances des questions ($\sum \sigma_q^2$) tend à diminuer. Ceci entraîne une diminution du rapport $\sum \sigma_q^2 / \sigma_t^2$ et une augmentation de l'alpha après soustraction à 1 et multiplication par le nombre de questions (*nq*) divisé par *nq* -1.

A titre d'exemple, voici les valeurs des $\sum \sigma_q^2$ et σ_t^2 ainsi que le détail du calcul des coefficients alpha aux seuils $Rs \geqslant 70$, $\geqslant 80$ et $\geqslant 90$ pour le test de physique (PHYSIQ) :

Test « PHYSIQ » (*nq* = 10)	Matrices binaires			Matrices spectrales		
	$Rs \geqslant 70$ (n = 1.724)	$Rs \geqslant 80$ (n = 1.009)	$Rs \geqslant 90$ (n = 275)	$Rs \geqslant 70$ (n = 1.724)	$Rs \geqslant 80$ (n = 1.009)	$Rs \geqslant 90$ (n = 275)
$\sum \sigma_q^2$	1,688	1,596 (↘)	1,449 (↘)	31.599	30.014 (↘)	27.418 (↘)
σ_t^2	2,936	2,892 (≈)	3,113 (↗)	58.053	61.907 (↗)	73.383 (↗)
$\sum \sigma_q^2 / \sigma_t^2$	0,5749	0,5519 (↘)	0,4655 (↘)	0,5443	0,4848 (↘)	0,3736 (↘)
$1 - (\sum \sigma_q^2 / \sigma_t^2)$	0,4251	0,4481 (↗)	0,5345 (↗)	0,4557	0,5152 (↗)	0,6264 (↗)
$\alpha = (10/9) \cdot 1 - (\sum \sigma_q^2 / \sigma_t^2)$	0,472	0,498 (↗)	0,594 (↗)	0,506	0,572 (↗)	0,696 (↗)

Signalons ici que si nous sélectionnons les sujets sur un autre critère : leurs taux d'exactitude à l'ensemble des questions de l'épreuve plutôt que leur réalisme, nous n'aboutirons pas à une augmentation de la fidélité, mais au contraire à une diminution de celle-ci.

En effet, si nous sélectionnons les sujets qui obtiennent les meilleurs scores, la dispersion de leurs notes sera forcément plus faible que si nous prenons tous les sujets quel que soit leur total. Dès lors, avec les sujets les plus performants la variance du total tendra à diminuer, ce qui entraînera une augmentation du rapport $\sum \sigma_q^2 / \sigma_t^2$ et une diminution de l'alpha. Voici les effets sur les étapes de calcul de l'alpha lorsque nous sélectionnons les données du test de physique en fonction des Taux d'Exactitude (*TE*).

Test « PHYSIQ » (*nq* = 10)	*TE* = 0 (*n* = 2.497)	*TE*⩾20 (*n* = 2.487)	*TE*⩾40 (*n* = 2.375)	*TE*⩾60 (*n* = 1.724)	*Rs*⩾80 (*n* = 745)
$\sum \sigma_q^2$	1,7729	1,7618 (↘)	1,6928 (↘)	1,4967 (↘)	0,9962 (↘)
σ_t^2	3,0959	2,960 (↘)	2,3827 (↘)	1,3852 (↘)	0,4895 (↘)
$\sum \sigma_q^2 / \sigma_t^2$	0,5745	0,5951 (↗)	0,7104 (↗)	1,0805 (↗)	2,0351 (↗)
$1 - (\sum \sigma_q^2 / \sigma_t^2)$	0,4254	0,4048 (↘)	0,2895 (↘)	-0,0805 (↘)	-1,0351 (↘)
$\alpha = (10/9) \cdot 1 - (\sum \sigma_q^2 / \sigma_t^2)$	0,4727	0,4498 (↘)	0,3217 (↘)	-0,0895 (↘)	-1,1501 (↘)

Remarquons que si les corrélations observées entre le réalisme des sujets (*Rs*) et leurs taux d'exactitude (*TE*) étaient plus fortes (voir p. 305), nous risquerions d'assister à une diminution de l'alpha aux paliers de turbo analyse élevés. En effet, dans l'hypothèse théorique où la corrélation serait totale entre *Rs* et *TE*, une sélection des données des sujets sur base du critère de réalisme reviendrait à sélectionner les données en fonction du critère des taux d'exactitude ce qui entraînerait les conséquences que l'on sait (voir tableau qui précède) sur le rapport $\sum \sigma_q^2 / \sigma_t^2$ et donc une diminution de l'alpha.

L'amélioration de la fidélité traduite par la hausse des valeurs α lorsque ces derniers sont calculés à partir des données des étudiants les plus réalistes observée pour les 10 épreuves dans le cas des matrices spectrales et pour 9 épreuves sur 10 dans le cas des matrices binaires est remarquable et constitue à notre avis un argument supplémentaire en faveur de l'utilisation des pourcentages de certitude dans le contexte de tests tels que les check up MOHICAN.

2. Facilité objective des tests (*pot*)

A quels niveaux de difficulté les tests MOHICAN permettent-ils de discriminer les étudiants ?

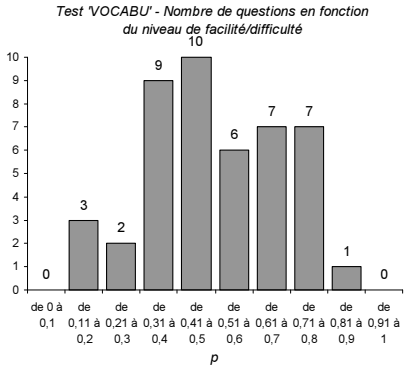
La somme des taux d'exactitude de toutes les questions du test divisée par le nombre de questions nous informe sur le niveau de facilité globale de l'épreuve (*pot*, voir p. 270) pour le groupe d'étudiants qui a passé l'épreuve.

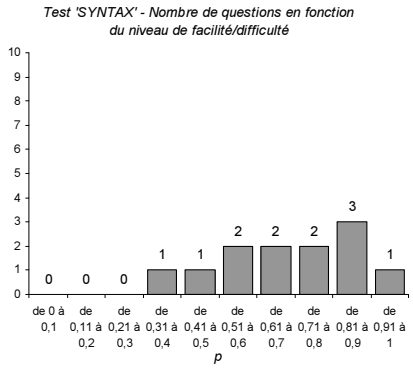
	VOCABU	SYNTAX	COMPRE	GEOGRA	HISTOI	ARTACT	MATHEM	BIOLOG	CHIMIE	PHYSIQ	Moyenne
pot	0,50	0,69	0,44	0,47	0,50	0,49	0,66	0,49	0,55	0,66	0,55

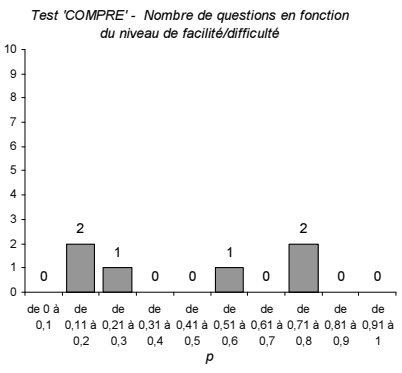
Nous remarquons que « COMPRE » est apparemment l'épreuve la plus difficile avec une moyenne de 44% de réponses correctes. La facilité moyenne de l'ensemble des épreuves MOHICAN, se situe à 0,55.

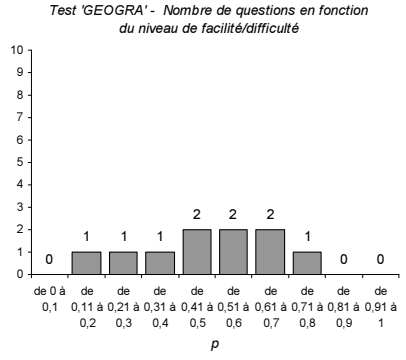
Nous pouvons représenter graphiquement la distribution des questions en fonction des valeurs obtenues par chaque question à l'indice de facilité objective.

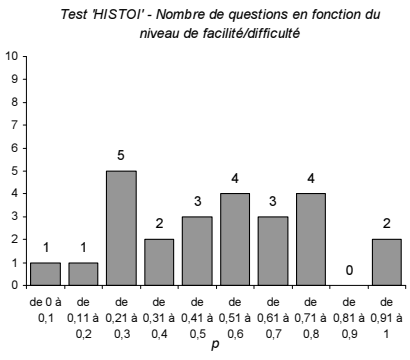
Sur les graphiques ci-dessous, l'abscisse reprend les valeurs de *p* en 10 intervalles de 0,1 en partant de 0 jusqu'à 1. Chaque bâton de l'histogramme représente le nombre de questions pour un intervalle de *p*. Ce type de graphique permet d'observer à quels niveaux de difficulté le test discrimine les sujets (voir p. 122).

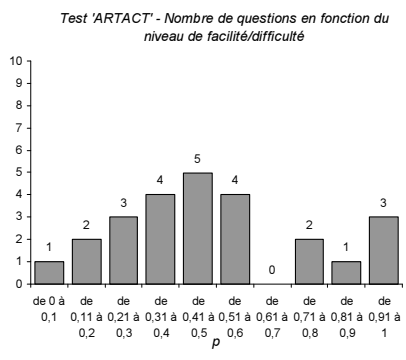


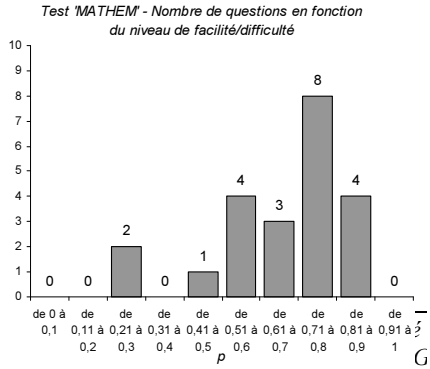


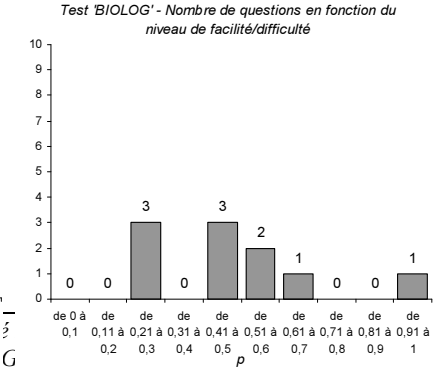


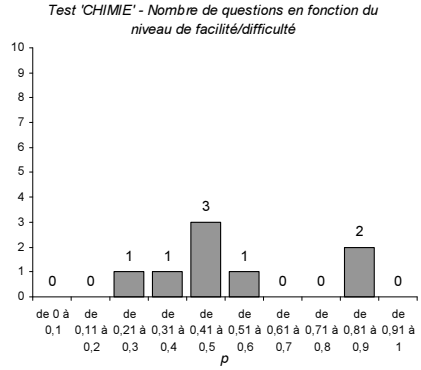


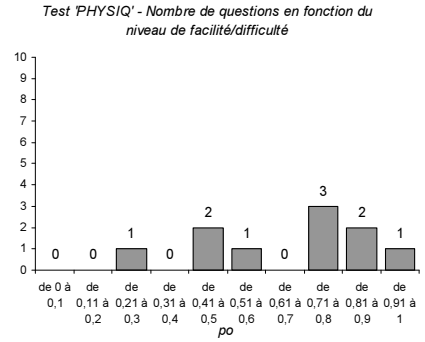












Les 10 graphiques précédents montrent des situations assez différentes. L'épreuve VOCABU couvre toutes les valeurs de p sauf aux deux extrémités. C'est le cas de GEOGRA aussi où il n'y a pas non plus de questions dont la facilité serait comprise entre p>0,81 et p ⩽0,90.

SYNTAX ne comporte pas de questions de grande difficulté et se caractérise par une répartition des questions à tous les niveaux de facilité élevé (10 questions sur 12 ont un p > 0,5). Il est dès lors logique que ce soit le test le mieux réussi (pot = 0,69).

Deux épreuves, HISTOI et ARTACT, se caractérisent par une répartition de questions difficiles à tous les niveaux p<0,5 l'absence de questions faciles à certains niveaux p>0,5.

Les autres épreuves comportent plusieurs « trous » (absence de questions) à différents niveaux de facilité/difficulté. C'est particulièrement le cas de COMPRE. Cette épreuve ne comportant que 6 questions il est impossible de couvrir tous les intervalles que nous avons définis (10 intervalles de 0,1 p).

Remarquons qu'aucune épreuve se caractérise par l'absence totale de questions à des niveaux de difficulté inférieurs à 0,5 ou à des niveaux de facilité supérieur à 0,5.

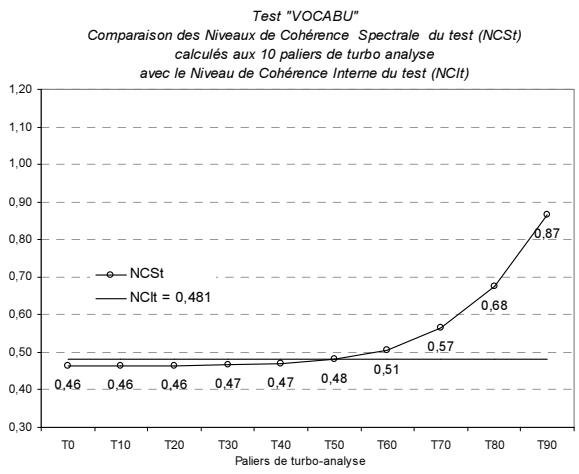
3. Niveau de Cohérence Interne des tests (*NCIt*)

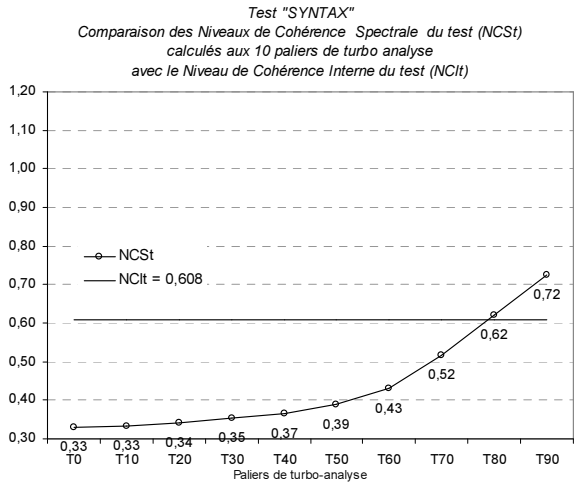
Lorsque nous avons comparé le NCIt et les NCSt obtenus aux 10 paliers de la turbo analyse (de T0 à T90 avec un pas de 10) pour l'épreuve de physique (n = 2.497), nous avons constaté qu'au palier T60 le NCSt était égal au NCIt et qu'à partir de T70 les valeurs du NCSt étaient supérieure à celles du NCIt (voir p. 263). Qu'en est-il dans les autres épreuves MOHICAN ? ([3.1], p. 263).

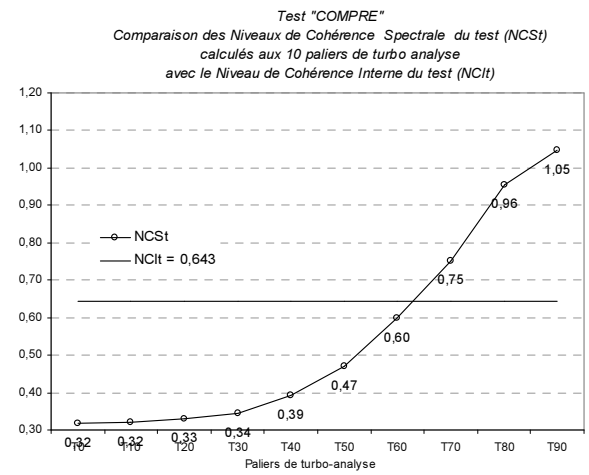
Précédemment, lorsque nous avons présenté la possibilité de calculer le Niveau de Cohérence Spectrale d'une question (*NCSq*, voir p. 231), nous avons aussi abordé la possibilité de calculer le Niveau de Cohérence Interne d'une question en soustrayant à la valeur du *rpbis* classique de la réponse correcte, la moyenne pondérée des *rpbis* classiques des réponses incorrectes (*NCIq*, voir p. 233).

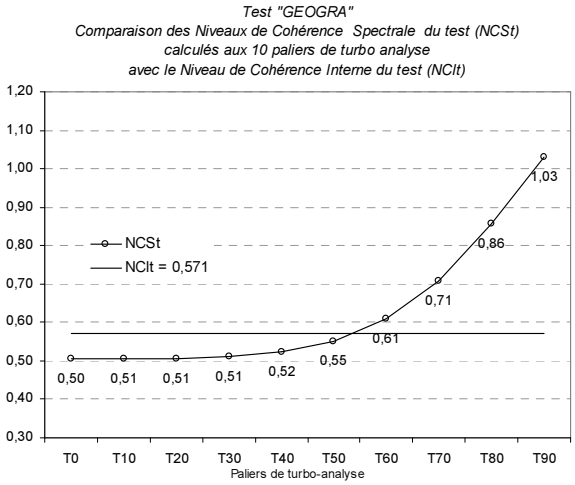
Nous pouvons calculer la moyenne des *NCIq* obtenus à toutes les QCM d'une épreuve et aboutir ainsi à un indice plus global que nous avons appelé le Niveau de Cohérence Interne d'un test (*NCIt*, voir p. 262).

Voici 10 graphes qui reprennent chacun les Niveaux de Cohérence Spectrale du test (*NCSt*) calculés aux 10 paliers de la turbo analyse (courbe) et le Niveau de Cohérence Interne du test (*NCIt*) (trait horizontal) pour chaque épreuve MOHICAN.

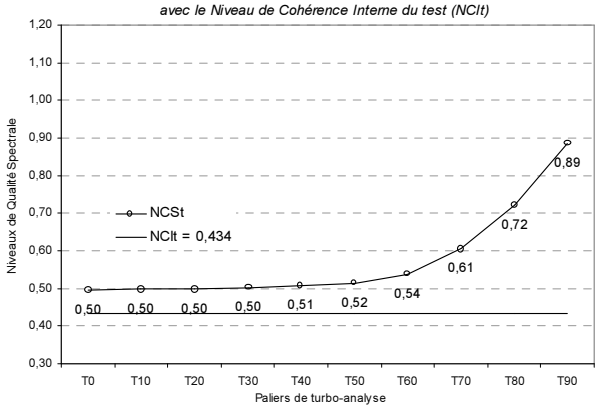




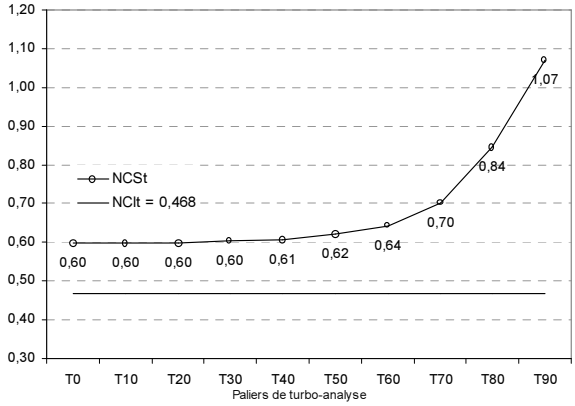




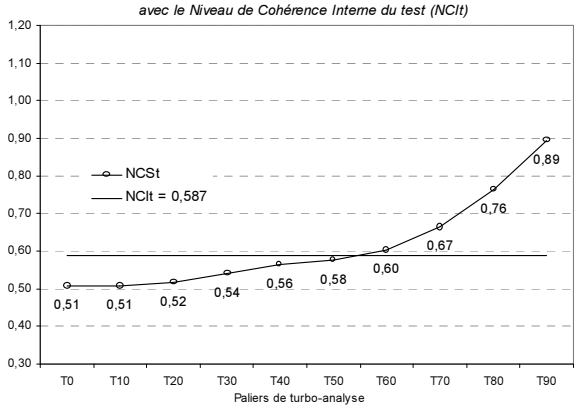
Test "HISTOI"
Comparaison des Niveaux de Cohérence Spectrale du test (NCSt)
calculés aux 10 paliers de turbo analyse
avec le Niveau de Cohérence Interne du test (NCIt)



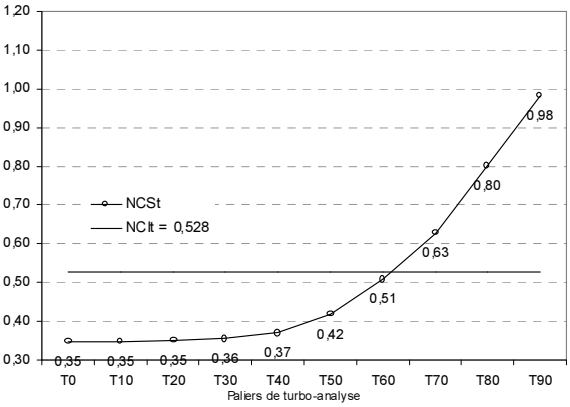
Test "ARTACT"
Comparaison des Niveaux de Cohérence Spectrale du test (NCSt)
calculés aux 10 paliers de turbo analyse
avec le Niveau de Cohérence Interne du test (NCIt)



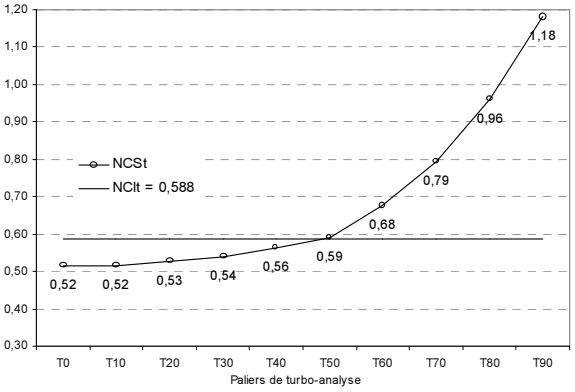
Test "MATHEM"
Comparaison des Niveaux de Cohérence Spectrale du test (NCSt)
calculés aux 10 paliers de turbo analyse
avec le Niveau de Cohérence Interne du test (NCIt)



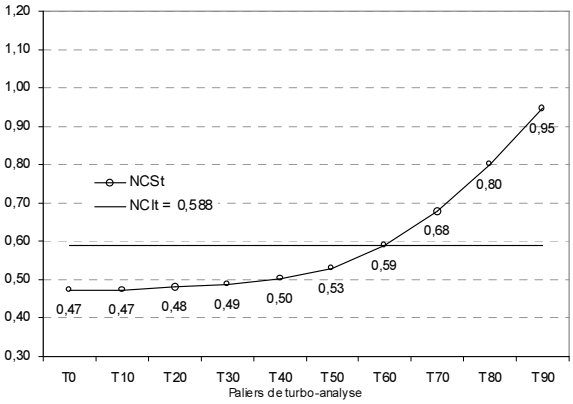
Test "BIOLOG"
Comparaison des Niveaux de Cohérence Spectrale du test (NCSt)
calculés aux 10 paliers de turbo analyse
avec le Niveau de Cohérence Interne du test (NCIt)



Test "CHIMIE"
Comparaison des Niveaux de Cohérence Spectrale du test (NCSt)
calculés aux 10 paliers de turbo analyse
avec le Niveau de Cohérence Interne du test (NCIt)



Test "PHYSIQ"
Comparaison des Niveaux de Cohérence Spectrale du test (NCSt)
calculés aux 10 paliers de turbo analyse
avec le Niveau de Cohérence Interne du test (NCIt)



Nous constatons que pour l'épreuve VOCABU, le trait horizontal représentant la valeur du NCIt coupe la courbe du NCSt obtenue au palier T50 de la turbo analyse, c'est également le cas pour CHIMIE.

Le trait du NCIt de l'épreuve PHYSIQ coupe la courbe des NCSt à T60. Pour deux autres épreuves, GEOGRA et MATHEM, les NCSt obtenu à T60 sont plus élevés que le NCIt.

Pour les tests COMPRE et BIOLOG, c'est à partir du palier turbo T70 que les valeurs des NCSt sont plus élevés. L'épreuve SYNTAX obtient un NCSt légèrement supérieur au NCIt au palier turbo T80. Enfin, nous remarquons que deux épreuves, HISTOI et ARTACT, obtiennent des NCIt toujours inférieurs au NCSt quelque soit le palier de la turbo analyse.

Pour les épreuves MOHICAN, à partir du palier T80 le NCST est donc toujours supérieur au NCIt.

E. Conclusions de l'analyse du niveau TEST

Dans ces conclusions partielles qui concernent le niveau TEST, nous allons synthétiser les observations que nous venons d'effectuer et répondre aux questions que nous nous étions posées précédemment lors de la présentation des instruments d'analyse spectrale où les données d'un seul test (PHYSIQ) avaient été utilisées. Ces questions et les pages où nous les étions posées sont rappelées par les mentions […] avec le numéro d'ordre et l'indication de la page.

1. A propos des performances en réalisme des sujets soumis aux épreuves MOHICAN

L'observation du graphique des effectifs (en pourcentages) des épreuves MOHICAN aux 10 paliers de turbo analyse (p. 219) montre que le seuil d'infléchissement des courbes se situe à T50 sauf pour la courbe de COMPRE qui commence à infléchir à T40 ([1.1], p. 219).

Les paliers de turbo analyse étant déterminés par les niveaux de performance en Réalisme des sujets (*Rs*), le graphique montre aussi que les pourcentages de sujets aux différents seuils de réalisme ne varient guère d'une épreuve à l'autre sauf pour COMPRE. Ce test obtient des pourcentages d'effectifs moins élevés que les autres épreuves pour les niveaux de performance en réalisme de $Rs \geqslant 50$ (T50) à $Rs \geqslant 80$ (T80). Donc, les performances en réalisme des étudiants ont été en moyenne moins bonnes lors de l'épreuve COMPRE. Le tableau des indices moyens de réalisme des sujets des épreuves MOHICAN (p. 307) montre aussi que le réalisme moyen du groupe (*Rg*) des 3.420 étudiants ayant participé au test COMPRE ($Rg = 65,2$) est le moins bon des 10 épreuves ([1.2], p. 219).

A l'aide du tableau des pourcentages d'effectifs observés à chacun des 10 paliers de la turbo analyse pour les 10 épreuves (p. 290) nous observons des pourcentages de sujets qui varient peu aux 5 premiers paliers turbo. En effet, de T0 à T50 le pourcentage moyen pour les 10 tests passe de 100% à 94%. Rappelons que cela signifie qu'en moyenne 6% des sujets seulement récoltent un score de réalisme (*Rs*) inférieur à 50 et donc qu'en moyenne 94% des étudiants obtiennent un *Rs* entre 50 et 100. Le tableau des indices moyens de réalisme des sujets des épreuves MOHICAN (p. 307) indique une moyenne générale pour l'ensemble des 10 épreuves à 74,1.

La courbe des fréquences moyennes en pourcentages des scores de réalisme des sujets pour les 10 épreuves MOHICAN et celle qui concerne les 28 examens ayant eu lieu à la FAPSE-ULg entre 1994 et 1996 sont très proches (p. 303) ([3.3], p. 276).

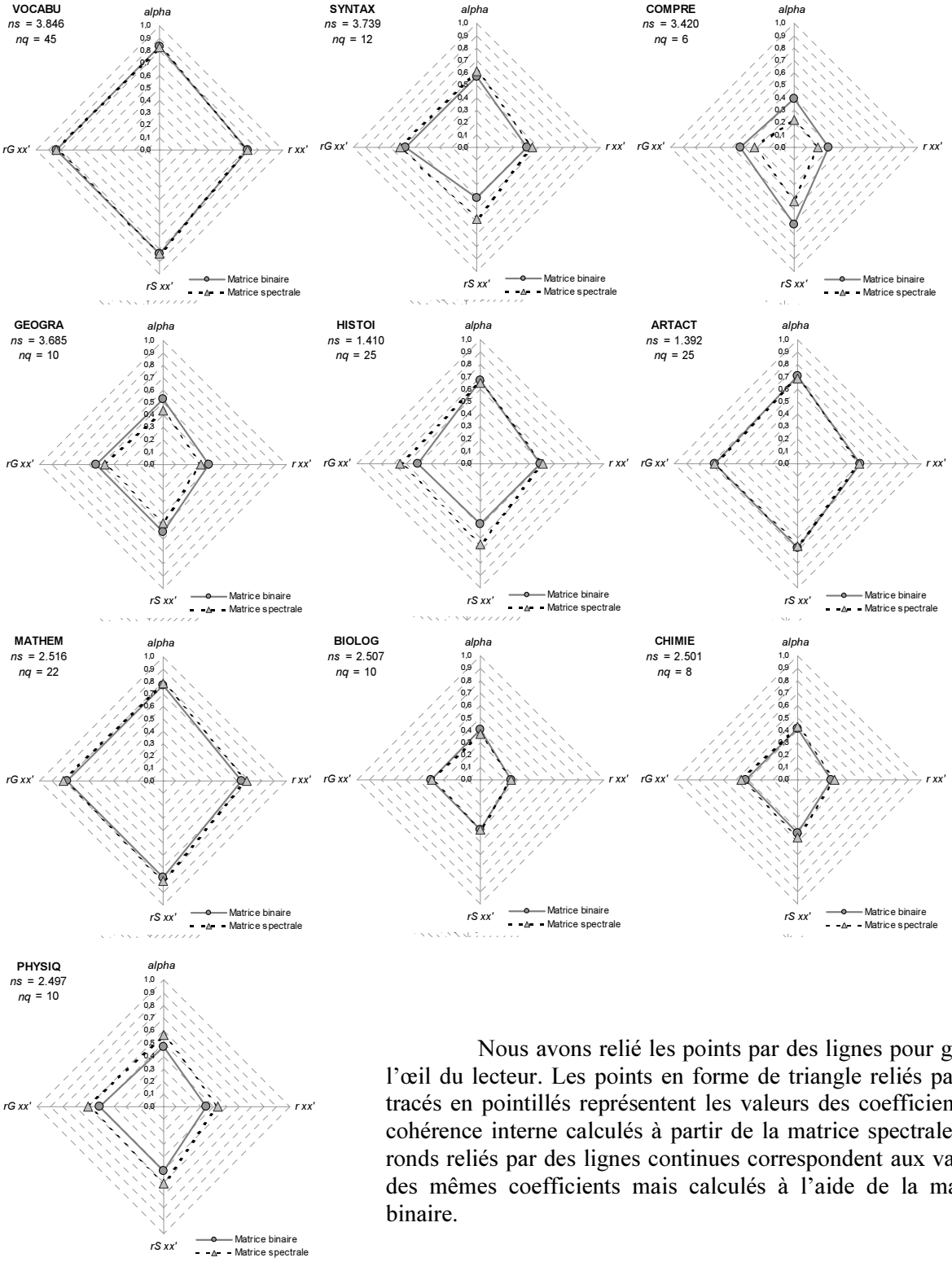
A partir du palier T60 les effectifs commencent à diminuer sensiblement, mais nous constatons qu'au palier de turbo analyse T80 il reste encore en moyenne pour les 10 tests 39% des effectifs totaux. A ce palier de turbo analyse, les étudiants sélectionnés dont le réalisme est supérieur ou égal à 80 commettent moins de 20% d'erreurs de certitude dans leurs auto-estimations. Etant donné cette qualité spectrale des données et les grands effectifs disponibles à ce palier turbo, la plupart des indices spectraux que nous calculerons par la suite le seront à l'aide des résultats obtenus à T80.

Nous nous étions également demandé si les étudiants les plus réalistes étaient aussi ceux qui obtenaient les meilleurs taux d'exactitude (*TE*) ([1.3], p. 219). Nous avons observé des corrélations positives mais peu élevées (de 0,13 à 0,25) entre *TE* et *Rs* pour les épreuves VOCABU, COMPRE, GEOGRA, ARTACT, BIOLOG, CHIMIE et PHYSIQ (p. 305). Deux autres épreuves possèdent des corrélations un peu plus élevées : MATHEM (0,40) et SYNTAX (0,33). Enfin, pour le test HISTOI on observe pas de corrélation entre *TE* et *Rs*. On peut donc conclure que ces corrélations positives existent sauf dans un cas et qu'elles sont peu élevées ($\leqslant 0,4$).

2. A propos de la cohérence interne des épreuves

Peut-on se doter d'un outil de visualisation des performances en cohérence interne des tests ?

Nous avons placé sur les axes d'ingénogrammes, graphiques polygonaux à coordonnées polaires, les valeurs récoltées par les 10 épreuves MOHICAN aux coefficients classiques de cohérence interne (voir tableau des valeurs récoltées par les indices α_{mb}, α_{ms}, $r_{xx'mb}$, $r_{xx'ms}$, rS_{mb}, rS_{ms}, rG_{mb} et rG_{ms}, p. 309).



Nous avons relié les points par des lignes pour guider l'œil du lecteur. Les points en forme de triangle reliés par des tracés en pointillés représentent les valeurs des coefficients de cohérence interne calculés à partir de la matrice spectrale. Les ronds reliés par des lignes continues correspondent aux valeurs des mêmes coefficients mais calculés à l'aide de la matrice binaire.

a) Disparité des performances des épreuves aux coefficients de cohérence interne

Lorsque nous comparons les ingénogrammes entre eux, nous visualisons l'ampleur des différences de cohérence interne entre les 10 épreuves (les valeurs figurent dans le tableau p. 309). VOCABU, MATHEM puis ARTACT obtiennent les meilleurs coefficients. BIOLOG, CHIMIE et COMPRE récoltent les valeurs les moins élevées. Nous pouvons regrouper les 4 épreuves qui restent dont la cohérence interne se situe de façon relativement intermédiaire entre les deux groupes précédents, il s'agit des tests HISTOI, PHYSIQ, SYNTAX et GEOGRA.

Les différences de cohérence interne peuvent s'expliquer par les nombres de questions très différents d'une épreuve à l'autre, VOCABU qui obtient les coefficients les plus élevés comporte 45 QCM contre seulement 6 pour COMPRE, 8 pour CHIMIE et 10 pour BIOLOG.

Un autre facteur peut aussi expliquer cette disparité des performances de cohérence interne : la diversité des contenus abordés au sein de certaines épreuves. Par exemple, le test HISTOI qui comporte 25 questions est intitulé « Connaissances en Histoire et Socio-économie » (deux contenus différents) et comporte des questions abordant des thèmes assez différents regroupés sous les sous-titres : « Institutions internationale » (8 QCM), « L'Union Européenne » (2 QCM), « Repères historiques » (3 QCM), « Géopolitique » (1 QCM), « Vocabulaire de français à référence historique » (5 QCM), « Médias » (2 QCM) et « Economie » (4 QCM).

b) Variabilité des performances en fonction des matrices spectrales et binaires

La visualisation des valeurs obtenues aux coefficients de cohérence interne selon qu'ils sont calculés à partir de matrices spectrales (traits pointillés) ou de matrices binaires (traits continus) à l'aide des ingénogrammes permet de mettre en évidence différents cas de figure.

Tous les coefficients sont quasi identiques avec matrice spectrale ou avec matrice binaire pour deux tests : VOCABU (l'épreuve obtenant les valeurs les plus élevées) et BIOLOG (une des trois épreuves obtenant les valeurs les moins élevées). Tous les coefficients calculés à partir de matrices spectrales sont plus élevés pour quatre tests : SYNTAX, MATHEM, CHIMIE et PHYSIQ.

Les coefficients calculés à partir de matrices binaires sont tous plus élevés pour deux tests : COMPRE et GEOGRA. Cependant, ces deux épreuves pourraient être éliminées par raisonnement car d'une part COMPRE ne contient que 6 questions et d'autre part GEOGRA est composée de deux sous-tests, l'un comportant 6 questions et l'autre seulement 4. Pour les tests HISTOI et ARTACT la situation est mitigée. Dans le cas de HISTOI les coefficients alpha et de bipartition ($r_{xx'}$) sont quasi identiques lorsqu'ils sont calculés avec une matrice binaire ou avec une matrice spectrale. Cependant, les calculs effectués à partir d'une matrice spectrale entraînent des valeurs plus élevées quand la correction de Spearman-Brown est appliquée au coefficient de bipartition (*rS*). Le coefficient de fidélité de Guttman (*rG*) est alors aussi plus élevé. En ce qui concerne ARTACT, Les valeurs obtenues par les coefficients alpha, *rS* et *rG* sont pratiquement les mêmes, mais pas pour le coefficient de bipartition pour lequel on observe des valeurs plus élevées lorsqu'il est calculé à partir d'une matrice binaire.

Nous avons calculé la cohérence interne à l'aide de 4 coefficients pour les 10 épreuves, ce qui nous fait 40 mesures effectuées avec les matrices binaires et 40 autres avec les matrices spectrales. Concernant ces 40 paires de mesures, signalons que dans 9 cas les valeurs obtenues sont plus élevées avec une matrice binaire, dans 11 cas elles sont quasi identiques (les différences sont inférieures à 0,015) et dans 19 cas les valeurs obtenues sont plus élevées lorsqu'elles sont calculées à l'aide d'une matrice spectrale.

On voit donc que dans une majorité de cas la fidélité est améliorée par l'utilisation de matrices spectrales, donc par le recours aux pourcentages de certitude qui permettent ce type de matrice de résultats.

c) Augmentation de la fidélité calculée à partir des données des sujets les plus réalistes

Nous avons vu (p. 310) que le coefficient alpha de Cronbach est systématiquement plus élevé lorsqu'il est calculé à partir des résultats des étudiants les plus réalistes ($Rs \geqslant 90$) à la fois dans le cas des matrices binaires et dans le cas des matrices spectrales.

Comme nous l'avons rappelé au point précédent, pour 8 épreuves il existe des différences entre les valeurs obtenues aux coefficients de fidélité lorsqu'ils sont calculés à l'aide d'une matrice spectrale ou à l'aide d'une matrice binaire. Pour certains tests ces coefficients sont plus élevés avec une matrice spectrale, pour d'autres c'est l'inverse. Parfois la situation est mitigée au sein d'une même épreuve, un type de matrice entraînant des valeurs plus élevés pour certains coefficients mais pas pour tous.

Qu'en est-il du point de vue d'une évolution éventuelle de ces différences observées entre les valeurs obtenues au coefficient alpha de Cronbach calculé à l'aide de matrices binaires (α_{mb}) ou à partir de matrices spectrales (α_{mb}) lorsqu'on prend en compte d'une part tous les sujets et d'autre part uniquement les sujets les plus réalistes ?

Nous avons comparé α_{mb} et α_{ms} calculé sur base de deux sélections contrastées de données en fonction du réalisme des sujets : l'une à $Rs \geqslant 0$ et l'autre à $Rs \geqslant 90$ (voir p. 312) et nous avons constaté :

- un *statu quo* au niveau des différences entre α_{mb} et α_{ms} quelles que soient les sélections de données ($Rs \geqslant 0$ ou $Rs \geqslant 90$) pour 7 épreuves : VOCABU, COMPRE, GEOGRA, HISTOI, ARTACT, MATHEM et CHIMIE ;

- une augmentation des différences entre α_{mb} et α_{ms} allant dans le sens d'une augmentation de la fidélité des matrices spectrales pour deux épreuves : SYNTAX et PHYSIQ ;

- une augmentation des différences entre α_{mb} et α_{ms} mais allant cette fois dans le sens d'une augmentation de la fidélité de la matrice binaire pour l'épreuve BIOLOG.

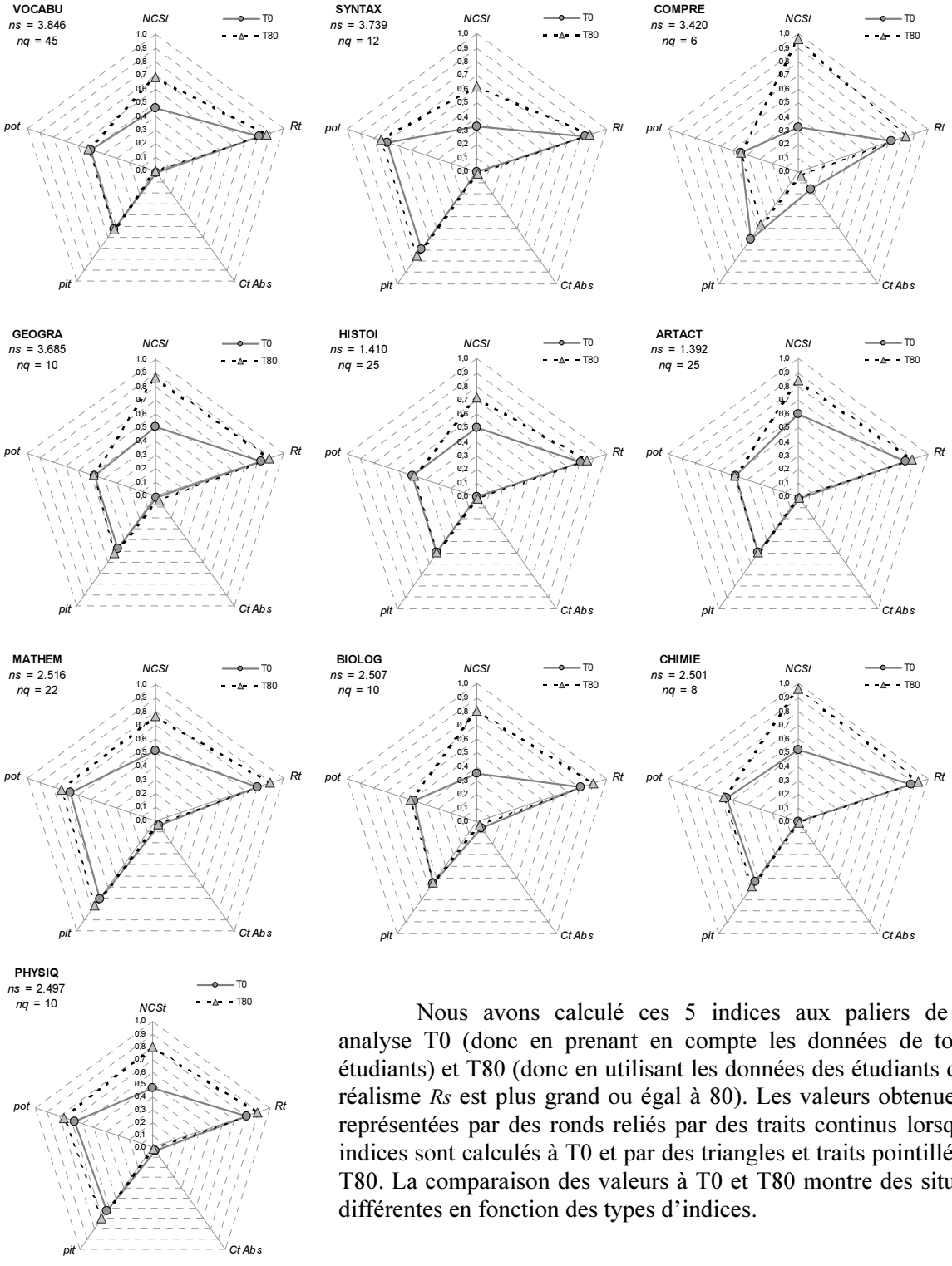
On constate donc que les différences observées entre l'alpha calculé à l'aide de matrices binaires (α_{mb}) ou à partir de matrices spectrale (α_{ms}) lorsqu'on prend en compte d'une part tous les sujets $Rs \geqslant 0$ et d'autre part les sujets $Rs \geqslant 90$ n'évoluent pas pour sept épreuves, ni dans le sens d'une augmentation ni dans le sens d'une diminution. Pour trois épreuves on assiste à une évolution des différences.

L'amélioration de la fidélité qui est observée lorsque l'alpha de Cronbach est calculé à partir des données des étudiants les plus réalistes pour les 10 épreuves dans le cas des matrices spectrales et pour 9 épreuves sur 10 dans le cas des matrices binaires est remarquable et constitue à notre avis un argument supplémentaire en faveur de l'utilisation des pourcentages de certitude dans le contexte de tests tels que les check up MOHICAN. Rappelons que les sujets soumis aux tests MOHICAN n'étaient pas entraînés à l'utilisation des pourcentages de certitude, dès lors on peut supposer que s'ils avaient pu bénéficier d'un entraînement préalable ils auraient probablement amélioré leurs scores de réalisme, nous aurions alors dans les résultats de l'épreuve une plus grande proportion de données de sujets commettant moins d'erreurs dans leurs auto-estimations et en conséquence on peut émettre l'hypothèse que la fidélité des tests en aurait été améliorée. Nous espérons pouvoir vérifier cette hypothèse dans nos prochains travaux de recherche.

3. A propos de la qualité spectrale des épreuves

Peut-on se doter d'un outil de visualisation de la qualité spectrale des tests ?

Les ingénogrammes de qualité spectrale reprennent les valeurs des indices : Niveau de Cohérence Spectrale du test (*NCSt*), Réalisation des prédictions par test (*Rt*), Centration par test calculée en valeur absolue (*Ct Abs*), facilité introspective du test (*pit*) et facilité objective du test (*pot*).



Nous avons calculé ces 5 indices aux paliers de turbo analyse T0 (donc en prenant en compte les données de tous les étudiants) et T80 (donc en utilisant les données des étudiants dont le réalisme *Rs* est plus grand ou égal à 80). Les valeurs obtenues sont représentées par des ronds reliés par des traits continus lorsque les indices sont calculés à T0 et par des triangles et traits pointillés pour T80. La comparaison des valeurs à T0 et T80 montre des situations différentes en fonction des types d'indices.

Les ingénogrammes de la qualité spectrale des tests nous permettent de tirer les constats suivants

En ce qui concerne le *NCSt* nous remarquons que les différences observées à T0 et T80 sont beaucoup plus élevées dans le cas du test COMPRE lorsqu'on compare les valeurs des 10 épreuves. A T0, COMPRE et SYNTAX sont les épreuves qui présentent les *NCSt* les moins élevés, respectivement : 0,32 et 0,33 (voir tableau p. 295). Par contre à T80 le *NCSt* de COMPRE vaut 0,96 (le plus élevé avec CHIMIE) tandis que SYNTAX obtient le *NCSt* le moins élevé des 10 épreuve (0,62).

Du point de vue de l'indice *Rt* nous remarquons que les valeurs à T0 et T80 sont très proches quelque soit l'épreuve.

Rappelons que pour l'indice de centration par test calculé en valeur absolue (*Ct Abs*), plus les valeurs sont proches de zéro (du centre du graphique) plus la tendance à la surestimation ou à la sous-estimation dans les résultats des questions est forte. Pour toutes les épreuves on observe à T80 des valeurs très proches de zéro ce qui est logique, les étudiants les plus réalistes (à $Rs \geqslant 80$) commettant peu d'erreurs d'auto-estimations. On remarque aussi qu'à T0 les valeurs sont assez proches de zéro sauf pour le test COMPRE. Le tableau des valeurs de *Ct* aux différents paliers turbo (p. 302) montre que lorsqu'on prend en compte les données de tous les étudiants (à T0) l'épreuve COMPRE présente une tendance relativement élevée à la surestimation.

Concomitamment avec ce qui vient d'être dit pour l'indice *Ct Abs* (pour rappel, *Ct* = *pit* - *pot*) , nous observons en ce qui concerne la facilité introspective (*pit*) et la facilité objective (*pot*), pour 9 épreuves (toutes sauf COMPRE), des valeurs assez proches à T0 et T80, ce qui donne des lignes en pointillés (T80) et en traits continus (T0) qui soit se recouvrent (VOCABU, HISTOI, ARTACT, BIOLOG), soit sont légèrement décalées (SYNTAX, GEOGRA, MATHEM, CHIMIE, PHYSIQ) et sont pratiquement parallèles. Seul le test COMPRE présente un non parallélisme marqué de la ligne en pointillés et de celle en trait continu dû au fait qu'à T0 *pit* = 0,60 et *pot* = 0,44 (voir tableau p. 299) d'où un décalage du rond et du triangle (d'où la valeur *Ct Abs* = 0,60 – 0,44 = 0,16) alors qu'à T80 rond et triangle sont pratiquement confondu (*pit* = 0,47 et *pot* = 0,44, donc *Ct Abs* = 0,03). A l'aide de l'ingénogramme nous visualisons la valeur plus élevée de pit par rapport à pot à T0, ce qui nous permet de déduire qu'il existe une tendance à la surestimation dans les résultats du test COMPRE.

Précédemment nous nous étions demandé ([3.2], p. 269) si la facilité introspective par test (*pit*) augmentait en fonction des paliers turbo dans toutes les épreuves comme on l'observe dans le cas du test de physique. Le tableau des valeurs *pit* et *pot* (p. 299) montre que l'indice *pit* monte légèrement pour les épreuves SYNTAX, MATHEM et PHYSIQ (augmentation d'environ 0,1), monte aussi très légèrement pour GEOGRA et CHIMIE (augmentation de 0,04), est quasi stationnaire pour VOCABU, ARTACT et diminue très légèrement à T90 pour HISTOI et BIOLOG (diminution de 0,05). Le test COMPRE se distingue des autres par une diminution plus forte et constante de T0 à T90 (diminution de 0,19), donc dans ce dernier cas plus les étudiants sont réalistes, plus le test leur paraît d'une difficulté proche de sa difficulté réelle (l'indice *Ct* nous renseigne sur l'écart entre la difficulté introspective et objective et à T80 cet écart est quasi nul [0,03]).

Globalement, les valeurs de *Rt*, *Ct Abs*, *pit* et *pot* évoluent peu entre T0 et T80 sauf pour l'épreuve COMPRE où une surestimation plus marquée apparaît à T0 mais est résorbée à T80.

L'écart entre les *NCSt T0* et *NCSt T80* se creuse plus fort pour COMPRE que pour les autres tests. Remarquons que du point de vue de la cohérence interne, les alpha (α_{mb} et α_{ms}) de l'épreuve COMPRE étaient les moins élevés (p. 309).

Les ingénogrammes de la qualité spectrale des tests permettent de visualiser rapidement les valeurs obtenues par les indices du Niveau de Cohérence Spectrale du test (NCSt), de Réalisation des prédictions par test (Rt), de Centration par test calculée en valeur absolue (Ct Abs), de facilité introspective du test (pit) et de facilité objective du test (pot). Nous exposerons dans la dernière partie de cette thèse nos projets d'interfaces de gestion de la qualité spectrale des tests qui intègrent ces ingénogrammes (p. 440).

4. Comparaison de la cohérence interne et de la qualité spectrale

Les indices classiques de cohérence interne et les indices de qualité spectrale mesurent des propriétés différentes des tests. Dans quelle mesure les résultats de ces analyses différentes se rejoignent-ils au niveau « TEST » ? Les épreuves les mieux classées en cohérence interne sont-elles aussi les mieux classées en qualité spectrale ?

Ordonnons les valeurs obtenues par les coefficients alpha calculés d'une part sur matrices spectrales (α_{ms}) ou binaires (α_{mb}), et, calculés d'autre part aux paliers turbo T0 et T80 ($\alpha_{mb\ T80}$) (voir graphiques p. 310). Les classements obtenus pour les 10 épreuves aux quatre indices α_{mb}, $\alpha_{mb\ T80}$, α_{ms} et $\alpha_{ms\ T80}$ figurent dans les tableaux ci-dessous. Nous remarquons que les deux premières épreuves sont toujours VOCABU et MATHEM, ensuite les classements changent peu sauf pour les épreuves COMPRE et SYNTAX qui échangent leurs dernière et cinquième places dans le classement lié à l'indice $\alpha_{mb\ T80}$.

		α_{mb}
1	VOCABU	0,833
2	MATHEM	0,769
3	ARTACT	0,707
4	HISTOI	0,668
5	SYNTAX	0,572
6	GEOGRA	0,530
7	PHYSIQ	0,473
8	CHIMIE	0,414
9	BIOLOG	0,410
10	COMPRE	0,393

		$\alpha_{mb\ T80}$
1	VOCABU	0,834
2	MATHEM	0,755
3	ARTACT	0,713
4	HISTOI	0,680
5	COMPRE	0,554
6	GEOGRA	0,543
7	CHIMIE	0,518
8	PHYSIQ	0,498
9	BIOLOG	0,477
10	SYNTAX	0,445

		α_{ms}
1	VOCABU	0,832
2	MATHEM	0,788
3	ARTACT	0,686
4	HISTOI	0,652
5	SYNTAX	0,615
6	PHYSIQ	0,565
7	GEOGRA	0,437
8	CHIMIE	0,430
9	BIOLOG	0,374
10	COMPRE	0,221

		$\alpha_{ms\ T80}$
1	VOCABU	0,840
2	MATHEM	0,760
3	HISTOI	0,658
4	ARTACT	0,670
5	SYNTAX	0,561
6	PHYSIQ	0,572
7	GEOGRA	0,442
8	CHIMIE	0,495
9	BIOLOG	0,403
10	COMPRE	0,317

Comparons ces classements relativement stables avec ceux ci-dessous obtenus pour les indices de qualité spectrale *Rt T0*, *Rt T80*, *NCSt T0* et *NCSt T80*.

		Rt T0
1	CHIMIE	0,870
2	SYNTAX	0,833
3	ARTACT	0,827
4	GEOGRA	0,824
5	VOCABU	0,809
6	HISTOI	0,802
7	BIOLOG	0,800
8	MATHEM	0,797
8	PHYSIQ	0,797
10	COMPRE	0,717

		Rt T80
1	CHIMIE	0,927
2	BIOLOG	0,902
3	MATHEM	0,896
4	PHYSIQ	0,894
5	GEOGRA	0,887
7	ARTACT	0,878
6	SYNTAX	0,875
8	VOCABU	0,861
9	HISTOI	0,848
10	COMPRE	0,830

		NCSt T0
1	ARTACT	0,60
2	CHIMIE	0,52
3	MATHEM	0,51
4	GEOGRA	0,50
4	HISTOI	0,50
7	PHYSIQ	0,47
6	VOCABU	0,46
8	BIOLOG	0,35
9	SYNTAX	0,33
10	COMPRE	0,32

		NCSt T80
1	CHIMIE	0,96
1	COMPRE	0,96
3	GEOGRA	0,86
4	ARTACT	0,84
5	BIOLOG	0,80
5	PHYSIQ	0,80
7	MATHEM	0,76
8	HISTOI	0,72
9	VOCABU	0,68
10	SYNTAX	0,62

Envisageons d'abord les différences dans les classements de chaque indice à T0 et T80. Nous remarquons qu'en ce qui concerne l'indice *Rt T0* et *Rt T80* seuls les premiers et derniers tests (CHIMIE et COMPRE) gardent leurs positions dans les classements des épreuves aux paliers T0 et T80. Pour les indices *NCSt T0* et *NCSt T80* tous les classements changent sauf pour CHIMIE et PHYSIQ et nous remarquons que COMPRE passe de la dernière à la première position (*ex æquo* avec CHIMIE). On observe donc des bouleversements importants lorsqu'on compare T0 à T80 dans les classements des épreuves avec les indices spectraux *Rt T0*, *Rt T80*, et *NCSt T0* et *NCSt T80.*

L'utilisation des données des sujets les plus réalistes entraîne des augmentations d'ampleurs différentes aux indices spectraux. La progression la plus forte étant celle du test COMPRE à l'indice *NCSt* (voir à ce sujet le graphique p. 295) qui passe de 0,32 à *NCSt T0* à 0,96 à *NCSt T80.* Ceci dit, les plus fortes progressions à l'indice *NCSt* doivent être relativisées étant donné que l'indice peut théoriquement varier

entre –2 et +2 alors que la plage de variation théorique de l'indice *Rt* varie, elle, entre 0 et 1 dans les tableaux comparatifs ci-dessus.

Lorsque nous comparons les classements réalisés avec les indices de qualité spectrale et ceux obtenus à l'aide de l'alpha, nous ne remarquons pas de similitudes sauf pour la dernière position attribuée à COMPRE dans 3 cas sur 4. On voit qu'une épreuve comme VOCABU comparée aux autres épreuves MOHICAN peut être considérée de bonne qualité en ce qui concerne la cohérence interne (quel que soit le type de matrice et le palier turbo T0 ou T80) et de qualité relativement moyenne pour l'indice spectral Rt T0, voire considérée comme figurant parmi les moins bien classées en ce qui concerne les autres indices spectraux Rt T0, NCSt T0 et NCSt T80. Les indices de qualité spectrale mesurant autre chose que la cohérence interne il est logique qu'on aboutisse à des classements différents.

Rappelons en ce qui concerne les indices de qualité spectrale que Rt étant calculé au départ de la moyenne des erreurs absolues de prédictions observées pour chacun des 6 pourcentages de certitude de chaque question, il donne une idée précise du taux de réalisation de ces prédictions. Ct, lui, nous donne une indication sur la tendance globale à la sur ou sous-estimation dans les résultats. Enfin, le NCSt montre la cohérence d'utilisation des certitudes dans les résultats, c'est-à-dire la tendance à obtenir des certitudes plus élevées pour les propositions correctes que pour les incorrectes.

Chapitre X :

Exploration du niveau « QCM »

Sommaire

A. Introduction

Précédemment nous avons proposé d'utiliser une série d'indices pour évaluer la qualité des épreuves MOHICAN (voir p. 286). Nous les avons classés dans trois catégories : (1) les indices spectraux, (2) les indices classiques calculés à l'aide des matrices binaires et (3) les indices classiques calculés à partir de matrices spectrales (voir p. 132).

Nous avons aussi proposé d'explorer les épreuves MOHICAN à l'aide de ces trois catégories d'indices, selon trois niveaux de profondeur d'analyse : TEST, QCM et PROPOSITION.

Après avoir exploré le niveau TEST dans la partie précédente, nous allons maintenant explorer le niveau QCM à l'aide des indices suivants :

	(1) INDICES SPECTRAUX	(2) INDICES CLASSIQUES SUR MATRICE BINAIRE	(3) INDICES CLASSIQUES SUR MATRICE SPECTRALE
NIVEAU QCM	• *NCSq* : *N*iveau de *C*ohérence *S*pectrale d'une *q*uestion (p. 231) ; • *Rq* : *R*éalisation des prédictions au niveau d'une *q*uestion (p. 242) ; • *piq* : *facilité i*ntrospective d'une *q*uestion (p. 251) ; • *Cq* : *C*entration moyenne d'une *q*uestion (p. 270).	• *poq mb* : indice de *facilité o*bjective de la *q*uestion calculé à partir de la matrice binaire (p. 254) ; • *NCIq* : Niveau de Cohérence Interne de la question (p. 233) ; • α-*q mb* : alpha calculé à partir d'une *m*atrice *b*inaire après suppression de la question *q* envisagée (p. 141) ; • $r_{qt\ mb}$: corrélation question-total calculée à partir d'une *m*atrice *b*inaire (p. 141).	• α-*q ms* = alpha calculé à partir d'une *m*atrice *s*pectrale après suppression de la question *q* envisagée (p. 141) ; • $r_{qt\ ms}$ = corrélation question-total calculé à partir d'une *m*atrice *s*pectrale (p. 141).

Le logiciel *SCANTEST 2.0 pour épreuves MOHICAN* (p. 199) nous a permis d'obtenir les valeurs des indices présentés dans le tableau ci-dessus pour chacune des questions des 10 épreuves. Nous avons ensuite calculé les corrélations entre ces indices (par exemple entre le coefficient alpha obtenu par le test lorsqu'on retire une question *q* [α-*q*] et le Niveau de Cohérence Interne de cette question [*NCIq*]).

B. Corrélations entre les indices d'évaluation de la qualité des QCM

Nous pensons qu'il existe des liaisons entre les différents indices d'évaluation de la qualité des QCM. Pour les indices de cohérence interne nous nous attendons à des corrélations très élevées, mais ces indices sont-ils corrélés aux autres indices de qualité spectrale ? Probablement moins étant donné qu'ils mesurent des propriétés différentes au sein des QCM. Dans cette section nous allons d'abord présenter les corrélations obtenues par les indices classiques, ensuite nous envisagerons les corrélations relatives aux indices spectraux. Enfin, nous examinerons les liaisons entre indices classiques et spectraux.

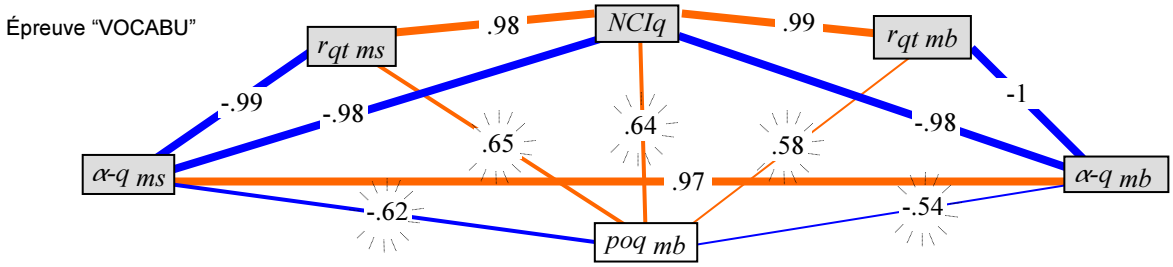
1. Corrélations des valeurs obtenues par les QCM aux indices classiques calculés à l'aide des matrices binaires et spectrales

Nous nous attendons à ce que les indices α-q $_{mb}$ (alpha calculé à partir d'une matrice binaire après suppression de la question q envisagée), $r_{qt\ mb}$ (corrélation question-total calculée à partir d'une matrice binaire), varient ensemble et soient même très fortement liés étant donné ce qu'ils mesurent (la cohérence interne). Nous nous attendons aussi à des corrélations très élevées entre les indices α-q $_{ms}$, $r_{qt\ ms}$ (ces deux indices étant quant à eux calculés à partir de matrices spectrales). Des laisons fortes devraient également exister entre ces indices et le Niveau de Cohérence Interne de la question (NCIq) qui est basé sur le rpbis classique de la réponse correcte auquel on soustrait la moyenne pondérée des rpbis classiques des solutions incorrectes (voir p. 233). Tous ces indices pourraient aussi être corrélés avec l'indice poq $_{mb}$ (facilité objective d'une question classiquement calculée sur des matrices binaires). Nous nous attendons à des corrélations entre la facilité des questions et les indices de cohérence interne tout en ignorant l'ampleur des liaisons.

Voici les corrélations obtenues par les indices classiques calculés sur matrice binaire pour l'épreuve VOCABU (épreuve comportant 45 questions soumises à 3.846 étudiants et dont les coefficients alpha étaient les plus élevés des 10 épreuves MOHICAN : α_{mb} = 0,833 et α_{ms} = 0,832, voir p. 309).

A l'aide de traits d'épaisseurs différentes nous avons mis en évidence le degré de corrélation des indices *poq $_{mb}$*, *NCIq*, *α-q $_{mb}$*, *$r_{qt\ mb}$*, *α-q $_{ms}$* et *$r_{qt\ ms}$*. Les quatre indices de cohérence interne sont mis en évidence à l'aide de boîtes grisées. Nous traçons des traits gras épais lorsque les corrélations en valeurs absolues sont supérieures à .8, des traits d'épaisseur moyenne lorsqu'elles sont comprises entre .6 et .79 et des traits plus fins lorsqu'elles sont inférieures à .6. Lorsqu'une corrélation est non significative nous la signalons par un trait fin en pointillés accompagné de la mention « ns ». Les corrélations positives sont signalées par des traits de couleur orange et les négatives par des traits bleus.

Quelle est l'ampleur des liaisons entre la facilité objective (poq $_{mb}$) et les indices de cohérence interne ?



Toutes les corrélations de cette épreuve VOCABU présentées sur le schéma sont significatives à p<0,05. Nous observons des corrélations moins élevées (nous avons mis en évidence leurs valeurs sur le graphique) avec la facilité/difficulté objective des questions (poq_{mb}) pour le *NCIq* et pour les deux autres indices calculés à l'aide des matrices spectrales ($r_{qt\ ms}$ et $\alpha\text{-}q_{ms}$). Ces corrélations existent aussi pour les indices $r_{qt\ mb}$ et $\alpha\text{-}q_{mb}$ calculés à l'aide des matrices binaires mais dans ce cas elles sont alors encore un peu moins élevées.

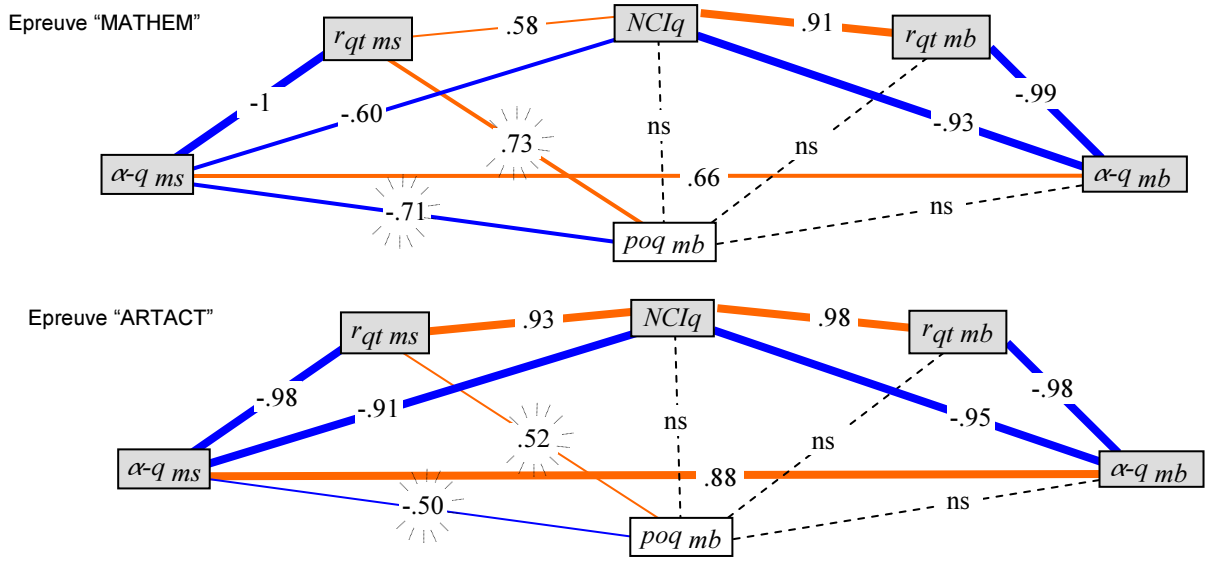
Donc, plus les questions de cette épreuve sont faciles (plus les proportions de réponses correctes sont élevées) et plus les indices de cohérence interne de ces questions augmentent (en valeur absolue pour α-q).

Nous remarquons aussi des corrélations très élevées (et attendues) entre les valeurs des niveaux de cohérence interne des questions (*NCIq*) calculés à l'aide des *rpbis* classiques (voir p. 233), les valeurs des corrélations question/total (r_{qt}, p. 141) et les valeurs des coefficients alpha après suppression de la question considérée ($\alpha\text{-}q$, p. 141). En ce qui concerne ces deux derniers indices, r_{qt} et $\alpha\text{-}q$, les corrélations sont pratiquement les mêmes qu'ils aient été calculés à l'aide de matrices binaires (…*mb*) ou à l'aide de matrices spectrales (…*ms*) (voir p. 132). Nous remarquons également une corrélation très élevée entre $\alpha\text{-}q_{mb}$ et $\alpha\text{-}q_{ms}$.Ces valeurs élevées sont logiques dans la mesure où il s'agit d'indices qui mesurent la même propriété des questions : leur cohérence interne.

Les très fortes corrélations entre r_{qt} et $\alpha\text{-}q$ sont négatives car si les scores d'une question sont très corrélés avec les scores totaux du test il est logique que le coefficient alpha du test diminue si on retire cette question. Donc lorsque les r_{qt} sont très élevés et positifs, les $\alpha\text{-}q$ ont tendance à diminuer. Il en va de même pour les niveaux de cohérence interne des questions : si pour une question donnée les étudiants qui fournissent la réponse correcte obtiennent des scores élevés au test et ceux qui choisissent les réponses incorrectes des scores totaux faibles (ce qui amène un *NCIq* élevé) il est logique que la suppression de cette question entraîne une diminution du coefficient alpha de l'épreuve.

Rappelons que la qualité des épreuves MOHICAN n'est pas égale. Du point de vue de la cohérence interne des questions, les trois tests qui obtiennent les meilleurs coefficients alpha (voir p. 309) sont VOCABU (α_{ms} 0,832), MATHEM (α_{ms} 0,788) et ARTACT (α_{mb} 0,707). Ensuite viennent les épreuves HISTOI (α_{ms} 0,668), SYNTAX (α_{ms} 0,615), PHYSIQ (α_{ms} 0,565), BIOLOG (α_{ms} 0,563), GEOGRA (α_{mb} 0,530). Enfin, deux épreuves obtiennent des scores de cohérence interne particulièrement faibles : CHIMIE (α_{ms} 0,430) et COMPRE (α_{mb} 0,393).

Les deux autres meilleures épreuves après VOCABU du point de vue de la cohérence interne mesurée à l'aide de l'alpha sont MATHEM (*nq* = 22) et ARTACT (*nq* = 25). Ces deux épreuves obtiennent les corrélations suivantes :



En ce qui concerne l'épreuve MATHEM toutes les corrélations sont significatives à p<0,05 sauf pour trois corrélations signalées par « ns » sur le schéma. En effet, nous ne retrouvons plus les corrélations significatives entre les indices de cohérence interne calculés à l'aide de matrices binaires et la facilité objective des questions (poq_{mb}). Par contre les corrélations sont significatives avec poq_{mb} lorsque r_{qt} et α-q sont calculés à partir des matrices spectrales. Par ailleurs, les corrélations restent très fortes entre $NCIq$, $r_{qt\ mb}$ et α-q_{mb}. La corrélation est maximale entre $r_{qt\ ms}$ et α-q_{ms}.

En ce qui concerne ARTACT, nous observons également des corrélations non significatives entre les indices de cohérence interne $NCIq$, $r_{qt\ mb}$, α-q_{mb} calculés à partir de matrices spectrales et la facilité objective des questions (poq_{mb}) (« ns » sur le schéma). En ce qui concerne la liaison entre la facilité objective et les indices de cohérence interne calculés à partir de matrices spectrales, les corrélations sont significatives à p<0,05 mais moins élevées que dans le cas des deux épreuves précédentes. Les autres corrélations calculées pour les indices $NCIq$, α-q_{mb} , $r_{qt\ mb}$, α-q_{ms} et $r_{qt\ ms}$ sont significatives à p<0,05 et très marquées (comme nous nous y attendions car il s'agit d'indices mesurant la cohérence interne).

Pour ces deux épreuves les corrélations entre les valeurs des coefficients alpha après suppression de question (α-q) calculé à partir de matrices binaire ou spectrale sont moins élevés que dans le cas du test VOCABU.

En résumé, pour ces trois épreuves dont la cohérence interne est supérieure aux sept autres (voir p. 309) deux types de relations très fortes auxquelles nous nous attendions sont observées : (1) une série de corrélations négatives très élevées (voire égales à -1) entre r_{qt} et α-q , que ces indices aient été calculés à l'aide de matrices binaires ou spectrales ainsi que (2) des corrélations en moyenne un peu moins élevées entre NCIq et les indices r_{qt} et α-q calculés sur matrices binaires ou spectrales. Les corrélations négatives obtenues avec l'alpha en cas de suppression de la question considérée (α-q) sont logiques car plus les indices r_{qt} et NCIq d'une question récoltent des valeurs élevées et plus le coefficient alpha diminue en cas de suppression de cette question performante en ce qui concerne sa cohérence interne avec les autres questions du test.

Les corrélations élevées, voire très élevées, obtenues par les NCIq avec r_{qt} et α-q calculés sur matrices binaires ou spectrales, montrent que l'indice NCIq basé sur le rpbis classique de la réponse correcte auquel on soustrait la moyenne pondérée des rpbis classiques des solutions incorrectes (voir p. 233) donne des valeurs fortement liées à celles obtenues par les indices plus classiquement utilisés pour calculer la cohérence des résultats d'une question par rapport aux résultats de l'ensemble des questions d'une épreuve. La liaison avec les indices classiques de cohérence interne des questions des trois épreuves envisagées est particulièrement élevée lorsqu'on utilise les matrices binaires pour les calculer. Pour l'épreuve MATHEM lorsqu'on se base sur la matrice spectrale des résultats pour calculer les indices r_{qt} et α-q, la liaison entre ces indices et le NCIq n'est plus aussi élevée.

On observe une liaison entre la facilité des questions (les proportions de réponses correctes par question) et les indices de cohérence interne mentionnés plus haut dans les trois tests, les corrélations se situent aux alentours de . 60 pour l'épreuve VOCABU, de .70 pour MATHEM et de .50 pour ARTACT. En ce qui concerne les épreuves MATHEM et ARTACT les corrélations calculées à partir des matrices binaires sont non significatives.

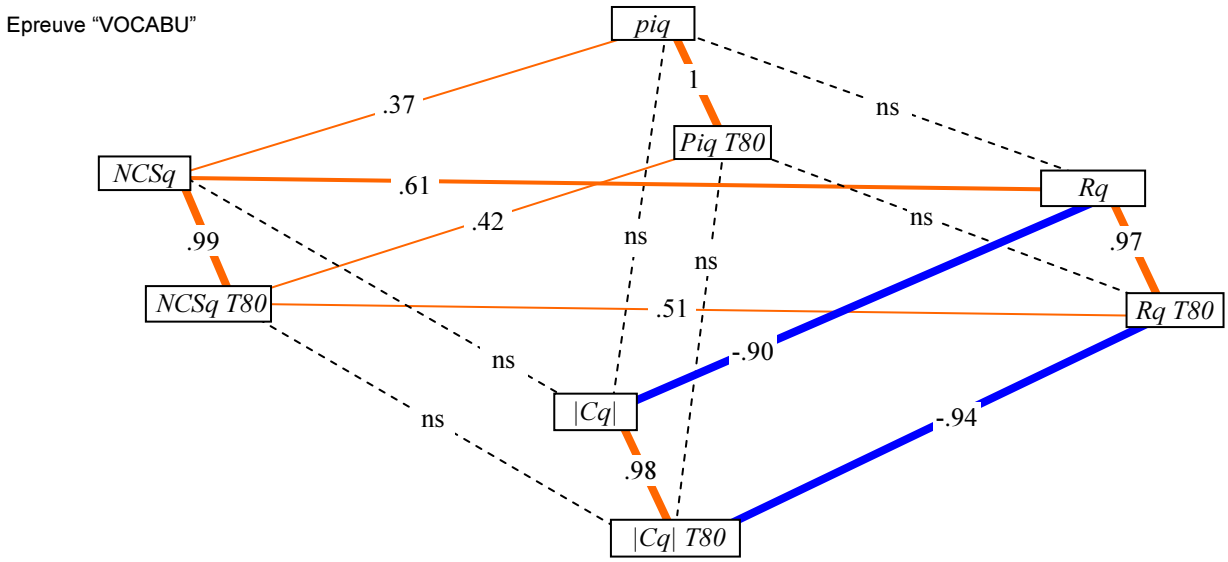
2. Corrélations entre les valeurs obtenues par les QCM aux indices spectraux

Dans cette sous-section consacrée aux corrélations observées entre indices spectraux, nous nous attendons à une liaison négative forte entre les valeurs obtenues à l'indice Rq (Réalisation des prédictions au niveau d'une question, voir p. 242) et à l'indice Cq (Centration moyenne d'une question, p. 270, plus Cq est proche de zéro moins il y a de sur ou sous-estimations) étant donné ce qu'ils mesurent (les erreurs d'auto-estimations dans les résultats des questions). Nous nous attendons aussi à des corrélations entre ces deux indices et le NCSq (Niveau de Cohérence Spectrale d'une question, p. 231) mais nous pensons qu'elles seront moins élevées. Le NCSq mesure en effet au niveau des résultats d'une question, une propriété spectrale voisine des erreurs d'auto-estimations mais cependant différente : la « cohérence spectrale » c'est-à-dire la propension à utiliser des certitudes plus élevées chez les sujets qui répondent correctement que chez les sujets qui répondent incorrectement. L'observation de corrélations positives sans être très fortes fortes entre NCSq et les indices Rq et Cq plaiderait en faveur d'une complémentarité de ces deux catégories d'indices spectraux.

En ce qui concerne la Centration par question (*Cq*), nous avons utilisé la valeur absolue de l'indice (|*Cq*|) pour le calcul des corrélations avec les autres indices spectraux. En effet, lorsqu'une question récolte –10 à l'indice de Centration par question (*Cq*, voir p. 254), le signe « - » indique une propension à la sous-estimation dans les réponses fournies, il est évident que cette erreur moyenne dans les auto-estimations n'est pas deux fois celle qui aurait été commise si la question avait récolté un score à l'indice *Cq* égal à +10, en fait il s'agit de la même « quantité » d'erreur d'auto-estimation mais dans un cas il s'agit de sous-estimations (-) et dans l'autre de surestimation (+).

Nous avons aussi corrélé les valeurs obtenues par ces indices spectraux au palier de turbo analyse T80 (les indices spectraux sont alors calculés à partir des données de sujets dont le réalisme est supérieur ou égal à 80).

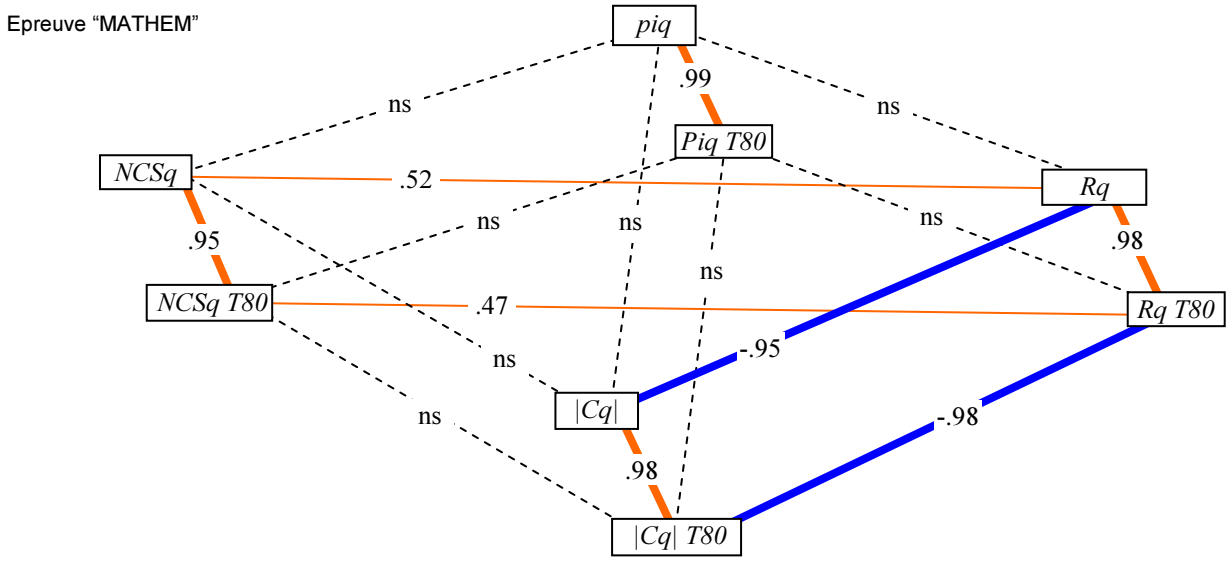
Voici les corrélations obtenues par les 45 QCM de l'épreuve VOCABU aux indices spectraux.



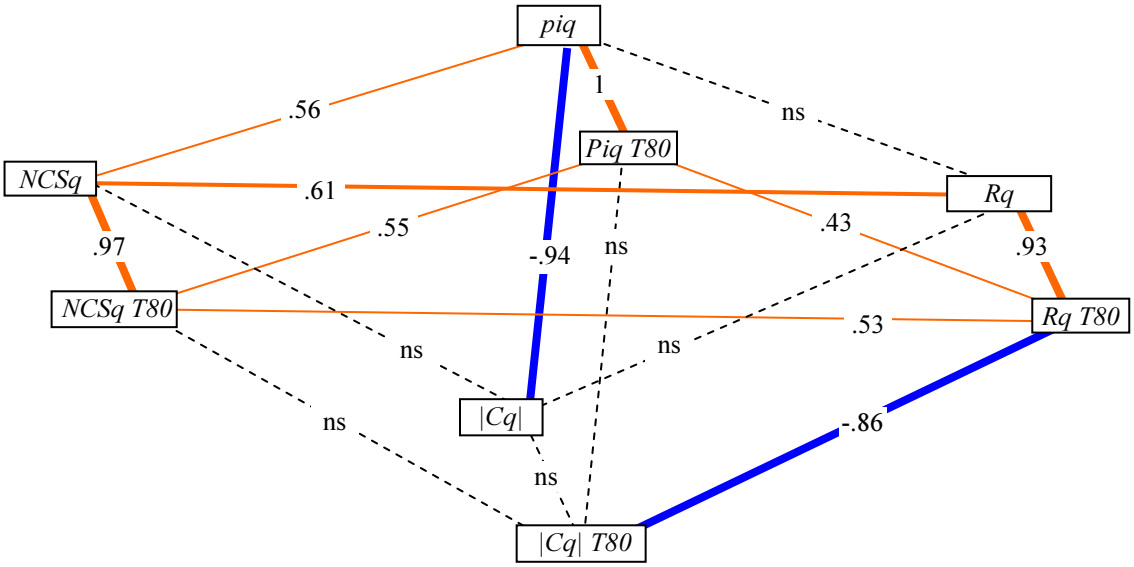
Nous observons des corrélations très élevées entre |*Cq*|, |*Cq*| *T80*, *Rq* et *Rq T80*. Des corrélations très marquées existent aussi entre chaque indice spectral et sa variante calculée à l'aide de données de sujets dont le réalisme est supérieur ou égal à 80. Nous remarquons également une corrélation assez élevée entre les valeurs obtenues par les questions à l'indice *NCSq* et les valeurs *Rq*.

Remarquons aussi l'absence de corrélation significative entre les indices *NCSq* et |*Cq*| ainsi qu'entre *piq* et d'une part *Rq* et d'autre part |*Cq*|.

Voici les corrélations des deux autres meilleures épreuves (MATHEM et ARTACT) après VOCABU du point de vue de la cohérence interne mesurée à l'aide de l'alpha :



On retrouve dans l'épreuve MATHEM les mêmes tendances que pour VOCABU en ce qui concerne les indices spectraux mais à la différence qu'ici les corrélations entre *NCSq* et *piq* disparaissent.



La situation est assez différente en ce qui concerne l'épreuve ARTACT. Des corrélations significatives présentes dans les deux épreuves précédentes disparaissent ici : entre |*Cq*| et |*Cq*| *T80*, ainsi qu'entre |*Cq*| et *Rq*. D'autres corrélations significatives apparaissent ici : très élevée entre |*Cq*| et *piq* et assez

élevée entre *piq T80* et *Rq T80.* Deux corrélations qui étaient présentes dans VOCABU et absentes dans MATHEM réapparaissent entre *NCSq* et *piq* ainsi qu'entre *NCSq T80* et *piq T80.*

Dans le cadre de cette épreuve ARTACT il semble exister une relation entre la facilité introspective des questions et les autres indices spectraux.

Dans les graphiques qui précèdent nous observons des corrélations très élevées entre chaque indice spectral et sa variante obtenue après turbo analyse à T80. Une exception cependant, pour l'indice *|Cq| qui* n'est pas significativement corrélé avec l'indice *|Cq|* T80 dans l'épreuve ARTACT. Nous remarquons également l'absence de corrélation significative entre *|Cq|* et *Rq* alors qu'on trouve des corrélations élevées entre ces indices dans les deux autres épreuves VOCABU et MATHEM.

Les corrélations négatives élevées entre *|Cq|* et *Rq* sont logiques dans la mesure où plus les prédictions se réalisent, c'est-à-dire plus les Taux d'Exactitude Annoncés (*TEA*) sont proches des Taux d'Exactitude Observés (*TEO*), moins les erreurs d'estimation seront fortes et fréquentes pour une question donnée. Donc, quand *Rq* est élevé on peut s'attendre à ce que *|Cq|* soit faible car il y aura peu de sous-estimations ou de surestimations. C'est ce que nous constatons pour VOCABU et MATHEM ainsi que pour ARTACT lorsque ces indices sont calculés à un palier turbo T80.

Nous voyons à partir des corrélations observées dans VOCABU, MATHEM et ARTACT que les valeurs obtenues aux indices spectraux par les questions de ces trois épreuves ne varient pas forcément de concert alors qu'une tendance à varier ensemble existe pour les trois indices classiques de cohérence interne : *NCIq*, r_{qt} et α-*q*. Dans ces épreuves nous n'observons pas de relation entre les *NCSq* et les *Cq*. Au palier de turbo analyse T80 les liens entre les *NCSq* et *Rq* ne sont pas aussi forts qu'entre *Rq* et *Cq*.

Ces observations renforcent l'idée d'une complémentarité entre les indices Rq, Cq et NCSq. L'indice de Réalisation des prédictions par question permet d'observer dans quelle mesure des erreurs d'auto-estimation ont été commises au niveau d'une QCM. L'indice Cq permet quant à lui de diagnostiquer si les erreurs d'estimation commises sont dans l'ensemble des sous-estimations (-) ou des surestimations (+). Notons qu'en cas de surestimations flagrantes pour une question très difficile, nous pourrions être amenés à envisager l'hypothèse d'une question qui fonctionnerait comme un piège, la majorité des sujets estimant l'item facile alors que les taux d'exactitude observés démontreraient le contraire.

Les Niveaux de Cohérence Spectrale par question (NCSq) permettent de mesurer la propension à utiliser des pourcentages de certitude plus élevés chez les sujets qui répondent correctement que chez les sujets qui se trompent. Il s'agit d'une mesure « voisine » du Rq car elle exploite aussi les informations liées à l'utilisation des pourcentages de certitude, mais dans un but différent : non plus pour calculer une quantité d'erreurs d'auto-estimations, mais bien pour évaluer une propension à l'utilisation cohérente des certitudes.

Les corrélations suivantes montrent qu'il existe en effet une liaison entre le NCSq et l'indice Rq mais comme nous nous y attendions, les corrélations ne sont pas fortement élevées (VOCABU r = 0,61, MATHEM r = 0,52 et ARTACT r = 0,61). Par ailleurs, nous constatons que ces corrélations faiblissent au palier T80 (VOCABU r T80 = 0,51, MATHEM r T80 = 0,47 et ARTACT r T80 = 0,53). Nous remarquons aussi qu'à T80 elles sont bien moins élevées que les corrélations (en valeurs absolues) qui relient l'indice Rq à l'indice |Cq| à ce palier turbo (VOCABU r T80 = -0,94, MATHEM r T80 = -0,98 et ARTACT r T80 = -0,86).

Nous voyons dans ces constats l'illustration d'une complémentarité entre les trois indices spectraux NCSq, Rq et Cq d'évaluation de la qualité spectrale d'une question. Le NCSq convenant bien pour mesurer la cohérence spectrale d'utilisation des degrés de certitude, le Rq étant particulièrement indiqué pour évaluer le taux d'erreurs d'auto-estimation et Cq, plus particulièrement lié à Rq, permettant de mesurer globalement le « sens » des erreurs d'auto-estimation, c'est-à-dire la tendance à la surestimation ou à la sous-estimation dans les résultats liés à une question.

3. Corrélations entre les valeurs obtenues aux indices classiques et aux indices spectraux

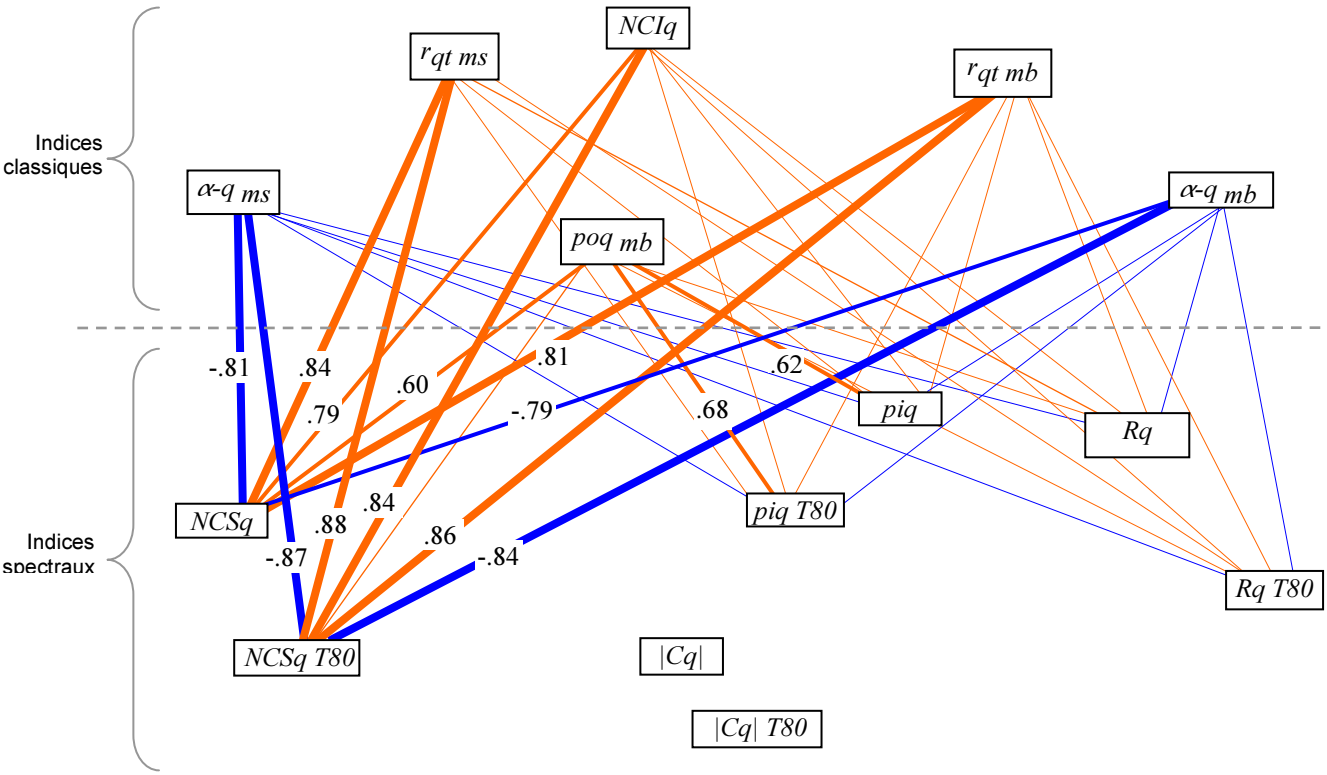
Existe-t-il des liens au niveau des QCM entre les nouveaux indices de qualité spectrale et les indices classiques de cohérence interne ? Nous ne nous attendons pas à des liaisons systématiques et fortes entre les indices classiques et les nouveaux indices de qualité spectrale étant donné que ces deux types d'instruments de contrôle de la qualité des questions mesurent des propriétés différentes : d'une part la cohérence avec les résultats globaux et d'autre part les erreurs d'auto-estimations ainsi que la cohérence spectrale. Qu'en est-il dans les trois épreuves VOCABU, MATHEM et ARTACT ?

Voici les corrélations obtenues entre les indices spectraux et les indices classiques pour l'épreuve VOCABU soumise à 3.846 étudiants. Dans cette épreuve, le nombre de questions (*nq*) s'élève à 45. Les corrélations significatives à p<0,05 figurent en gras dans le tableau ci-dessous.

	piq	*Piq T80*	*NCSq*	*NCSq T80*	*Rq*	*Rq T80*	*|Cq|*	*|Cq| T80*
poq mb	**,62**	**,68**	**,60**	**,59**	**,50**	**,36**	ns (-,28)	ns (-,23)
NCIq	**,50**	**,55**	**,79**	**,84**	**,47**	**,42**	ns (-,21)	ns (-,19)
r_{qt} mb	**,49**	**,54**	**,81**	**,86**	**,44**	**,39**	ns (-,17)	ns (-,14)
α-q mb	**-,48**	**-,53**	**-,79**	**-,84**	**-,42**	**-,37**	ns (,14)	ns (,12)
r_{qt} ms	**,45**	**,51**	**,84**	**,88**	**,52**	**,46**	ns (-,24)	ns (-,21)
α-q ms	**-,41**	**-,47**	**-,81**	**-,87**	**-,51**	**-,45**	ns (,25)	ns (,22)

Nous pouvons représenter ces relations entre indices classiques et indices spectraux à l'aide du schéma ci-dessous. Les indices classiques sont placés au-dessus du trait en pointillés gris et en-dessous les indices spectraux. Comme sur les schémas précédents, la grosseur des traits représente la force des corrélations. Les corrélations qui ne sont pas significatives à p<0,05 ne figurent pas sur le schéma.

Epreuve "VOCABU"
Corrélations entre indices classiques et indices spectraux



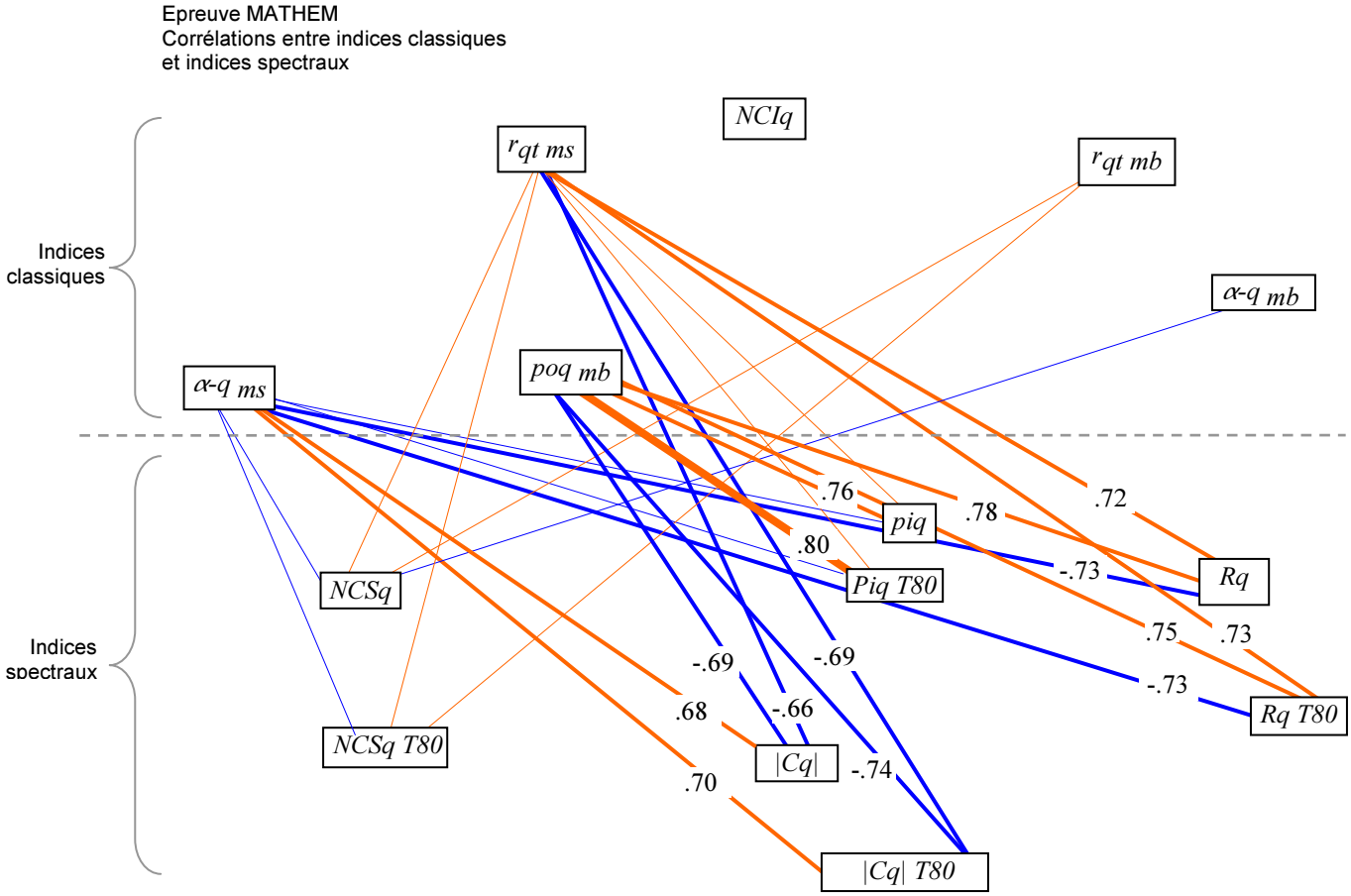
Rappelons que lorsqu'une corrélation qui unit un indice classique à un indice spectral est supérieure ou égale à 0,8 nous traçons un trait épais. Lorsqu'elle est inférieure à 0,8 et supérieure ou égale à 0,6 nous utilisons un trait de grosseur moyenne. Quand une corrélation est inférieure à 0,6 nous utilisons un trait fin. Les corrélations positives sont représentées en orange, les négatives en bleu. Nous voyons sur le schéma qui précède que dans le contexte de l'épreuve VOCABU (nq = 45, α_{mb} = 0,833 et α_{ms} = 0,832) les corrélations les plus fortes se situent entre d'une part le Niveaux de Cohérence Spectrale des questions (*NCSq*) et sa version calculée à T80 et, d'autre part, tous les indices classiques de cohérence interne calculés à l'aide de matrices spectrales ou non.

Nous remarquons aussi que l'indice *Rq* et sa version calculée à T80 sont corrélés avec les indices de cohérence interne, mais de façon moins marquée que le *NCSq*. Nous remarquons aussi des corrélations entre la facilité/difficulté introspective par question (*piq*, *piq T80*) et la facilité/difficulté objective (*poq*).

Par contre il n'y a pas de corrélation significative entre les indices de centration par question en valeurs absolues (*|Cq|*) même au palier T80 et les indices de cohérence interne classiques.

Les corrélations négatives observées entre tous les indices spectraux et les indices classiques alpha de Cronbach obtenus en cas de suppression de la question considérée lorsque l'alpha est calculé à partir d'une matrice binaire (α-q $_{mb}$) ou à partir d'une matrice spectrale (α-q $_{mb}$) s'expliquent par le fait que plus une question est de bonne qualité en ce qui concerne la cohérence interne classique, plus l'alpha diminue si on supprime cette question dans le test. Dès lors les corrélations négatives montrent que les indices spectraux ont tendance à varier avec les indices classiques de cohérence interne : lorsque les indices spectraux des questions sont bons, les indices de cohérence interne ont tendance à l'être aussi et vice versa, lorsque les indices spectraux sont mauvais, ils le sont en général aussi pour les indices de cohérence interne.

Voyons maintenant la représentation schématique des corrélations entre les indices classiques et les indices spectraux pour le test MATHEM. Nous présentons le tableau des corrélations après le schéma.



Voici le tableau des corrélations liées à l'épreuve MATHEM soumise à 2.516 étudiants (*nq* = 22, α_{mb} = 0,769 et α_{ms} = 0,788). Les corrélations significatives à p<0,05 figurent en gras.

	piq	*piq T80*	*NCSq*	*NCSq T80*	*Rq*	*Rq T80*	*|Cq|*	*|Cq| T80*
poq mb	**,76**	**,80**	ns (,18)	ns (,29)	**,78**	**,75**	**-,69**	**-,74**
NCIq	ns (,16)	ns (,17)	ns (,23)	ns (,22)	ns (,08)	ns (,16)	ns (-,10)	ns (-,12)
$r_{qt\ mb}$	ns (,09)	ns (,12)	**,46**	**,43**	ns (,22)	ns (,24)	ns (-,21)	ns (-,21)
α-q_{mb}	ns (-,12)	ns (-,15)	**-,43**	ns (-,40)	ns (-,22)	ns (-,24)	ns (,22)	ns (,21)
$r_{qt\ ms}$	**,49**	**,55**	**,44**	**,49**	**,72**	**,73**	**-,66**	**-,69**
α-q_{ms}	**-,46**	**-,52**	**-,42**	**-,47**	**-,73**	**-,73**	**,68**	**,70**

Sur le schéma, nous remarquons une configuration de corrélations très différente du test VOCABU. Ici, dans le cas de MATHEM, on ne retrouve plus les corrélations très élevées entre *NCSq*, *NCSq T80* et indices de cohérence interne classiques, des corrélations existent mais sont plus faibles que pour VOCABU. L'indice des Niveaux de Cohérence Interne des questions (*NCIq*) n'est plus corrélé significativement avec les indices spectraux.

Par contre on observe dans le cadre de cette épreuve MATHEM des corrélations entre les indices de centration par question en valeurs absolues (*|Cq|* et *|Cq| T80*) et les indices de cohérence interne classiques calculés à l'aide de matrices spectrales (α-q_{ms} et $r_{qt\ ms}$).

On observe aussi des corrélations plus élevées entre les indices de Réalisation des prédictions par question (*Rq* et *Rq T80*) et les indices de cohérence interne par question calculés à l'aide de matrices spectrales (α-q_{ms} et $r_{qt\ ms}$).

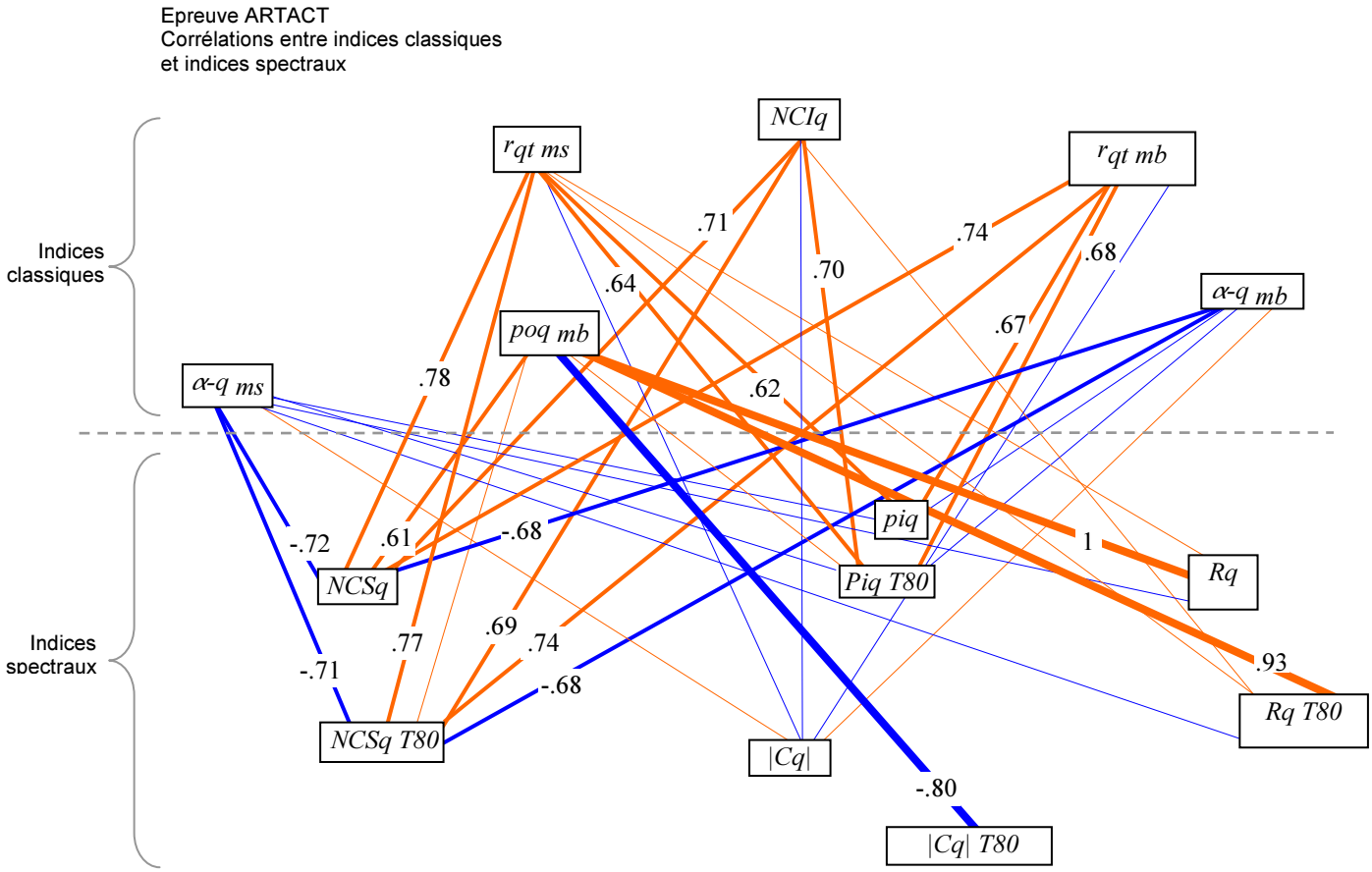
Les corrélations sont aussi plus élevées entre l'indice de facilité/difficulté objective (poq_{mb}) et les indices *Rq*, *Rq T80* et *|Cq|*, *|Cq| T80*. Nous remarquons aussi des corrélations élevées, plus fortes que dans VOCABU, entre la facilité/difficulté introspective par question (*piq*, *piq T80*) et la facilité/difficulté objective (*poq*).

Contrairement à l'épreuve VOCABU où tous les indices de cohérence interne classiques des questions étaient très corrélés avec les indices des niveaux de cohérence spectrale (*NCSq* et *NCSq T80*), ici, dans le cadre de l'épreuve MATHEM, ce sont les indices de réalisation des prédictions par question (*Rq* et *Rq T80*) et de centration par question (*|Cq|* et *|Cq| T80*) qui sont les plus corrélés avec certains indices de cohérence interne classiques : ceux qui ont été calculés à l'aide de matrice spectrales (α-q_{ms} et $r_{qt\ ms}$).

Voici les corrélations liées à l'épreuve ARTACT soumise à 1.392 étudiants (*nq* = 25, α_{mb} = 0,707 et α_{ms} = 0,686). Les corrélations significatives à p<0,05 figurent en gras dans ce tableau.

	piq	*piq T80*	*NCSq*	*NCSq T80*	*Rq*	*Rq T80*	*|Cq|*	*|Cq| T80*
poq mb	ns (,39)	**,41**	**,61**	**,58**	**1**	**,93**	-,05	**-,80**
NCIq	**,68**	**,70**	**,71**	**,69**	ns (,38)	**,48**	**-,59**	ns (-,09)
r_{qt} mb	**,67**	**,68**	**,74**	**,74**	ns (,36)	**,42**	**-,58**	ns (-,02)
α-q mb	**-,55**	**-,56**	**-,68**	**-,68**	ns (-,26)	ns (-,34)	**,49**	ns (-,03)
r_{qt} ms	**,62**	**,64**	**,78**	**,77**	**,52**	**,57**	**-,47**	ns (-,18)
α-q ms	**-,56**	**-,59**	**-,72**	**-,71**	**-,50**	**-,55**	**,41**	ns (,18)

Voici la représentation schématique des corrélations entre les indices classiques et les indices spectraux pour le test ARTACT.



Nous constatons que la configuration des corrélations est encore différente des deux précédentes.

Ici, dans le cadre de l'épreuve ARTACT nous observons des corrélations entre les *NCSq, NCSq T80* et les indices de cohérence interne classiques des questions, mais aussi des corrélations très élevées entre les indices *Rq*, *Rq T80*,*|Cq| T80* et les indices de facilité/difficulté des questions (*poq mb*).

Comme dans les épreuves précédentes, nous remarquons aussi des corrélations assez élevées entre la facilité/difficulté introspective par question (piq, piq T80) et la facilité/difficulté objective (poq). Nous remarquons également qu'en comparaison avec les deux précédentes épreuves, les tendances générales de liaison diffèrent aussi pour ce troisième test.

4. Conclusions à propos des corrélations observées entre les indices d'évaluation de la qualité des QCM pour les épreuves VOCABU, MATHEM et ARTACT

Précédemment nous avons constaté que les meilleures épreuves MOHICAN du point de vue de la cohérence interne mesurée à l'aide du coefficient alpha de Cronbach étaient les trois tests VOCABU, MATHEM et ARTACT (voir p. 309). Nous avons utilisé ces trois épreuves dans l'observation des corrélations entre les indices classiques de cohérence interne (*NCIq*, $r_{qt\ mb}$, α-$q_{\ mb}$, $r_{qt\ ms}$, α-$q_{\ ms}$) et les indices d'analyse spectrale questions (*NCSq*, *NCSq T80*, *Rq*, *Rq T80*, |*Cq*| et |*Cq*| *T80*).

En ce qui concerne les indices classiques de cohérence interne *NCIq*, $r_{qt\ mb}$, α-$q_{\ mb}$, $r_{qt\ ms}$, α-$q_{\ ms}$ nous observons deux relations très fortes : d'une part une série de corrélations négatives très élevées (voire égales à -1) entre r_{qt} et α-*q* que ces indices aient été calculés à l'aide de matrices binaires ou spectrales.

Pour ce qui est de la liaison de l'indice *NCIq* avec les indices r_{qt} et α-*q*, nous remarquons que les corrélations obtenues dans le cadre de l'épreuve MATHEM à partir de matrices spectrales sont moins fortes que dans les autres épreuves où elles sont très élevées.

Lorsque les corrélations entre NCIq et les indices r_{qt} et α-q sont calculées à partir de matrices binaires nous constatons qu'elles sont très élevées, et ce dans les trois épreuves. Comme nous nous y attendions, dans le contexte des tests VOCABU, MATHEM et ARTACT l'indice NCIq basé sur les rpbis classiques donne des valeurs fortement liées à celles obtenues par les indices classiquement utilisés pour calculer la cohérence des résultats d'une question par rapport aux résultats du test.

En ce qui concerne les corrélations entre les indices d'analyse spectrale des questions (*NCSq*, *NCSq T80*, *Rq*, *Rq T80*, |*Cq*| et |*Cq*| *T80*), nous constatons principalement que les valeurs obtenues ne varient pas forcément de concert alors qu'une tendance à varier ensemble existe pour les trois indices de cohérence interne (*NCIq*, r_{qt} et α-*q*). Nous n'observons pas de corrélation significative entre les *NCSq*, *NCSq T80* et les |*Cq*|, |*Cq*| *T80*. Nous remarquons aussi qu'au palier de turbo analyse *T80* les liens entre les *NCSq* et *Rq* ne sont pas aussi forts qu'entre *Rq* et *Cq*.

Les corrélations moins élevées observées entre NCSq et les deux indices Rq et Cq nous paraît aussi intéressante à relever car elles montrent que le NCSq bien qu'ayant tendance à varier avec les Rq, Rq T80, |Cq| et |Cq| T80 apporte en complément de la mesure des erreurs d'auto-estimations une information supplémentaire : la cohérence d'utilisation des pourcentages de certitude, c'est-à-dire la propension à utiliser des certitudes plus élevées chez les sujets qui répondent correctement que chez les sujets qui se trompent.

Le lien entre *Rq* et *Cq* nous paraît logique dans la mesure où ces indices spectraux mesurent des quantités d'erreurs d'auto-estimations dans les résultats des questions.

En ce qui concerne les corrélations entre indices classiques et indices spectraux, nous remarquons que des tendances à varier ensemble existent entre ces deux types d'indices dans chacune des trois épreuves mais les configurations globales des liaisons entre indices spectraux et indices classiques sont assez différentes d'un test à l'autre.

Pour le test VOCABU nous remarquons d'une part des corrélations très élevées entre les *NCSq*, *NCSq T80* et tous les autres indices de cohérence interne classiques et d'autre part des corrélations moins élevées entre les indices *Rq*, *Rq T80*, |*Cq*| et |*Cq*| *T80* et les indices de cohérence interne classiques. Dans ce test, facilité objective des questions ($poq_{\ mb}$) et facilité introspective des questions (*piq* et *piq T80*) sont assez corrélés (respectivement 0,62 et 0,68).

Pour MATHEM, les corrélations sont d'une part plus élevées entre *Rq*, *Rq T80*, |*Cq*|, |*Cq*| *T80* et les indices de cohérence interne classiques calculés à partir des matrices spectrales $r_{qt\ ms}$ et α-$q_{\ ms}$, et d'autre part, on n'observe plus les corrélations élevées entre les *NCSq*, *NCSq T80* et tous les autres indices de cohérence interne classiques. Dans ce test, facilité objective des questions ($poq_{\ mb}$) et facilité introspective des questions (*piq* et *piq T80*) sont très corrélés (respectivement 0,76 et 0,80).

Enfin pour le test ARTACT, nous observons des corrélations assez élevées d'une part entre *NCSq*, *NCSq T80* et tous les autres indices de cohérence interne classiques et d'autre part entre les indices *Rq*, *Rq T80,*|*Cq*| *T80* et les indices de facilité/difficulté des questions ($poq_{\ mb}$). Dans ce test, en ce qui concerne les corrélations observées entre facilité objective des questions ($poq_{\ mb}$) et facilité introspective des questions nous constatons d'une part une corrélation non significative entre *piq* et $poq_{\ mb}$ et d'autre part une plus faible corrélation que dans les épreuves précédentes entre *piq T80* et $poq_{\ mb}$ (0,41).

De ces observations sur les trois épreuves MOHICAN obtenant les meilleurs scores de cohérence interne, nous pouvons conclure qu'il existe entre indices spectraux et indices classiques portant sur les questions des tendances à varier ensemble qui ne se présentent pas de la même manière d'une épreuve à l'autre. Dans certaines épreuves des relations fortes existeront entre une série d'indices spectraux et classiques qui n'apparaîtront plus ou de façon beaucoup moins marquée dans les autres tests.

C. Comparaison des performances des questions se situant aux extrêmes des continuums de qualité spectrale et classique

Un outil qui permette de visualiser les niveaux de performance à la fois en cohérence interne et en qualité spectrale des questions est-il envisageable ?

1. Classement des QCM et choix de questions se situant aux extrêmes des continuums de qualité

Dans cette partie nous allons comparer des questions contrastées par les scores qu'elles obtiennent aux indices classiques et aux indices spectraux que nous avons présentés. Notre but est de mettre en évidence des profils de questions qui se situent aux extrêmes des continuums de qualité spectrale et classiques pour ensuite les comparer. Afin que le lecteur puisse se rendre compte par lui-même des difficultés des contenus des questions des dix tests, nous avons repris l'ensemble des 173 questions en annexe (voir p. 482). Voici la méthode que nous proposons pour visualiser les niveaux de performances classiques et spectrales des questions.

a) Choix des trois indices de classement

Nous allons classer les 173 QCM des 10 épreuves MOHICAN selon trois critères : cohérence interne classique, cohérence spectrale et réalisation des prédictions. Nous effectuerons ce classement en fonction de trois indices, un par critère, dont nous justifierons le choix dont les paragraphes suivants. Nous identifierons aux extrêmes du classement de qualité réalisé selon ces trois indices, les questions les mieux classées ainsi que celles qui figurent en queue de classement. Ces niveaux de performances de ces QCM se situant aux extrêmes des continuums de qualité spectrale et classique constitueront des repères pour l'évaluation de la qualité des autres questions des épreuves MOHICAN.

En ce qui concerne le critère de cohérence interne classique, dans la partie précédente nous avons montré les corrélations très élevées entre les trois indices de cohérence interne *NCIq*, r_{qt} et α-*q* pour les épreuves VOCABU, MATHEM et ARTACT (voir p. 331). Nous n'utiliserons pas ici le coefficient alpha après suppression de la question considérée (α-*q*) car les valeurs obtenues à cet indice dépendent de l'alpha de l'épreuve qui est très différent d'un test à l'autre (voir p. 309). Nous préférerons utiliser l'indice de corrélation $r_{qt\ mb}$ (corrélation des scores à la question calculés à partir de la matrice binaire avec le total de l'épreuve) comme indicateur de la qualité de cohérence interne, cet indice étant le plus corrélé avec les coefficients alpha après suppression de la question considérée.

En ce qui concerne l'évaluation de la qualité spectrale, nous utiliserons deux indices, un premier pour évaluer la réalisation des prédictions par question et un second pour la cohérence d'utilisation des degrés de certitude.

Pour ce qui est de l'évaluation de la réalisation des prédictions mesurée au niveau des questions à l'aide de *Rq* et *Cq*, les corrélations très élevées observées entre ces deux indices lorsqu'ils sont calculés à un palier de turbo analyse élevé ainsi que leur complémentarité (*Rq* permet d'évaluer le degré de réalisation des prédictions et *Cq* donne l'indication d'une tendance à la sur ou sous-estimation au sein de la question) nous amène à utiliser ici l'indice *Rq T80* comme critère de classement des 173 QCM du point de vue de la réalisation des prédictions. Nous emploierons l'indice *Cq T80* dans un second temps après avoir procédé au classement.

Enfin, en ce qui concerne la cohérence d'utilisation des pourcentages de certitude (*NCSq*) nous avons observé des corrélations moins élevées entre les indices *NCSq T80* et *Rq T80* qu'entre *Rq T80* et

|Cq| T80. Les corrélations moins élevées entre le degré de réalisation des prédictions par question (*Rq*) et le niveau de cohérence spectrale par question (*NCSq*) sont logiques dans la mesure où ces deux indices spectraux mesurent des propriétés liées à l'utilisation des pourcentages de certitude qui sont différentes : *Rq* permet de chiffrer la concordance des taux d'exactitude annoncés par rapport aux taux d'exactitude observés et *NCSq* la propension à utiliser des pourcentages de certitude plus élevés dans le cas des réponses correctes que dans le cas des solutions incorrectes. Nous utiliserons ici comme troisième critère de classement le niveau de cohérence spectrale par question calculé au palier de turbo analyse T80 (*NCSq T80*).

b) Sélection des QCM se situant aux extrémités du continuum de qualité

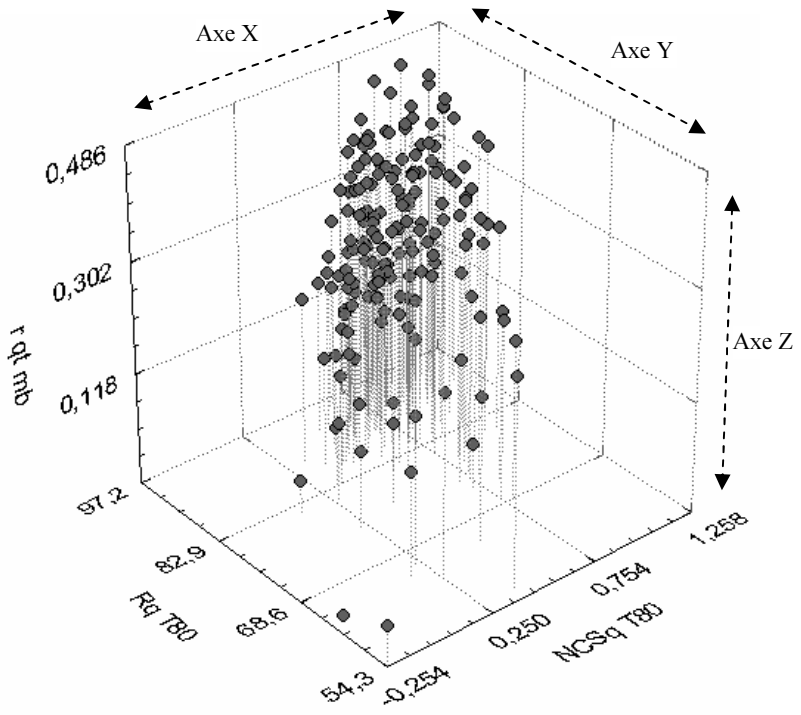
Nous allons donc choisir parmi les 173 QCM des épreuves MOHICAN les deux questions qui obtiennent les meilleurs scores aux indices $r_{qt\ mb}$, *Rq T80* et *NCSq T80* ainsi que deux autres questions qui obtiennent les moins bons scores à ces indices.

Voici un graphique en trois dimensions que nous intitulons « *3D classico-spectral* » où : l'axe des X reprend les valeurs à l'indice *NCSq T80*, l'axe des Y correspond aux valeurs de *Rq T80* et l'axe Z (vertical) nous donne les valeurs des corrélations $r_{qt\ mb}$.

Nous obtenons à l'aide de ce graphe un nuage de 173 points représentant les questions des 10 épreuves.

Pour chaque axe X, Y et Z nous avons défini les valeurs minimales et maximales ainsi que les graduations intermédiaires en prenant respectivement : la valeur minimum observée, la valeur maximum et en divisant l'étendue par 3 afin d'obtenir les pas de graduation qui figurent sur le graphique.

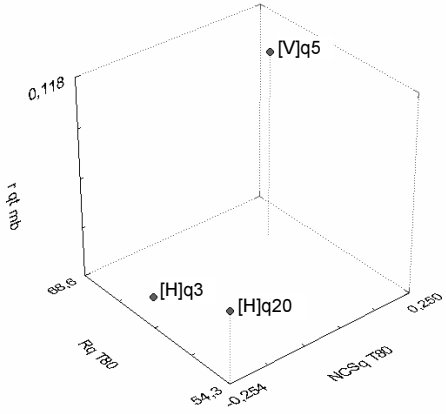
« 3D classico-spectral » des 173 QCM des 10 épreuves MOHICAN tracé en fonction des valeurs aux indices NCSq T80, Rq T80 et $r_{qt\ mb}$.



Nous remarquons que les questions sont plus nombreuses pour les valeurs maximales en haut au fond du graphique que pour les valeurs minimales en bas vers l'avant.

Nous allons maintenant agrandir une zone de ce graphique, celle qui contient les points des questions qui obtiennent les moins bons scores aux trois indices utilisés.

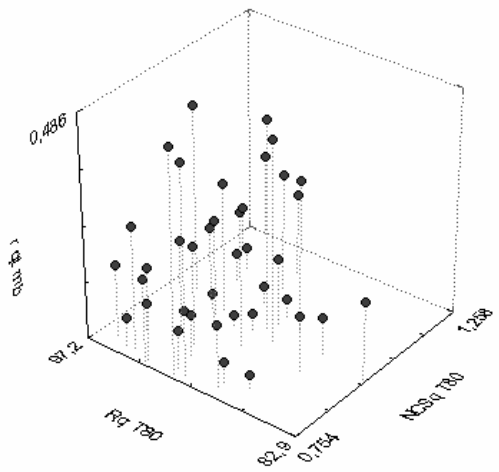
Cette zone est définie par les valeurs suivantes :
- axe X : *NCSq T80* de *-0,254* à *0,250 ;*
- axe Y : *Rq T80* de *54,3* à *68,6 ;*
- axe Z : $r_{qt\ mb}$ de *-0,066* à *0,118*.



Nous constatons qu'elle contient trois questions : la question 5 du test VOCABU ([V]q5) ainsi que les questions 3 et 20 du test

HISTOI ([H]q3, [H]q20). De toutes les 173 QCM des épreuves MOHICAN ces deux dernières obtiennent les moins bons scores aux indices *NCSq T80*, *Rq T80* et $r_{qt\ mb}$.

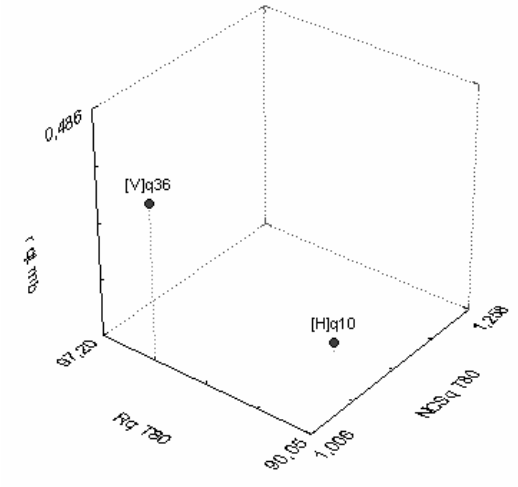
A l'opposé de cette zone de faibles performances se trouve une autre zone délimitée par les valeurs :
- axe X : *NCSq T80* de *0,754* à *1,258 ;*
- axe Y : *Rq T80* de *82,9* à *97,2 ;*
- axe Z : $r_{qt\ mb}$ de *0,302* à *0,486*.



Cette zone contient une série de questions qui se caractérisent par les valeurs les plus élevées observées aux indices *NCSq T80*, *Rq T80* et $r_{qt\ mb}$. Elle contient 39 points représentant 39 QCM.

Nous allons maintenant délimiter une nouvelle zone au sein de cette dernière :
- axe X : *NCSq T80* de *1,006* à *1,258 ;*
- axe Y : *Rq T80* de *90,05* à *97,2 ;*
- axe Z : $r_{qt\ mb}$ de *0,394* à *0,486*.

Ce nouveau découpage aboutit au graphique suivant.



Deux questions apparaissent dans cette zone de hautes performances aux indices *NCSq T80*, *Rq T80* et $r_{qt\ mb}$: il s'agit des questions 36 du test VOCABU ([V]q36) et 10 du test HISTOI ([H]q10).

Nous constatons que ces QCM font partie des deux mêmes épreuves (VOCABU et HISTOI) que les trois questions les moins performantes sélectionnées plus haut.

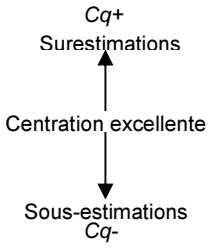
Nous avons intitulé ces instruments de visualisation du niveau de qualité spectrale des questions et de leur niveau de participation à la cohérence interne des épreuves, les « 3D classico-spectraux ». Nous voyons qu'à l'aide de tels instruments nous pouvons identifier les questions qui présentent de faibles performances. Nous exposerons dans la dernière partie de cette thèse une exploitation de ces graphiques dans le cadre de nos projets d'interfaces de gestion de la qualité spectrale des tests (p. 440).

D. Comparaison des performances de l'ensemble des 173 questions des 10 épreuves MOHICAN

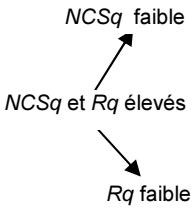
Peut-on se doter d'un outil « tableau de bord d'ensemble » faisant apparaître les éventuelles anomalies spectrales et leur ampleur ainsi que les problèmes de cohérence interne pour les 173 QCM appartenant aux 10 épreuves MOHICAN ?

1. Ingénogrammes de qualité spectrale et de cohérence interne des questions

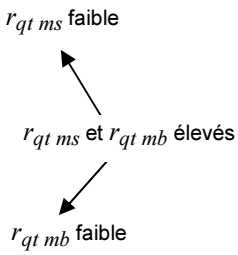
Pour visualiser ces performances nous avons représenté sur les axes d'ingénogrammes (un par QCM) les valeurs obtenues aux 6 indices que nous avons sélectionnés : quatre indices spectraux et deux indices de cohérence interne (voir ci-dessous). Notre but est de montrer à l'aide d'un même graphe les qualités spectrales et de cohérence interne d'une question de manière à mettre en évidence les questions qui mériteraient une exploration plus approfondie dans l'un ou/et l'autre champs d'analyse (spectrale ou/et de cohérence interne). Les ingénogrammes nous permettront également de comparer les cinq questions mises en évidence dans la partie précédente ([V]q5, [V]q36, [H]q3, [H]q10 et [H]q20) aux 168 autres des 10 épreuves. Les ingénogrammes nous permettront également de comparer les cinq questions mises en évidence dans la partie précédente ([V]q5, [V]q36, [H]q3, [H]q10 et [H]q20) aux 168 autres des 10 épreuves.



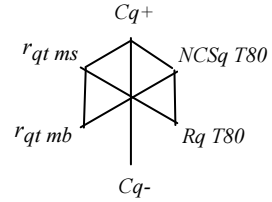
Nous avons sélectionné l'indice spectral de Centration par question (*Cq*). Afin de mieux visualiser l'ampleur des surestimations ou des sous-estimations, nous allons en extraire deux sous-indices : l'indice des surestimations ($Cq+ = Cq > 0$) et l'indice des sous-estimations ($Cq- = Cq < 0$). Dès lors, nous utiliserons la moitié supérieure de l'axe vertical (Nord) pour mettre en évidence les surestimations et la moitié inférieure pour les sous-estimations (Sud).



Les deux autres indices spectraux sont le Niveau de Cohérence Spectrale des questions (*NCSq*) et l'indice de Réalisation des prédictions par question (*Rq*), calculés au palier de turbo analyse T80. Nous utiliserons le côté droit des ingénogrammes pour visualiser ces indices. L'axe oblique Nord-Est contiendra les *NCSq*. Nous avons transformé les valeurs de façon à ce que les moins élevées soient les plus éloignées du centre. L'axe Sud-Est sera quant à lui réservé aux valeurs obtenues à l'indice *Rq.* Comme pour l'indice précédent, les valeurs les moins élevées sont également les plus éloignées du centre.



Enfin, nous utiliserons deux indices de cohérence interne : la corrélation question-total calculée d'une part à l'aide d'une matrice binaire ($r_{qt\ mb}$) et d'autre part à l'aide d'une matrice spectrale ($r_{qt\ ms}$). Rappelons que ces deux indices sont fortement corrélés (pour trois épreuves au moins) avec les autres coefficients de cohérence interne (α-q_{mb}, α-q_{ms} et *NCIq*, voir p. 331). Nous emploierons la partie gauche des ingénogrammes pour les indices de cohérence interne. L'axe Nord-Ouest représentera les $r_{qt\ ms}$ et l'axe Sud-Ouest les $r_{qt\ mb}$. Comme pour les deux indices spectraux précédents, nous avons transformé les valeurs de façon à ce que les moins élevées soient les plus éloignées du centre.



Cette disposition spatiale des données donne des ingénogrammes en étoile tel que ci-contre. Nous avons relié les points sur les axes par des lignes afin de guider l'œil du lecteur. L'exemple montre le profil d'une question dont les indices de cohérence interne ($r_{qt\ mb}$ et $r_{qt\ ms}$) ainsi que les indices de qualité spectrale (*NCSq T80* et *Rq T80*) sont très peu élevés. Les résultats de la question de cet

exemple sont également accompagnés d'une forte tendance à la surestimation. C'est théoriquement le type de profil que nous devrions obtenir pour les questions [V]q5 et [H]q20 peu performantes et dont nous avons analysé les scores aux indices de cohérence interne et de qualité spectrale dans la partie précédente.

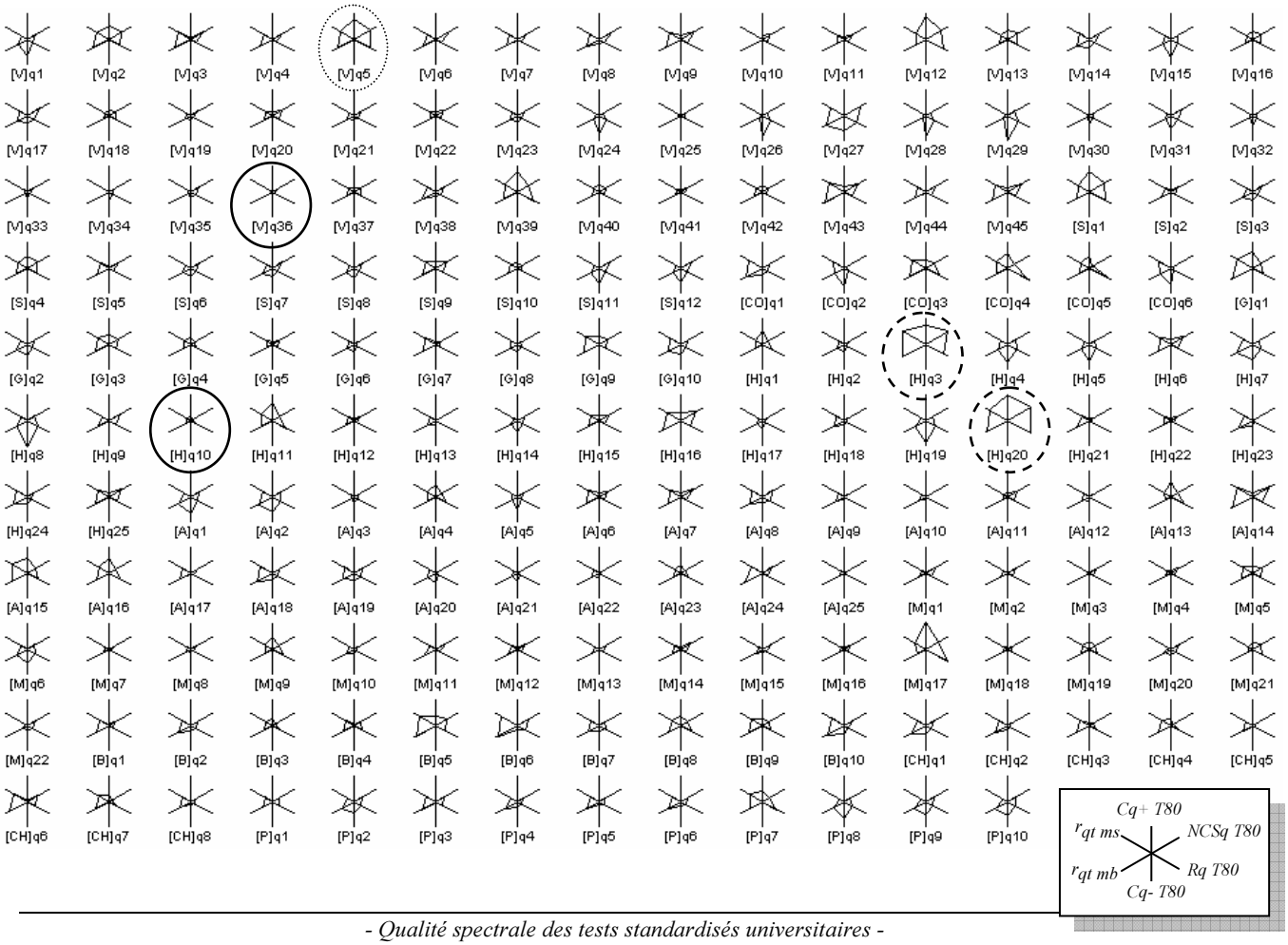
2. Visualisation des performances des 173 QCM

Nous avons identifié chaque question par un code constitué de la première lettre du test (ou les deux premières lettres dans le cas de COmpréhension et CHimie) placée entre crochets, le tout suivi de la lettre « q » et du numéro d'ordre de la question dans l'épreuve. Nous obtenons ainsi les codes d'identification suivants :

- [V]q1 à [V]q45 (V = VOCABU = épreuve de vocabulaire soumise à 3.846 étudiants) ;
- [S]q1 à [S]q12 (S = SYNTAX = épreuve de syntaxe et articulation logique, 3.739 étudiants) ;
- [CO]q1 à [CO]q6 (CO = COMPRE = épreuve de compréhension, 3.420 étudiants) ;
- [G]q1 à [G]q10 (G = GEOGRA = épreuve de lecture de documents et géographie, 3.688 étudiants) ;
- [H]q1 à [H]q25 (H = HISTOI = épreuve de connaissances en histoire et socio éco., 1.410 étudiants) ;
- [A]q1 à [A]q25 (A = ARTACT = épreuve de connaissances artistiques, 1.392 étudiants) ;
- [M]q1 à [M]q22 (M = MATHEM = épreuve de mathématiques, 2.516 étudiants) ;
- [B]q1 à [B]q10 (B = BIOLOG = épreuve de biologie, 2.507 étudiants) ;
- [CH]q1 à [CH]q8 (CH = CHIMIE = épreuve de chimie, 2.501 étudiants) ;
- [P]q1 à [P]q10 (P = PHYSIQ = épreuve de physique, 2.497 étudiants).

Voici les profils observés des questions des 10 épreuves MOHICAN. En guise de points de repère nous avons entouré les deux questions les plus performantes ([V]q36 et [H]q10) par un trait continu et les trois moins performantes ([V]q5, [H]q3 et [H]q20) par des pointillés (voir p. 344). La légende dans le coin inférieur droit rappelle la signification des axes.

Ingénogrammes de qualité spectrale et de cohérence interne des 173 QCM MOHICAN



Les ingénogrammes présentés ci-dessus permettent de visualiser les anomalies spectrales et de cohérence interne des 173 questions des 10 épreuves MOHICAN. Nous y avons mis en évidence les questions [H]q3, [H]q20 et [V]q5 épinglées précédemment pour leurs faibles performances. Nous remarquons aussi d'autres questions qui présentent des surfaces qui signalent des problèmes en qualité spectrale et/ou en cohérence interne. Dans la sous-section qui suit nous allons établir un inventaire des questions en fonction de leurs niveaux de performances.

3. Répartition des questions en fonction de la qualité spectrale et de la cohérence interne

Quelle est la répartition des questions en fonction de la qualité spectrale et de la cohérence interne ? Précédemment nous nous sommes demandé combien de QCM enfermaient une tendance à la surestimation élevée dans les 10 tests ([2.4], p. 256). Quelle sont les performances spectrales et classiques des questions en cas de surestimation élevée ? Et en cas de sous-estimation élevée ?

a) Catégorisation

Voici un tableau reprenant pour chaque indice utilisé dans l'élaboration des ingénogrammes, les valeurs minimum, maximum ainsi que l'étendue des valeurs observées dans le cadre des 10 épreuves MOHICAN. Nous avons aussi calculé les valeurs de deux bornes situées à 1/3 et 2/3 de l'étendue pour chaque indice.

	Min.	Borne 1/3	Borne 2/3	Max.	Etendue
NCSq T80	-0,254	0,250	0,754	1,257	1,511
Rq T80	54,3	68,6	82,9	97,2	42,9
Cq+ T80	0	0,144	0,287	0,431	0,431
Cq- T80	-0,237	-0,158	-0,079	0	0,237
$r_{qt\ mb}$	-0,066	0,117	0,301	0,485	0,551
$r_{qt\ ms}$	-0,290	-0,028	0,235	0,497	0,787

A l'aide de ces valeurs nous créons trois catégories de performances pour chaque indice. Nous notons ces catégories : « + », « ≃ » et « - ». Voici les valeurs correspondantes :

	+	≃	–
NCSq T80	1,257 ⩽ x < 0,754	0,754 ⩽ x < 0,250	0,250 ⩽ x < -0,254
Rq T80	97,2 ⩽ x < 82,9	82,9 ⩽ x < 68,6	68,6 ⩽ x <54 ,3
Cq+ T80	0 ⩽ x <0,144	0,144 ⩽ x <0,287	0,287 ⩽ x <0,432
Cq- T80	0 > x >-0,079	-0,079 ⩾ x > -0,158	-0,158 ⩾ x > -0,237
$r_{qt\ mb}$	0,485 ⩽ x < 0,301	0,301 ⩽ x < 0,117	0,117 ⩽ x < -0,066
$r_{qt\ ms}$	0,497 ⩽ x < 0,235	0,235 ⩽ x < -0,028	-0,028 ⩽ x < -0,290

Voici le comptage des nombres de questions dont les valeurs récoltées aux indices spectraux et de cohérence interne se situent dans ces trois catégories « +, ≃ ou – ».

	+	≃	–	Total
NCSq T80	99	67	7	173
Rq T80	140	25	8	173
Cq+ T80	58	15	7	80
Cq- T80	57	27	9	93
$r_{qt\ mb}$	58	94	21	173
$r_{qt\ ms}$	98	70	5	173

Pour les indices spectraux *NCSq T80*, *Rq T80*, *Cq+ T80* et *Cq- T80* nous observons une majorité de questions dans la catégorie « + ». Les nombres de questions sont moins élevés dans la catégorie « ≃ » et encore moins élevés dans la catégorie « – ».

Nous constatons qu'un peu plus de questions contiennent une tendance à la sous-estimation : 93 contre 80 pour les surestimations. Nous savons que l'étendue de l'indice *Cq+ T80* (0,431) vaut presque le double de l'étendue de *Cq- T80* (0,237) (voir 1er tableau) et bien que le nombre des sous-estimations soit un peu plus élevé, les tendances aux surestimations sont plus marquées (nous pouvons l'observer sur les ingénogrammes, p. 347).

En ce qui concerne la cohérence interne, lorsque la corrélation question-total est calculée à l'aide d'une matrice binaire ($r_{qt\ mb}$) nous rencontrons plus de questions qui récoltent un « ≃ ». Ce n'est pas le cas lorsqu'une matrice spectrale est utilisée ($r_{qt\ ms}$), on observe alors une majorité de questions qui obtiennent un « + ».

b) Tableau des performances des 173 QCM

Voyons maintenant dans quelles catégories de performances « +, ≃ ou – » se situe chacune des 173 questions. Rappelons qu'un « + » pour les indices *Cq+* ou *Cq-* signifie que la valeur est plus proche du zéro que du maximum (elle se situe dans le 1er tiers de l'étendue), donc il y a relativement peu de surestimation ou de sous-estimation. Remarquons que pour chaque question il n'y a pas de valeur dans une des deux cases de ces indices car soit *Cq* est ⩾ 0 et la case *Cq-* est alors inoccupée, soit *Cq* est < 0 et c'est alors la case *Cq+* qui est inoccupée.

Tableau des performances en qualité spectrale et en cohérence interne des 173 QCM des épreuves MOHICAN check up '99

	NCSq/80	Rq T80	Cq+ T80	Cq- T80	rqt mb	rqt ms
[V]q1	≃	+		≃	≃	+
[V]q2	≃	≃	≃		-	≃
[V]q3	≃	+	+		-	≃
[V]q4	≃	+		+	≃	≃
[V]q5	-	-	-		-	≃
[V]q6	≃	+		+	≃	≃
[V]q7	≃	+		+	≃	≃
[V]q8	≃	+		+	≃	≃
[V]q9	≃	+	+		-	≃
[V]q10	+	+	+		+	+
[V]q11	≃	+	+		+	+
[V]q12	-	-	-		≃	≃
[V]q13	≃	≃	+		≃	+
[V]q14	≃	+		≃	≃	≃
[V]q15	+	≃		-	≃	+
[V]q16	≃	+	+		+	+
[V]q17	≃	+		+	≃	≃
[V]q18	+	+	+		+	+
[V]q19	+	+		+	+	+
[V]q20	≃	+	+		≃	+
[V]q21	≃	+		+	+	+
[V]q22	+	+	+		+	+
[V]q23	+	+		+	+	+
[V]q24	+	≃		-	+	+
[V]q25	+	+	+		+	+
[V]q26	+	≃		-	+	+
[V]q27	-	≃		≃	-	≃
[V]q28	+	≃		-	+	+
[V]q29	+	≃		-	+	+
[V]q30	+	+		+	+	+
[V]q31	≃	+		≃	+	+
[V]q32	+	+		≃	+	+
[V]q33	+	+		+	+	+
[V]q34	+	+		≃	+	+
[V]q35	≃	+		+	+	+
[V]q36	+	+		+	+	+
[V]q37	+	+	+		+	+
[V]q38	≃	+		+	≃	≃
[V]q39	≃	-	-		≃	≃
[V]q40	+	+	+		+	+
[V]q41	+	+	+		+	+
[V]q42	+	+	+		≃	+
[V]q43	≃	+	+		-	-
[V]q44	≃	+		+	≃	≃
[V]q45	≃	+	+		-	≃

	NCSq/80	Rq T80	Cq+ T80	Cq- T80	rqt mb	rqt ms
[S]q1	-	-	-		≃	≃
[S]q2	+	+	+		≃	+
[S]q3	≃	+		≃	≃	+
[S]q4	≃	≃	≃		≃	≃
[S]q5	≃	+	+		≃	≃
[S]q6	≃	+		≃	≃	+
[S]q7	≃	+		+	≃	+
[S]q8	≃	+		≃	≃	+
[S]q9	≃	+	+		≃	≃
[S]q10	+	+	+		≃	+
[S]q11	≃	+		≃	≃	+
[S]q12	≃	+		≃	≃	+
[Co]q1	+	+		+	-	≃
[Co]q2	+	+		-	≃	≃
[Co]q3	≃	+	+		-	≃
[Co]q4	+	≃	≃		≃	≃
[Co]q5	+	≃	≃		≃	≃
[Co]q6	+	+		≃	≃	≃
[G]q1	≃	≃	≃		-	≃
[G]q2	+	+		≃	≃	+
[G]q3	≃	+	+		≃	≃
[G]q4	+	+	+		+	+
[G]q5	+	+	+		+	+
[G]q6	+	+		+	+	+
[G]q7	+	+	+		≃	≃
[G]q8	+	+		+	+	+
[G]q9	≃	+	+		-	≃
[G]q10	+	+		≃	≃	≃
[H]q1	+	≃	≃		+	+
[H]q2	+	+		+	+	+
[H]q3	-	-	-		-	-
[H]q4	+	≃		-	≃	+
[H]q5	+	+		≃	≃	+
[H]q6	≃	+	+		≃	≃
[H]q7	+	+		≃	-	≃
[H]q8	+	≃		-	≃	+
[H]q9	≃	+	+		≃	+
[H]q10	+	+	+		+	+
[H]q11	+	≃	≃		≃	≃
[H]q12	+	+	+		≃	+
[H]q13	+	+		+	≃	+
[H]q14	+	+		≃	+	+
[H]q15	≃	+	+		≃	≃
[H]q16	≃	≃	+		-	≃
[H]q17	+	+		≃	+	+

	NCSq/80	Rq T80	Cq+ T80	Cq- T80	rqt mb	rqt ms
[H]q18	+	+		+	≃	≃
[H]q19	+	≃		-	≃	+
[H]q20	-	-	-		-	-
[H]q21	+	+	+		≃	+
[H]q22	+	+	+		+	+
[H]q23	+	+		+	≃	≃
[H]q24	+	+		+	≃	≃
[H]q25	≃	+	+		≃	≃
[A]q1	≃	+		≃	≃	+
[A]q2	≃	+		≃	≃	≃
[A]q3	+	+		+	+	+
[A]q4	+	≃	≃		≃	≃
[A]q5	+	+		≃	+	+
[A]q6	≃	+	+		≃	≃
[A]q7	≃	+	+		≃	≃
[A]q8	≃	+		+	≃	≃
[A]q9	+	+		+	+	+
[A]q10	+	+		+	+	+
[A]q11	≃	+	+		+	+
[A]q12	+	+		+	+	+
[A]q13	+	≃	≃		+	+
[A]q14	-	≃	+		-	≃
[A]q15	≃	≃	≃		≃	≃
[A]q16	+	≃	≃		≃	≃
[A]q17	+	+		+	≃	+
[A]q18	+	+		+	≃	≃
[A]q19	≃	+		+	≃	≃
[A]q20	+	+		≃	≃	+
[A]q21	+	+		+	+	+
[A]q22	+	+		+	+	+
[A]q23	+	+	+		+	+
[A]q24	≃	+		+	-	≃
[A]q25	+	+		+	+	+
[M]q1	≃	+	+		+	+
[M]q2	≃	+		+	+	+
[M]q3	+	+	+		+	+
[M]q4	+	+	+		+	+
[M]q5	≃	+	+		≃	≃
[M]q6	+	+		≃	≃	+
[M]q7	+	+		+	+	+
[M]q8	+	+		+	+	+
[M]q9	+	≃	≃		+	+
[M]q10	+	+		+	≃	+
[M]q11	≃	+		+	≃	+
[M]q12	≃	+	+		≃	+

	NCSq/80	Rq T80	Cq+ T80	Cq- T80	rqt mb	rqt ms
[M]q13	+	+		+	+	+
[M]q14	≃	+	+		≃	+
[M]q15	≃	+		+	+	+
[M]q16	+	+	+		+	+
[M]q17	≃	-	-		≃	≃
[M]q18	+	+	+		+	+
[M]q19	+	+	+		≃	+
[M]q20	≃	+		+	+	+
[M]q21	≃	+	+		+	+
[M]q22	≃	+		+	+	+
[B]q1	+	+	+		≃	≃
[B]q2	+	+		+	≃	≃
[B]q3	+	+	+		≃	+
[B]q4	+	+	+		≃	≃
[B]q5	≃	≃	≃		-	-
[B]q6	≃	+		+	-	-
[B]q7	+	+		+	≃	≃
[B]q8	+	+	≃		≃	≃
[B]q9	+	+	+		≃	≃
[B]q10	≃	+		+	-	≃
[Ch]q1	+	+		+	≃	≃
[Ch]q2	+	+		+	≃	+
[Ch]q3	+	+	+		≃	≃
[Ch]q4	+	+	+		≃	≃
[Ch]q5	+	+		+	≃	+
[Ch]q6	≃	+	+		-	≃
[Ch]q7	+	+	+		≃	≃
[Ch]q8	+	+		+	≃	≃
[P]q1	+	+		+	≃	+
[P]q2	+	+		≃	≃	+
[P]q3	≃	+		+	≃	≃
[P]q4	+	+		+	≃	+
[P]q5	+	+		+	≃	≃
[P]q6	+	+		+	≃	≃
[P]q7	+	-	≃		≃	≃
[P]q8	+	+		≃	≃	+
[P]q9	≃	+		≃	≃	+
[P]q10	≃	+		≃	≃	+

Nous avons mis en évidence par un fond gris clair les 33 (19%) de questions qui récoltent uniquement des « + » aux indices utilisés, donc des questions qui se situent parmi les meilleures du point de vue de la qualité spectrale et de la cohérence interne.

Deux autres questions sont mises en évidence à l'aide d'un fond gris foncé, elles récoltent systématiquement des « - ». Il s'agit des deux questions les moins performantes précédemment épinglées ([H]q3 et [H]q20, voir p. 344).

Nous remarquons une tendance à la surestimation dans les résultats des questions [H]q3 et [H]q20 aux performances les moins élevées aux indices de qualité spectrale et de cohérence interne.

c) Répartition des questions en fonction de la qualité spectrale

Le tableau ci-contre nous montre la répartition des nombres de questions en fonction des catégories de performances « +, ≃ ou - » pour les indices spectraux *NCSq T80* et *Rq T80*.

		NCSq T80 +	≃	-
	+	82	58	0
Rq T80	≃	16	7	2
	-	1	2	5

Nous constatons que la grosse majorité des questions (82) obtiennent un « + » pour les deux indices. On rencontre plus souvent des questions avec des « + » à l'indice *Rq T80* accompagnées de *NCSq T80* moyens (58) que des questions où l'inverse se produit (16). Peu de questions (5) obtiennent des valeurs à la fois peu élevées pour *Rq T80* et pour *NCSq T80* : [V]q5, [V]q12, [S]q1, [H]q3 et [H]q20.

Une question obtient un « - » à l'indice *Rq T80* et un « + » pour *NCSq T80*, il s'agit de [P]q7, la 7ème QCM du test de physique. Lors de la présentation des indices spectraux à l'aide des données de l'épreuve de physique soumise à 2.497 étudiants, nous avions relevé cette situation paradoxale ainsi qu'une tendance à la surestimation dans les résultats liés à cette QCM (voir p. 245) et nous nous étions demandé si d'autres questions des épreuves MOHICAN pouvaient se trouver dans le même cas. En ce qui concerne l'association d'une mauvaise performance à l'indice *Rq T80* avec une bonne performance *NCSq T80*, nous voyons à l'aide de cette répartition qu'il s'agit d'un cas isolé dans le contexte MOHICAN check up '99.

Nous pouvons conclure à l'aide de ce tableau que [P]q7 est la seule question qui présente un contraste élevé entre Rq T80 et NCSq T80 et que cinq questions ([V]q5, [V]q12, [S]q1, [H]q3 et [H]q20) présentent des valeurs peu élevées à la fois à l'indice Rq T80 et NCSq T80.

d) Répartition des questions en fonction de la cohérence interne

Voyons maintenant la répartition des performances en ce qui concerne les deux indices de cohérence spectrale.

On ne rencontre pas ici de question où l'un des deux indices obtiendrait un « + » alors que l'autre serait « - ».

Notre attention est attirée par les 5 questions qui récoltent un « - » aux deux indices. Il s'agit de [V]q43, [H]q3, [B]q5, [B]q6 et [H]q20. Parmi ces 5 QCM nous retrouvons [H]q3 et [H]q20 obtenant des « - » pour tous les indices (et donc aussi pour NCSq T80 et Rq T80).

		$r_{qt\ mb}$ +	≃	-
	+	58	40	0
$r_{qt\ ms}$	≃	0	54	16
	-	0	0	5

e) Répartition des questions lorsqu'on combine qualité spectrale et cohérence interne

Dans quelle mesure les questions qui présentent des anomalies en qualité spectrale sont-elles aussi celles qui présentent des problèmes de cohérence interne ?

Jusqu'à présent nous avons examiné de façon séparée, d'une part deux indices de qualité spectrale et d'autre part deux indices de cohérence interne.

Nous allons maintenant envisager la répartition des questions dans une série de tableaux qui combinent à la fois la qualité spectrale et la cohérence interne. Pour y parvenir nous proposons les catégories « ++, +≃, ≃≃, ≃–, – – et +– » qui recouvrent les cas décrits dans la table ci-contre. Ainsi, en ce qui concerne la qualité spectrale, une question qui récolterait un « + » pour l'indice *NCSq T80* et un autre « + » pour l'indice *Rq T80* obtiendrait un « ++ » en qualité spectrale (1ère case en haut à gauche du tableau). Si l'indice *NCSq T80* = « + » et *Rq T80* = « ≃ », alors la qualité spectrale vaut « +≃ », etc. Nous utilisons le même principe pour la cohérence interne des questions mesurée à l'aide des indices $r_{qt\ mb}$ et $r_{qt\ ms}$.

Rq T80 \ *NCSq T80*	+	≃	-
+	++	+≃	+–
≃	+≃	≃≃	≃–
-	+–	≃–	– –

Voici ce que donne cette répartition en fonction de la qualité spectrale et de la cohérence interne. La première case en haut à gauche contient « 39 », donc 39 questions obtiennent des performances élevées à la fois aux deux indices de qualité spectrale et aux deux indices de cohérence interne.

Répartition des 173 QCM MOHICAN en fonction de la qualité spectrale et de la cohérence interne des questions

Cohérence interne \ Qualité spectrale	++	+≃	≃≃	-≃	--	+-
++	39	19	-	-	-	
+≃	20	19	1	-	-	
≃≃	21	26	2	2	2	1
-≃	2	8	3	2	1	
--	-	2	1	-	2	

L'intitulé de la colonne « ++ » signifie que les QCM obtiennent un *Rq T80* « + » ainsi qu'un *NCSq T80* « + ». L'intitulé de la ligne « ++ » signifie qu'elles obtiennent un « + » pour chaque indice de cohérence interne.

La deuxième case de la première ligne du tableau contient « 19 ». L'intitulé de la deuxième colonne « +≃ » contenant cette case signifie qu'un des deux indices spectraux obtient « + » tandis que l'autre obtient « ≃ ». Donc, 19 QCM dont les valeurs aux indices de cohérence interne sont élevées (« ++ »), obtiennent des valeurs élevées « + » à un des deux indices spectraux et des valeurs plus moyennes « ≃ » à l'autre indice spectral.

Dans le tableau ci-dessus, nous avons traité la catégorie « +– » de façon séparée car d'une part, une seule question (seulement dans le cas des indices spectraux) est concernée, il s'agit de [P]q7, et d'autre part, nous pouvons ainsi hiérarchiser les autres catégories de « ++ » à « -- ». Nous constatons qu'en ce qui concerne [P]q7 la cohérence interne récolte «≃≃ ». Précédemment lors de la présentation de l'indice *Rq* (p. 242), nous avons épinglé question 7 du test de physique et nous nous étions demandé combien de questions obtiendraient des performances opposées aux indices *Rq* et *NCSq* dans le cadre des 10 épreuves MOHICAN ([2.3], p. 249). Nous pouvons constater à l'aide du tableau que [P]q7 est un cas isolé, nous ne rencontrons pas d'autres QCM pour lesquelles les performances aux indices *Rq* et *NCSq* sont opposées.Nous avons grisé les cases de la partie inférieure droite du tableau afin de mettre en évidence les cas de figure où les questions sont à la fois peu performantes du point de vue de la qualité spectrale et de la cohérence interne.

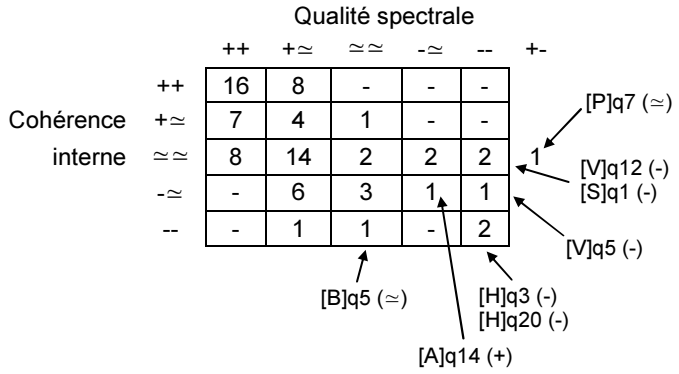
Nous remarquons que 8 QCM sont concernées. Nous identifierons ces questions dans les tableaux suivants.

4. Qualité des questions et tendance à la sur ou sous-estimation dans les résultats

Observe-t-on les anomalies spectrales et les problèmes de cohérence interne plus fréquemment chez les questions qui se caractérisent par la présence de surestimation dans les résultats ou chez les questions qui contiennent de la sous-estimation ?

Précédemment nous avons observé pour les deux QCM dont les valeurs sont les moins élevées aux indices spectraux et classiques ([H]q3 et [H]q20) une tendance prononcée à la surestimation (*Cq+* obtient « - »). Nous retrouvons ces questions dans la case inférieure droite du tableau ci-dessous qui reprend les catégories de qualité spectrale et de cohérence interne pour les 80 questions dont l'indice *Cq* à T80 est supérieur à 0, donc où on observe les tendances à la surestimation.

Répartition des 80 QCM MOHICAN dont la Centration (Cq T80) est supérieure à 0



Nous remarquons dans les cases grisées de la dernière ligne correspondant à « -- » en cohérence interne, 3 QCM qui obtiennent des valeurs peu élevées aux indices $r_{qt\ mb}$ et $r_{qt\ ms}$ et aux indices *NCSq T80* et *Rq T80.* Les codes de ces questions sont placés en bordure du tableau ([B]q5, [H]q3 et [H]q20), nous avons indiqué entre parenthèses la catégorie de performance correspondant à l'indice *Cq+ T80*. Nous remarquons que [B]q5 possède une tendance moyenne à la surestimation. Dans la dernière colonne des questions obtenant « -- » du point de vue de la qualité spectrale nous retrouvons les 5 QCM déjà identifiées pour leurs indices *Rq T80* et *NCSq T80* peu élevés : [V]q5, [V]q12, [S]q1, [H]q3 et [H]q20. Nous constatons que pour ces 5 questions la tendance à la surestimation est forte (-). Enfin, nous remarquons que la case grisée correspondant à une qualité spectrale et une cohérence interne « -≃ » contient la question [A]q14 dont la surestimation est peu marquée (+).

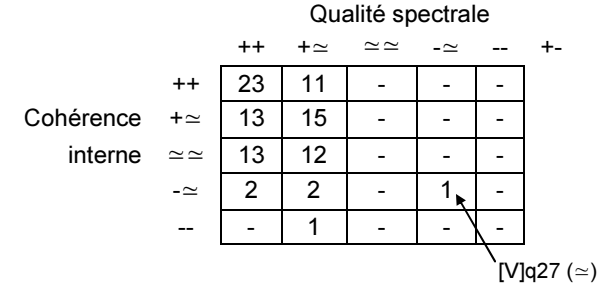
Donc, sur les sept questions figurant dans les cases grisées, cinq ([V]q5, [V]q12, [S]q1, [H]q3 et [H]q20) possèdent une tendance élevée à la surestimation dans leurs résultats. Ces 5 questions obtiennent des scores moyens voire faibles en cohérence interne ainsi que des scores faibles en qualité spectrale.

Voici maintenant le tableau de répartition des 93 questions qui obtiennent un *Cq T80* inférieur à 0.

En ce qui concerne la qualité spectrale, nous constatons que les questions se cantonnent dans les colonnes « ++ » et « +≃ ».

Une question obtient « -- » pour la cohérence interne, il s'agit de [B]q6 (dernière ligne du tableau). Après consultation du tableau des performances des QCM nous remarquons que comparée aux autres, elle ne présente pas de tendance particulière à la sous-estimation.

Répartition des 93 QCM MOHICAN dont la Centration (Cq T80) est inférieure à 0



Dans le dernier tableau nous remarquons une seule question qui se trouve dans la zone des cases grisées, cette question obtient « –≃ » en qualité spectrale et en cohérence interne, il s'agit de [V]q27, sa tendance à la sous-estimation n'est pas particulièrement élevée et vaut « ≃ ».

Dans le contexte des épreuves MOHICAN les 8 questions qui obtiennent les valeurs les moins élevées à la fois du point de vue de la qualité spectrale et de la cohérence interne (voir cases grisées du 1er tableau) enferment une tendance plus élevée à la surestimation dans leurs résultats dans 5 cas sur 8. Une autre QCM ([B]q5) contient une tendance à la surestimation qui est moyenne, une autre ([A]q14) ne présente pratiquement pas de surestimation. Enfin, dans le cas de la dernière question [V]q27, nous observons une tendance à la sous-estimation qui, relativement aux autres questions où on observe de la sous-estimation, est moyenne « ≃ ».

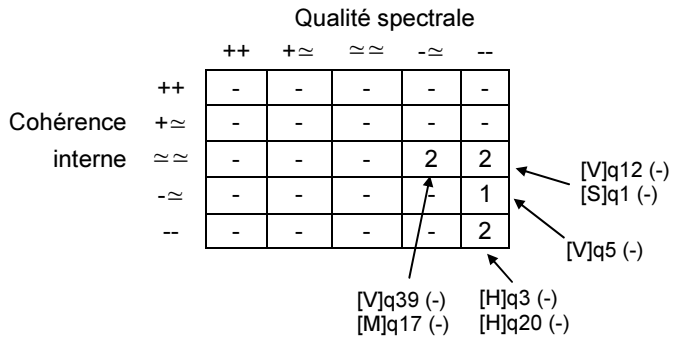
Dans la sous-section qui suit nous allons envisager les seules questions qui enferment les tendances les plus élevées à la sur ou sous-estimation dans leurs résultats.

a) Quelle répartition en cas de surestimation élevée ?

Précédemment, lors de la présentation de l'indice de centration par question (p. 254), nous nous étions demandé combien de QCM enfermaient une tendance à la surestimation élevée dans les 10 tests ([2.4], p. 256).

Observons maintenant quelle est la répartition des questions lorsqu'on ne prend en compte seulement les surestimations les plus élevées (*Cq+ T80* = « - »).

Répartition des 7 QCM MOHICAN dont la Centration positive (Cq+ T80) obtient « - » (Cq+ T80 ⩾ 0,287)



Nous constatons que sur les 80 QCM, sept sont concernées (leurs codes figurent en bordure du tableau) et qu'aucune obtient « ++, +≃ ou ≃≃ » en qualité spectrale ni « ++ ou +≃ » en cohérence interne. Cinq QCM figurent dans les cases grisées et 2 en bordure de ces cases.

Les cinq questions concernées obtiennent donc toutes des valeurs peu élevées « - » ou moyennes « ≃ » aux indices spectraux ainsi qu'aux indices de cohérence interne.

b) Quelle répartition en cas de sous-estimation élevée ?

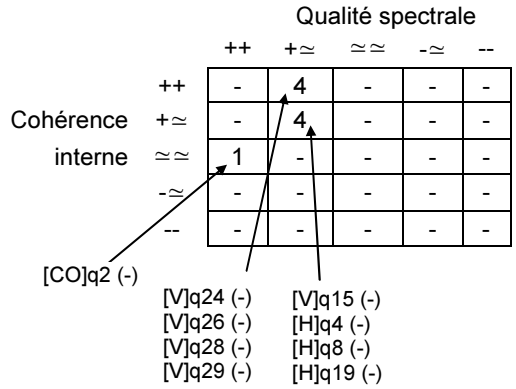
Nous nous sommes aussi demandé combien de QCM enfermaient les tendances les plus élevées à la sous-estimation ([2.5], p. 256). Voyons quelle est la répartition des questions lorsqu'on ne prend en compte que les 9 questions où la sous-estimation est la plus prononcée (*Cq- T80* « - ») parmi les 93 QCM où *Cq T80* est inférieur à zéro.

D'une part, rappelons que les catégories de performances « +, ≃ ou – » que nous avons définies sont liées aux valeurs obtenues par chaque indice dans le contexte des 173 QCM. Ainsi, lorsque nous envisageons ici les questions dont la sous-estimation est la plus élevée, cette sous-estimation « - » recouvre des valeurs dont les nombres absolus sont à peu près deux fois moins élevés que les valeurs des surestimations notées « - » (voir les deux premiers tableaux de cette partie). Donc, lorsqu'il y a sous-estimation dans une question MOHICAN, cette sous-estimation n'atteint jamais les valeurs qu'on observe pour les surestimations les plus élevées.

D'autre part, nous remarquons dans le tableau que les 9 questions où la sous-estimation est la plus prononcée se cantonnent plutôt dans la partie supérieure gauche à l'inverse des questions avec surestimation forte qui figurent dans la partie inférieure droite du tableau précédent.

Nous constatons donc que les 9 questions qui présentent une tendance à la sous-estimation relativement plus élevée obtiennent des valeurs élevées ou moyennes aux indices de qualité spectrale et de cohérence interne.

Répartition des 9 QCM MOHICAN dont la Centration négative (Cq-) obtient « - » (Cq- ⩽ -0,158)



c) Conclusions

A l'issue de ces observations sur la répartition des questions en fonction de la qualité spectrale et de la cohérence interne, nous constatons une tendance forte à la surestimation dans les résultats des deux questions [H]q3 et [H]q20 qui obtiennent les valeurs les moins élevées à la fois aux indices de qualité spectrale (*Rq T80* et *NCSq T80*) et aux indices de cohérence interne ($r_{qt\ mb}$ et $r_{qt\ ms}$).

A l'aide du tableau des performances des 173 QCM (p. 350), d'une part nous remarquons que lorsqu'il y a surestimation élevée non seulement la qualité spectrale n'est jamais élevée, ce à quoi nous pouvions nous attendre étant donné les liaisons observées entre les indices spectraux (p. 334), mais en plus la cohérence interne elle aussi n'est jamais élevée. D'autre part, nous constatons qu'il n'existe que deux questions qui contiennent une tendance élevée à la surestimation dans leurs résultats et dont le score de cohérence spectrale (NCSq T80) est moyen, il s'agit de [V]q39 et [M]q17, pour ces deux QCM la cohérence interne est elle aussi moyenne.

En ce qui concerne les sous-estimations la situation est très différente, les questions qui présentent les Cq- T80 les plus élevés font partie des questions qui récoltent des valeurs moyennes, voire élevées, aux indices de qualité spectrale et de cohérence interne, et, jamais les valeurs les moins élevées à ces indices comme c'est le cas lorsqu'il y a surestimation. Une seule question qui obtient des scores spectraux et de cohérence interne moyens ou peu élevés possède une tendance à la sous-estimation relativement moyenne (≃), il s'agit de [V]q27.

En synthèse, voici les performances des huit questions qui obtiennent les valeurs les moins élevées en qualité spectrale et en cohérence interne (les questions contenues dans les cases grisées). Nous avons ajouté les performances de [P]q7 qui obtient un « + » à l'indice *NCSq T80* et un « - » à l'indice *Rq T80*.

Répartition des 173 QCM MOHICAN en fonction de la qualité spectrale et de la cohérence interne des questions

Cohérence interne \ Qualité spectrale	++	+≃	≃≃	-≃	--	+-
++	39	19	-	-	-	
+≃	20	19	1	-	-	
≃≃	21	26	2	2	2	1
-≃	2	8	3	2	1	
--	-	2	1	-	2	

Tableau des performances des 8 QCM qui obtiennent les valeurs les moins élevées en qualité spectrale et en cohérence interne ainsi que des performances de [P]q7

		NCSq T80	*Rq T80*	*Cq+ T80*	*Cq- T80*	$r_{qt\ mb}$	$r_{qt\ ms}$
[H]q3		−	−	−		−	−
[H]q20		−	−	−		−	−
[V]q5		−	−	−		−	≃
[V]q12		−	−	−		≃	≃
[S]q1		−	−	−		≃	≃
[V]q27		−	≃		≃	−	≃
[A]q14		−	≃	+		−	≃
[B]q5		≃	≃	≃		−	−
[P]q7		+	−	≃		≃	≃

Le dernier tableau des performances des 8 QCM qui obtiennent les valeurs les moins élevées en qualité spectrale et en cohérence interne reprend aussi les ingénogrammes (voir détails p. 346) de ces questions signalées par un fond grisé. Pour [H]q3, [H]q20, [V]q5, [V]q12 et [S]q1, le fait que le haut des ingénogrammes présente une surface plus grande montre une tendance à la surestimation dans les résultats de ces questions. Pour [V]q27 on remarque une plus grande surface en bas de l'ingénogramme, ce qui montre une tendance à la sous-estimation. En ce qui concerne [A]q14 l'ingénogramme présente une forme de diabolo et la question présente ni trop de sous-estimation, ni trop de surestimation dans ses résultats. Enfin, pour [B]q5 on remarque que les contours sont plus proches du centre avec cependant une surface plus orientée vers la gauche, donc les performances en cohérence interne sont plus faibles que les performances spectrales.

On voit bien tout le parti que pourraient retirer les professeurs des outils de visualisation que nous avons proposés dans cette section (ingénogrammes et graphiques 3D classico-spectraux) car ils permettent notamment de visualiser rapidement les éventuels problèmes liés aux questions. Nous comptons introduire ces outils dans nos interfaces de gestion de la qualité des tests standardisés, des projets sont en cours de réalisation et nous livrons quelques exemples dans la partie « Perspectives » de nos conclusions détaillées (voir p 440).

E. Brins Spectraux (BSq) de questions sélectionnées sur la base de leurs performances globales très élevées ou très faibles

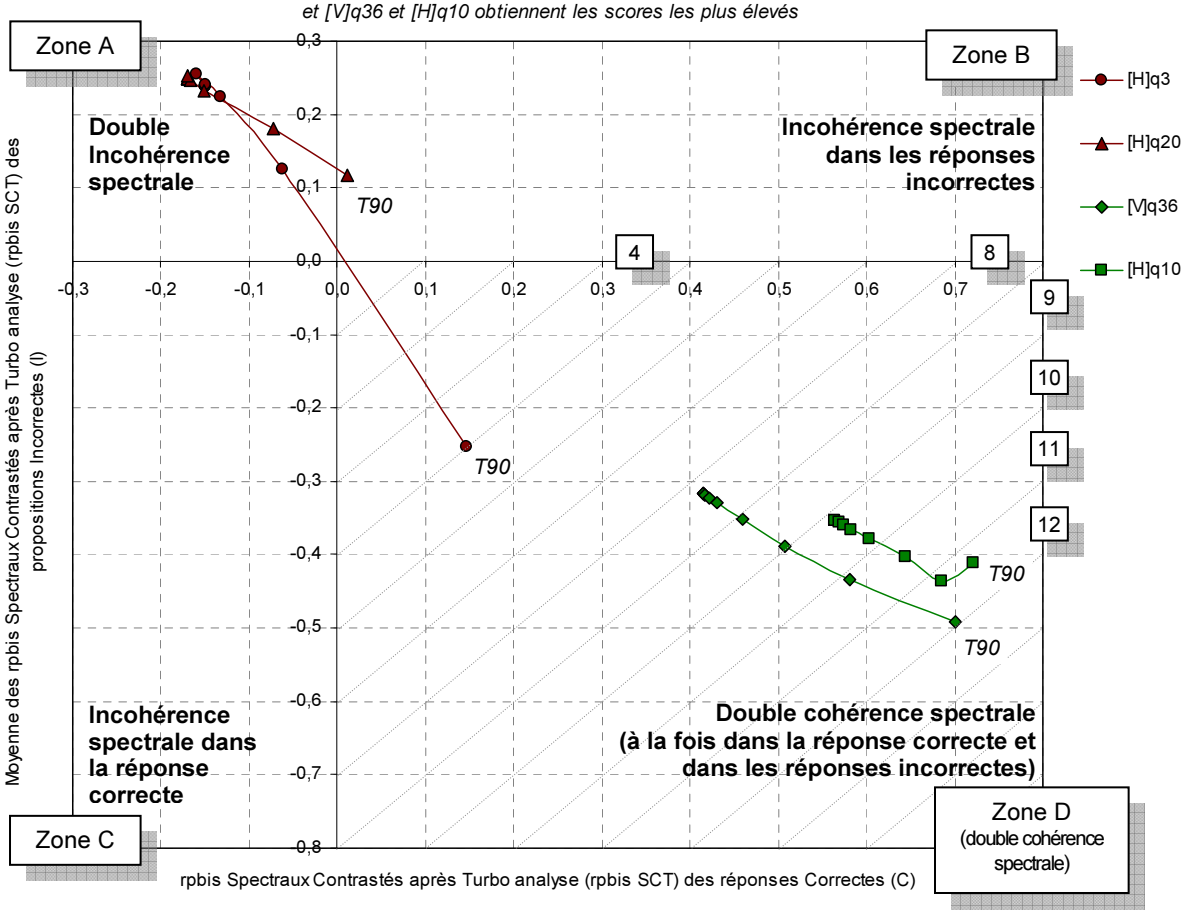
Comment visualiser les niveaux de cohérence spectrale (indice NCSq) pour les QCM appartenant aux 10 épreuves MOHICAN ?

Nous avons sélectionné dans cette section 10 questions en fonction de leurs performances spectrales et de cohérence interne. Notre sélection comporte huit questions qui obtiennent les valeurs les moins élevées en qualité spectrale et en cohérence interne ([V]q5, [V]q12, [V]q27, [H]q3, [H]q20, [S]q1, [A]q14 et [B]q5) ainsi que deux questions qui obtiennent les performances les plus élevées ([H]q10 et [V]q36).

1. *BSq* de deux questions performantes ([H]q10, [V]q36) comparés aux *BSq* de deux autres moins performantes ([H]q3, [H]q20)

Nous allons d'abord comparer les Brins Spectraux (*BSq*, voir p. 228) des deux questions les plus performantes ([H]q10 et [V]q36) avec ceux des deux questions les moins performantes ([H]q3 et [H]q20) selon le classement du tableau des performances des 173 QCM (voir p. 350). Le graphique ci-dessous reprend ces quatre questions contrastées du point de vue de leurs performances. Nous avons modifié les échelles des axes du graphique afin d'en améliorer la lisibilité. Le minimum est fixé à -0,8 sur l'axe des ordonnées (Y) et à –0,3 sur l'axe des abscisses (X). Quant au maximum, il est fixé à 0,3 sur l'axe Y et à 0,8 sur l'axe X (normalement les maximum et minimum sont respectivement fixés à 1 et –1).

Zoom sur la Gerbe Spectrale d'une sélection de 4 qcm (parmi les 173 des 10 épreuves MOHICAN) :
[H]q3 et [H]q20 obtiennent les moins bons scores en qualité spectrale et en cohérence interne
et [V]q36 et [H]q10 obtiennent les scores les plus élevés



Rappelons la signification des quatre zones A, B, C et D. Les points situés dans la zone A de « Double incohérence spectrale » sont représentatifs d'une situation problématique à la fois pour les distracteurs et pour la réponse correcte : la moyenne pondérée des *rpbis SCT* des réponses incorrectes est positive et le *rpbis SCT* de la réponse correcte est négatif. La zone B est révélatrice des situations où une ou plusieurs réponses incorrectes présentent des problèmes d'incohérence spectrale, la moyenne pondérée des *rpbis SCT* des réponses incorrectes étant positive. Dans cette zone B, le *rpbis SCT* de la réponse correcte est positif. A l'inverse de la zone B, la zone C permet de visualiser les QCM présentant des problèmes d'incohérence spectrale au niveau de la réponse correcte : le *rpbis SCT* de la réponse correcte est négatif. La moyenne pondérée des *rpbis SCT* des réponses incorrectes est aussi négative. Enfin, la zone D est une zone de double cohérence spectrale dans la mesure où les points qui figurent dans cette zone représentent, à différents niveaux de turbo analyse, les situations où les questions récoltent pour la réponse correcte un *rpbis SCT* positif et pour les propositions incorrectes une moyenne pondérée des *rpbis SCT* négative.

Les nombres placés en pourtour de la zone D dans les cadres ombrés désignent une série de niveaux de cohérence spectrale représentés par les espaces entre les diagonales. Plus un point se situe près du coin inférieur droit du graphique, meilleure est la performance en cohérence spectrale.

Nous constatons que les *BSq* des deux questions dont les performances globales en qualité spectrale et en cohérence interne sont les moins bonnes, [H]q3 et [H]q20, obtiennent aux paliers T0 et T80 de turbo analyse des valeurs qui les amènent dans la zone A. Rappelons que cette zone A est révélatrice de problèmes d'incohérence spectrale qui apparaissent tant du point de vue de la réponse correcte que du point de vue d'un ou plusieurs distracteurs de la QCM. En effet, lorsque les points d'une question apparaissent dans la zone A, d'une part la moyenne pondérée des *rpbis SCT* des propositions incorrectes est positive et d'autre part le *rpbis SCT* de la réponse correcte est négatif. Cette double incohérence spectrale paraît d'autant plus anormale lorsqu'elle est observée à partir des données des étudiants dont les scores de réalisme (*Rs*) sont particulièrement élevés, par exemple à T80, quand $Rs \geqslant 80$. Ces étudiants commettant en moyenne moins de 20% d'erreurs d'auto-estimations, les questions qui ne se trouvent pas dans la zone de cohérence spectrale à ce palier T80 doivent particulièrement attirer notre attention car elles contiennent peut être une(des) anomalie(s) qui explique(nt) l'incohérence spectrale observée.

Nous voyons qu'à T90 la question [H]q3 entre dans la zone de cohérence spectrale où la moyenne pondérée des *rpbis SCT* des réponses incorrectes est négative et où le *rpbis SCT* de la réponse correcte est positif. Donc, pour [H]q3, les étudiants dont le score de réalisme *Rs* est égal ou supérieur à 90 et qui choisissent la réponse correcte ont tendance à accompagner celle-ci de pourcentages de certitude plus élevés que les étudiants qui se sont trompés (et dont le réalisme *Rs* est aussi égal ou supérieur à 90). Ces choix sont cohérents et, à ce niveau de turbo analyse T90, la question ne pose pas de problèmes de cohérence spectrale au sens où nous l'avons défini précédemment (voir p. 227). Cependant, nous remarquons que le niveau de cohérence spectrale atteint par la question [H]q3 à T90 est peu élevé, il s'agit du niveau « 4 » (cadre ombré en pourtour) et on voit que 8 niveaux séparent cette question à T90 des deux QCM les plus performantes ([V]q36 et [H]q10).

Pour la question [H]q20, la situation se présente différemment au palier de turbo analyse T90 où la question entre dans la zone B. Un point dans cette zone signifie que des problèmes d'incohérence spectrale apparaissent au niveau des réponses incorrectes, la réponse correcte ne présentant quant à elle pas de problème d'incohérence spectrale. Au palier T90, les étudiants particulièrement réalistes ($Rs \geqslant 90$) lorsqu'ils choisissent une réponse incorrecte ont donc tendance à l'accompagner de pourcentages de certitude plus élevés. Cette question doit donc être examinée plus attentivement et nous le ferons dans la partie suivante consacrée à l'analyse du niveau « propositions ».

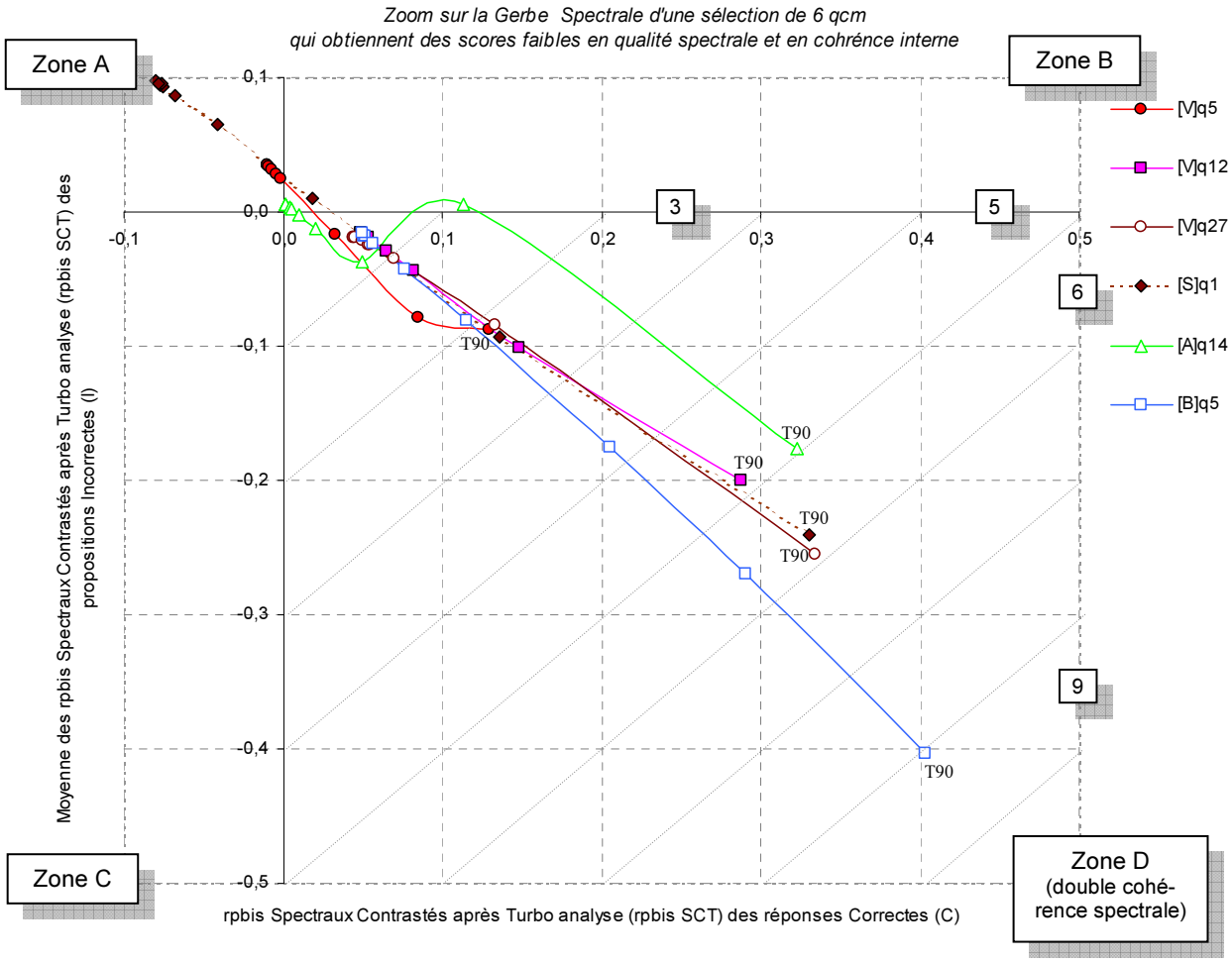
Pour des deux questions qui obtiennent les performances globales les plus élevées, [V]q36 et [H]q10, contrairement à [H]q20, tous les points des *BSq* se trouvent dans la zone D de cohérence spectrale. Les points dans cette zone signifient que la moyenne pondérée des *rpbis SCT* des réponses incorrectes est négative et que le *rpbis SCT* de la réponse correcte est positif, c'est le type de situation à laquelle on s'attend en cas d'utilisation cohérente des pourcentages de certitude. En ce qui concerne ces deux questions globalement plus performantes, nous remarquons qu'elles atteignent le niveau de cohérence spectrale

« 12 » à T90. Pour [H]q10 nous constatons que le niveau « 12 » est déjà atteint à T80. Le Brin Spectral (*BSq*) de la question [H]q10 est plus court que celui de [V]q36 : au palier T0, [H]q10 atteint le niveau « 10 » tandis qu'au même palier nous trouvons [V]q36 à deux niveaux inférieurs (c'est-à-dire le « 8 »).

2. Comparaison des six autres questions aux moins bonnes performantes globales : [V]q5, [V]q12, [V]q27, [S]q1, [A]q14, et [B]q5

Nous allons maintenant envisager la gerbe spectrale des six autres questions qui ont obtenu les moins bonnes performances globales[67] en qualité spectrale et cohérence interne (voir tableau p. 356).

Pour améliorer la lisibilité du graphique ci-après nous avons modifié les échelles des axes. Sur ce graphique le minimum est fixé à -0,5 sur l'axe Y et à –0,1 sur l'axe X. Quant au maximum, il est fixé à 0,1 sur l'axe Y et à 0,5 sur l'axe X.



Nous voyons qu'aux paliers turbo les moins élevés les *BSq* des questions [S]q1 et [V]q5 se positionnent dans la zone A où à la fois les solutions incorrectes et la réponse correcte posent un problème au niveau de la cohérence spectrale. Cependant, aux paliers turbo les plus élevés ces questions entrent dans la zone de cohérence spectrale. A T90, [V]q5 atteint le niveau de cohérence spectrale « 3 » et [S]q1 le niveau « 6 ».

[67] C'est à dire les 6 questions classées après les deux QCM les moins performantes [H]q3 et [H]q20 que nous avons discutées au début de cette section.

Nous constatons que trois questions sur les six, [V]q27, [B]q5 et [V]q12 se positionnent dès le palier turbo T0 dans la zone A de cohérence spectrale.

A T90, [B]q5 atteint le niveau « 9 » ; [V]q27, [S]q1 atteignent le niveau « 6 » ; [V]q12, [A]q14 le niveau « 5 » et [V]q5 le niveau « 3 ».

La question [A]q14 présente un *BSq* qui aux paliers turbo les plus bas est à la limite de la zone B et de la zone A et qui entre dans la zone B à T80 pour en ressortir et se positionner finalement dans la zone D au palier T90. A ce palier turbo elle atteint le niveau de cohérence spectrale « 5 ».

Lorsque nous comparons les niveaux de cohérence spectrale atteints par les six questions de cette gerbe spectrale avec les niveaux des quatre questions précédentes (voir gerbe spectrale p. 357), nous constatons que les niveaux atteints à T90 sont tous inférieurs au niveau « 12 » des deux questions globalement les plus performantes [H]q10 et [V]q36 au même palier turbo (une question au niveau « 9 », deux à « 6 », deux à « 5 » et une à « 3 »). Remarquons aussi qu'une des deux questions globalement les moins performantes, [H]q3 (sur la gerbe spectrale p. 357), atteint le niveau 4 à T90 et dépasse la [V]q5 à ce palier de turbo analyse.

En conclusion, dans le contexte des épreuves MOHICAN, au palier de turbo analyse T90, parmi les huit questions globalement les moins performantes que nous avons identifiées précédemment (voir tableau p. 356), une seule ([H]q20) ne figure pas dans la zone A de cohérence spectrale, mais dans la zone B qui révèle des problèmes d'incohérence spectrale au niveau des solutions incorrectes. Toutes les autres obtiennent à T90 des niveaux situés entre « 3 » et « 9 ». Les deux questions globalement les plus performantes, [H]q10 et [V]q36, obtiennent « 12 » à T90.

Au palier turbo T80, deux questions se situent dans la zone A de double incohérence spectrale, il s'agit de [H]q20 et [H]q3.

F. Conclusions de l'analyse du niveau « QCM »

1. Liaison des indices de qualité spectrale et indices de cohérence interne

Au début de cette exploration du niveau QCM dans le contexte des 10 épreuves MOHICAN, nous avons examiné les liaisons entre les indices classiques et les indices spectraux.

Nous avons sélectionné les trois tests qui obtiennent les meilleurs coefficients de fidélité : VOCABU, MATHEM et ARTACT (voir tableau p. 309 et ingénogrammes p. 320) et corrélé les valeurs obtenues aux indices spectraux et classiques. Dans ces trois épreuves nous avons observé des corrélations élevées entre les différents indices classiques de cohérence interne (voir p. 331).

En ce qui concerne les indices spectraux, nous constatons qu'ils ne possèdent pas des liaisons aussi fortes que celles observées entre les indices de cohérence interne (p. 334). Pour les trois épreuves nous ne remarquons pas de corrélation entre le Niveau de Cohérence Spectrale par question (*NCSq*) et la Centration par question (*Cq*). Par contre nous observons une liaison marquée, surtout au palier de turbo analyse T80, entre les indices *Cq* et de Réalisation des prédictions par question (*Rq*). Ces corrélations élevées sont négatives ce qui est logique dans la mesure où l'indice *Rq* nous donne une information sur les erreurs moyennes d'auto-estimations qui ont été commises et l'indice *Cq* de Centration par question une information complémentaire sur la tendance à la sous-estimation [-] ou à la surestimation [+] et sur l'ampleur de cette tendance enfermée dans les résultats de la question. Dès lors, un *Rq* élevé (peu d'erreurs d'auto-estimation) aura tendance à être associé à une *Cq* (calculée en valeur absolue) faible (les sur et sous-estimations étant peu élevées).

Nous observons des corrélations moins élevées entre *Rq* et *NCSq* et nous constatons aussi que ces corrélations faiblissent au palier de turbo analyse T80.

Nous remarquons aussi de fortes corrélations entre chaque indice spectral sans turbo analyse et sa variante calculée au palier turbo T80, sauf pour l'épreuve ARTACT entre les indices *Cq* et *Cq T80*.

Il existe aussi des tendances à varier ensemble entre indices de qualité spectrale et indices de cohérence interne (p. 337), mais ces tendances ne se présentent pas de la même manière d'une épreuve à l'autre. Pour une épreuve, des corrélations fortes existeront entre certains indices spectraux et certains indices classiques qui n'apparaîtront plus dans un autre test.

2. Lien entre qualité spectrale et cohérence interne peu élevées et surestimation prononcée

Nous avons d'abord visualisé sur les axes de 173 ingénogrammes (p. 347) correspondant aux 173 QCM, les valeurs obtenues par six indices : quatre spectraux (*Rq T80*, *NCSq T80*, *Cq+ T80* et *Cq- T80*) et deux de cohérence interne ($r_{qt\ mb}$ et $r_{qt\ ms}$). Nous avons repéré les graphiques correspondant aux deux questions les plus performantes et aux deux les moins performantes et avons remarqué que d'autres questions possédaient des ingénogrammes dont les caractéristiques étaient proches.

Nous avons aussi établi une catégorisation en trois niveaux de performance « +, ≃ ou – » pour les six indices envisagés. Nous avons attribué à chacune des questions la catégorie de performance qui lui correspond pour chacun des six indices. Nous avons ainsi abouti à un tableau des performances des 173 QCM (p. 350). Les questions qui récoltent que des « + », donc les plus performantes, ont été mises en évidence et nous en dénombrons 32 parmi lesquelles figurent les deux questions les plus performantes [V]q36 et [H]q10 précédemment repérées à l'aide des 3 indices *Rq T80*, *NCSq T80* et $r_{qt\ mb}$. Les questions les moins performantes récoltant que des « – » ont aussi été mises en évidence, nous en trouvons deux : [H]q3 et [H]q20 et remarquons que [V]q5 précédemment repérée ne compte pas parmi elles, cette dernière question obtenant un « ≃ » à l'indice $r_{qt\ ms}$. Nous remarquons à l'aide du tableau que la question peu

performante [H]q3 comporte (comme [H]q20) une tendance prononcée à la surestimation (*Cq+ T80* = « – »).

Ensuite, nous avons réalisé une série de tableaux de répartition des QCM en fonction des valeurs obtenues aux différents indices.

Le tableau de répartition des questions en fonction de la qualité spectrale (p. 351) montre qu'une seule question parmi les 173 obtient un score élevé « + » à l'indice spectral *NCSq T80* et un score faible « - » à l'autre indice spectral *Rq T80*, il s'agit de la QCM [P]q7. Nous avons aussi relevé cinq questions de moins bonne qualité spectrale, obtenant un score faible « – » aux deux indices (*NCSq T80* et *Rq T80*), il s'agit de : [V]q5, [V]q12, [S]q1, [H]q3 et [H]q20. Nous retrouvons parmi ces questions les deux QCM qui globalement sont les moins performantes : [H]q3 et [H]q20.

La répartition des questions en fonction de la cohérence interne ($r_{qt\ mb}$ et $r_{qt\ ms}$) ne montre pas de cas contradictoire où pour une question un indice de cohérence interne récolterait un « – » et l'autre indice un « + » (voir p. 351). Cinq questions récoltent un « - » aux deux indices de cohérence interne : [V]q43, [H]q3, [B]q5, [B]q6 et [H]q20.

Lorsque nous n'envisageons que les questions qui enferment une tendance élevée à la surestimation (*Cq+* = « - ») dans leurs résultats (voir tableau p. 354), nous constatons que 7 QCM sont concernées : [V]q5, [V]q12, [V]q39, [H]q3, [H]q20, [S]q1 et [M]q17. Elles obtiennent toutes des valeurs faibles « - » ou moyennes « ≃ » aux indices spectraux ainsi qu'aux indices de cohérence interne. Sur ces 7 questions à forte surestimation, 5 figurent dans les cases grisées désignant les plus mauvaises performances en qualité spectrale et en cohérence interne du tableau récapitulatif de la page 356.

Lorsque nous envisageons seulement les questions qui enferment la plus forte tendance à la sous-estimation dans leurs résultats (*Cq-* = « - », voir tableau p. 354), nous en dénombrons 9. Aucune de ces 9 questions avec sous-estimation élevée n'obtient des performances faibles en qualité spectrale et cohérence interne. Parmi les 93 QCM qui enferment une tendance à la sous-estimation, seule une question ([V]q27) obtient des scores spectraux et de cohérence interne moyens ou peu élevés et présente une tendance à la sous-estimation plutôt moyenne (≃) dans ses résultats.

Nous concluons de ces constats qu'il existe un lien entre tendance élevée à la surestimation dans les résultats des questions et performances faibles à la fois en qualité spectrale et en cohérence interne. Cependant, ce lien n'est pas absolu dans la mesure où il existe des cas de questions présentant une forte surestimation et qui obtiennent des scores aux indices de cohérence interne et spectrale qui ne figurent pas parmi les plus faibles. Il s'agit des questions [V]q39 et [M]q17 dont les indices *NCSq T80*, $r_{qt\ mb}$ et $r_{qt\ ms}$ récoltent des valeurs moyennes (« ≃ »).

3. Conclusions des observations sur les Brins Spectraux des questions (*BSq*)

Nous avons montré les Brins Spectraux (*BSq*, voir p. 228) d'une sélection de 10 QCM. Parmi ces questions, deux obtenaient dans notre classement (p. 343) les performances les plus élevées en cohérence spectrale et en cohérence interne : [H]q10 et [V]q36. Huit autres obtenaient les performances les moins élevées (p. 356). Il s'agit de [V]q5, [V]q12, [V]q27, [H]q3, [H]q20, [S]q1, [A]q14 et [B]q5.

Dans un premier temps nous avons prélevé parmi ces dix QCM, les quatre les plus contrastées du point de vue de leurs performances : [H]q10, [V]q36 (performances les plus élevées) et [H]q3, [H]q20 (performances les plus faibles) et avons observé leurs *BSq* (voir graphique p. 357).

Nous remarquons que les plus performantes atteignent le niveau de cohérence spectrale « 12 ». Rappelons que les niveaux de cohérence spectrale se situent dans la zone D du graphique et vont de 1 à 20. Une des deux questions les moins performantes, [H]q3, n'atteint pas cette zone D de cohérence spectrale et se cantonne aux paliers turbo T10 à T80 dans la zone A révélatrice de problèmes à la fois au niveau de la proposition correcte et des propositions incorrectes ; à T90 elle se situe dans la zone B, mais juste en

bordure de A. L'autre question la moins performante, [H]q20, se cantonne aussi dans la zone A de T10 à T80 mais « pénètre » dans la zone D de cohérence spectrale à T90 et atteint alors le niveau de cohérence spectrale « 3 ».

Nous avons ensuite observé les *BSq* des six autres questions peu performantes (voir graphique p. 360). Nous constatons que dès le palier turbo T0, quatre questions se positionnent dans la zone D de cohérence spectrale ([V]q27, [B]q5 et [V]q12). A T90 les niveaux de cohérence atteints sont les niveaux : « 9 » pour [B]q5, « 6 » pour [V]q27 et « 5 » pour [V]q12. Pour ce qui est des trois autres questions qui se positionnent dans la zone A aux paliers de turbo analyse les moins élevés, [S]q1, [A]q14 et [V]q5, A T80 ces trois questions entrent dans la zone D de cohérence interne. A T90, [S]q1 atteint le niveau « 6 », [A]q14 le niveau « 5 » et [V]q5 le niveau « 3 ».

Nous constatons que parmi les 8 questions les moins performantes des 173 QCM, une seule ne figure pas à T90 (et *a fortiori* à T80) dans la zone de cohérence interne D (il s'agit de [V]q20) et aucune ne franchit le niveau « 9 ».

Lorsque nous avons présenté les Brins Spectraux par question (p. 228), nous nous sommes demandé si tous les 173 *BSq* des QCM des 10 épreuves MOHICAN figuraient dans la zone D de cohérence spectrale ([2.1], p. 234). Après analyse des *BSq* des 10 questions sélectionnées nous pouvons répondre qu'il en existe deux qui au palier de turbo T80 ne figurent pas dans la zone D mais dans la zone A : [H]q3 et [H]q20. Lorsque nous observons les brins spectraux des autres questions sur les Gerbes Spectrales des tests (*GSt*) des épreuves MOHICAN en annexes (pp. 511-521), nous remarquons que toutes atteignent la zone de cohérence spectrale à T80 et *a fortiori* à T90. Signalons qu'au palier de turbo analyse T0, seulement trois questions qui n'ont pas été signalée précédemment, [CO]q3, [H]q16 et [B]q6 ne figurent pas dans la zone D. Donc, en ce qui concerne la cohérence spectrale, seules deux questions, [H]q3 et [H]q20, posent des problèmes de double incohérence lorsqu'on prend en compte les données des étudiants dont le réalisme est supérieur ou égal à 80.

Lors de la présentation des *BSq* à l'aide des données du test de physique (p. 229) nous avons remarqué un regroupement des points à la base des *BSq*. Nous nous étions demandé si ces regroupements à la base des brins, aux paliers T0 à T40, étaient une caractéristique présente dans les *BSq* des 9 autres épreuves ([2.2], p. 235).

L'observation des *GSt* en annexes montre que le regroupement des points à la base des brins constitue une caractéristique commune aux 173 QCM. Comment l'interpréter ? C'était une des questions que nous nous sommes posé lors de la présentation des *BSq* ([2.2], voir p. 235).

Les points des brins spectraux sont placés sur le graphique à l'aide des coordonnées provenant de la valeur du *rpbis SCT* de la solution correcte (coordonnée X) et de la moyenne pondérée des *rpbis SCT* des propositions incorrectes (coordonnée Y). Si de T0 à T40 ces points sont proches les uns des autres, c'est que les différences entre les *rpbis SCT* sont faibles à ces paliers turbo peu élevés.

Or, nous savons qu'entre T0 et T40, pour toutes les épreuves, les effectifs d'étudiants varient peu (p. 290). Entre T0 et T40 on observe dans les effectifs une faible différence de l'ordre de 3% seulement de sujets (100% - 97%). Les résultats calculés à ces différents paliers turbo peu élevés ne varient guère, les données étant *grosso modo* les mêmes. Donc, si les *rpbis SCT* des propositions varient peu à ces paliers turbo, c'est parce que les données à l'aide desquelles nous les calculons sont pratiquement identiques.

Par contre, aux paliers de turbo analyse élevés T70, T80 et T90 les effectifs diminuent fortement à chaque palier. Il reste en moyenne pour les 10 épreuves : 69% de sujets à T70 contre 39% à T80 et 9% à T90. De plus, nous savons que plus on monte dans les paliers de la turbo analyse, plus les étudiants sont réalistes et moins ils commettent d'erreurs dans leurs auto-estimations. Les *rpbis SCT* étant calculé sur des données de plus en plus fiables du point de vue de l'utilisation des pourcentages de certitude, l'écart entre les *rpbis SCT* calculés à deux paliers turbo élevés voisins se creuse. Dès lors, on peut s'attendre à ce que les différences qui séparent les *rpbis SCT80* des *rpbis SCT90* puissent être bien plus élevées que celles qui séparent un *rpbis SCT0* d'un *rpbis SCT10* et c'est ce que nous observons sur les 173 *BSq*.

Chapitre XI :

Exploration du niveau « PROPOSITIONS »

Sommaire

A. Introduction

Les statistiques suivantes correspondent à un 3ème niveau de profondeur pour l'analyse des épreuves MOHICAN, elles fournissent des informations qualitatives spécifiques pour chaque proposition d'une QCM. Ces statistiques ont été détaillées dans les sections précédentes. Nous les rappelons brièvement (les pages où ces indices sont présentés sont mentionnées entre parenthèses) :

(1) INDICES SPECTRAUX	(2) INDICES CLASSIQUES SUR MATRICE BINAIRE	(3) INDICES CLASSIQUES SUR MATRICE SPECTRALE[68]
• *rpbis SC : Coefficient de corrélation p*oint *bis*érial *S*pectral *C*ontrasté (p. 178 et 216) • *rpbis SCT : Coefficient de corrélation p*oint *bis*érial *S*pectral *C*ontrasté calculé après *T*urbo analyse (pp. 184 et 217) • *CMp : C*ertitude *M*oyenne par *p*roposition (pp. 217 et 252) • *N Rép. T : n*ombre de *Rép*onses calculé après *T*urbo analyse pour une proposition (p. 217) • *% Rép. T : pourcentage* de *Rép*onses calculé après *T*urbo analyse pour une proposition (p. 217) ;	• *N Rép. : n*ombre de *Rép*onses pour une proposition (p. 216) • *% Rép. : pourcentage* de *Rép*onses pour une proposition (p. 216) • *rpbis classique : Coefficient de corrélation p*oint *bis*érial classique (pp. 171 et 216)	

Dans la partie précédente nous avons identifié parmi les 173 questions des 10 épreuves MOHICAN, parmi huit questions aux performances faibles, deux QCM qui globalement présentent les valeurs les moins élevées en qualité spectrale et en en cohérence interne (voir pp. 350 et 357). Il s'agit de [H]q3 et [H]q20. Sur le graphique p. 357, nous remarquons que le brin spectral de [H]q3 n'atteint pas la zone D de cohérence spectrale et se cantonne à T90 dans la zone B révélatrice d'un problème de cohérence d'utilisation des pourcentages de certitude pour les solutions incorrectes. [H]q20, atteint le niveau de cohérence spectrale « 4 » à T90 (rappelons que la zone D compte 20 niveaux de cohérence spectrale).

Nous avons aussi identifié la question qui obtient les valeurs les plus élevées aux indices classiques et spectraux : [V]q36 (p. 344). Cette QCM atteint le niveau « 12 » à T90 dans la zone D de double cohérence spectrale (voir graphique p. 357). Dans cette partie nous allons d'abord commenter et analyser les valeurs obtenues par les propositions de cette question [V]q36. Ensuite nous analyserons les valeurs obtenues aux indices spectraux et classiques par les propositions des huit questions dont les performances globales sont faibles (voir tableau récapitulatif, p. 356).

Nous avons aussi systématiquement analysé les *rpbis classiques*, *rpbis SC*, *rpbis SCT80* et *rpbis SCT90* des 173 questions des 10 épreuves (voir tableau comparatif en annexe, p. 544) et avons épinglé huit autres questions dont les configurations des rpbis posent problèmes. Les propositions de ces huit QCM seront analysées dans l'avant dernière section. Enfin, dans les conclusions de ce chapitre nous comparerons les performances spectrales et classiques des questions dont nous avons analysé les indices des propositions et confronterons les valeurs observées aux avis des experts de l'équipe de réalisation des épreuves, avis exprimés lors de deux réunions de débriefing des tests organisées au siège du CIUF[69] les 3 et 7 février 2000.

[68] Cette zone est grisée car qu'il s'agisse d'une matrice binaire ou d'une matrice spectrale, les *N Rép.* et *% Rép.* sont identiques. En ce qui concerne les *rpbis classiques*, leur calcul n'est envisagé qu'avec des matrices binaires.

[69] Conseil interuniversitaire de la Communauté française de Belgique

B. Interprétation des indices des propositions de la question la plus performante : [V]q36

1. Indices spectraux après turbo analyse

Voici d'abord un extrait du protocole *SCANTEST 2.0* d'analyse des propositions de la question [V]q36 (voir en annexe, p. 523) posée à 3.846 étudiants.

Rappelons que les performances en qualité spectrale et en cohérence interne de cette QCM sont globalement les meilleures des 173 questions. Nous avons sélectionné la partie du protocole qui concerne les *rpbis SCT* obtenus aux paliers de turbo analyse T10 à T90 avec un pas de 10.

Les tableaux ci-contre reprennent une série d'informations pour les trois propositions (P1 à P3) et l'omission (OM) de cette 36ème question du test de vocabulaire. La proposition correspondant à la réponse correcte est signalée en gras (P1).

Dans chaque tableau, la ligne dont l'intitulé est « N Rép. » suivi du palier de turbo analyse « T… » reprend les effectifs d'étudiants correspondant à chacune des trois propositions ainsi qu'à l'omission, et ce, pour le palier turbo mentionné. La ligne « % Rép. T… » reprend ces effectifs mais exprimés en pourcentages du nombre d'étudiants dont les données ont été sélectionnées à ce palier turbo T.

La ligne « C. Moy. T… » contient la moyenne des pourcentages de certitude qui ont accompagné le choix de la proposition ou de l'omission.

Enfin, la ligne « rpbis SCT… » contient la valeur du rpbis spectral contrasté calculé au palier turbo T.

Nous remarquons que de T10 à T20 les effectifs ne varient pas. Nous nous attendons à ce que les effectifs varient peu jusqu'à T50 étant donné nos observations sur les effectifs des 10 tests aux 10 paliers turbo (voir tableaux p. 290 et graphique p. 291).

Nous remarquons qu'un petit pourcentage d'étudiants omettent (voir colonne OM, les % Rép. varient entre 4% et 9%).

En ce qui concerne la certitude moyenne de la réponse correcte (P1) nous constatons qu'elle augmente très peu de T10 à T80 (elle passe de 68% à 72%). La certitude moyenne augmente encore de 6 % entre T80 et T90 où elle vaut 78%.

Pour les réponses incorrectes P2 et P3 les certitudes moyennes diminuent de 5% pour P2 en passant de 42% (T10) à 37% (T90). Pour P3 la certitude moyenne passe de 44% (T10) à 39% (T90).

1. Palier de Turbo analyse : T10

	OM	**P1**	P2	P3
1.1 N Rép. T10	157	**2080**	323	1252
1.2 % Rép. T10	4%	**54%**	8%	33%
1.3 C. Moy. T10	9%	**68%**	42%	44%
1.4 rpbis SC T10	-0,35	**0,41**	-0,22	-0,35

2. Palier de Turbo analyse : T20

	OM	**P1**	P2	P3
2.1 N Rép. T20	157	**2080**	323	1252
2.2 % Rép. T20	4%	**54%**	8%	33%
2.3 C. Moy. T20	9%	**68%**	42%	44%
2.4 rpbis SC T20	-0,35	**0,41**	-0,22	-0,35

3. Palier de Turbo analyse : T30

	OM	**P1**	P2	P3
3.1 N Rép. T30	157	**2078**	323	1252
3.2 % Rép. T30	4%	**54%**	8%	33%
3.3 C. Moy. T30	9%	**68%**	42%	44%
3.4 rpbis SC T30	-0,35	**0,42**	-0,22	-0,35

4. Palier de Turbo analyse : T40

	OM	**P1**	P2	P3
4.1 N Rép. T40	157	**2073**	321	1250
4.2 % Rép. T40	4%	**54%**	8%	33%
4.3 C. Moy. T40	9%	**69%**	41%	44%
4.4 rpbis SC T40	-0,35	**0,42**	-0,23	-0,35

5. Palier de Turbo analyse : T50

	OM	**P1**	P2	P3
5.1 N Rép. T50	156	**2054**	318	1242
5.2 % Rép. T50	4%	**54%**	8%	33%
5.3 C. Moy. T50	9%	**69%**	41%	44%
5.4 rpbis SC T50	-0,36	**0,43**	-0,23	-0,36

6. Palier de Turbo analyse : T60

	OM	**P1**	P2	P3
6.1 N Rép. T60	152	**2003**	307	1194
6.2 % Rép. T60	4%	**54%**	8%	32%
6.3 C. Moy. T60	7%	**70%**	41%	42%
6.4 rpbis SC T60	-0,37	**0,46**	-0,24	-0,38

7. Palier de Turbo analyse : T70

	OM	**P1**	P2	P3
7.1 N Rép. T70	133	**1789**	252	1033
7.2 % Rép. T70	4%	**56%**	8%	32%
7.3 C. Moy. T70	6%	**71%**	39%	41%
7.4 rpbis SC T70	-0,40	**0,51**	-0,26	-0,42

8. Palier de Turbo analyse : T80

	OM	**P1**	P2	P3
8.1 N Rép. T80	103	**1199**	135	612
8.2 % Rép. T80	5%	**58%**	7%	30%
8.3 C. Moy. T80	3%	**72%**	37%	40%
8.4 rpbis SC T80	-0,48	**0,58**	-0,28	-0,47

9. Palier de Turbo analyse : T90

	OM	**P1**	P2	P3
9.1 N Rép. T90	41	**274**	24	113
9.2 % Rép. T90	9%	**60%**	5%	25%
9.3 C. Moy. T90	2%	**78%**	37%	39%
9.4 rpbis SC T90	-0,67	**0,70**	-0,28	-0,49

Pour ce qui est des *rpbis SCT*, nous remarquons que pour la réponse correcte (P1) il augmente très peu de T10 (0,41) à T40 (0,42) puis de plus en plus de T50 à T90 (0,43, 0,46, 0,51, 0,58 et 0,70).

Pour les propositions incorrectes P2 et P3, la situation est différente : les valeurs sont négatives et même de plus en plus négatives à partir de T50. P2 passe de –0,22 (T10, T20, T30) à –0,23 (T40, T50), puis à -0,24 (T60), -0,26 (T70) et à -0,28 (T80 et T90). Pour P3, les *rpbis SCT*, sont plus négatifs : -0,35 (T10, T20, T30, T40), -0,36 (T50), -0,38 (T60), -0,42 (T70), -0,47 (T80) et -0,49 (T90).

2. Le *rpbis SCT* de P2 est-il significativement différent de zéro lorsqu'il est calculé au palier de turbo analyse T90 ?

Contrairement au *rpbis classique* où les données de tous les sujets sont impliquées dans le calcul des corrélations (voir problématique du *rpbis classique*, p. 171), tous les couples de mesures ne sont pas utilisés pour le calcul de la corrélation bisériale de point spectrale contrastée des propositions incorrectes (voir problématique du *rpbis SC*, p. 178). Dès lors, certains se demanderont si à T90, la valeur du *rpbis SCT* de la proposition incorrecte P2 (-0,28) est significativement différente de zéro étant donné que P2 a été choisie par 24 sujets seulement (nous allons voir qu'en fait il y a bien plus que 24 couples de mesures appareillées qui sont impliquées dans le calcul du *rpbis SCT90* de la proposition P2).

Nous savons aussi et nous l'avons présentée précédemment, qu'il existe une formule proposée par Fisher (voir p. 164) qui permet la transformation du *r* en *t* de Student pour effectuer un test de signification $r \neq 0$.

Mais avant d'appliquer ce test, interrogeons-nous d'abord à propos du nombre de couples de mesures impliquées dans les calculs de la corrélation bisériale de point spectrale contrastée après turbo analyse au palier T90 (*rpbis SCT90*). Nous avons montré précédemment que les sujets impliqués dans le calcul du *rpbis SCT90* d'une proposition incorrecte sont parmi tous ceux dont le réalisme est supérieur ou égal à 90, d'une part ceux qui ont choisi la proposition incorrecte envisagée mais d'autre part aussi ceux qui ont choisi la réponse correcte (d'où l'appellation « contrasté »). Dès lors, dans le cas de la proposition P2 de [V]q36, il faut ajouter aux 24 sujets ayant opté pour P2, les 274 autres qui ont choisi la réponse correcte P1. Donc, nous obtenons 298 (24 + 274) couples de mesures à partir desquelles la corrélation sera calculée.

Voici ce que donne l'application de la formule de Fisher, étant donné *ns* = 298 :

$$t_c = \frac{|r|\sqrt{ns-2}}{\sqrt{1-r^2}} = \frac{0{,}28\sqrt{296}}{\sqrt{1-0{,}0784}} = \frac{4{,}817}{0{,}96} = 5{,}018$$

Lorsque nous comparons le *t* calculé (t_c) avec les valeurs théoriques (t_t) de la table du *t* de Student (voir annexe, p. 543), nous constatons que pour un nombre de degrés de liberté (ddl) qui tend vers ∞, on a t_t = 3,291 à 0,001. Donc, t_c est largement supérieur à t_t et il y a donc moins de 1 chance sur 1000 que le coefficient de corrélation ne soit pas différent de zéro à T90. Pour P2, le *rpbis SCT90* qui vaut –0,28 est donc significativement différent de zéro à T90.

3. Résumé des performances spectrales des propositions de [V]q36

Donc, pour cette question, globalement très performante, nous remarquons un *rpbis SCT* positif et très élevé pour la réponse correcte à T90 (0,70). Rappelons qu'à T90 les calculs sont effectués à partir des données des étudiants dont le réalisme est très élevé ($Rs \geqslant 90$) [70]. On voit également que pour la proposition incorrecte P3, le *rpbis SCT90* est négatif (-0,49). Le *rpbis SCT90* est aussi négatif pour P2 (-0,28), mais de façon moins marquée que le *rpbis SCT90* de P3.

[70] Ces étudiants commettent moins de 10% d'erreurs dans leurs auto-estimations.

Nous pouvons en conclure que l'utilisation des pourcentages de certitude est cohérente même aux paliers de turbo analyse les moins élevés : les étudiants qui répondent correctement ont tendance à fournir des certitudes plus élevées que celles fournies par ceux qui répondent incorrectement.

Enfin, en ce qui concerne les pourcentages moyens de certitude des différentes solutions et concomitamment avec ce qui vient d'être dit à propos des *rpbis SCT*, nous constatons, que la proposition correcte P1 obtient un pourcentage moyen de certitude presque deux fois plus élevé (72% à T80) que les propositions incorrectes P2 (37% à T80) et P3 (40% à T80).

4. indices classiques

Voici les informations du protocole d'analyse des propositions relatives aux indices classiques liés aux propositions de la question performante [V]q36.

Nous remarquons que pour la réponse correcte (P1) le *rpbis classique* est positif et vaut 0,51. En ce qui concerne les propositions incorrectes il est négatif pour P2 (-0,16) et P3 (-0,30).

	OM	**P1**	P2	P3
1) N Rép.	157	**2080**	323	1252
2) % Rép.	4%	**54%**	8%	33%
3) rpbis	-0,27	**0,51**	-0,16	-0,30

Dans le contexte des *rpbis classiques* nous sommes amené à calculer une valeur repère pour contrer le problème du recouvrement entre le score de la question et le score total du test (voir p. 176). Rappelons que dans le cadre des épreuves que nous corrigeons à l'aide du Système Méthodologique d'Aide à la Réalisation de Tests (SMART) nous utilisons une solution qui consiste à calculer une valeur repère pour la réponse correcte. L'application de cette solution dans le cas du test de vocabulaire qui comporte 45 questions donne :

$$repère\ \ rpbis\ \ classique\ \ = \ 1/\sqrt{nq} = 1/\sqrt{45} = 0{,}15$$

Nous remarquons que le *rpbis classique* de la réponse correcte (0,51) est largement au-dessus de ce repère.

Les étudiants qui répondent correctement ont donc tendance à récolter des scores totaux élevés (calculés sur l'ensemble des questions du test) et ceux qui répondent incorrectement des score totaux faibles.

En ce qui concerne l'attractivité (« N Rép. » et « % Rép. ») des distracteurs (P2 et P3), nous constatons que P2 est environ quatre fois moins choisie (8%) que P3 (33%). La question aurait pu encore être améliorée par un meilleur équilibrage de la difficulté des distracteurs.

5. Discussion des analyses des propositions de [V]q36

Notons d'abord qu'en ce qui concerne les rpbis spectraux contrastés (*rpbis SC*), nous ne les avons pas commenté ici car nous constatons que les valeurs obtenues au palier de turbo analyse T10 sont identiques. Cette équivalence des *rpbis SC* (où les données de tous les sujets sont prises en compte) et des *rpbis SCT10* (seules les données des étudiants dont le réalisme est supérieur ou égal à 10 sont utilisées) est due aux effectifs qui sont quasi les mêmes. En effet, ces effectifs ne varient pratiquement pas aux paliers de turbo analyse T0 et T10 (voir tableau p. 290 et graphique p. 291), ainsi, calculer le *rpbis SCT0* revient à calculer le *rpbis SC* (voir schéma de la turbo analyse, p. 187). Notons également que les effectifs varient en général très peu jusqu'au palier T50 et dès lors pour l'analyse spectrale des propositions de certaines questions nous avons choisi de n'utiliser que les *rpbis SC T50* à *T90* (avec un pas 10).

Ensuite, nous constatons que dans le cas de cette question [V]q36, en ce qui concerne l'exploration des propositions à l'aide des *rpbis classiques* et des *rpbis SCT*, les analyses se rejoignent et permettent de conclure qu'il n'y a pas de problème particulier à signaler au niveau des propositions en ce qui concerne la cohérence interne ou la cohérence spectrale. Ce constat effectué au niveau « PROPOSITIONS » confirme donc la qualité excellente de cette question mise en évidence au niveau d'exploration « QCM ». De plus, cette question n'a pas soulevé de commentaire particulier de la part des experts lors de la réunion de débriefing de l'épreuve de vocabulaire. Il s'agit donc bien d'une « bonne question ».

C. Analyse des propositions des questions dont les performances globales en cohérence interne et en qualité spectrale sont faibles

Parmi les 173 questions des 10 épreuves MOHICAN nous avons mis en évidence huit questions dont les performances globales en cohérence interne et en qualité spectrale analysées au niveau « QCM » se sont révélées faibles (voir tableau récapitulatif p. 356). Quelles sont les propositions au sein de ces QCM qui expliquent leur faible niveau de performance ? Comment expliquer les faibles performances ?

1. Analyse des indices des propositions des deux questions dont les performances globales sont les plus faibles

Dans le tableau de répartition des 173 questions en fonction de leur qualité spectrale et de leur cohérence interne (voir détails p. 352), nous avons mis en évidence huit questions aux performances globales faibles (elles figurent dans les cases grisées). Parmi ces huit questions, deux obtiennent les valeurs les plus faibles, il s'agit de [H]q3 et [H]q20 dont nous avons signalé la case grisée en bas à droite dans le tableau ci-contre.

Répartition des 173 QCM MOHICAN en fonction de la qualité spectrale et de la cohérence interne des questions

Cohérence interne \ Qualité spectrale	++	+≃	≃≃	-≃	--	+-
++	39	19	-	-	-	
+≃	20	19	1	-	-	
≃≃	21	26	2	2	2	1
-≃	2	8	3	2	1	
--	-	2	1	-	2	

[H]q3 & [H]q20

Ces deux questions récoltent donc des valeurs « - » aux deux indices de cohérence interne ($r_{qt\ mb}$ et $r_{qt\ ms}$) ainsi qu'aux deux indices de qualité spectrale (*NCSq* et *Rq*) (voir tableau p. 350).

a) [H]q3

(1) Indices spectraux après turbo analyse

Voici un extrait du protocole *SCANTEST 2.0* d'analyse des propositions de la 3ème QCM du test de connaissances en Histoire et Socio Economie, [H]q3 (voir protocole en annexe, p. 524). Cette question est globalement la moins performante (avec la [H]q20) des 173 QCM des différentes épreuves MOHICAN (voir p. 344). Nous avons sélectionné la partie qui concerne les informations spectrales après turbo analyse.

Nous remarquons qu'au palier turbo T10 la proposition correcte P6 est choisie par 128 étudiants (9%) avec un pourcentage de certitude moyen peu élevé (24%). Le pourcentage de sujets est quasi identique jusque T80. A T90, il passe de 8% à 5%.

En ce qui concerne les pourcentages des effectifs des

1. Palier de Turbo analyse : T10

	OM	P1	P2	P3	P4	P5	**P6**	P7
1.1 N Rép. T10	86	501	132	196	104	258	**128**	4
1.2 % Rép. T10	6%	36%	9%	14%	7%	18%	**9%**	0%
1.3 C. Moy. T10	4%	60%	30%	31%	30%	42%	**24%**	10%
1.4 rpbis SC T10	-0,14	0,49	0,05	0,07	0,04	0,19	**-0,15**	-0,02

2. Palier de Turbo analyse : T20

	OM	P1	P2	P3	P4	P5	**P6**	P7
2.1 N Rép. T20	86	501	132	196	104	258	**128**	4
2.2 % Rép. T20	6%	36%	9%	14%	7%	18%	**9%**	0%
2.3 C. Moy. T20	4%	60%	30%	31%	30%	42%	**24%**	10%
2.4 rpbis SC T20	-0,14	0,49	0,05	0,07	0,04	0,19	**-0,15**	-0,02

3. Palier de Turbo analyse : T30

	OM	P1	P2	P3	P4	P5	**P6**	P7
3.1 N Rép. T30	86	499	132	196	104	257	**128**	4
3.2 % Rép. T30	6%	35%	9%	14%	7%	18%	**9%**	0%
3.3 C. Moy. T30	4%	60%	30%	31%	30%	42%	**24%**	10%
3.4 rpbis SC T30	-0,14	0,49	0,05	0,07	0,04	0,19	**-0,15**	-0,02

4. Palier de Turbo analyse : T40

	OM	P1	P2	P3	P4	P5	**P6**	P7
4.1 N Rép. T40	85	498	132	196	103	257	**128**	4
4.2 % Rép. T40	6%	35%	9%	14%	7%	18%	**9%**	0%
4.3 C. Moy. T40	3%	60%	30%	31%	30%	42%	**24%**	10%
4.4 rpbis SC T40	-0,14	0,49	0,05	0,07	0,04	0,19	**-0,15**	-0,02

5. Palier de Turbo analyse : T50

	OM	P1	P2	P3	P4	P5	**P6**	P7
5.1 N Rép. T50	85	495	130	195	103	254	**128**	4
5.2 % Rép. T50	6%	36%	9%	14%	7%	18%	**9%**	0%
5.3 C. Moy. T50	3%	60%	30%	31%	30%	42%	**24%**	10%
5.4 rpbis SC T50	-0,15	0,49	0,05	0,07	0,04	0,19	**-0,15**	-0,02

propositions incorrectes, P1 à P5, ils sont pratiquement identiques de T10 à T90 avec de légères fluctuations (moins de 5%).

6. Palier de Turbo analyse : T60

	OM	P1	P2	P3	P4	P5	**P6**	P7
6.1 N Rép. T60	84	483	124	186	100	247	**124**	4
6.2 % Rép. T60	6%	36%	9%	14%	7%	18%	**9%**	0%
6.3 C. Moy. T60	3%	61%	31%	31%	28%	42%	**23%**	10%
6.4 rpbis SC T60	-0,14	0,52	0,06	0,07	0,04	0,20	**-0,16**	-0,02

7. Palier de Turbo analyse : T70

	OM	P1	P2	P3	P4	P5	**P6**	P7
7.1 N Rép. T70	79	416	114	164	88	209	**103**	2
7.2 % Rép. T70	7%	35%	10%	14%	7%	18%	**9%**	0%
7.3 C. Moy. T70	2%	61%	30%	31%	28%	40%	**26%**	0%
7.4 rpbis SC T70	-0,17	0,49	0,04	0,05	0,01	0,16	**-0,13**	xxxx

8. Palier de Turbo analyse : T80

	OM	P1	P2	P3	P4	P5	**P6**	P7
8.1 N Rép. T80	55	218	67	97	52	111	**54**	2
8.2 % Rép. T80	8%	33%	10%	15%	8%	17%	**8%**	0%
8.3 C. Moy. T80	0%	58%	31%	34%	30%	36%	**33%**	0%
8.4 rpbis SC T80	-0,27	0,38	0,00	0,02	-0,03	0,06	**-0,06**	xxxx

9. Palier de Turbo analyse : T90

	OM	P1	P2	P3	P4	P5	**P6**	P7
9.1 N Rép. T90	25	36	12	13	14	17	**6**	0
9.2 % Rép. T90	20%	29%	10%	11%	11%	14%	**5%**	0%
9.3 C. Moy. T90	1%	52%	29%	40%	42%	28%	**53%**	xxxx
9.4 rpbis SC T90	-0,70	-0,03	-0,18	-0,17	-0,14	-0,27	**0,15**	xxxx

Les pourcentages de certitude moyens changent très peu de T10 à T70. A partir de T80 on observe pour la réponse correcte P6 une augmentation de 7%, puis à T90 de 20% (on passe de 33% à 53%). Pour la solution incorrecte P1 le pourcentage moyen de certitude diminue de T80 à T90 (on passe de 58% à 52%). P2 diminue légèrement de 31% à 29%, P3 augmente de 34% à 40%, P4 augmente aussi en passant de 30% à 42%, P5 diminue de 36% à 28%.

Pour ce qui est des *rpbis SCT*, nous remarquons que la situation est quasi identique de T10 à T50 : on observe un *rpbis SCT* positif pour la 1ère proposition incorrecte P1, des *rpbis SCT* très proches de zéro pour les autres propositions incorrectes P2 à P5 et des *rpbis SCT* négatifs pour l'omission et la réponse correcte. Cette configuration des *rpbis SCT* est révélatrice d'une utilisation incohérente des pourcentages de certitude à ces paliers de turbo analyse, à la fois pour les propositions incorrectes et pour la proposition correcte. Nous avions déjà visualisé cette situation précédemment à l'aide du graphique montrant une partie du brin spectral de la question [H]q3 dans la zone A (voir graphique p. 357).

Au palier turbo T60 la situation évolue un peu, le *rpbis SCT60* de P1 augmente légèrement et passe de 0,49 à 0,52, le *rpbis SCT60* de P6 passe de –0,16 à –0,15.

A T70 le *rpbis SCT* de P1 diminue et passe de 0,52 à 0,49, tandis que le *rpbis SCT* de P6 continue à augmenter et passe de –0,16 à –0,13.

L'augmentation du *rpbis SCT* de la solution correcte P6 continue à T80 (-0,06) et à T90 où le *rpbis SCT* devient positif et passe à 0,15. En ce qui concerne P1, la diminution s'accentue : *rpbis SCT80* = 0,38 et *rpbis SCT90* = -0,03.

Cependant au niveau de turbo analyse T90 les effectifs et les *rpbis SCT90* sont peu élevés, en particulier pour les propositions incorrectes P2, P3, P4, P5 et P6. Le tableau ci-dessous synthétise les informations liées à l'application de la formule proposée par Fisher (voir p. 164) qui permet la transformation du *r* en *t* de Student en vue d'un test de signification « $r \neq 0$ » pour chacune des propositions au palier T90.

Palier T90	P1	P2	P3	P4	P5	P6 (RC)
ns	36 + 6 = 42	12 + 6 = 18	13 + 6 = 19	14 + 6 = 20	17 + 6 = 23	6 + 36 + 12 + 13 + 14 + 17 = 98
rpbis SCT90	-0,03	-0,18	-0,17	-0,14	-0,27	0,15
t_c	0,190	0,732	0,711	0,600	1,285	1,487
ddl *(ns* – 2)	40	16	17	18	21	96
	$t_c < t_t$ à 0,10	$t_c < t_t$ à 0,10	$t_c < t_t$ à 0,10	$t_c < t_t$ à 0,10	$t_c < t_t$ à 0,10	$t_c < t_t$ à 0,10

Comme le montre le tableau ci-dessous, les corrélations obtenues à T90 par toutes les propositions ne sont pas significativement différentes de zéro.

Lorsque nous consultons les *rpbis SCT80* (donc au palier turbo inférieur à T90) de P2, P3, P4 et P6 nous remarquons d'une part que les effectifs sont importants et que d'autre part, les *rpbis SCT* de ces quatre propositions sont quasi égaux à zéro. En ce qui concerne P1, nous remarquons que le *rpbis SCT80* est positif et vaut 0,38 (*ns* = 218).

Etant donné ces faibles effectifs à T90, nous utiliserons les données calculées au palier turbo T80 pour conclure qu'en ce qui concerne les propositions de cette question [H]q3 (globalement peu performante, voir p. 343) il existe un problème de cohérence spectrale à la fois au niveau des propositions incorrectes et de la proposition correcte. En ce qui concerne les propositions incorrectes, d'une part P1 récolte un *rpbis SCT80* positif, donc les étudiants réalistes ($Rs \geqslant 80$) ont tendance à l'accompagner de degrés de certitude qui sont en moyennes plus élevés que les certitudes qui accompagnent la réponse correcte alors qu'on s'attend à l'inverse, et d'autre part, les solutions P2 à P5 récoltent des *rpbis SCT80* qui sont équivalents à zéro alors qu'on s'attend à ce qu'ils soient plus marqués et négatifs. Pour ce qui est de la réponse correcte P3, nous constatons qu'elle pose aussi un problème étant donné la valeur du *rpbis SCT80* très proche de zéro : on observe pas de corrélation entre les choix de certitudes plus élevées et les choix de cette proposition correcte.

(2) indices classiques

Voyons maintenant les données des indices classiques du protocole d'analyse des propositions.

Nous remarquons un *rpbis classique* positif pour la proposition incorrecte P1 tandis que tous les autres *rpbis classiques* sont négatifs ou très proches de zéro.

	OM	P1	P2	P3	P4	P5	**P6**
1) N Rép.	86	501	132	197	104	258	**128**
2) % Rép.	6%	36%	9%	14%	7%	18%	**9%**
3) rpbis	-0,21	0,36	-0,02	-0,10	-0,06	-0,17	**0,01**

Cette question [H]q3 pose un problème au niveau de la cohérence interne lorsqu'on analyse les informations fournies par les *rpbis classiques* de 3 propositions incorrectes. On remarque en effet que la 1ère proposition récolte un *rpbis classique* positif alors qu'on s'attend à ce qu'il soit négatif et que deux autres propositions incorrectes, P2 et P4, discriminent peu les sujets dans la mesure où leurs *rpbis classiques* sont proches de zéro.

La question pose aussi un problème en ce qui concerne la réponse correcte P6 dont le *rpbis classique* est négatif alors qu'on s'attend à ce qu'il soit positif et supérieur à la valeur repère[71].

En ce qui concerne l'attractivité des propositions, nous remarquons que P1 est la plus choisie (36%), P5 est deux fois moins choisie (18%), viennent ensuite : P3 (14%), P2 et P6 (9%) et P4 (7%). Cette question est donc très difficile (réussie par seulement 9% des étudiants) et une de ses propositions incorrectes est particulièrement attractive : P1 (36% des choix).

(3) Discussion des analyses des propositions de [H]q3

Pour cette question [H]q3 nous constatons que les diagnostics sur les propositions à l'aide des *rpbis classiques* et à l'aide des *rpbis SCT80* se rejoignent. A l'aide de ces indices nous serions tenté de dire que les propositions P1 et P6 posent un problème. Nous pourrions même suspecter une erreur d'encodage de la réponse correcte étant donné les valeurs positives obtenues par P1 et les valeurs négatives liées à P6.

Comment expliquer que de tels problèmes d'incohérence spectrale et de cohérence interne au niveau des propositions au sein de cette QCM ?

[71] Valeur repère calculée pour contrer le problème du recouvrement entre le score de la question et le score total au test, voir p. 6.

A ce stade, étant donné les problèmes observés, nous pensons qu'une analyse du contenu de la question et de ses propositions est nécessaire. Voici le libellé de la question [H]q3 extraite du questionnaire de l'épreuve « Connaissance en Histoire et Socio Economie » (voir annexe, p. 494).

Q3. Quel est le Secrétaire général actuel de l'OTAN ?

1. Javier Solana
2. Boutros Boutros Ghali
3. Jacques Santer
4. Antonio Samaranch
5. Kofi Anan
6. Aucune

Nous constatons qu'il s'agit d'une question factuelle de connaissance de l'actualité socio-politique. L'amorce de cette QCM est très claire et ne peut pas porter à confusion. Les solutions proposées sont elles aussi très claires. Il s'agit de noms de personnalités médiatisées.

Dès lors, pourquoi plus d'un étudiant sur trois a-t-il choisi la proposition incorrecte « 1. Javier Solana » plutôt que la « 6. Aucune » (en octobre 1999 c'est George Robertson qui était Secrétaire général de l'OTAN) ?

Pour comprendre ce qui s'est passé nous pensons qu'il faut remonter le fil du temps et nous placer dans le contexte des étudiants soumis à l'épreuve de « Connaissances en Histoire et Socio Economie » administrée entre le 4 et le 8 octobre 1999.

Lors de l'année scolaire 1998-1999 qui a précédé, ces étudiants ont très probablement été amenés à traiter des questions d'actualité qui fut malheureusement riche en évènements en ce qui concerne le conflit yougoslave où sont impliquées les troupes de l'OTAN. A l'époque, le Secrétaire général de l'OTAN est Javier Solana . Tout au long de cette année scolaire 1998-1999 son nom sera régulièrement prononcé et son rôle de patron de l'OTAN régulièrement évoqué lors de reportages télévisés ou radiodiffusés. Probablement en parlera-t-on aussi dans le contexte scolaire. Or, Javier Solana occupera son poste de Secrétaire général jusqu'au mercredi 4 août 1999, date où George Robertson lui succèdera. Cette passation de pouvoir a lieu au mois d'août, période de vacances où les questions d'actualité ne sont plus traitées par les étudiants dans le contexte scolaire. Dès lors, il est très probable qu'en octobre 1999 la mémoire des étudiants soit encore très imprégnée du nom de Javier Solana, Secrétaire général de l'OTAN lorsqu'ils étaient rhétoriciens.

On comprend dès lors la difficulté de cette question qui confine au piège dans la mesure où posée trois mois auparavant la réponse correcte eut été celle qui fut considérée comme incorrecte (et à raison) en octobre 1999.

Signalons ici que les questions des tests MOHICAN ont été discutées lors de deux réunions de débriefing des épreuves qui ont eu lieu au siège du CIUF les 3 et 7 février 2000 en présence des auteurs. Nous étions présent lors de ces deux réunions et nous avons eu l'occasion d'entendre les commentaires des auteurs du test « Connaissances en Histoire et Socio Economie » (D. Leclercq et F. Georges) qui contient la question [H]q3. Dans les notes que nous avions prises nous avons relevé les deux commentaires suivants.

D'abord, de façon plus générale en ce qui concerne le test, les auteurs ont fait remarquer que beaucoup de questions portaient sur l'actualité ou l'histoire proche et que les connaissances relatives aux organisations telles que l'ONU, l'OTAN, l'UNESCO, … faisaient partie de la matière enseignée dans le cadre du cours d'Histoire de dernière année du secondaire.

Ensuite, à propos de la question 3, mais aussi de la question 2 fort similaire (voir questionnaire en annexe, p. 494), les auteurs ont signalé que ces deux QCM non contextualisées d'un point de vue temporel ne pouvaient « durer dans le temps ». Dès lors, le type de formulation utilisée devait être revue si ce genre de question était posée lors d'un prochain *check up*.

Après analyse de la qualité des questions du test doit-on garder cette QCM ? Comptabiliser ses scores dans les résultats ? Supprimer la question serait injuste vis-à-vis de ceux qui, méritants, se sont intéressés de près à l'actualité même après juin 1999. Garder la question serait cependant aussi quelque peu injuste vis à vis de ceux qui ont répondu Javier Solana, une erreur certes, mais qui après analyse n'est pas à mettre sur le même pied que les autres propositions incorrectes.

Une solution, dans le cas de [H]q3 aurait pu consister à valoriser la proposition « 1. Javier Solana », tout en continuant évidemment à considérer la solution « 6. Aucune » comme étant LA réponse correcte et les autres propositions 2 à 5 comme étant incorrectes.

b) [H]q20

(1) Indices spectraux après turbo analyse

Voici maintenant l'extrait du protocole *SCANTEST 2.0* d'analyse des propositions de la 20ème QCM du test de connaissances en Histoire et Socio Economie, [H]q20 (voir protocole en annexe, p. 525). Cette question est globalement (avec la [H]q3) la moins performante des 173 QCM des différentes épreuves MOHICAN (voir p. 344). Nous avons sélectionné la partie qui concerne les informations spectrales après turbo analyse.

En ce qui concerne les effectifs (N Rép. T… et %Rép. T…) nous remarquons que la proposition incorrecte P6 est la plus attractive (53% des choix de T10 à T60, puis diminution : 52% à T70, 49% à T80 et 43% à T90). Ensuite vient une autre proposition incorrecte, la proposition P1 (19% de T10 à T70 avec une légère diminution à T60 : 18%, 21% à T80 et 24% à T90). P3, la réponse correcte, est un peu moins attractive (18% de T10 à T70, 17% à T80 et 13% à T90). Les autres propositions P2, P4 et P5 ne sont très peu attractives, les pourcentages de choix sont très faibles à tous les paliers de turbo analyse (de 0% à 3%). En ce qui concerne la proposition P5, remarquons que dès T10 les effectifs en nombres sont très faibles (N Rép. T10 = 7).

Lorsque nous comparons les certitudes moyennes (C. Moy. T…), nous remarquons que pour P1 les valeurs sont relativement stables et peu élevées autour de 30%. Pour P3 (réponse correcte) elles sont plus élevées de T10 à T70 avec environ 47%, la certitude moyenne augmente ensuite à T80 (53%) puis diminue à T90 pour revenir à 47%. Pour P6, les étudiants semblent particulièrement convaincu qu'il s'agit du bon choix, de T10 à T80 la certitude moyenne se situe à 82% ou à 83%, à T90 elle diminue à 68%. En ce qui concerne les autres propositions peu attractives, la certitude moyenne de P2 diminue

1. Palier de Turbo analyse : T10

	OM	P1	P2	**P3**	P4	P5	P6
1.1 N Rép. T10	58	263	39	**256**	20	7	743
1.2 % Rép. T10	4%	19%	3%	**18%**	1%	0%	53%
1.3 C. Moy. T10	7%	30%	24%	**46%**	38%	37%	82%
1.4 rpbis SC T10	-0,20	-0,16	-0,10	**-0,17**	-0,03	-0,02	0,46

2. Palier de Turbo analyse : T20

	OM	P1	P2	**P3**	P4	P5	P6
2.1 N Rép. T20	58	263	39	**256**	20	7	743
2.2 % Rép. T20	4%	19%	3%	**18%**	1%	0%	53%
2.3 C. Moy. T20	7%	30%	24%	**46%**	38%	37%	82%
2.4 rpbis SC T20	-0,20	-0,16	-0,10	**-0,17**	-0,03	-0,02	0,46

3. Palier de Turbo analyse : T30

	OM	P1	P2	**P3**	P4	P5	P6
3.1 N Rép. T30	58	262	39	**256**	20	6	742
3.2 % Rép. T30	4%	19%	3%	**18%**	1%	0%	53%
3.3 C. Moy. T30	7%	30%	24%	**46%**	38%	27%	83%
3.4 rpbis SC T30	-0,20	-0,16	-0,10	**-0,17**	-0,03	-0,03	0,46

4. Palier de Turbo analyse : T40

	OM	P1	P2	**P3**	P4	P5	P6
4.1 N Rép. T40	58	261	39	**256**	20	6	740
4.2 % Rép. T40	4%	19%	3%	**18%**	1%	0%	53%
4.3 C. Moy. T40	7%	30%	24%	**47%**	38%	27%	83%
4.4 rpbis SC T40	-0,20	-0,17	-0,10	**-0,17**	-0,03	-0,03	0,46

5. Palier de Turbo analyse : T50

	OM	P1	P2	**P3**	P4	P5	P6
5.1 N Rép. T50	58	261	38	**254**	20	6	734
5.2 % Rép. T50	4%	19%	3%	**18%**	1%	0%	53%
5.3 C. Moy. T50	7%	30%	22%	**47%**	38%	27%	83%
5.4 rpbis SC T50	-0,20	-0,17	-0,11	**-0,17**	-0,03	-0,03	0,46

6. Palier de Turbo analyse : T60

	OM	P1	P2	**P3**	P4	P5	P6
6.1 N Rép. T60	58	250	36	**247**	19	6	715
6.2 % Rép. T60	4%	18%	3%	**18%**	1%	0%	53%
6.3 C. Moy. T60	7%	29%	20%	**47%**	39%	27%	83%
6.4 rpbis SC T60	-0,21	-0,17	-0,11	**-0,17**	-0,02	-0,03	0,47

7. Palier de Turbo analyse : T70

	OM	P1	P2	**P3**	P4	P5	P6
7.1 N Rép. T70	54	226	34	**209**	18	4	611
7.2 % Rép. T70	5%	19%	3%	**18%**	2%	0%	52%
7.3 C. Moy. T70	4%	30%	20%	**48%**	44%	20%	83%
7.4 rpbis SC T70	-0,24	-0,18	-0,12	**-0,15**	-0,02	-0,04	0,46

8. Palier de Turbo analyse : T80

	OM	P1	P2	**P3**	P4	P5	P6
8.1 N Rép. T80	38	137	18	**113**	16	3	320
8.2 % Rép. T80	6%	21%	3%	**17%**	2%	0%	49%
8.3 C. Moy. T80	0%	31%	13%	**53%**	40%	20%	82%
8.4 rpbis SC T80	xxxx	-0,24	-0,17	**-0,07**	-0,05	-0,06	0,40

9. Palier de Turbo analyse : T90

	OM	P1	P2	**P3**	P4	P5	P6
9.1 N Rép. T90	14	30	3	**16**	5	1	53
9.2 % Rép. T90	11%	24%	2%	**13%**	4%	1%	43%
9.3 C. Moy. T90	0%	29%	7%	**47%**	20%	60%	68%
9.4 rpbis SC T90	xxxx	-0,22	-0,18	**0,01**	-0,13	0,03	0,35

de 24% à 7% de T10 à T90. Pour P4 elle augmente de 38% à 44% de T10 à T70 et diminue ensuite : à T80 elle vaut 40% et à T90 20%. P5 diminue de 37% à 20% de T10 à T80 et s'élève à 60% à T90, mais à ce palier turbo il s'agit de la certitude (moyenne) d'un seul étudiant.

Pour ce qui est des valeurs observées aux *rpbis SCT* nous remarquons qu'aux différents paliers de turbo analyse toutes les valeurs sont négatives ou très proches de zéro sauf pour la proposition incorrecte P6 où elles sont positives et se situent aux alentours de 0,46 de T10 à T70 puis diminuent à T80 (0,40) et à T90 (0,35). Le *rpbis SCT* de la proposition correcte est négatif de T10 à T80 (il passe de –0,17 à –0,15 à T80) et très proche de zéro à T90 (0,01).

Nous voyons dans le tableau qui précède qu'à T90 les effectifs sont très faibles, dès lors, nous préférons prendre en compte les *rpbis SCT80* pour les propositions dont le *rpbis SCT90* est proche de zéro. Nous observons à T80 un faible *rpbis SCT80* négatif pour la proposition P2 (-0,17) et nous nous demandons si nous pouvons l'interpréter comme étant différent de zéro. L'application du test $r \neq 0$ à l'aide de la formule de Fisher de transformation du *r* en *t* de Student (voir p. 165) donne (*ns* = 113 + 18 = 131) :

$$t_c = \frac{|r|\sqrt{ns-2}}{\sqrt{1-r^2}} = \frac{0{,}17\sqrt{129}}{\sqrt{1-0{,}0289}} = \frac{1{,}931}{0{,}985} = 1{,}96$$

Lorsque nous consultons les valeurs théoriques de la table du *t* de Student (voir annexe, p. 543), nous constatons que pour un nombre de degrés de liberté qui tend vers ∞, le t_c (1,96) est égal au t_t (1,960) à 0,05. Il y a donc 95% de chances que la valeur du *rpbis SCT80* de P2 (-0,17) soit différente de zéro.

Comme le montre le graphique (p. 357, les points de [H]q20 se trouvent dans la zone A ou en bordure de celle-ci à T90) du point de vue de la cohérence spectrale (au sens où nous l'avons définie p. 227), le problème de cette question se situe à la fois au niveau de la réponse correcte (P3) et de la plupart des propositions incorrectes.

En effet à T80, il n'existe pas de corrélation négative entre le choix des pourcentages de certitude et le choix des propositions incorrectes P4 et P5 contrairement aux propositions P1 (-0,24) et P2 (-0,17). En ce qui concerne la dernière proposition incorrecte, P6, le *rpbis SCT80* est, à l'inverse de nos attentes, positif et relativement élevé (0,40). Les étudiants, pourtant réalistes ($Rs \geqslant 80$) à ce palier turbo, ont eu tendance à accompagné cette proposition P6 incorrecte de certitudes en moyenne plus élevées que les certitudes qui ont été fournies par les étudiants qui ont choisi la réponse correcte.

Nous remarquons aussi qu'au palier de turbo analyse T80 il n'existe pas de corrélation positive marquée entre le choix des pourcentages de certitude et le choix de la proposition correcte P3 alors qu'une question qui fonctionne bien du point de vue de la cohérence spectrale devrait présenter à ce palier turbo une corrélation positive élevée. En effet, les sujets réalistes qui choisissent la réponse correcte devraient normalement accompagner celle-ci de certitudes en moyenne plus élevées que les certitudes qui ont accompagné les réponses incorrectes.

(2) indices classiques

Voyons maintenant les données des indices classiques du protocole d'analyse des propositions de [H]q20.

	OM	P1	P2	**P3**	P4	P5	P6
1) N Rép.	58	263	39	**256**	20	7	744
2) % Rép.	4%	19%	3%	**18%**	1%	0%	53%
3) rpbis	-0,07	-0,18	-0,06	**0,10**	0,00	-0,09	0,14

Pour trois propositions incorrectes nous remarquons l'absence de corrélation entre les choix incorrects et les scores totaux du test. Les *rpbis classiques* sont égaux, ou quasi égaux, à zéro pour P2 (-0,06), P4 (0,00) et P5 (-0,09) alors que nous nous attendons à des corrélations négatives. Nous devrions en effet observer chez les étudiants qui obtiennent les scores les moins élevés au total de l'épreuve une tendance à choisir les propositions incorrectes. Si c'était le cas, ces propositions P2, P4 et P5

permettraient alors de discriminer les sujets qui obtiennent en moyenne un nombre de réponses correctes plus élevé des autres qui obtiennent en moyenne un nombre de réponses correctes moins élevé.

C'est d'ailleurs ce que nous constatons pour la proposition P1 qui fonctionne correctement étant donné son *rpbis classique* négatif (-0,18).

Par contre, la proposition P6 fonctionne à l'inverse dans la mesure où elle obtient un *rpbis classique* positif (0,14) alors qu'elle est incorrecte.

Enfin, en ce qui concerne la proposition correcte P3, nous constatons que la corrélation n'est que très légèrement positive (0,10).

Le *rpbis classique* de P3 est aussi inférieur à la valeur repère calculée en vue de contrer le problème du recouvrement entre le score de la question et le score total du test (voir p. 176). Cette valeur repère vaut dans le cas de cette épreuve « Connaissances en Histoire et Socio Economie » qui compte 25 questions :

$$repère\ \ rpbis\ \ classique\ \ = \ 1/\sqrt{nq} = \ 1/\sqrt{25} = 0{,}2$$

(3) Discussion des analyses des propositions de [H]q20

Nous constatons que pour cette seconde question globalement peu performante (voir p. 343) les analyses des propositions établies d'une part à l'aide des *rpbis SCT* et d'autre part à l'aide des *rpbis classiques* convergent[72]. Pour les deux types d'indices de discrimination, les propositions P3 (réponse correcte) et P6 fonctionnent de façon incohérente. Pour les deux types d'indices nous remarquons aussi l'absence de corrélation pour les propositions P4 et P5. En ce qui concerne la proposition P1, les deux types d'indices convergent aussi en montrant un fonctionnement cohérent de cette proposition. Par contre, pour la proposition P2, nous observons un fonctionnement correct du point de vue de la discrimination spectrale (*rpbis SCT80* = -0,17) et incorrect en ce qui concerne la discrimination classique (*rpbis classique* = -0,06).

Comment expliquer ces disfonctionnements dans les propositions ? Voici la question (voir questionnaire en annexe, p. 496). Lors de la réunion de débriefing de l'épreuve en février 2000, nous avons relevé les deux commentaires suivants émanant des deux auteurs (D. Leclercq et F. Georges) et des personnes présentes.

Q20. Quelle chaîne de TV crée et diffuse la séquence « No Comment » ?

1. La BBC
2. FR3
3. CNN
4. RTL
5. La RAI
6. Aucune

D'abord, cette QCM n'est pas une question d'histoire mais plutôt une question d'actualité socioculturelle et probablement n'a-t-elle pas sa place dans cette épreuve.

Ensuite, il semble que la séquence « No Comment » soit aussi diffusée par la chaîne EuroNews, ce qui expliquerait que les étudiants aient répondu en majorité (53%) : « 6. Aucune ».

Nous avons depuis vérifié cette dernière suggestion, et c'était en effet le cas : l'émission était diffusée sur « EuroNews Fr ». La question présente donc une erreur d'encodage de la réponse correcte qui n'est pas P3 et ne figure donc pas parmi les différentes propositions. Dès lors, il est logique que 53% des étudiants aient choisi P6 « Aucune ».

[72] En ce qui concerne le *rpbis classique* de la réponse correcte P3, bien que la valeur récoltée soit positive (0,1), celle-ci se situe sous le seuil repère qui vaut 0,2.

2. Analyse de la question [V]q5 dont les indices de cohérence spectrale sont faibles et qui récolte des indices mitigés en cohérence interne

Cette cinquième question du test de vocabulaire fait partie des 8 QCM épinglées pour leurs performances faibles en cohérence spectrale ou/et en cohérence interne (voir tableau des performances, p. 356).

Les indices de cohérence interne sont mitigés dans la mesure où la question récolte un « - » à l'indice $r_{qt\ mb}$ et un « ≃ » à l'indice $r_{qt\ ms}$.

Les différentes propositions de [V]q5 présentent des rpbis spectraux qui, jusqu'au palier de turbo analyse T70, sont très proches de zéro. A ces paliers turbo inférieurs à T70, les propositions discriminent donc peu en ce qui concerne l'utilisation des pourcentages de certitude. Les étudiants qui fournissent la réponse correcte (P6) n'ont pas tendance à utiliser des degrés de certitudes en moyenne plus élevés que les degrés de certitude choisit par ceux qui se trompent.

A partir du palier de turbo analyse T80, ceux qui choisissent les propositions incorrectes P1, P3 et P4 ont une légère tendance à utiliser des pourcentages de certitude plus faibles (*rpbis SC T80* de P1 = -0,15, de P3 = -0,11 et de P4 = -0,12)[73]. Par contre la réponse correcte (P6) récolte un *rpbis SC T80* très proche de 0 (0,08).

Lorsque les *rpbis SC* sont calculés à partir des données des étudiants très réalistes (dont $Rs \geqslant 90$) au palier de turbo T90, on observe une tendance plus marquée à

1. Palier de Turbo analyse : T10

	OM	P1	P2	P3	P4	P5	**P6**	P7
1.1 N Rép. T10	135	761	1999	78	101	229	**455**	88
1.2 % Rép. T10	4%	20%	52%	2%	3%	6%	**12%**	2%
1.3 C. Moy. T10	8%	44%	51%	34%	37%	51%	**46%**	50%
1.4 rpbis SC T10	-0,23	-0,03	0,09	-0,06	-0,05	0,04	**-0,01**	0,02

2. Palier de Turbo analyse : T20

	OM	P1	P2	P3	P4	P5	**P6**	P7
2.1 N Rép. T20	135	761	1999	77	101	229	**455**	88
2.2 % Rép. T20	4%	20%	52%	2%	3%	6%	**12%**	2%
2.3 C. Moy. T20	8%	44%	51%	34%	37%	51%	**46%**	50%
2.4 rpbis SC T20	-0,23	-0,03	0,09	-0,06	-0,05	0,04	**-0,01**	0,02

3. Palier de Turbo analyse : T30

	OM	P1	P2	P3	P4	P5	**P6**	P7
3.1 N Rép. T30	135	760	1999	76	101	229	**454**	88
3.2 % Rép. T30	4%	20%	52%	2%	3%	6%	**12%**	2%
3.3 C. Moy. T30	8%	44%	51%	33%	37%	51%	**46%**	50%
3.4 rpbis SC T30	-0,23	-0,03	0,08	-0,06	-0,05	0,04	**-0,01**	0,02

4. Palier de Turbo analyse : T40

	OM	P1	P2	P3	P4	P5	**P6**	P7
4.1 N Rép. T40	135	753	1995	75	101	229	**453**	88
4.2 % Rép. T40	4%	20%	52%	2%	3%	6%	**12%**	2%
4.3 C. Moy. T40	8%	43%	51%	32%	37%	52%	**46%**	50%
4.4 rpbis SC T40	-0,24	-0,04	0,08	-0,07	-0,05	0,04	**-0,01**	0,02

5. Palier de Turbo analyse : T50

	OM	P1	P2	P3	P4	P5	**P6**	P7
5.1 N Rép. T50	134	745	1983	75	98	225	**449**	88
5.2 % Rép. T50	4%	20%	52%	2%	3%	6%	**12%**	2%
5.3 C. Moy. T50	8%	43%	51%	32%	38%	52%	**47%**	50%
5.4 rpbis SC T50	-0,24	-0,05	0,08	-0,07	-0,05	0,04	**0,00**	0,02

6. Palier de Turbo analyse : T60

	OM	P1	P2	P3	P4	P5	**P6**	P7
6.1 N Rép. T60	131	721	1921	72	93	215	**439**	87
6.2 % Rép. T60	4%	20%	52%	2%	3%	6%	**12%**	2%
6.3 C. Moy. T60	8%	43%	51%	31%	36%	52%	**47%**	50%
6.4 rpbis SC T60	-0,25	-0,05	0,08	-0,07	-0,06	0,05	**0,00**	0,02

7. Palier de Turbo analyse : T70

	OM	P1	P2	P3	P4	P5	**P6**	P7
7.1 N Rép. T70	119	627	1676	63	73	200	**387**	78
7.2 % Rép. T70	4%	19%	52%	2%	2%	6%	**12%**	2%
7.3 C. Moy. T70	6%	42%	50%	30%	32%	51%	**49%**	49%
7.4 rpbis SC T70	-0,29	-0,09	0,03	-0,09	-0,09	0,03	**0,03**	0,00

8. Palier de Turbo analyse : T80

	OM	P1	P2	P3	P4	P5	**P6**	P7
8.1 N Rép. T80	92	403	1032	37	40	137	**262**	53
8.2 % Rép. T80	4%	20%	50%	2%	2%	7%	**13%**	3%
8.3 C. Moy. T80	4%	42%	50%	30%	30%	53%	**52%**	48%
8.4 rpbis SC T80	-0,37	-0,15	-0,04	-0,11	-0,12	0,01	**0,08**	-0,02

9. Palier de Turbo analyse : T90

	OM	P1	P2	P3	P4	P5	**P6**	P7
9.1 N Rép. T90	32	84	225	10	8	33	**58**	7
9.2 % Rép. T90	7%	18%	49%	2%	2%	7%	**13%**	2%
9.3 C. Moy. T90	0%	41%	49%	43%	20%	50%	**53%**	54%
9.4 rpbis SC T90	xxxx	-0,18	-0,06	-0,09	-0,15	-0,01	**0,13**	0,01

[73] Ces *rpbis SCT80* sont significativement différents de zéro.

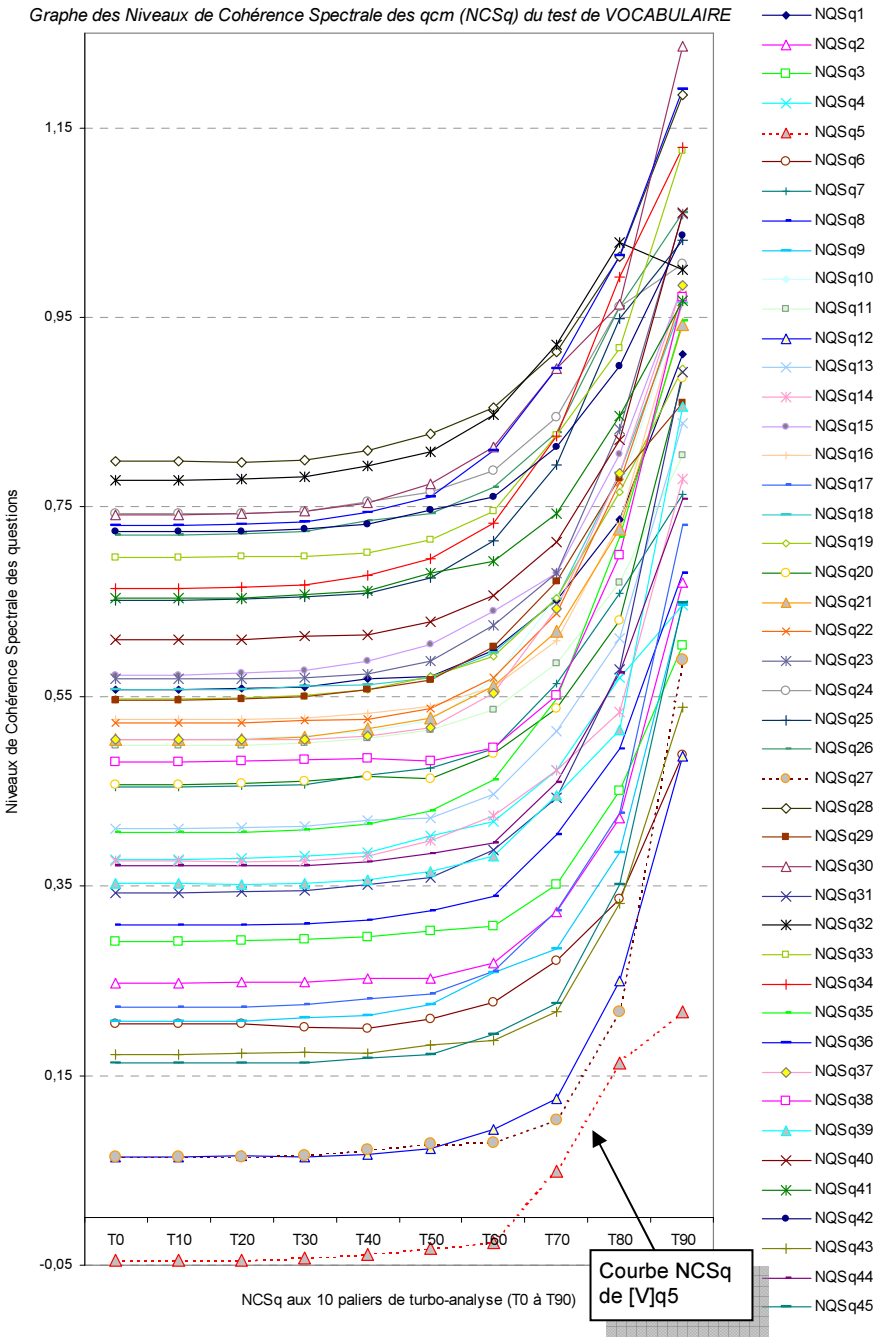
utiliser des pourcentages de certitude peu élevés pour les propositions incorrectes P1 et P4. Par contre, pour P3 le *rpbis SCT90* se rapproche de zéro (la différence n'est plus significativement différente). En ce qui concerne la réponse correcte, le *rpbis SCT90* est légèrement positif (0,13) et significativement différent de zéro (à p=0,01).

Du point de vue de la cohérence d'utilisation des pourcentages de certitude, la proposition correcte P6 et les propositions incorrectes P1 et P4 sont donc peu cohérentes. Pour les autres propositions incorrectes on observe pas de corrélation entre les pourcentages de certitude et le fait d'avoir répondu incorrectement.

Lorsqu'on examine les courbes des Niveaux de Cohérence Spectrale des questions (*NCSq*) calculés aux différents paliers de turbo analyse (de T0 à T90 avec un pas de 10) et qu'on compare [V]q5 aux autres QCM, on constate que cette question :

1) récolte les indices *NCSq* les moins élevés du test (courbe en pointillés en bas du graphique) ;

2) présente des *NCSq* inférieurs à zéro jusqu'à T70 où ils deviennent positifs tout en restant peu élevés.



Donc, du point de vue de la cohérence spectrale, les résultats de la cinquième question du test de vocabulaire montrent une utilisation cohérente des certitudes chez les étudiants qui commettent le moins d'erreurs dans leurs auto-estimations, à partir de $Rs \geqslant 70$. Lorsqu'ils atteignent ce niveau de réalisme, ceux qui fournissent la réponse correcte accompagnent leur choix de certitudes en moyenne plus élevées que les certitudes de ceux qui fournissent une réponse incorrecte.

Ajoutons qu'à partir de T70, cette utilisation des pourcentages de certitude cohérente, n'atteint cependant pas les niveaux de cohérence spectrale observés pour les autres questions du test de vocabulaire.

	OM	P1	P2	P3	P4	P5	**P6**	P7
1) N Rép.	135	761	1999	78	101	229	**455**	88
2) % Rép.	4%	20%	52%	2%	3%	6%	**12%**	2%
3) rpbis	-0,15	-0,07	0,01	-0,07	-0,10	0,12	**0,15**	0,02

En ce qui concerne la cohérence interne, on remarque que les *rpbis classiques* sont peu marqués. Pour la réponse correcte (P6) nous constatons que le *rpbis classique* est égal à la valeur repère (voir p. 176) : $1/\sqrt{45} = 0{,}15$.

En ce qui concerne les *rpbis classiques* des propositions incorrectes P1, P2, P3 et P7, nous remarquons qu'ils sont proches de zéro. Pour ce qui est de P4, le *rpbis classique* est un peu plus négatif (-0,10). Pour P5 il est positif (0,12). Pour la proposition P5 on observe donc une légère tendance à choisir cette solution incorrecte chez les étudiants qui obtiennent en moyenne des scores plus élevés au total du test, légère tendance qui n'est pas logique.

Cette question 5 pose donc des problèmes de cohérence interne plus particulièrement au niveau de la proposition incorrecte P5. En ce qui concerne la cohérence spectrale, elle pose aussi des problèmes car les rpbis SCT sont faibles et la QCM n'atteint pas les niveaux de cohérence spectrale observés pour les autres questions du test.

Une analyse plus qualitative du contenu de cette question permet-elle d'expliquer ces problèmes ?

Cette question a été discutée avec les auteurs du test lors de la réunion de débriefing de l'épreuve en février 2000, mais aucune explication n'a pu être trouvée qui pourrait expliquer les rpbis insatisfaisants.

En ce qui concerne le choix de la proposition incorrecte P2 « anciennes » par 53% des étudiants, il semble que ces derniers fassent un amalgame entre l'idée d'une haine qui est latente, qui ne se manifeste pas mais qui peut à tout moment le faire, et une haine qui serait ancienne.

Pour ce qui est de la proposition P5 « patentes » il faut relever la similitude phonologique avec « latentes » qui pourrait expliquer une légère tendance chez les étudiants obtenant les meilleurs scores à choisir cette proposition, d'où le rpbis classique légèrement positif (0,12).

Parmi les cinq mots proposés à la suite de chaque énoncé, choisissez CELUI QUI REPREND LE PLUS EXACTEMENT POSSIBLE le sens que le mot souligné a dans la phrase donnée

…

Q5. Des haines <u>latentes</u> opposaient les deux hommes

1. profondes
2. anciennes
3. farouches
4. douloureuses
5. patentes
6. Aucune
7. Toutes

Suite à l'analyse de cette question [V]q5, globalement, il se dégage une impression qu'elle ne fonctionne pas correctement tout en ne montrant pas un défaut particulièrement évident comme dans le cas des deux questions précédentes [H]q3 et [H]q20.

3. Analyse de deux questions [S]q1 et [V]q12 dont les indices de qualité spectrale sont faibles tandis que la cohérence interne est moyenne

a) [S]q1

Cette 1ère question du test « Syntaxe et articulation logique » fait partie des 8 QCM épinglées pour leurs performances faibles en cohérence spectrale ou/et en cohérence interne (voir tableau des performances, p. 356).

[S]q1 récolte des « - » aux indices spectraux *NCSq T80*, *Rq T80* et *Cq+ T80* et des « ≃ » aux indices de cohérence interne $r_{qt\ mb}$ et $r_{qt\ ms}$.

Lorsque nous analysons les *rpbis SC T80* des propositions nous remarquons pour la solution correcte P6 une valeur légèrement positive (0,14), ce qui témoigne d'une tendance chez les étudiants qui répondent correctement à accompagner leur choix de certitudes en moyenne plus élevées que les certitudes de ceux qui répondent incorrectement.

Pour ce qui est des propositions incorrectes P3, P4, P5 et P7 nous remarquons que les valeurs des *rpbis SC T80* sont négatives mais assez proches de zéro.

L'application du test $r \neq 0$ à l'aide de la formule de Fisher de transformation du *r* en *t* de Student (voir p. 165) montre qu'à *T80* les *rpbis SCT* de P4 et P7 ne sont pas significativement différents de zéro.

5. Palier de Turbo analyse : T50

	OM	P1	P2	P3	P4	P5	**P6**	P7
5.1 N Rép. T50	47	54	31	417	72	1780	**1152**	8
5.2 % Rép. T50	1%	2%	1%	12%	2%	50%	**32%**	0%
5.3 C. Moy. T50	19%	51%	45%	63%	59%	66%	**60%**	68%
5.4 rpbis SC T50	-0,17	-0,04	-0,05	0,03	0,00	0,12	**-0,07**	0,01

6. Palier de Turbo analyse : T60

	OM	P1	P2	P3	P4	P5	**P6**	P7
6.1 N Rép. T60	43	47	25	384	70	1679	**1080**	7
6.2 % Rép. T60	1%	1%	1%	12%	2%	50%	**32%**	0%
6.3 C. Moy. T60	17%	51%	41%	63%	58%	67%	**62%**	63%
6.4 rpbis SC T60	-0,19	-0,05	-0,08	0,01	-0,02	0,09	**-0,04**	0,00

7. Palier de Turbo analyse : T70

	OM	P1	P2	P3	P4	P5	**P6**	P7
7.1 N Rép. T70	36	40	20	320	58	1393	**900**	7
7.2 % Rép. T70	1%	1%	1%	12%	2%	50%	**32%**	0%
7.3 C. Moy. T70	17%	51%	41%	64%	57%	67%	**65%**	57%
7.4 rpbis SC T70	-0,22	-0,07	-0,08	-0,02	-0,05	0,03	**0,02**	0,00

8. Palier de Turbo analyse : T80

	OM	P1	P2	P3	P4	P5	**P6**	P7
8.1 N Rép. T80	24	27	13	191	37	847	**594**	3
8.2 % Rép. T80	1%	2%	1%	11%	2%	49%	**34%**	0%
8.3 C. Moy. T80	13%	47%	40%	64%	60%	66%	**71%**	33%
8.4 rpbis SC T80	-0,28	-0,13	-0,11	-0,09	-0,06	-0,09	**0,14**	-0,06

9. Palier de Turbo analyse : T90

	OM	P1	P2	P3	P4	P5	**P6**	P7
9.1 N Rép. T90	8	8	4	47	17	234	**214**	1
9.2 % Rép. T90	2%	2%	1%	9%	3%	44%	**40%**	0%
9.3 C. Moy. T90	0%	45%	50%	67%	65%	67%	**79%**	20%
9.4 rpbis SC T90	xxxx	-0,19	-0,12	-0,14	-0,12	-0,27	**0,33**	-0,12

On voit aussi qu'à *T90* les *rpbis SCT* sont plus marqués. A ce palier de turbo analyse, après application du test $r \neq 0$ de Fisher, nous avons pu vérifier qu'ils sont tous significativement différents de zéro. Du point de vue de la cohérence d'utilisation des pourcentages de certitude cette question fonctionne donc correctement (chez les étudiants les plus réalistes [$Rs \geqslant 90$] le choix de la réponse correcte a tendance à être accompagné de certitudes en moyenne plus élevées que les certitudes qui ont accompagné le choix des propositions incorrectes). Dès lors, comment expliquer qu'elle récolte un « - » à l'indice *NCSq* basé sur la différence entre le *rpbis SC T80* de la solution correcte et la moyenne pondérée des *rpbis SC T80* des propositions incorrectes ?

Rappelons d'abord que dans le cadre du classement des 173 QCM des 10 épreuves MOHICAN, une question obtient un « - » lorsque son *NCSq T80* est compris entre -0,254 et 0,250 (voir détails et

tableaux p. 349). Ce qui signifie qu'une question qui relève de cette catégorie « - » peut obtenir un *NCSq T80* positif entre 0 et 0,250 (rappelons que l'indice *NCSq* varie entre –2 et +2) et donc être légèrement cohérente du point de vue de l'utilisation des pourcentages de certitude et obtenir un « - ». En fait, sept questions parmi les 173 des 10 tests MOHICAN obtiennent un score *NCSq T80* inférieur ou égal à 0,250 et parmi elles seules deux QCM peuvent être qualifiées d'incohérentes du point de vue de l'utilisation des degrés de certitude dans la mesure où leur score *NCSq T80* est inférieur à zéro. Il s'agit des questions [H]q3 et [H]q20 qui obtiennent respectivement un *NCSq T80* négatif égal à –0,187 et –0,254.

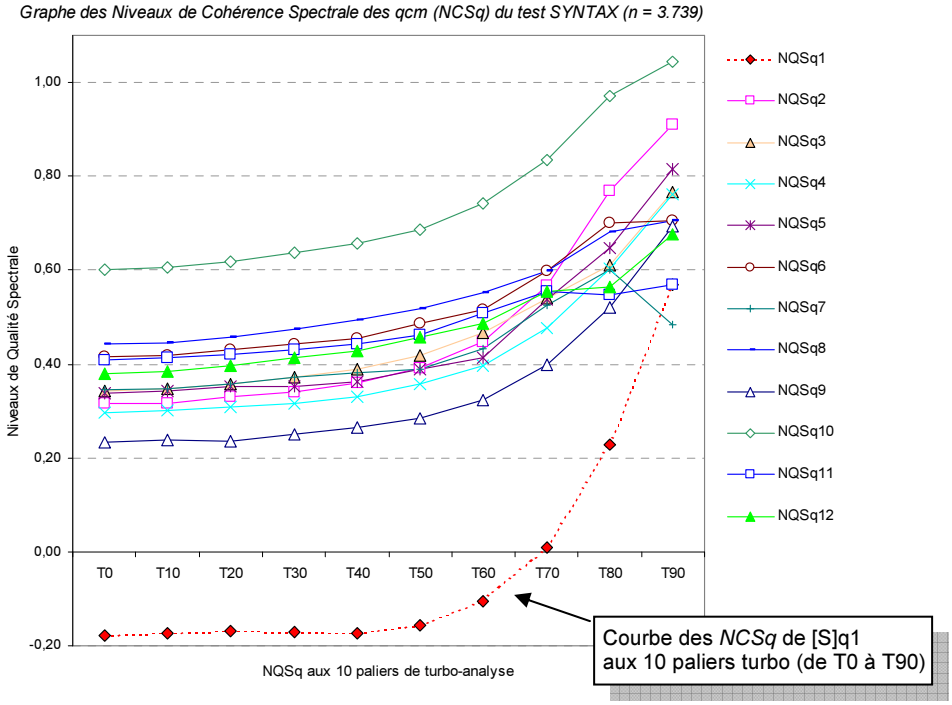
Donc, [S]q1 fait partie des questions qui obtiennent un *NCSq T80* peu élevé tout en étant supérieur à zéro à ce palier de turbo analyse *T80*. Le graphique ci-dessous permet de comparer les scores *NCSq* de cette question avec ceux des autres QCM du test « Syntaxe et articulation logique » et ce, aux 10 paliers de turbo analyse.

La courbe des *NCSq* de [S]q1 montre que :

1) cette question se démarque des 11 autres du test, elle est la moins cohérente d'un point de vue spectral ;

2) jusqu'à T70 on observe une incohérence spectrale dans l'utilisation des pourcentages de certitude (*NCSq* inférieurs à zéro) ;

3) lorsqu'on ne prend en compte que les données des étudiants dont le réalisme dépasse 70, l'utilisation des pourcentages de certitude devient cohérente.



4) Le *NCSq T80* de [S]q1, bien que positif, est le moins élevé des 12 questions du test et celui du *NCSq T90* figure en avant dernière position.

Dès lors, on voit que cette question pose un problème de cohérence spectrale lorsqu'on prend en compte les étudiants dont le réalisme est inférieur à 70 mais plus du tout lorsqu'on sélectionne les données de ceux dont le réalisme est supérieur à 80.

En ce qui concerne la cohérence interne, l'analyse des *rpbis classiques* montre que cette question [S]q1 fonctionne correctement du point de vue de la discrimination en fonction du nombre de réponses correctes obtenues au total du test.

1) N Rép.
2) % Rép.
3) rpbis

OM	P1	P2	P3	P4	P5	**P6**	P7
54	59	39	453	82	1848	**1194**	10
1%	2%	1%	12%	2%	49%	**32%**	0%
-0,07	-0,15	-0,11	-0,15	-0,06	-0,16	**0,38**	-0,03

Le *rpbis classique* de la réponse <u>correcte</u> (P6) est supérieur à la valeur repère 0,29 ($1/\sqrt{12}$).

Les *rpbis classiques* des propositions <u>incorrectes</u> sont quant à eux négatifs. Sauf pour les *rpbis classiques* de P4 et P7 proches de zéro.

La question pourrait être améliorée par un meilleur équilibrage de l'attractivité des propositions incorrectes P1, P2 et P4 très peu choisies par les étudiants dans la version actuelle.

Une analyse qualitative du contenu de [S]q1 ne montre pas d'anomalies particulières dans les propositions de cette question.

Q1. Cet étudiant fait des efforts considérables pour suivre la conférence du professeur …….. on voit qu'il ne comprend que quelques bribes.

1. que l'
2. qu'
3. duquel
4. laquelle
5. pour laquelle
6. aucune
7. toutes

Les experts présents lors de la réunion de débriefing ont signalé que le choix massif de P5 (49% des étudiants) s'explique par le fait que cette proposition est la seule qui contienne une préposition devant le pronom relatif.

Un autre facteur qui rend cette question relativement difficile est lié au fait qu'elle comporte plusieurs subordonnées ce qui rend sa construction assez lourde.

Les réponses correctes sont « à laquelle » ou « dont » (ces dernières n'ont pas été proposées, d'où la réponse correcte P6 « aucune »).

En conclusion les constats tirés de l'analyse des *rpbis classiques*, *rpbis SCT80* et *rpbis SCT90* sont confirmés par l'avis des experts qui estiment cette question difficile mais sans anomalie majeure. Par contre, en ce qui concerne les informations livrées par les *rpbis SC* calculés à des paliers turbo inférieurs à T70 nous constatons qu'elles ne permettent pas de confirmer le diagnostic des experts.

b) [V]q12

Comme la précédente, cette 12ème question du test de vocabulaire récolte des « - » aux indices spectraux *NCSq T80*, *Rq T80* et *Cq+ T80.* Par contre, aux indices de cohérence interne $r_{qt\ mb}$ et $r_{qt\ ms}$ elle récolte des « ≃ ».

P6 est la réponse correcte à cette question dont « aucune » des propositions est correcte.

Au palier de turbo analyse T80, nous observons pour la proposition correcte un *rpbis SCT* positif (0,15). Les *rpbis SCT* des propositions incorrectes P1, P2, P3 et P5 sont négatifs. En ce qui concerne P4, le *rpbis SCT* est positif mais pratiquement égal à zéro. Pour P7, il est négatif mais aussi pratiquement égal à zéro.

Au palier T90, les *rpbis SCT* sont plus marqués et la proposition incorrecte P4 passe de 0,01 à –0,12. Quant à P7, il reste pratiquement égal à zéro (-0,06).

5. Palier de Turbo analyse : T50

	OM	P1	P2	P3	P4	P5	**P6**	P7
5.1 N Rép. T50	98	242	791	343	1317	279	**702**	25
5.2 % Rép. T50	3%	6%	21%	9%	35%	7%	**18%**	1%
5.3 C. Moy. T50	11%	45%	55%	49%	68%	49%	**61%**	49%
5.4 rpbis SC T50	-0,25	-0,13	-0,08	-0,11	0,10	-0,10	**0,05**	-0,03

6. Palier de Turbo analyse : T60

	OM	P1	P2	P3	P4	P5	**P6**	P7
6.1 N Rép. T60	95	233	760	327	1283	273	**684**	24
6.2 % Rép. T60	3%	6%	21%	9%	35%	7%	**19%**	1%
6.3 C. Moy. T60	8%	45%	54%	48%	68%	49%	**62%**	47%
6.4 rpbis SC T60	-0,27	-0,13	-0,10	-0,12	0,09	-0,11	**0,06**	-0,04

7. Palier de Turbo analyse : T70

	OM	P1	P2	P3	P4	P5	**P6**	P7
7.1 N Rép. T70	88	204	663	279	1112	241	**616**	20
7.2 % Rép. T70	3%	6%	21%	9%	35%	7%	**19%**	1%
7.3 C. Moy. T70	7%	45%	54%	48%	67%	48%	**63%**	47%
7.4 rpbis SC T70	-0,30	-0,14	-0,12	-0,13	0,08	-0,12	**0,08**	-0,04

8. Palier de Turbo analyse : T80

	OM	P1	P2	P3	P4	P5	**P6**	P7
8.1 N Rép. T80	69	116	410	167	702	164	**416**	12
8.2 % Rép. T80	3%	6%	20%	8%	34%	8%	**20%**	1%
8.3 C. Moy. T80	4%	42%	53%	46%	66%	49%	**66%**	45%
8.4 rpbis SC T80	-0,38	-0,18	-0,17	-0,18	0,01	-0,16	**0,15**	-0,06

9. Palier de Turbo analyse : T90

	OM	P1	P2	P3	P4	P5	**P6**	P7
9.1 N Rép. T90	24	19	104	42	123	38	**102**	5
9.2 % Rép. T90	5%	4%	23%	9%	27%	8%	**22%**	1%
9.3 C. Moy. T90	0%	48%	48%	41%	61%	48%	**69%**	52%
9.4 rpbis SC T90	xxxx	-0,15	-0,27	-0,29	-0,12	-0,19	**0,29**	-0,06

D'un point de vue spectral, l'utilisation des pourcentages de certitude est donc assez cohérente lorsqu'on prend en compte les données des étudiants les plus réalistes. Les étudiants dont $Rs \geqslant 80$ ont donc tendance à accompagner la réponse correcte de pourcentages de certitude en moyenne plus élevés que les pourcentages de certitude de ceux qui choisissent une proposition incorrecte.

On peut dès lors se demander (comme dans le cas de [S]q1) comment cette question [V]q12 peut récolter un « - » à l'indice *NCSq T80* ? En fait, le *NCSq T80* de cette question est positif et vaut 0,250. Peu élevé, il est donc aussi à la limite supérieure de la catégorie « - » qui reprend les questions dont le *NCSq T80* varie entre –0,254 et 0,250. Comme dans le cas de [S]q1, il s'agit donc d'une question dont le *NCSq T80* est peu élevé mais cependant positif et, dès lors, on ne décèle pas d'incohérence d'utilisation des pourcentages de certitude lors d'une analyse spectrale approfondie des propositions de la QCM.

En ce qui concerne la cohérence interne le tableau ci-dessous montre que la proposition correcte P6 discrimine bien les étudiants du point de vue du nombre de réponses correctes fournies au total du test.

	OM	P1	P2	P3	P4	P5	**P6**	P7
1) N Rép.	100	246	799	349	1335	280	**712**	25
2) % Rép.	3%	6%	21%	9%	35%	7%	**19%**	1%
3) rpbis	-0,19	-0,13	-0,08	-0,06	-0,01	0,03	**0,29**	-0,05

Le *rpbis classique* de P6 est en effet supérieur à la valeur repère qui vaut 0,15 ($1/\sqrt{45}$). En ce qui concerne les propositions incorrectes on constate qu'elles sont fort proches de zéro pour P2, P3, P4, P5 et P7 tandis que le *rpbis classique* de P1 est légèrement négatif (-0,13). La plupart des distracteurs discriminent donc moins bien les étudiants du point de vue du nombre de réponses correctes fournies au total du test.

Une analyse plus qualitative du contenu de la question permet de mieux comprendre pourquoi 35% des étudiants ont choisi la solution P4.

En effet, selon les experts consultés lors du débriefing de l'épreuve, probablement ont-ils confondu l'expression « à l'instar de » et « à l'insu de » qui sont phonétiquement proches bien que très différentes d'un point de vue sémantique.

Ceci dit, comme le soulignait un des intervenants lors de la réunion de février 2000, on ne peut se tromper si on a vu la pièce de théâtre...

Q12. Le valet de don Juan agit <u>à l'instar de</u> son maître, ce qui crée un décalage comique.

Le valet de don Juan…

1. désobéit effrontément à son maître
2. fait tout le contraire de ce que fait son maître
3. seconde et aide maladroitement son maître
4. cache à son maître tout ce qu'il fait
5. exagère ce que fait son maître
6. aucune
7. toutes

4. Analyse des questions [V]q27 et [A]q14 qui présentent un « - » et un « ≃ » en cohérence interne ainsi qu'en qualité spectrale

a) [V]q27

Cette 27ème question du test de vocabulaire fait partie des 8 QCM épinglées pour leurs performances faibles en cohérence spectrale ou/et en cohérence interne. Dans le tableau des performances (voir p. 356) nous remarquons que cette question est la seule qui présente un « ≃ » à l'indice *Cq- T80*, c'est-à-dire une tendance à la sous-estimation (qui est moyenne lorsqu'on la compare aux 172 autres questions MOHICAN).

Au palier de turbo analyse T80 nous remarquons un *rpbis SCT* positif mais peu élevé pour la proposition correcte P3 (0,13).

En ce qui concerne les propositions incorrectes, nous remarquons que P1 obtient un *rpbis SCT80* peu élevé mais positif (0,12), ce qui témoigne d'une utilisation incohérente des pourcentages de certitude. Pour P2, nous observons un *rpbis SCT80* négatif (-0,13) ce qui est cohérent d'un point de vue spectral.

Au palier de turbo analyse T90 les *rpbis SCT90* de la réponse correcte (0,33) et de P2 (-0,21) sont plus marqués. Pour ce qui est de P1 la corrélation n'est plus positive et très proche de zéro (-0,03).

Qu'en est-il de la cohérence interne ?

5. Palier de Turbo analyse : T50

	OM	P1	P2	P3
5.1 N Rép. T50	320	704	867	**1873**
5.2 % Rép. T50	8%	19%	23%	**49%**
5.3 C. Moy. T50	3%	42%	28%	**32%**
5.4 rpbis SC T50	-0,27	0,13	-0,06	**0,05**

6. Palier de Turbo analyse : T60

	OM	P1	P2	P3
6.1 N Rép. T60	310	683	846	**1811**
6.2 % Rép. T60	8%	19%	23%	**49%**
6.3 C. Moy. T60	2%	43%	27%	**32%**
6.4 rpbis SC T60	-0,27	0,14	-0,07	**0,05**

7. Palier de Turbo analyse : T70

	OM	P1	P2	P3
7.1 N Rép. T70	275	611	741	**1572**
7.2 % Rép. T70	9%	19%	23%	**49%**
7.3 C. Moy. T70	2%	43%	27%	**33%**
7.4 rpbis SC T70	-0,29	0,14	-0,08	**0,07**

8. Palier de Turbo analyse : T80

	OM	P1	P2	P3
8.1 N Rép. T80	207	413	488	**938**
8.2 % Rép. T80	10%	20%	24%	**46%**
8.3 C. Moy. T80	1%	45%	28%	**36%**
8.4 rpbis SC T80	-0,37	0,12	-0,13	**0,13**

9. Palier de Turbo analyse : T90

	OM	P1	P2	P3
9.1 N Rép. T90	75	99	92	**189**
9.2 % Rép. T90	16%	22%	20%	**41%**
9.3 C. Moy. T90	0%	43%	34%	**49%**
9.4 rpbis SC T90	-0,61	-0,03	-0,21	**0,33**

Le *rpbis classique* de la réponse correcte (P3) vaut 0,13 et est inférieur à la valeur repère ($1/\sqrt{45} = 0{,}15$). Cette proposition correcte discrimine donc mal les étudiants en fonction du nombre de réponses correctes fournies au total du test.

La proposition incorrecte P1 obtient la même valeur positive que le *rpbis classique* de la réponse correcte P3 (0,13).

	OM	P1	P2	P3
1) N Rép.	321	718	874	**1892**
2) % Rép.	8%	19%	23%	**49%**
3) rpbis	-0,24	0,13	-0,09	**0,13**

En ce qui concerne la seconde proposition incorrecte (P2) on observe un *rpbis classique* négatif mais très proche de zéro (donc absence de discrimination).

L'analyse de la question [V]q27 par les experts présents lors de la réunion de débriefing de février 2000 n'a pas permis de mettre en évidence une anomalie particulière dans le contenu de cette question.

Q27. …déterminez le rapport de sens entre <u>injonction</u> et <u>invective</u> en écrivant

1. si <u>dans certains contextes</u>, les deux mots peuvent avoir à peu près le MEME sens.
2. si <u>dans certains contextes</u>, les deux mots peuvent avoir des sens à peu près CONTRAIRES.
3. si <u>dans aucun contexte</u>, les deux mots n'entretiennent l'un des deux rapports précédents.

b) [A]q14

Cette 14ème question du test de connaissances artistique (voir annexe p. 498) fait partie des 8 QCM épinglées pour leurs performances faibles en cohérence spectrale ou/et en cohérence interne (elle présente un « - » et un « ≃ » en cohérence interne ainsi qu'en qualité spectrale). Dans le tableau des performances (voir p. 356) nous remarquons qu'elle est aussi la seule qui présente un « + » à l'indice *Cq+ T80*, ce qui signifie que l'indice de Centration par question révèle globalement pour cette question un niveau de surestimation très peu élevé.

La proposition P3 est la réponse correcte attendue à cette question. Au palier de turbo analyse T90, P3 obtient un *rpbis SCT* de 0,32 (p = .001). La réponse correcte fonctionne donc logiquement : les étudiants qui répondent P3 accompagnent leur choix d'un pourcentage plus élevé que ceux qui ne la choisissent pas.

5. Palier de Turbo analyse : T50

	OM	P1	P2	**P3**	P4	P5	P6	P7
5.1 N Rép. T50	207	160	112	**292**	311	128	150	13
5.2 % Rép. T50	15%	12%	8%	**21%**	23%	9%	11%	1%
5.3 C. Moy. T50	0%	28%	28%	**25%**	38%	25%	19%	35%
5.4 rpbis SC T50	-0,30	0,04	0,03	**0,01**	0,19	0,00	-0,06	0,04

6. Palier de Turbo analyse : T60

	OM	P1	P2	**P3**	P4	P5	P6	P7
6.1 N Rép. T60	205	155	109	**287**	303	124	147	13
6.2 % Rép. T60	15%	12%	8%	**21%**	23%	9%	11%	1%
6.3 C. Moy. T60	0%	27%	28%	**25%**	37%	23%	19%	35%
6.4 rpbis SC T60	-0,31	0,03	0,04	**0,02**	0,18	-0,02	-0,07	0,04

7. Palier de Turbo analyse : T70

	OM	P1	P2	**P3**	P4	P5	P6	P7
7.1 N Rép. T70	196	145	98	**256**	274	110	124	12
7.2 % Rép. T70	16%	12%	8%	**21%**	23%	9%	10%	1%
7.3 C. Moy. T70	0%	27%	29%	**26%**	36%	22%	18%	32%
7.4 rpbis SC T70	-0,34	0,01	0,03	**0,05**	0,16	-0,05	-0,09	0,02

8. Palier de Turbo analyse : T80

	OM	P1	P2	**P3**	P4	P5	P6	P7
8.1 N Rép. T80	156	89	62	**147**	174	67	74	9
8.2 % Rép. T80	20%	11%	8%	**19%**	22%	9%	10%	1%
8.3 C. Moy. T80	0%	26%	30%	**30%**	37%	24%	20%	27%
8.4 rpbis SC T80	xxxx	-0,04	-0,01	**0,11**	0,10	-0,05	-0,10	-0,01

9. Palier de Turbo analyse : T90

	OM	P1	P2	**P3**	P4	P5	P6	P7
9.1 N Rép. T90	54	20	10	**29**	47	22	17	2
9.2 % Rép. T90	27%	10%	5%	**14%**	23%	11%	8%	1%
9.3 C. Moy. T90	0%	28%	36%	**47%**	37%	24%	28%	60%
9.4 rpbis SC T90	xxxx	-0,20	-0,08	**0,32**	-0,13	-0,28	-0,23	-0,06

En ce qui concerne les propositions incorrectes, nous remarquons à T90 seule la proposition P5 obtient un *rpbis SCT* significativement différent de zéro (-0,28, p = .05). Pour cette proposition P5, les étudiants ont accompagné leur choix de pourcentages de certitude plus faibles que ceux qui ont répondu correctement. Le pourcentage des omissions est très élevé (27%).

Pour les propositions incorrectes P1, P2, P6 et P7, nous remarquons à T80 des *rpbis SCT* négatifs mais proches de zéro. On observe donc pas de tendance marquée à accompagner ces propositions incorrectes de certitudes plus faibles lorsque nous les contrastons avec les certitudes de la réponse correcte.

En ce qui concerne P4, la très légère corrélation positive à T80 (0,10) doit être interprétée comme non significativement différente de zéro (nous avons appliqué le test $r \neq 0$ de Fisher). C'est aussi le cas pour le *rpbis SCT90* de P4 (0,13) qui n'est pas significativement différent de zéro. Nous remarquons également que quelque soit le niveau de turbo analyse, cette proposition P4 est la plus attractive. Beaucoup d'étudiants choisissent P4 (environ un sur quatre) mais lorsqu'on ne prend en considération que les données des étudiants les plus réalistes (à partir de T80) on constate que ces derniers n'ont pas tendance à accompagner leurs choix de certitudes plus élevées que ceux qui répondent correctement (les *rpbis SCT80* et *rpbis SCT90* de P4 ne sont pas significativement différents de zéro).

L'analyse des *rpbis classiques* montre pour toutes les propositions des

	OM	P1	P2	**P3**	P4	P5	P6	P7
1) N Rép.	210	162	113	**296**	316	130	152	13
2) % Rép.	15%	12%	8%	**21%**	23%	9%	11%	1%
3) rpbis	-0,18	0,03	-0,01	**0,09**	0,08	0,01	-0,06	0,02

valeurs très proches de zéro. Les propositions de cette question ne discriminent donc pas entre les étudiants du point de vue du nombre de réponses correctes obtenues au total du test, y compris en ce qui concerne la réponse correcte P3.

Remarquons également le taux élevé des omissions (15%).

Comment peut-on expliquer les faibles indices de cohérence interne et de cohérence spectrale et la forte attractivité de P4 ?

Il convient d'abord de rappeler le sens du mot décimation. Le dictionnaire[74] renseigne : « *Dans l'antiquité romaine, Action de décimer* » ; décimer ayant par ailleurs deux sens : « *1. Dans l'antiquité, mettre à mort une personne sur dix désignée par le sort. (...) 2. Faire périr un grand nombre de personnes, ... détruire, exterminer* ». Remarquons que « décimation » est dès lors un terme plutôt connoté « armée », « guerre », « violence »...

Q14. Lequel de ces 5 films traite de la « décimation » ?

1. Orange mécanique
2. 2001 Odyssée de l'espace
3. Les sentiers de la gloire
4. Full Metal Jacket
5. Barry Lindon
6. aucune
7. toutes

Dans l'enseignement secondaire, le concept de décimation pris dans son premier sens est souvent abordé au cours d'histoire dans le cadre de la matière ayant trait à l'antiquité et aux conquêtes romaines. Les étudiants peuvent bien sûr aussi rencontrer le mot décimation pris dans son second sens lors d'autres lectures (« …*cette population fut décimée par la peste*… »).

Envisageons maintenant les différentes solutions proposées.

En ce qui concerne le film « Orange mécanique », il ne s'agit pas à proprement parler d'un film qui illustre l'idée de la décimation. Cependant, il comporte de nombreuses scènes de violence, et on peut comprendre, sans pour autant accepter cette première proposition comme étant correcte, l'amalgame qui a été fait entre le terme « décimation » et la sauvagerie de certaines scènes du film par une partie des étudiants.

Pour ce qui est de « 2001 Odyssée de l'espace » on comprend moins bien ce qui a pu pousser 8% des étudiants à choisir cette solution. Même remarque en ce qui concerne le dernier film proposé : « Barry Lindon » choisi par 9 % des examinés.

Qu'en est-il de la réponse correcte P3 « Les sentiers de la gloire » ?

Rappelons que ce film est un pamphlet antimilitariste et qu'il raconte une histoire vraie qui eut lieu en 1916 pendant la première guerre mondiale. Il décrit les conséquences d'une offensive suicidaire lancée contre une position allemande imprenable par un général en manque d'avancement. L'échec est attribué aux hommes, trois soldats sont désignés au sort et exécutés pour « couardise ». Cette façon de désigner les victimes au hasard parmi les « coupables » constitue une illustration du premier sens du mot « décimation » mais une illustration peu conforme au rapport 1 sur 10. Il y a bien l'aspect « …*désignée par le sort* » mais on est loin du rapport 1/10 (dans le film on désigne « seulement » 3 soldats sur quelques centaines). En ce qui concerne le second sens « *Faire périr un grand nombre de personnes, ... détruire, exterminer* », on peut dire que bien des scènes du film en sont une illustration.

La quatrième proposition concerne le film intitulé « Full Metal Jacket ». Il s'agit aussi d'un film de guerre mais l'histoire se rapporte au conflit du Vietnam. On n'y aborde pas le thème de la « décimation » au premier sens du terme (« *mettre à mort une personne sur dix désignée par le sort* »). Par contre on y montre de nombreuses scènes où un grand nombre de personnes sont exterminées. C'est donc plutôt de

[74] *Le nouveau petit ROBERT* ; dictionnaire alphabétique et analogique de la langue française, Dictionnaires Le Robert, Paris, 1993.

« décimation » au sens « *Faire périr un grand nombre de personnes, ... détruire, exterminer* » dont il est question dans ce film. C'est probablement ce second sens du terme qui explique le choix de P4 effectué par 23% des étudiants.

L'analyse du contenu de la question montre que pour la proposition P3, considérée au départ comme correcte, ne peut pas être interprétée comme étant une illustration du sens classique et premier du mot « décimation ». L'analyse du contenu de P4 nous permet aussi de comprendre qu'un pourcentage un peu plus élevé d'étudiants s'oriente vers le choix de cette quatrième proposition (considérée au départ comme incorrecte) car on y trouve un rapport avec le second sens attribué au mot « décimation ».

Lors de la réunion de débriefing du 7 février 2000, la difficulté de cette question fut relevée (un étudiant sur cinq fournit la réponse attendue). Les experts reconnaissent que cette question est mal construite : « *dans les sentiers de la gloire, la moitié de la définition est respectée : 'mettre à mort x personnes désignées par le sort', mais l'autre moitié de la définition, l'idée des 10%, est complètement non respectée : dans le film on n'a fusillé que 3 soldats sur des centaines. En outre, le sens métaphorique 'décimer = en tuer beaucoup' a pris en français, le pas sur le sens strict. Pour toutes ces raisons il faut accepter la réponse 4 comme correcte* ».

Pour cette question [A]q14, nous constatons que les *rpbis classiques* montrent un manque de discrimination étant donné des valeurs toutes assez proches de zéro. La question discrimine mal les étudiants ayant obtenu en moyenne un nombre supérieur de réponses correctes au total du test des étudiants ayant obtenu un nombre inférieur de réponses correctes. Dès lors on peut dire que les *rpbis classiques* ne détectent pas clairement les problèmes inhérents aux propositions P3 (à moitié correcte) et P4 (à moitié incorrecte).

En ce qui concerne la cohérence spectrale, les *rpbis SCT80* des distracteurs P1, P2, P4, P5 et P7 sont non significativement différents de zéro. Le distracteur P6 est légèrement négatif et significativement différent de zéro. La réponse considérée comme étant correcte est légèrement positive et significativement différente de zéro. Donc, les *rpbis SCT80* ne permettent pas de détecter clairement les problèmes relevés par les experts dans les propositions P3 et P4.

Pour ce qui est des *rpbis SCT90*, l'aspect à moitié correct de P3 n'est pas non plus détecté dans la mesure où son *rpbis SCT90* est positif et plus élevé (0,32). En ce qui concerne le problème du distracteur P4 (à moitié incorrect) le *rpbis SCT90* récolté par cette proposition, bien que négatif, n'est pas significativement différent de zéro. Le problème de P4 n'est donc pas détecté.

5. Analyse de la question [B]q5 : cohérence spectrale moyenne et cohérence interne faible

Cette cinquième question du test de biologie fait partie des 8 QCM épinglées pour leurs performances faibles en cohérence spectrale ou/et en cohérence interne (voir tableau des performances, p. 356).

5. Palier de Turbo analyse : T50

	OM	P1	P2	**P3**	P4		P6	P7
5.1 N Rép. T50	135	1115	263	**498**	202		137	18
5.2 % Rép. T50	6%	47%	11%	**21%**	9%		6%	1%
5.3 C. Moy. T50	3%	40%	29%	**39%**	29%		26%	36%
5.4 rpbis SC T50	-0,27	0,02	-0,11	**0,08**	-0,10		-0,11	-0,01

6. Palier de Turbo analyse : T60

	OM	P1	P2	**P3**	P4		P6	P7
6.1 N Rép. T60	124	993	235	**435**	172		124	16
6.2 % Rép. T60	6%	47%	11%	**21%**	8%		6%	1%
6.3 C. Moy. T60	3%	40%	28%	**42%**	29%		27%	33%
6.4 rpbis SC T60	-0,30	-0,03	-0,14	**0,11**	-0,12		-0,12	-0,03

7. Palier de Turbo analyse : T70

	OM	P1	P2	**P3**	P4		P6	P7
7.1 N Rép. T70	103	777	178	**302**	120		108	12
7.2 % Rép. T70	6%	48%	11%	**19%**	7%		7%	1%
7.3 C. Moy. T70	1%	39%	26%	**46%**	25%		25%	25%
7.4 rpbis SC T70	-0,38	-0,13	-0,22	**0,21**	-0,20		-0,18	-0,06

8. Palier de Turbo analyse : T80

	OM	P1	P2	**P3**	P4		P6	P7
8.1 N Rép. T80	67	410	96	**151**	64		53	10
8.2 % Rép. T80	8%	48%	11%	**18%**	7%		6%	1%
8.3 C. Moy. T80	0%	37%	23%	**52%**	22%		25%	24%
8.4 rpbis SC T80	-0,47	-0,24	-0,32	**0,29**	-0,26		-0,21	-0,10

9. Palier de Turbo analyse : T90

	OM	P1	P2	**P3**	P4		P6	P7
9.1 N Rép. T90	25	95	24	**27**	18		13	2
9.2 % Rép. T90	12%	46%	12%	**13%**	9%		6%	1%
9.3 C. Moy. T90	0%	36%	18%	**58%**	29%		20%	0%
9.4 rpbis SC T90	xxxx	-0,42	-0,47	**0,40**	-0,33		-0,35	-0,07

Nous remarquons que pour la réponse correcte (P3) les *rpbis SCT* dépassent 0,2 à partir du niveau de turbo analyse T70.

En ce qui concerne les propositions incorrectes (il n'y avait pas de 5ème proposition, d'où l'absence de P5) nous constatons que le *rpbis SCT* de P1 est négatif et assez marqué à partir de T80. Les *rpbis SCT* de P2 et P4 sont négatifs et assez marqués à partir de T70. Les *rpbis SCT* de P6 sont négatifs et assez marqués à partir de T70. La proposition P7 est également très peu choisie (1%) et les rpbis SCT sont très proches de zéro.

En ce qui concerne les *rpbis classiques*, nous remarquons ci-dessous que la réponse correcte (P3) récolte 0,24. Il existe donc une tendance chez les étudiants qui la choisissent à obtenir un nombre de réponses correctes au total du test plus élevé que ceux qui ne choisissent pas P3. Les valeurs récoltées à P1 (très attractive, 47% des choix) et P7 sont très proches de zéro, ces propositions incorrectes ne discriminent donc pas les étudiants en fonction du nombre de réponses correctes fournies au total du test. En ce qui concerne P2 et P4, la discrimination existe mais est peu élevée.

	OM	P1	P2	**P3**	P4		P6	P7
1) N Rép.	142	1175	277	**533**	213		139	20
2) % Rép.	6%	47%	11%	**21%**	8%		6%	1%
3) rpbis	-0,13	0,02	-0,10	**0,24**	-0,12		-0,05	0,00

Cette question fut l'objet de commentaires lors de la réunion de débriefing des épreuves de sciences (le 3 février 2000). Après discussion, entre les experts présents, l'auteur du test en est arrivé à la conclusion que cette question n'avait pas sa place dans le test : « *elle ne mesure pas la même chose que les autres sans pour autant être mauvaise* ». Elle concerne en fait un concept écologique trop nouveau, pas encore, ou très peu

Q5. La *niche écologique* est

1. la place de la population dans un ensemble de variables physiques, chimiques et climatiques
2. l'ensemble des adaptations morphologiques de la population à son environnement
3. la place de la population dans une communauté, c'est-à-dire l'ensemble des relations biologiques
4. la place de chaque individu au sein de sa population
6. aucune
7. toutes

expliqué dans les cours (il y a 20 ans, l'écologie n'était pas au programme de biologie).

Nous remarquons que beaucoup d'étudiants ont répondu P1 (47%) ce qui selon les spécialistes présents lors du débriefing témoigne d'une confusion entre habitat et niche.

Du point de vue des problèmes mis en évidence par les différents types de *rpbis*, d'une part on voit que le *rpbis classique* de la proposition incorrecte P1 est proche de zéro (0,02). Il indique donc que P1 ne discrimine pas les étudiants alors qu'on s'attend à une valeur négative indiquant que ceux qui obtiennent un nombre moins élevé de bonnes réponses ont tendance à choisir ce distracteur.

D'autre part, nous remarquons que les *rpbis SCT80* et *rpbis SCT90* de P1 révèlent quant à eux un fonctionnement spectral logique pour cette proposition incorrecte : les étudiants dont le réalisme est élevé ($Rs \geqslant 80$) et qui choisissent P1, le font en étant en moyenne beaucoup moins sûrs de leur réponse que ceux qui à ce niveau de turbo analyse choisissent la réponse correcte, d'où les valeurs négatives (*rpbis SCT80* de P1 = -0,24). Ce constat est encore plus marqué à T90 (*rpbis SCT90* de P1 = -0,42).

Dès lors, on peut dire que les valeurs obtenues par les *rpbis SCT80* et *rpbis SCT80* confirment l'analyse des experts (P1 est incorrecte et témoigne d'une confusion entre habitat et niche) alors que les valeurs observées pour les *rpbis classiques* ne permettent pas de confirmer l'avis des experts.

6. Analyse de la question [P]q7 aux performances opposées aux indices spectraux NCSq T80 et Rq T80

La 7ème question du test de physique a été mise en évidence pour les valeurs opposées obtenue à l'indice *NCSq T80* « + » et à l'indice *Rq T80* « - » (voir tableau des performances, p. 356).

5. Palier de Turbo analyse : T50

	OM	P1	**P2**	P3	P4	P5	P6	P7
5.1 N Rép. T50	68	176	**573**	1131	55	121	220	3
5.2 % Rép. T50	3%	7%	**24%**	48%	2%	5%	9%	0%
5.3 C. Moy. T50	13%	57%	**74%**	57%	40%	33%	44%	0%
5.4 rpbis SC T50	-0,30	-0,13	**0,28**	-0,25	-0,15	-0,27	-0,26	xxxx

6. Palier de Turbo analyse : T60

	OM	P1	**P2**	P3	P4	P5	P6	P7
6.1 N Rép. T60	62	163	**530**	1041	51	106	203	3
6.2 % Rép. T60	3%	8%	**25%**	48%	2%	5%	9%	0%
6.3 C. Moy. T60	12%	58%	**77%**	58%	40%	34%	44%	0%
6.4 rpbis SC T60	-0,33	-0,15	**0,32**	-0,28	-0,17	-0,28	-0,29	xxxx

7. Palier de Turbo analyse : T70

	OM	P1	**P2**	P3	P4	P5	P6	P7
7.1 N Rép. T70	46	126	**469**	838	35	75	167	2
7.2 % Rép. T70	3%	7%	**27%**	48%	2%	4%	9%	0%
7.3 C. Moy. T70	14%	58%	**79%**	58%	45%	37%	43%	0%
7.4 rpbis SC T70	-0,32	-0,18	**0,36**	-0,32	-0,16	-0,26	-0,33	xxxx

8. Palier de Turbo analyse : T80

	OM	P1	**P2**	P3	P4	P5	P6	P7
8.1 N Rép. T80	25	65	**330**	470	19	36	108	1
8.2 % Rép. T80	2%	6%	**31%**	45%	2%	3%	10%	0%
8.3 C. Moy. T80	19%	58%	**83%**	59%	43%	39%	44%	0%
8.4 rpbis SC T80	-0,31	-0,20	**0,45**	-0,40	-0,17	-0,26	-0,38	xxxx

9. Palier de Turbo analyse : T90

	OM	P1	**P2**	P3	P4	P5	P6	P7
9.1 N Rép. T90	15	17	**142**	96	2	11	29	0
9.2 % Rép. T90	5%	5%	**46%**	31%	1%	4%	9%	0%
9.3 C. Moy. T90	15%	52%	**89%**	59%	80%	38%	41%	xxxx
9.4 rpbis SC T90	-0,51	-0,25	**0,60**	-0,42	-0,12	-0,30	-0,41	xxxx

En ce qui concerne les *rpbis SCT*, nous remarquons qu'à T80 celui de la réponse correcte P2 est positif et marqué (0,45). A T90 le *rpbis SCT* de P2 (0,60) est encore plus élevé.

En ce qui concerne les propositions incorrectes P1 et P3 à P6, nous remarquons que les *rpbis SCT* sont tous négatifs. Remarquons aussi la très forte attractivité de P3 à T50 (48% des choix). Cette attractivité diminue plus les étudiants sélectionnés sont réalistes (à T90 P3 est choisi par 31% des étudiants).

La proposition P7 présente un cas particulier dans la mesure où elle est choisie par très peu d'étudiants. A T80 un seul étudiant choisit cette proposition et les *rpbis SCT* sont incalculables (d'où les xxxx).

En ce qui concerne les *rpbis classiques*, nous remarquons que P2 récolte une valeur positive assez marquée (0,43). Les propositions incorrectes P1, P4, P6 et P7 obtiennent des valeurs proches de zéro, ces quatre propositions ne permettent donc pas de discriminer les sujets lorsqu'on utilise comme variable critère le nombre de réponses correctes fournies au total du test.

	OM	P1	**P2**	P3	P4	P5	P6	P7
1) N Rép.	86	188	**602**	1200	57	135	225	4
2) % Rép.	3%	8%	**24%**	48%	2%	5%	9%	0%
3) rpbis	-0,14	-0,08	**0,43**	-0,20	-0,09	-0,12	0,01	-0,06

Pour ce qui est de P3, nous remarquons qu'elle est choisie par 48% des étudiants (alors que 24% choisissent la réponse correcte) ainsi qu'un *rpbis classique* plus marqué (-0,20). Pour P5, la valeur récoltée est négative mais peu élevée (-0,12).

Lors de la réunion de débrieffing du 3 février 2000 les auteurs du test de physique ont souligné la difficulté de cette question mais n'y ont pas décelé un problème particulier. Les auteurs ont cependant proposé d'améliorer la question en y introduisant une explication de la signification de "g".

Hormis la très forte attractivité de la proposition P3, les autres propositions de cette question ne posent pas de problèmes particulièrement criants.

Q7. On élève une masse m à une hauteur h du sol et on l'abandonne librement. Déterminer l'expression littérale de la vitesse v lorsque la masse retombe sur le sol.

1. $v=\frac{1}{2}mv^2$
2. $v=\sqrt{2gh}$
3. $v=mgh$
4. $V=mc^2$
5. $V=\frac{(mh)}{g}$
6. Aucune
7. Toutes

D. Comparaisons des configurations des rpbis des questions sélectionnées

Précédemment nous avons analysé les propositions d'une question particulièrement performante, la [V]q36. Cette QCM obtient les meilleurs scores aux indices de cohérence spectrale et de cohérence interne classique (voir p. 343).

Ensuite nous avons analysé les propositions des 8 QCM épinglées pour leurs performances faibles en cohérence spectrale ou/et en cohérence interne (voir tableau des performances, p. 356).

Enfin, nous avons analysé les performances des propositions de la question [P]q7, la seule parmi les 173 questions des 10 épreuves MOHICAN qui présente des performances opposées aux indices spectraux *NCSq T80* et *Rq T80*.

Les performances spectrales et classiques des propositions de ces 10 questions ont été étudiées à l'aide des protocoles d'analyse des propositions générés par le programme *SCANTEST 2.0 pour épreuves MOHICAN* (p. 195). Nous allons maintenant nous focaliser sur la comparaison des configurations de quatre type de rpbis : d'une part l'indice classique de cohérence interne, le *rpbis classique,* et d'autre part, trois indices de cohérence spectrale, les indices *rpbis SC* (la cohérence spectrale calculée à partir des données de tous les étudiants quelque soit leur niveau de réalisme *Rs*), *rpbis SCT80* (calculée à partir des données des étudiants dont *Rs* est ⩾ à 80) et *rpbis SCT90* (*Rs* ⩾ 90). Nous comparerons également les diagnostics de la qualité des propositions qui ont été établis à l'aide des valeurs obtenues à ces indices avec les analyses du contenu des questions effectuées par les experts et parfois approfondies par nous-même.

1. Configuration des *rpbis* d'une question qui fonctionne particulièrement bien du point de vue de la cohérence spectrale et de la cohérence interne

Pour la question performante [V]q36 (p. 369) nous remarquons qu'à tous les paliers de turbo analyse les propositions incorrectes fonctionnent convenablement du point de vue de la cohérence spectrale (au sens où nous l'avons définie, p. 227) : les étudiants qui choisissent les distracteurs ont tendance à les accompagner de certitudes moins élevées que celles qui accompagnent les réponses correctes. La cohérence spectrale est également excellente pour la solution correcte : les étudiants qui la choisisse ont tendance à l'accompagner de pourcentages de certitude plus élevés.

L'indice classique de discrimination (*rpbis classique*) montre également pour chaque proposition un fonctionnement cohérent : les étudiants qui choisissent les propositions incorrectes sont ceux qui ont tendance à obtenir des scores moins élevés au total de l'épreuve. Par contre, les sujets qui répondent correctement, récoltent des scores totaux en moyenne plus élevés.

[V]q36	OM	**P1**	P2	P3
rpbis SC	-0,35	**0,41**	-0,22	-0,35
rpbis SCT80	-0,48	**0,58**	-0,28	-0,47
rpbis SCT90	-0,67	**0,70**	-0,28	-0,49
rpbis classique	-0,27	**0,51**	-0,16	-0,30

On observe donc pour le niveau d'analyse « propositions » de cette question [V]q36, une connexité des valeurs obtenues par les indices de cohérence spectrale et de discrimination classique. Nous remarquons aussi que les valeurs obtenues par les rpbis SCT80 et rpbis SCT90 sont plus marquées que celles du rpbis classique. En comparaison avec le rpbis classique, la valeur obtenue par le rpbis SC de la réponse correcte est moins élevée. En ce qui concerne les solutions incorrectes, les rpbis SC sont plus négatifs.

2. Comparaison des configurations de *rpbis* des 8 QCM dont les performances en cohérence spectrale ou/et en cohérence interne sont les plus faibles

a) [H]q3

En ce qui concerne [H]q3 (p. 373), nous constatons que la situation est radicalement différente de [V]q36. A T90, les corrélations peu élevées et les très faibles effectifs nous amènent à devoir considérer les *rpbis SCT90* des 6 propositions que compte la question comme étant non significativement différents de zéro (nous avons indiqué « *ns≠0* » dans les tableaux tels que celui ci-dessous après avoir appliqué la formule de Fisher de transformation de la valeur *r* en *t* de Student, voir détails p. 165).

A T80, les effectifs sont plus élevés et nous constatons que pour 5 propositions, dont la réponse correcte P6, les *rpbis SCT80* valent pratiquement zéro. Donc, d'une part nous n'observons pas de corrélations positives entre le choix de la réponse correcte et l'utilisation de pourcentages de certitude en moyenne plus élevés que ceux qui accompagnent les distracteurs (rpbis SCT80 de la réponse correcte P6 = -0,06). D'autre part pour les propositions incorrectes P2 à P4 nous ne remarquons pas de corrélations négatives. Pour la proposition incorrecte P1, nous observons une corrélation positive (0,38), les sujets ont donc eu tendance à accompagner cette proposition de certitudes plus élevées que celles qui accompagnent la réponse correcte, ce qui est anormal.

Les *rpbis SC* montrent également un fonctionnement spectral incohérent pour P1 (0,49) et P6 (-0,15), mais aussi pour P5 (0,19).

[H]q3	OM	P1	P2	P3	P4	P5	**P6**
rpbis SC	-0,14	0,49	0,05	0,07	-0,04	0,19	**-0,15**
rpbis SCT80	-0,27	0,38	0,00	0,02	-0,03	0,06	**-0,06**
rpbis SCT90	-0,70	-0,03 *ns≠0*	-0,18 *ns≠0*	-0,17 *ns≠0*	-0,14 *ns≠0*	-0,27 *ns≠0*	**0,15** *ns≠0*
rpbis classique	-0,21	0,36	-0,02	-0,10	-0,06	-0,17	**0,01**

Lorsque nous analysons les *rpbis classiques* nous remarquons que P1 récolte aussi une valeur positive (0,36), les étudiants qui ont choisi cette proposition incorrecte ont tendance à obtenir des scores totaux plus élevés que les étudiants qui répondent correctement. Pour deux autres propositions incorrectes, P2 et P4, nous remarquons un manque de liaison avec les scores totaux, de même pour la réponse correcte P6 (les *rpbis classiques* sont quasi égaux à zéro).

Quelle est la cause de ces incohérences de fonctionnement classique et spectral des propositions ? L'analyse du contenu de [H]q3 et les avis des experts présents lors de la réunion de débriefing de l'épreuve montrent que le choix de la proposition P1 par des étudiants qui en moyenne obtiennent plus de réponses correctes au total du test et qui sont aussi en moyenne plus sûrs de cette réponse incorrecte peut se comprendre si on se replace dans le contexte des examinés en octobre 1999 (voir détails p. 375). L'analyse du contenu de la question a montré que P1 était une solution très attractive pour les étudiants, le caractère plausible de P1 (étant donné le contexte) a joué en défaveur de P6, la réponse correcte attendue, qui elle n'a pas permis de discriminer les sujets.

Pour [H]q3, le caractère incorrect de la solution P1 a été remis en question par les experts étant donné le contexte des étudiants en octobre 1999 (détails p. 375). Cette anomalie liée à la première proposition a été détectée par les rpbis classiques, rpbis SC et rpbis SCT80. Nous constatons que pour la réponse correcte (P6) le rpbis SCT80 (-0,06) est plus proche de celui du rpbis classique (0,01) que ne l'est le rpbis SC (-0,15). Remarquons que le rpbis SCT80 de P1 est légèrement plus élevé (0,38) que le rpbis classique (0,36) et que P5 obtient un rpbis classique négatif (-0,17) alors que son rpbis SCT80 est proche de zéro (0,06). En ce qui concerne les rpbis SCT90, les corrélations peu élevées et les très faibles effectifs nous amènent à devoir considérer les valeurs récoltées par les 6 propositions comme étant non significativement différentes de zéro. Les rpbis SCT90 ne détectent donc pas les problèmes liés à cette question (ce que font les autres types de rpbis).

b) [H]q20

En ce qui concerne la seconde question suspecte [H]q20, nous constatons que les *rpbis SCT80* de la proposition correcte (P3) et de deux propositions incorrectes (P4 et P5) valent (quasi) zéro. Nous n'observons pas de liaison entre le choix des pourcentages de certitude moins élevées et le choix des propositions incorrectes P4 et P5 ni entre le choix de la réponse correcte P3 et des certitudes plus élevées, ce qui est anormal et témoigne d'un problème au niveau de la cohérence d'utilisation des certitudes.

Nous remarquons aussi un *rpbis SCT80* positif relativement élevé pour la proposition incorrecte P6, ce qui est aussi anormal (ceux qui ont choisi cette proposition incorrecte ont eu tendance à donner des certitudes plus élevées que ceux qui ont choisi la réponse correcte). Comme dans le cas de la question précédente ([H]q3), ce genre de situation nous interpelle car au palier de turbo analyse T80 nous utilisons les données d'étudiants particulièrement réalistes ($Rs \geqslant 80$), donc qui commettent peu d'erreurs (moins de 20%) dans leurs auto-estimations.

Seules deux propositions incorrectes P1 et P2 obtiennent des *rpbis SCT80* négatifs. En effet, les sujets dont le réalisme est supérieur ou égal à 80 et qui ont choisi P1 ou P2 ont eu tendance a accompagner leurs choix de certitudes plus faibles que ceux qui à ce niveau T80 ont choisi la réponse correcte P3. Parallèlement à l'incohérence spectrale observée pour P6, on constate donc une cohérence spectrale dans l'utilisation des pourcentages de certitude pour P1 et P2.

L'analyse des *rpbis classiques* de [H]q20 montre que pour trois propositions incorrectes (P2, P4 et P5) il n'existe pas de liaison avec les scores totaux. Ces propositions ne discriminent pas. Par contre P1 contribue à discriminer les sujets en fonction du nombre de réponses correctes fournies au total du test, le *rpbis classique* est égal à –0,18. En ce qui concerne la proposition P6 nous observons chez ceux qui la choisissent une légère tendance (0,14) à récolter des scores totaux plus élevés, ce qui est anormal pour une solution incorrecte (on s'attend à une corrélation négative). Enfin, pour la réponse correcte P3 nous remarquons une légère liaison positive avec les scores totaux (0,10), mais cette corrélation est inférieure à la valeur repère calculée pour contrer le problème du recouvrement entre le score de la question et le score total (voir p. 176) et qui vaut 0,2 dans le cas de l'épreuve qui contient cette question.

[H]q20	OM	P1	P2	**P3**	P4	P5	P6
rpbis SC	-0,20	-0,16	-0,10	**-0,17**	-0,03	-0,02	0,46
rpbis SCT80	xxxx	-0,24	-0,17	**-0,07**	-0,05	-0,06	0,40
rpbis SCT90	xxxx	-0,22 *ns≠0*	-0,18 *ns≠0*	**0,01** *ns≠0*	-0,13 *ns≠0*	0,03 *ns≠0*	0,35
rpbis classique	-0,07	-0,18	-0,06	**0,10**	0,00	-0,09	0,14

Pour les propositions P2, P3 et dans une certaine mesure pour P6, les valeurs obtenues par les *rpbis SCT80* et les *rpbis classiques* ne se rejoignent pas. Dans le cas de la proposition incorrecte P2, le choix de cette solution s'accompagne d'une faible propension à récolter des pourcentages de certitude moins élevés (*rpbis SCT80* de P2 = -0,17) mais ne s'accompagne pas d'une tendance à récolter des scores totaux plus faibles (*rpbis classique* de P2 quasi nul : -0,06).

Lors de l'analyse du contenu de cette question nous avons mis en évidence une erreur d'encodage de la réponse correcte (voir détails p. 380). En fait après analyse du contenu, il s'avère que la proposition qui doit être considérée comme correcte est P6 et non P3. L'analyse spectrale des propositions rejoint clairement cette analyse du contenu car lorsque les sujets dont le réalisme est élevé ($Rs \geqslant 80$) choisissent la proposition P6, ils ont tendance à accompagner ce choix de pourcentages de certitude élevés. Par contre du point de vue de la discrimination en fonction du nombre de réponses correctes fournies au total du test et calculée à l'aide du *rpbis classique*, la mise en évidence du problème est moins nette dans la mesure où d'une part le *rpbis classique* du distracteur P6 (0,14) se situe sous le seuil calculé pour contrer le problème du recouvrement de la question dans les scores totaux (ici ce seuil vaut 0,20) et où d'autre part le *rpbis classique* de la réponse correcte P3 est positif bien que sous le seuil de 0,20.

Pour [H]q20 nous constatons que le rpbis SCT80 met en évidence de façon plus nette que le rpbis classique l'erreur d'encodage de la réponse correcte. En effet, nous observons un rpbis SCT80 légèrement négatif (-0,07) pour la réponse P3 (considérée erronément correcte) et un rpbis SCT80 très positif (0,40) pour la proposition P6 (qui après analyse du contenu se révèle être la réponse correcte à cette question) alors que le rpbis classique de P3 est positif (0,10) et de P6 est peu élevé (0,14). La mise en évidence de l'erreur d'encodage liée à [H]q20 est bien plus claire encore à l'aide des rpbis SC qui sont plus marqués que les rpbis SCT80 (rpbis SC de P3 = -0,17 contre –0,07 pour le rpbis SCT80 et rpbis SC de P6 = 0,46 contre 0,40 pour le rpbis SCT80).

c) [V]q5

Voici la synthèse des quatre types de *rpbis* pour cette cinquième question du test de vocabulaire :

[V]q5	OM	P1	P2	P3	P4	P5	**P6**	P7
rpbis SC	-0,23	-0,03	0,09	-0,06	-0,05	0,04	**-0,01**	0,02
rpbis SCT80	-0,37	-0,15	-0,04	-0,11	-0,12	0,01	**0,08**	-0,02
rpbis SCT90	xxxx	-0,18	-0,06 *ns≠0*	-0,09 *ns≠0*	-0,15	-0,01 *ns≠0*	**0,13**	0,01 *ns≠0*
rpbis classique	-0,15	-0,07	0,01	-0,07	-0,10	0,12	**0,15**	0,02

Comme nous l'avons souligné précédemment (p. 373), le distracteur P5 de cette QCM pose un problème de discrimination des étudiants en fonction de leurs performances globales au test. En effet, le *rpbis classique* positif (0,12), bien que légèrement sous le seuil de 0,15 ($1/\sqrt{45}$), indique une légère tendance à choisir P5 chez les étudiants qui obtiennent un nombre de réponses correctes plus élevés au total du test.

L'analyse du contenu de la question a montré que les propositions ne comportaient pas d'anomalies. En ce qui concerne P5, seule une similitude phonologique entre « latente » et « patente » pourrait expliquer le fait qu'un certain nombre d'étudiants obtenant plus de réponses correctes au total du test aient choisi cette proposition incorrecte (voir détails, p. 383). Ceci dit, les *rpbis SC*, *rpbis SCT80* et *rpbis SCT90* n'indiquent pas quant à eux que P5 soit particulièrement problématique, les valeurs obtenues à ces rpbis spectraux étant proches de zéro.

Les avis des experts et l'analyse du contenu de [V]q5 ne mettent pas d'anomalie particulière en évidence. Dès lors, si nous considérons que la question est « bonne », le rpbis classique positif de la proposition incorrecte P5 provoque une « fausse alerte » (idéalement, la valeur du rpbis classique devrait être négative, ce qui n'est pas le cas ici). Lorsque nous analysons les rpbis spectraux de P5 nous remarquons qu'ils sont tous très proches de zéro. Ils ne montrent pas d'utilisation incohérente des certitudes (les corrélations obtenues sont inférieures à 0,05). Si cette proposition incorrecte P5 obtenait des rpbis spectraux positifs élevés on devrait alors conclure que ceux-ci déclenchent aussi une « fausse alerte » en contradiction avec l'analyse du contenu, mais ce n'est pas le cas. Nous sommes donc devant un cas de figure où pour une proposition incorrecte (P5), les valeurs observées aux rpbis spectraux sont plus proches des conclusions des experts et de l'analyse du contenu que la valeur obtenue par le rpbis classique. En ce qui concerne la réponse correcte P6, c'est le rpbis classique qui est le plus marqué (0,15) et le rpbis SCT90 qui est le plus proche de cette valeur (0,13). Cette convergence des analyses classique et spectrale lorsqu'on prend en compte les sujets les plus réalistes tend à montrer que la réponse correcte fonctionne relativement bien du point de vue de la discrimination « classique » et de la cohérence spectrale. Ce constat à propos de P6 est en accord avec le jugement des experts qui estiment que la QCM ne contient pas d'anomalie majeure.

d) [S]q1

L'analyse de la question lors de la réunion de débriefing n'a pas permis de mettre en évidence un problème particulier lié à [S]q1. Selon les experts consultés lors de la réunion de débriefing de l'épreuve, les propositions de la première question du test de syntaxe ne contiennent pas d'anomalies.

Calculés sur la base des données des étudiants les plus réalistes à T80 et à T90, les rpbis spectraux ne révèlent pas non plus de problèmes particuliers.

[S]q1	OM	P1	P2	P3	P4	P5	**P6**	P7
rpbis SC	-0,17	-0,03	-0,06	0,06	0,00	0,13	**-0,08**	0,01
rpbis SCT80	-0,28	-0,13	-0,11	-0,09	-0,06	-0,09	**0,14**	-0,06
rpbis SCT90	xxxx	-0,19	-0,12 *ns≠0*	-0,14	-0,12 *ns≠0*	-0,27	**0,33**	-0,12 *ns≠0*
rpbis classique	-0,07	-0,15	-0,11	-0,15	-0,06	-0,16	**0,38**	-0,03

Lorsqu'on compare les *rpbis SCT80* et *rpbis SCT90* avec les *rpbis classiques*, on constate que les résultats des analyses convergent.

En ce qui concerne les informations liées à la cohérence spectrale recueillies aux paliers turbo inférieurs à T70 (et donc aussi les *rpbis SC*) nous constatons qu'elles ne permettent pas de confirmer les analyses précédentes, les valeurs des *rpbis SC* étant toutes proches de zéro sauf pour le distracteur P5 qui est légèrement positif (0,13).

Pour [S]q1, les avis des experts ainsi que les informations livrées par les rpbis classiques, les rpbis SCT80 et rpbis SCT90 convergent vers un même constat : la question ne comporte pas d'anomalie particulière. Par contre, les rpbis SC ne permettent pas de confirmer ce diagnostic dans la mesure où la valeur récoltée par P5 est légèrement positive (ce qui déclenche une fausse alerte) et celles obtenues par les autres propositions (y compris la réponse correcte) sont très proches de zéro.

e) [V]q12

D'un point de vue spectral l'utilisation des pourcentages de certitude est assez cohérente lorsqu'on prend en compte les *rpbis SCT80* et *rpbis SCT90* calculés à l'aide des données des étudiants les plus réalistes.

[V]q12	OM	P1	P2	P3	P4	P5	**P6**	P7
rpbis SC	-0,25	-0,11	-0,07	-0,10	0,09	-0,10	**0,05**	-0,03
rpbis SCT80	-0,38	-0,18	-0,17	-0,18	0,01	-0,16	**0,15**	-0,06
rpbis SCT90	xxxx	-0,15 *ns≠0*	-0,27	-0,29	-0,12 *ns≠0*	-0,19	**0,29**	-0,06 *ns≠0*
rpbis classique	-0,19	-0,13	-0,08	-0,06	-0,01	0,03	**0,29**	-0,05

Du point de vue de la discrimination basée sur le critère du nombre de réponses correctes récoltées au total du test, les *rpbis classiques* montrent que la question fonctionne relativement bien : le *rpbis classique* de la réponse correcte est assez marqué (0,29) et les valeurs sont négatives ou proches de zéro pour les propositions incorrectes.

Par contre, en ce qui concerne les *rpbis SC*, nous remarquons que la valeur de la proposition correcte est quasi égale à zéro (0,05) et que celles récoltées par les propositions incorrectes sont proches de zéro ou peu marquées.

Pour [V]q12, les avis des experts ainsi que les informations livrées par les rpbis classiques, les rpbis SCT80 et rpbis SCT90 convergent vers un même constat : la question ne comporte pas d'anomalie particulièrement grave. Par contre, les rpbis SC ne permettent pas de confirmer ce diagnostic dans la mesure où les valeurs récoltées par toutes les propositions (y compris la réponse correcte) sont très proches de zéro. D'où une fausse alerte du rpbis SC en ce qui concerne la réponse correcte P6 dont la valeur n'est pas assez élevée.

f) [V]q27

Les experts présents lors de la réunion de débriefing de l'épreuve n'ont pas détecté d'anomalie particulière liée à la 27ème question du test de vocabulaire .

[V]q27	OM	P1	P2	**P3**
rpbis SC	-0,27	0,14	-0,06	**0,04**
rpbis SCT80	-0,37	0,12	-0,13	**0,13**
rpbis SCT90	-0,61	-0,03 *ns≠0*	-0,21	**0,33**
rpbis classique	-0,24	0,13	-0,09	**0,13**

Nous remarquons que le *rpbis classique* de la proposition incorrecte P1 est positif (0,13) et que celui de la réponse correcte P3 (0,13) se situe sous le seuil calculé pour contrer le problème du recouvrement de la question dans le total du test ($\sqrt{1/45} = 0{,}15$). Quant à P2, son *rpbis classique* est proche de zéro (-0,09).

La configuration des *rpbis SCT80* est très proche de la configuration des *rpbis classiques*. En ce qui concerne les *rpbis SC*, nous remarquons une différence au niveau de la valeur récoltée par la réponse correcte : le *rpbis SC* de P3 est quasi égal à zéro (0,04).

Par contre, les *rpbis SCT90* montrent une situation beaucoup plus proche de l'avis des experts dans la mesure où P1 récolte un *rpbis SCT90* non significativement différent de zéro (*ns≠0*) ainsi qu'un *rpbis SCT90* pour la réponse correcte plus élevé (0,33) que dans le cas des autres types de rpbis. Remarquons aussi que c'est le *rpbis SCT90* qui est le plus marqué pour la seconde proposition incorrecte P2 (-0,21).

Pour [V]q27 les experts qui ont analysé le contenu des propositions de cette question n'ont pas détecté d'anomalie majeure. Dès lors, les rpbis classiques, rpbis SC et rpbis SCT80 déclenchent de fausses alertes en signalant la solution incorrecte P1 comme étant problématique (valeur positive alors qu'il s'agit d'une réponse incorrecte). Par contre, les rpbis SCT90, eux, sont en accord avec l'analyse du contenu effectuée par les experts dans la mesure où ils indiquent une corrélation qui n'est plus positive pour P1, une autre plus négative pour P2 ainsi qu'une valeur positive plus marquée pour P3 (la réponse correcte).

g) [A]q14

Les experts présents lors de la réunion de débriefing de l'épreuve de Connaissance artistiques considèrent cette 14ème question du test comme étant mal construite. L'analyse du contenu des propositions montre en effet d'une part que la proposition P3, considérée au départ comme étant correcte, ne l'est qu'à moitié et d'autre part que le distracteur P4 est quant à lui à moitié incorrect (voir détails, p. 390).

[A]q14	OM	P1	P2	**P3**	P4	P5	P6	P7
rpbis SC	-0,30	0,04	0,03	**0,00**	0,20	0,01	-0,04	0,04
rpbis SCT80	xxxx	-0,04 *ns≠0*	-0,01 *ns≠0*	**0,11**	0,10 *ns≠0*	-0,05 *ns≠0*	-0,10	-0,01 *ns≠0*
rpbis SCT90	xxxx	-0,20 *ns≠0*	-0,08 *ns≠0*	**0,32**	-0,13 *ns≠0*	-0,28	-0,23 *ns≠0*	-0,06 *ns≠0*
rpbis classique	-0,18	0,03	-0,01	**0,09**	0,08	0,01	-0,06	0,02

En résumé, pour cette question [A]q14 :

- les *rpbis SC* indiquent que lorsqu'on prend en compte les données de tous les étudiants, il n'existe pas de tendance à utiliser des certitudes plus élevées pour la réponse correcte et des certitudes en moyenne moins élevées pour les distracteurs P1, P2, P5, P6 et P7. De même en ce qui concerne la proposition correcte P3, ce qui est anormal. Pour P4 on remarque une tendance à utiliser des certitudes plus élevées (*rpbis SC* = 0,20) ce qui pour une proposition incorrecte n'est pas cohérent d'un point de vue spectral. Il y a donc détection des propositions problématiques ;

- les *rpbis classiques* indiquent que la proposition correcte et les distracteurs ne discriminent guère les étudiants lorsqu'on prend en compte le critère du nombre de réponses correctes obtenues au total du test. Les valeurs sont toutes assez proches de zéro. Il n'y a pas de détection des propositions problématiques P3 et P4 ;

- les *rpbis SCT80* montrent qu'il existe une légère tendance chez les étudiants qui répondent correctement à accompagner P3 de certitudes plus élevées (*rpbis SCT80* de P3 = 0,11). Les distracteurs sont non significativement différents de zéro ou légèrement négatifs. Il n'y a donc pas de détection claire des propositions problématiques P3 et P4 ;

- les *rpbis SCT90* indiquent d'une part de façon plus marquée que la proposition correcte fonctionne logiquement du point de vue de la cohérence spectrale (*rpbis SCT90* de P3 = 0,32) et d'autre part qu'il n'existe pas de tendance à accompagner les distracteurs de certitudes plus faibles (*rpbis SCT90* non significativement différents de zéro) sauf pour P5 (-0,28). Les propositions problématiques P3 et P4 ne sont donc pas détectées.

Pour cette question, <u>seuls les rpbis SC parviennent à mettre en évidence les deux propositions P3 et P4 que les experts désignent comme étant problématiques</u>. Les rpbis classiques ainsi que les rpbis SCT80 et rpbis SCT90 ne détectent pas clairement ces deux propositions comme étant suspectes. Donc, dans le cas de [A]q14, seuls les rpbis SC convergent avec les analyses réalisées sur base des avis des experts et de l'analyse approfondie du contenu.

h) [B]q5

Les experts présents lors de la réunion de débriefing de l'épreuve ont signalé que [B]q5 ne comportait pas d'anomalie particulière si ce n'est qu'elle concerne une matière peu enseignée. Selon les experts, le choix de la 1ère proposition par un grand nombre d'étudiants (47%, alors que la réponse correcte est choisie par 21%) témoigne d'une confusion entre « niche » et « habitat » (voir détails p. 392). Cette question concerne un concept assez récent en biologie et qui est encore peu enseigné dans l'enseignement secondaire.

[B]q5	OM	P1	P2	**P3**	P4	P6	P7
rpbis SC	-0,26	0,05	-0,08	**0,05**	-0,08	-0,09	-0,02
rpbis SCT80	-0,47	-0,24	-0,32	**0,29**	-0,26	-0,21	-0,10
rpbis SCT90	xxxx	-0,42	-0,47	**0,40**	-0,33	-0,35	-0,07
rpbis classique	-0,13	0,02	-0,10	**0,24**	-0,12	-0,05	0,00

Le *rpbis classique* de P3 est positif (0,24) mais inférieur à la valeur repère (0,32). En ce qui concerne les distracteurs, deux propositions obtiennent des *rpbis classiques* très proches de zéro : P1 (0,02), P6 (-0,05) et une est égale à zéro : P7 (0,00).

La situation est plus tranchée en ce qui concerne les *rpbis SCT80* et *rpbis SCT90* dont les valeurs sont d'une part négatives et plus marquées pour les distracteurs et d'autre part positives et plus élevées pour la réponse correcte.

Par contre en ce qui concerne les *rpbis SC* calculés sur la base des données de tous les étudiants, (qu'ils commettent peu ou beaucoup d'erreurs dans leurs auto-estimations) nous remarquons que les valeurs obtenues par les propositions sont toutes proches de zéro. Cette configuration de *rpbis SC* nous semble donc plus proche des avis des experts que les autres.

Dans le cas de [B]q5 les experts considèrent que cette question tout en étant valable ne devrait pas figurer dans ce questionnaire car elle recouvre un concept encore peu enseigné dans l'enseignement secondaire. Lorsqu'on analyse les différents types de rpbis, on remarque que le rpbis classique de la réponse correcte (P3) déclenche une fausse alerte en récoltant une valeur positive mais située sous le seuil repère. Les autres rpbis classiques sont proches de zéro ou peu marqués. En ce qui concerne les rpbis SC nous constatons qu'ils sont tous très proches de zéro, donc, lorsque les données de tous les étudiants sont prises en compte, ceux qui répondent correctement ne sont pas plus sûr de leur réponse que ceux qui répondent incorrectement. Les rpbis SCT80 et rpbis SCT90 sont quant à eux plus marqués et indiquent une situation normale.

3. Configuration des *rpbis* de [P]q7, la seule question qui obtient des valeurs opposées aux indices de cohérence spectrale

Nous avons repris cette question dans les analyses du niveau « proposition » car est la seule parmi les 173 QCM des 10 épreuves MOHICAN à obtenir des valeurs contradictoires aux indices spectraux *NCSq T80* et *Rq T80.*

Comme nous l'avons évoqué précédemment (p. 394), lors de la réunion de débriefing de l'épreuve, les experts ont souligné la difficulté de [P]q7 (24% de réussites) et la forte attractivité de la proposition P2 (choisie par 48% des étudiants). Par ailleurs, les experts n'ont pas décelé d'anomalie grave dans cette question.

[P]q7	OM	P1	**P2**	P3	P4	P5	P6	P7
rpbis SC	-0,31	-0,12	**0,26**	-0,22	-0,12	-0,26	-0,24	-0,05
rpbis SCT80	0,04	-0,18	**0,45**	-0,40	-0,16	-0,25	-0,38	-0,06
rpbis SCT90	xxxx	-0,22	**0,61**	-0,44	-0,17	-0,26	-0,40	xxxx
rpbis classique	-0,14	-0,08	**0,42**	-0,20	-0,09	-0,12	0,01	-0,05

Les différents rpbis montrent que cette question ne pose pas de problème particulier du point de vue de la cohérence interne ni de la cohérence spectrale.

Pour [P]q7, les rpbis classiques ainsi que les rpbis spectraux ne permettent pas de relever des problèmes particuliers au niveau des propositions. Lors de la réunion de débriefing de l'épreuve, les experts n'ont pas signalé d'anomalie majeure dans cette question. Les valeurs récoltées aux différents types de rpbis convergent avec les avis des experts.

E. Qualité des propositions de huit autres questions épinglées pour leurs configurations de rpbis anormales

Précédemment, parmi les 173 questions des 10 épreuves MOHICAN nous avons mis en évidence huit questions dont les performances globales en cohérence interne et en qualité spectrale analysées au niveau « QCM » se sont révélées faibles (voir tableau récapitulatif p. 356). Dans le cadre de cette exploration du niveau « PROPOSITIONS » nous avons aussi systématiquement examiné les valeurs des rpbis classiques, rpbis SC, rpbis SCT80 et rpbis SCT90 des autres questions des 10 épreuves (voir annexe, p. 544). Suite à cette analyse nous avons épinglé huit autres QCM ([V]q43, [B]q6, [G]q1, [A]q24, [H]q16, [M]q17, [B]q10 et [Ch]q1) qui présentent des valeurs anormales aux rpbis. L'analyse des propositions de ces huit autres questions permet-elle de mettre en évidence des problèmes ?

1. Configurations des *rpbis* de [V]q43 et [B]q6, deux questions aux indices de cohérence interne $r_{qt\ mb}$ et $r_{qt\ ms}$ relativement faibles

Répartition des 173 QCM MOHICAN en fonction de la qualité spectrale et de la cohérence interne des questions

Cohérence interne \ Qualité spectrale	++	+≃	≃≃	-≃	--	+-
++	39	19	-	-	-	
+≃	20	19	1	-	-	
≃≃	21	26	2	2	2	1
-≃	2	8	3	2	1	
--	-	2	1	-	2	

[V]q43 & [B]q6

Dans le tableau de répartition des 173 questions en fonction de leur qualité spectrale et de leur cohérence interne (voir p. 352), les deux QCM [V]q43 et [B]q6 occupent la case que nous avons mise en évidence ci-contre. Ces deux questions récoltent donc des valeurs « - » aux deux indices de cohérence interne ($r_{qt\ mb}$ et $r_{qt\ ms}$) et un « ≃ » ainsi qu'un « + » aux indices de qualité spectrale (*NCSq* et *Rq*) (voir tableau p. 350).

a) [V]q43

Lorsque nous examinons les configurations des rpbis de [V]q43 présentés ci-dessous (extrait du protocole en annexe, p. 533), nous remarquons à l'aide des *rpbis classiques* que les propositions ne discriminent pas les étudiants lorsqu'on prend en compte le critère du nombre de réponses correctes récoltées à l'ensemble des questions du test alors que les configurations des *rpbis spectraux* ne montrent pas d'incohérence dans l'utilisation des pourcentages de certitude.

[V]q43	OM	P1	P2	**P3**	P4	P5	P6	P7
rpbis SC	-0,31	0,04	0,00	**0,11**	-0,12	-0,08	-0,19	0,00
rpbis SCT80	-0,42	-0,04	-0,02	**0,21**	-0,19	-0,11	-0,24	-0,04
rpbis SCT90	-0,61	-0,07	-0,10	**0,32**	-0,27	-0,13	-0,29	-0,09
rpbis classique	-0,21	0,06	0,03	**0,09**	-0,03	-0,05	-0,05	0,02

Les experts consultés lors de la réunion de débriefing n'ont pas trouvé d'anomalie dans le contenu de la question (question en annexe, p. 486).

En ce qui concerne cette QCM [V]q43, <u>les rpbis classiques déclenchent une fausse alerte</u> dans la mesure où ils indiquent une situation anormale pour la réponse correcte (la valeur du rpbis classique de P3 se situe sous le seuil de la valeur repère) alors que de l'avis des experts la question est « bonne ». Par contre, les rpbis spectraux offrent une image de la qualité des propositions qui est en phase avec le diagnostic des experts.

b) [B]q6

En ce qui concerne [B]q6 (protocole en annexe, p. 534), la configuration des *rpbis classiques* montre des valeurs négatives ou proches de zéro pour les distracteurs, et, une valeur positive mais inférieure à la valeur du repère (0,32) pour la réponse correcte. Les *rpbis SC* sont peu marqués, par contre les valeurs des *rpbis SCT80* et *rpbis SCT90* indiquent une situation tout à fait normale du point de vue de la cohérence spectrale.

[B]q6	OM	P1	**P2**	P3	P4	P6	P7
rpbis SC	-0,27	-0,07	**0,06**	0,09	-0,12	-0,06	-0,07
rpbis SCT80	xxxx	-0,22	**0,38**	-0,13	-0,24	-0,17	-0,11
rpbis SCT90	xxxx	-0,28	**0,53**	-0,21	-0,31	-0,30	-0,20
rpbis classique	-0,18	-0,16	**0,29**	-0,05	-0,11	-0,07	-0,07

En ce qui concerne l'analyse du contenu de la question (annexe, p. 506), les experts ont indiqué lors de la réunion de débriefing de l'épreuve que la proposition P2 (considérée comme correcte) est en lien avec la thèse du lamarckisme (action modelante du milieu sur les caractéristiques des individus) tandis que P3 est plutôt liée au darwinisme (extension par sélection naturelle des individus d'une population porteurs d'un caractère avantageux pour elle). Dès lors, pour les tenants de la thèse darwiniste, P3 pourrait être considérée comme correcte. En ce qui concerne P2, nous constatons que du point de vue des *rpbis SCT80* et *rpbis SCT90*, cette proposition fonctionne normalement lorsqu'on la suppose correcte. Pour ce qui est de P3, nous remarquons que lorsqu'on la présente comme une proposition incorrecte, du point de vue des *rpbis SCT80* et *rpbis SCT90,* elle fonctionne normalement (les étudiants qui ont choisi P3 ont eu tendance à accompagner cette proposition de certitudes moins élevées que ceux qui ont choisi P2). Par contre, du point de vue des *rpbis SC* nous remarquons que les différentes propositions sont assez proches de zéro et que le *rpbis SC* de P3 (0,09, $r \neq 0$ à p 0,001) est légèrement plus élevé que celui de P2 (0,06, $r \neq 0$ à p 0,01). Du point de vue des *rpbis classiques*, nous constatons que la réponse correcte P2 récolte une valeur positive (0,29) mais située légèrement sous le seuil de la valeur repère ($1/\sqrt{10} = 0{,}32$).

Pour cette question [B]q6, si nous considérons que P2 et P3 posent problème dans la mesure où l'une et l'autre pourraient être considérées comme étant correcte (en fonction de la thèse que l'on défend, le lamarckisme ou le darwinisme) alors seuls les rpbis SC signalent un problème pour P2 et P3 (valeurs proches de zéro). Le rpbis classique ne détecte pas P3. Les rpbis SCT80 et rpbis SCT90 ne détectent pas de problèmes (ni dans P2, ni dans P3).

2. Configurations des *rpbis* de cinq questions dont les rpbis classiques des réponses correctes sont positifs mais inférieurs à la valeur repère

Nous avons remarqué dans le tableau récapitulatif des rpbis des 173 QCM (voir annexe, p. 544), les questions [G]q1, [A]q24, [H]q16, [B]q10 et [Ch]q1 qui ont en commun des configurations de *rpbis classiques* assez similaires : la réponse correcte obtient une valeur positive mais qui se situe sous le seuil de la valeur repère et les distracteurs sont proches de zéro.

a) [G]q1

En ce qui concerne [G]q1, la valeur repère vaut 0,32 ($1/\sqrt{10}$) et le *rpbis classique* de la réponse correcte (P4) vaut 0,20. Comme le montre le tableau récapitulatif ci-dessous (protocole détaillé en annexe, p. 535), cette faiblesse du *rpbis classique* de la réponse correcte, est accompagnée par des *rpbis classiques* proches de zéro pour les distracteurs, la question discrimine donc peu les étudiants du point de vue du nombre de réponses correctes fournies à l'ensemble des questions du test.

[G]q1	OM	P1	P2	P3	**P4**	P5	P6	P7
rpbis SC	-0,46	-0,09	-0,15	-0,13	**0,17**	-0,10	-0,15	-0,13
rpbis SCT80	-0,60	-0,27	-0,28	-0,26	**0,33**	-0,26	-0,45	-0,23
rpbis SCT90	-0,68	-0,30	-0,29	-0,34	**0,39**	-0,30	-0,59	-0,18
rpbis classique	-0,19	0,03	-0,07	-0,04	**0,20**	-0,02	0,05	-0,07

Ceci dit, lorsqu'on examine la ventilation des effectifs dans les différentes propositions (voir tableau ci-dessous), on remarque que la question est très difficile (12% des étudiants ont choisi la réponse correcte) et qu'un grand nombre de sujets ont opté pour la réponse P6 « aucune » (46%).

[G]q1	OM	P1	P2	P3	**P4**	P5	P6	P7
N Rép.	399	439	226	223	**444**	196	1698	63
% Rép.	11%	12%	6%	6%	**12%**	5%	46%	2%

Lors de la réunion de débriefing, l'analyse du contenu n'a pas révélé d'anomalie dans la question (voir annexe, p. 490).

Etant donné les avis des experts qui considèrent que la question ne comporte pas d'anomalie et les valeurs récoltées par les rpbis spectraux, nous pensons que les rpbis classiques de [G]q1 reflètent moins bien le diagnostic des experts que les rpbis spectraux. Dans le cas de cette question, le rpbis classique de la réponse correcte situé sous la valeur repère constitue une fausse alerte.

b) [A]q24

Pour ce qui est de [A]q24, la valeur repère du *rpbis classique* vaut 0,2 ($1/\sqrt{25}$). La réponse correcte vaut 0,16 et se situe donc sous ce seuil. Comme pour la question précédente, cette faiblesse du *rpbis classique* de la réponse correcte est accompagnée de *rpbis classiques* proches de zéro pour les distracteurs (protocole en annexe, p. 537).

[A]q24	OM	**P1**	P2	P3	P4	P5	P6	P7
rpbis SC	-0,33	**0,13**	-0,10	-0,06	-0,10	0,03	-0,14	0,01
rpbis SCT80	-0,56	**0,28**	-0,20	-0,26	-0,26	-0,13	-0,27	-0,03
rpbis SCT90	xxxx	**0,40**	-0,32	-0,40	-0,52	-0,17	-0,30	-0,18
rpbis classique	-0,16	**0,16**	-0,03	0,05	0,00	-0,03	0,00	-0,04

Nous remarquons qu'en ce qui concerne les rpbis spectraux, les valeurs obtenues par les *rpbis SC*, *rpbis SCT80* et *rpbis SCT90* ne montrent pas d'incohérence spectrale.

Les experts consultés lors de la réunion de débriefing n'ont pas relevé d'anomalie dans cette question. Dès lors, nous considérons, comme dans le cas de la QCM précédente, que le rpbis classique de la réponse correcte sous la valeur repère constitue une fausse alerte. Pour cette question, les valeurs observées aux rpbis spectraux sont plus en phase avec les avis des experts.

c) [H]q16

En ce qui concerne [H]q16, le tableau récapitulatif des différents types de rpbis (voir protocole en annexe, p. 536) est repris ci-dessous. Le seuil repère pour les *rpbis classiques* vaut 0,2 ($1/\sqrt{25}$).

[H]q16	OM	P1	P2	P3	**P4**	P5	P6
rpbis SC	-0,27	0,19	-0,03	-0,05	**0,02**	-0,04	-0,13
rpbis SCT80	-0,44	0,03 *ns≠0*	-0,09 *ns≠0*	-0,14	**0,16**	-0,12 *ns≠0*	-0,20
rpbis SCT90	xxxx	-0,11 *ns≠0*	-0,16 *ns≠0*	-0,27 *ns≠0*	**0,29**	-0,14 *ns≠0*	-0,32 *ns≠0*
rpbis classique	-0,12	0,06	-0,07	-0,07	**0,17**	0,00	-0,08

Pour ce qui est de la configuration des *rpbis classiques*, comme pour [G]q1 et [A]q24, les distracteurs sont tous proches de zéro et la réponse correcte récolte une valeur positive (0,17) mais sous le seuil repère (0,2).

Au palier turbo T80, les rpbis spectraux des distracteurs P1, P2, et P5 ne sont pas significativement différents de zéro (*ns≠0*). A T90 aucun distracteur n'est significativement différent de zéro. Dans le cas des *rpbis SC*, lorsqu'on prend en compte les données de tous les étudiants quel que soit leur réalisme, en plus des valeurs proches de zéro pour P2, P3 et P5, nous remarquons que P1 présente une incohérence spectrale. Le *rpbis SC* positif de P1 (0,19) montre que les sujets ont utilisé des pourcentages de certitude plus élevés pour ce distracteur que pour la réponse correcte dont la valeur du *rpbis SC* est quasi égale à zéro (0,02).

Comme le montre le tableau ci-dessous, P1 est aussi très attractive, 29% des étudiants l'ont choisie alors que la réponse correcte P4 est choisie par 23%.

[H]q16	OM	P1	P2	P3	**P4**	P5	P6
N Rép.	108	405	221	137	**328**	79	131
% Rép.	8%	29%	16%	10%	**23%**	6%	9%

Nous rappelons ci-contre le contenu de cette question (en annexe, p. 495). Pour les experts présents lors de la réunion de débriefing, P4, d'un point de vue strictement historique, est bien la réponse correcte mais beaucoup trop complexe car le sens s'est généralisé, ce que le *rpbis SC* positif (0,19) et la forte attractivité (29%) de P1 confirment. Selon les experts, « P1 est acceptable ».

Q16. Iconoclaste fait référence à :

1. Quelqu'un qui détruit les images pieuses, les statues… à l'époque de la Réforme protestante au 16ème Siècle
2. Un adorateur d'Icônes dans le monde de la religion orthodoxe depuis le 11ème Siècle
3. Un Vandale, appartenant à ce peuple, lors des grandes invasions qui détruisirent l'Empire Romain d'Occident au 5ème Siècle et qui détruisait tout sur son passage
4. Un partisan d'une idéologie née dans l'empire byzantin au 9ème Siècle et qui refuse l'adoration des images pieuses, allant jusqu'à les détruire
5. Un destructeur des peintures religieuses dans les églises russes durant la période communiste (XXème Siècle)
6. Aucune

La référence qui est faite à la Réforme protestante dans P1 peut aussi nous aider à comprendre l'attractivité de cette proposition. Le désir des protestants de revenir à une église purifiée et leur réaction contre la politique de grandeur de Rome ayant amené une tendance à l'iconoclasme chez les réformés au 16ème Siècle.

Donc, <u>seuls les rpbis SC détectent un problème au niveau du distracteur P1 que les experts considèrent « acceptable »</u>. Le rpbis SC de P4 étant pratiquement à zéro (alors qu'on s'attend à une valeur positive élevée pour une réponse correcte) déclenche ce que nous appelons une « fausse alerte » dans la mesure où les experts ne contestent pas cette proposition. On remarque aussi une fausse alerte pour la réponse correcte en ce qui concerne le rpbis classique dont la valeur est située sous le seuil repère de 0,2. Enfin, signalons que dans le cadre de cette question [H]q16 les rpbis SCT80 et rpbis SCT90 ne détectent pas d'anomalie au niveau de la proposition P1.

d) [B]q10

Pour ce qui est de la quatrième question, [B]q10, le tableau récapitulatif ci-dessous (extrait du protocole en annexe, p. 538) montre pour le *rpbis classique* de la réponse correcte une valeur positive (0,18) située sous le seuil repère ($1/\sqrt{10} = 0{,}32$) et pour les distracteurs des valeurs toutes proches de zéro.

[B]q10	OM	P1	P2	**P3**	P4	P5	P6	P7
rpbis SC	-0,20	-0,09	-0,03	**0,23**	-0,07	-0,05	-0,08	-0,06
rpbis SCT80	xxxx	xxxx	xxxx	**0,50**	-0,17	-0,14	xxxx	-0,10
rpbis SCT90	xxxx	xxxx	xxxx	**0,62**	-0,21	xxxx	xxxx	xxxx
rpbis classique	-0,13	0,06	-0,05	**0,18**	-0,06	-0,05	-0,04	-0,05

En ce qui concerne les *rpbis spectraux*, les configurations des valeurs se présentent normalement.

Voici la ventilation des effectifs dans les différentes propositions, les chiffres du tableau montrent l'extrême facilité de cette question.

[B]q10	OM	P1	P2	**P3**	P4	P5	P6	P7
N Rép.	23	8	11	**2433**	8	11	5	8
% Rép.	0,9%	0,3%	0,4%	**97%**	0,3%	0,4%	0,2%	0,3%

Selon les experts présents lors de la réunion de débriefing, le taux de réussite très élevé de cette QCM pourrait s'expliquer par le fait que le contenu (lié à la prévention du SIDA, voir question en annexe p. 506) est largement relayé par les médias. Le pourcentage élevé de sujets (97%) qui ont choisi la proposition P3 intervient dans l'explication de la valeur peu élevée du *rpbis classique* de la réponse correcte pour cette question qui ne contient pas d'anomalie dans son contenu.

Pour cette question [B]q10, nous considérons que les rpbis spectraux reflètent mieux la qualité de la question que le rpbis classique qui déclenche une fausse alerte au niveau de la réponse correcte (valeur située sous le seuil repère).

e) [Ch]q1

Enfin, en ce qui concerne [Ch]q1, le tableau des configurations des différents types de rpbis montre que le *rpbis classique* de la réponse correcte (P4) vaut 0,32, valeur située sous le seuil repère qui vaut 0,35 ($1/\sqrt{8}$), tandis que pour trois distracteurs les valeurs sont proches de zéro : P1 = -0,06, P2 = -0,1 et P5 = -0,08. Pour deux autres distracteurs, les *rpbis classiques* sont un peu plus marqués : P3 = -0,16 et P6 = -0,19.

[Ch]q1	OM	P1	P2	P3	**P4**	P5	P6	P7
rpbis SC	-0,08	-0,07	-0,10	-0,07	**0,23**	-0,06	-0,20	0,02
rpbis SCT80	-0,25	-0,14	-0,19	-0,21	**0,51**	-0,06	-0,37	xxxx
rpbis SCT90	xxxx	-0,19	xxxx	-0,24	**0,70**	-0,19	-0,44	xxxx
rpbis classique	-0,13	-0,06	-0,10	-0,16	**0,32**	-0,08	-0,19	-0,02

Nous remarquons que les valeurs récoltées par les *rpbis spectraux* montrent une bonne cohérence dans l'utilisation des pourcentages de certitude.

Les experts ont souligné la facilité de cette question (voir annexe, p. 507) dont l'analyse du contenu n'a pas révélé d'anomalie.

Comme le montre le tableau ci-dessous, le pourcentage élevé d'étudiants (89,7%) qui ont choisi P4 intervient dans l'explication du *rpbis classique* de la réponse correcte (0,32) situé légèrement sous la valeur repère (0,35).

[Ch]q1	OM	P1	P2	P3	**P4**	P5	P6	P7
N Rép.	16	32	18	81	**2244**	26	82	2
% Rép.	0,6%	1,3%	0,7%	3,2%	**89,7%**	1%	3,3%	0,1%

Pour cette question [Ch]q1, les rpbis SCT80 et les rpbis SCT90 sont plus en phase avec les avis des experts que le rpbis classique qui déclenche une fausse alerte au niveau de la réponse correcte P4 (valeur située sous le seuil repère).

3. Configuration des *rpbis* de [M]q17

La 17ème question du test de mathématique présente une configuration de *rpbis classiques* particulière dans la mesure où un des distracteurs (P5) récolte une valeur positive (0,18). Contrairement aux questions épinglées précédemment dans cette section, la valeur du *rpbis classique* de la réponse correcte (0,37) est supérieure à la valeur repère ($1/\sqrt{22} = 0{,}21$). Voici le tableau récapitulatif (extrait du protocole en annexe, p. 540) :

[M]q17	OM	**P1**	P2	P3	P4	P5	P6	P7
rpbis SC	-0,23	**0,09**	-0,08	-0,10	-0,09	0,02	-0,17	-0,02
rpbis SCT80	-0,29	**0,23**	-0,25	-0,16	-0,17	-0,06	-0,27	-0,04
rpbis SCT90	-0,33	**0,40**	-0,37	-0,16	-0,10	-0,23	-0,28	-0,09
rpbis classique	-0,05	**0,37**	-0,36	-0,09	-0,01	0,18	-0,08	-0,02

Les configurations des valeurs récoltées par les propositions de [M]q17 aux *rpbis SCT80* et *rpbis SCT90* montrent que la question fonctionne correctement du point de vue de la cohérence spectrale. Les *rpbis SC* montrent une cohérence spectrale moins marquée.

Pour cette question [M]q17, le rpbis classique du distracteur P5 déclenche une fausse alerte. Les valeurs obtenues aux rpbis spectraux sont plus en phase avec les avis des experts qui n'ont pas détecté de problème majeur dans le contenu de la question et de P5 (annexe, p. 504).

F. Conclusions

Dans les sections précédentes nous avons été amené à analyser deux séries de questions qui posent problèmes. D'une part **un premier lot de huit questions** mises en évidence lors de l'exploration du niveau « QCM » pour leurs faibles performances globales aux indices de cohérence interne ($r_{qt\ mb}$ et $r_{qt\ ms}$) ainsi qu'aux indices qualité spectrale (*NCSq* et *Rq*). Il s'agit de [H]q3, [H]q20, [V]q5, [S]q1, [V]q12, [V]q27, [A]q14 et [B]q5. D'autre part, nous avons analysé les propositions d'**un second lot de huit autres questions** épinglées parce qu'elles présentent des valeurs anormales aux *rpbis classiques* et aussi, mais dans deux cas seulement, aux *rpbis spectraux*. Il s'agit de [V]q43, [B]q6, [G]q1, [A]q24, [H]q16, [M]q17, [B]q10 et [Ch]q1. Notons que les propositions des **157 autres questions** (173 - 16 = 157) des 10 épreuves MOHICAN obtiennent toutes des valeurs normales aux *rpbis classiques* et *rpbis spectraux* (voir annexe, p. 544).

Nous avons par ailleurs également analysé les indices des propositions de la question la plus performante [V]q36 et ceux de [P]q7, la seule question qui présente des performances opposées aux indices spectraux *NCSq* et *Rq.*

Lors des analyses des 16 questions suspectes mentionnées précédemment (« suspectes » car elles obtiennent des *rpbis classiques* ou/et *rpbis spectraux* anormaux), nous avons remarqué des différences notables entre les valeurs récoltées par les propositions aux indices *rpbis classiques* et aux indices spectraux (*rpbis SC*, *rpbis SCT80* et *rpbis SCT90*). Des différences apparaissent aussi au sein des différents types de rpbis spectraux, plus particulièrement entre les *rpbis SC* d'une part, et, d'autre part, les *rpbis SCT80* et *rpbis SCT90*. De plus, les avis des experts ne coïncident pas forcément avec les propositions mises en évidence après interprétation des valeurs observées aux différents types de rpbis. Nous synthétisons dans le tableau ci-dessous.

		Avis experts	*rpbis classique*	*rpbis SC*	*rpbis SCT80*	*rpbis SCT90*
8 **QCM** dont les indices de cohérence interne et de qualité spectrale sont les moins élevés	[H]q3	✗P1 ✗P6	✗P1 ✗P6	✗P1 ✗P5 FA ✗P6	✗P1 ✗P6	×
	[H]q20	✗P3 ✗P6	✗P3 ✗P6	✗P3 ✗P6	✗P3 ✗P6	✗P3 ✗P6
	[V]q5	ok	✗P5 FA	×	✓	✓
	[S]q1	ok	✓	✗P5 FA ✗P6 FA	✓	✓
	[V]q12	ok	✓	✗P6 FA	✓	✓
	[V]q27	ok	✗P1 FA	✗P1 FA ✗P3 FA	✗P1 FA	✓
	[A]q14	✗P3 ✗P4	×	✗P3 ✗P4	✓	✓
	[B]q5	ok	✗P3 FA	×	✓	✓
8 autres QCM épinglées après analyse des rpbis des **propositions**	[V]q43	ok	✗P3 FA	✓	✓	✓
	[B]q6	✗P2 ✗P3	✗P2	✗P2 ✗P3	✓	✓
	[G]q1	ok	✗P4 FA	✓	✓	✓
	[A]q24	ok	✗P1 FA	✓	✓	✓
	[H]q16	✗P1	✗P4 FA	✗P1 ✗P4 FA	✓	✓
	[B]q10	ok	✗P3 FA	✓	✓	✓
	[Ch]q1	ok	✗P4 FA	✓	✓	✓
	[M]q17	ok	✗P5 FA	✓	✓	✓
QCM aux meilleures performances	[V]q36	ok	✓	✓	✓	✓
Performances opposées aux indices *NCSq* et *Rq*	[P]q7	ok	✓	✓	✓	✓

Légende :

- « ✓ » : *globalement, le fonctionnement des propositions est cohérent ;*
- « ✓ » : *mise en évidence d'un fonctionnement global cohérent, mais de façon peu marquée ;*
- « ✗P… » : *(suivi d'un n°, nous soulignons en pointillés la proposition correcte) désigne une proposition qui fonctionne de manière incohérente ;*
- « ✗P… » : *mise en évidence du fonctionnement incohérent d'une proposition, mais de façon peu marquée ;*
- « × » : *les valeurs des indices des propositions sont proches de zéro;*
- « FA » : *Fausse Alerte ;*
- : *Non détection de propositions que les experts désignent comme étant problématiques.*

Notons qu'en ce qui concerne les avis des experts, nous n'avons pas relevé d'autres questions que [H]q3, [H]q20, [A]q14, [B]q6 et [H]q16 qui à leurs yeux posent particulièrement problème.

Examinons d'abord les huit premières questions du tableau, celles dont les indices de cohérence interne et de qualité spectrale sont les moins élevés.

En ce qui concerne [H]q3 et [H]q20, nous remarquons une forte convergence des problèmes mis en évidence à l'aide des différents types de rpbis. Rappelons qu'au niveau d'exploration « QCM » ces deux questions ont récolté les valeurs les plus faibles en cohérence interne et en qualité spectrale.

Pour [H]q3 les *rpbis classiques*, les *rpbis SC* et les *rpbis SCT80* désignent les mêmes propositions que celles que les experts considèrent comme étant problématiques (P1 et P6). Les *rpbis SC* déclenchent aussi une fausse alerte au niveau de P5 dont la vérification du contenu montre qu'elle ne pose pas de problème. Les *rpbis SCT90* ne détectent pas P1 ni P6 dans la mesure où les valeurs récoltées par toutes les propositions sont proches de zéro.

Pour [H]q20, l'erreur d'encodage de la réponse correcte est pointée par tous les types de rpbis qui mettent en évidence le fonctionnement incohérent de deux propositions : P3 et P6. Cependant, les *rpbis classiques* mettent de façon peu marquée les deux propositions problématiques en évidence alors que d'un point de vue spectral, d'une part les *rpbis SC* désignent clairement à la fois P3 et P6 comme étant toutes deux problématiques et d'autre part les *rpbis SCT80* et *rpbis SCT90* montrent la présence d'un problème de façon très nette pour P6 et de façon moins marquée pour P3.

Pour cinq autres QCM : [V]q5, [S]q1, [V]q12, [V]q27 et [B]q5 de ce premier lot de huit questions, les experts n'ont pas détecté d'anomalie dans les propositions. Nous constatons que dans 2 cas sur 5 ([S]q1 et [V]q12) les *rpbis classiques* sont en phase avec les avis des experts. Les *rpbis SC* ne le sont pas du tout. Les *rpbis SCT80* le sont dans 4 cas sur 5 ([V]q5, [S]q1, [V]q12 et [B]q5). En ce qui concerne les *rpbis SCT90*, ils sont pour les cinq questions les plus en phase avec les avis des experts, dans 5 cas sur 5.

En ce qui concerne [A]q14, les experts ont signalé des problèmes au niveau de deux propositions (P3 et P4). Seuls les *rpbis SC* détectent ces deux propositions.

Envisageons maintenant les huit questions suivantes : [V]q43, [B]q6, [G]q1, [A]q24, [H]q16, [B]q10, [Ch]q1 et [M]q17 épinglées après analyse des rpbis de chaque proposition.

Nous constatons que parmi ces huit questions suspectes, les experts en signalent deux dont le contenu de certaines propositions pose problème : [B]q6 et [H]q16.

Pour [B]q6, les *rpbis SC* désignent P2 et P3 aussi considérées comme étant problématiques par les experts. Les *rpbis classiques* mettent en évidence une des deux propositions et ne détectent pas l'autre. Les *rpbis SCT80* et *rpbis SCT90* n'en détectent aucune.

Pour [H]q16, les experts signalent que le distracteur P1 est « acceptable ». Seul le *rpbis SC* détecte cette proposition mais déclenche aussi une fausse alerte à propos de la réponse correcte P4.

Pour cinq autres questions de ce deuxième lot, [V]q43, [G]q1, [A]q24, [B]q10 et [Ch]q1 nous observons des valeurs situées sous le seuil repère pour les rpbis classiques des réponses correctes, ce qui déclenche de fausses alertes. Les rpbis spectraux, eux, ne mettent pas en évidence de problèmes dans ces questions et sont en phase avec les avis des experts qui considèrent ces questions comme ne présentant pas d'anomalies particulières.

En ce qui concerne [M]q17, les experts et les rpbis spectraux ne détectent pas de problèmes. Par contre le *rpbis classique* positif du distracteur P5 déclenche une nouvelle fausse alerte.

L'avant dernière question du tableau, [V]q36 était celle qui parmi les 173 QCM des 10 épreuves MOHICAN présentait les meilleures performances. Comme nous pouvions nous y attendre étant donné les très bons scores en cohérence interne et en qualité spectrale, les propositions de [V]q36 ne posent pas de

problèmes du point de vue de la discrimination classique et de la cohérence spectrale, ce qui correspond aussi à l'avis des experts qui ne décèlent pas d'anomalies.

Enfin, en ce qui concerne la dernière question du tableau, [P]q7 (la seule parmi les 173 qui présentait deux indices spectraux contradictoires au niveau d'exploration « QCM »), nous observons des valeurs normales aux rpbis spectraux et aux *rpbis classiques* et constatons que les experts n'ont pas émis de griefs par rapport à cette question.

Les divergences qui peuvent apparaître entre des valeurs livrées d'une part par les indices spectraux et d'autre part par les *rpbis classiques* (par exemple [M]q17) montrent bien que même si il existe dans certains cas des liaisons entre ces indices spectraux et classiques (par exemple pour les deux questions les plus problématiques : [H]q3 et [H]q20), ils ne mesurent pas les mêmes phénomènes. En fait, à partir d'un principe de calcul similaire, celui du rpbis (auquel nous avons greffé le principe du traitement contrasté et de la turbo analyse dans le cas des rpbis spectraux turbo), ils mesurent pour un même type d'objets, les propositions au sein d'une question à choix multiple, des propriétés différentes. Dans le cas du *rpbis classique* on mesure la propension d'une proposition à discriminer les étudiants qui obtiennent un nombre élevé de réponses correctes à l'ensemble des questions du test des étudiants qui obtiennent un nombre moins élevé de réponses correctes (voir l'exposé détaillé de la problématique du *rpbis classique*, p. 171), et, dans le cas des *rpbis SC*, *rpbis SCT80* et *rpbis SCT90* nous mesurons la cohérence spectrale, c'est-à-dire la propension d'une proposition à recueillir des pourcentages de certitude moins élevés quand elle est incorrecte et plus élevés lorsqu'elle est correcte (voir l'exposé détaillé de la problématique du *rpbis Spectral Contrasté*, p. 178 et du *rpbis Spectral Contrasté avec Turbo analyse*, p. 184). Ces deux éclairages différents, de cohérence interne d'une part et de cohérence spectrale d'autre part, convergent pour les deux questions les plus problématiques ([H]q3 et [H]q20).

Au total, lorsque nous examinons les synthèses des 16 questions suspectes, nous remarquons que les analyses réalisées sur base des *rpbis SCT80* et *rpbis SCT90* sont le plus souvent en phase avec les avis des experts : *grosso modo* dans 12 cas sur 16. En ce qui concerne les *rpbis SC,* les valeurs observées rejoignent les avis des experts dans 10 cas sur 16. Pour ce qui est des *rpbis classiques*, nous remarquons une convergence avec les avis des experts dans environ 4 cas sur 16.

En ce qui concerne les fausses alertes (notées « FA » en rouge dans le tableau de synthèse), nous en dénombrons dix pour les *rpbis classiques*, sept pour les *rpbis SC*, une pour les *rpbis SCT80* et aucune pour les *rpbis SCT90*.

Enfin, pour ce qui est des « non détections » (fond jaune dans le tableau) de propositions problématiques signalées par les experts, nous en relevons trois pour les *rpbis classiques*, aucune pour les *rpbis SC*, trois pour les *rpbis SCT80* et quatre pour les *rpbis SCT90*.

En conclusion, lorsque nous comparons les performances des différents types de rpbis, nous constatons que :

- les *rpbis SCT80* et *rpbis SCT90* (calculés à l'aide des données des étudiants qui commettent moins d'erreurs dans leurs auto-estimations) déclenchent moins de fausses alertes. Par contre, ils ne permettent pas de détecter toutes les propositions désignées comme étant problématiques par les experts. les *rpbis SCT80* détectent ces propositions pour deux questions (sur cinq questions pointées par les experts) et les *rpbis SCT90* les détectent pour seulement une question ;

- les *rpbis SC* (calculés à l'aide des données de tous les étudiants) déclenchent plus de fausses alertes mais par contre ils présentent l'avantage de détecter toutes les propositions problématiques signalées par les experts ;

- les *rpbis classiques* déclenchent le plus de fausses alertes et ne détectent que dans le cas de deux questions (sur cinq) les propositions désignées comme étant problématiques par les experts.

Enfin, terminons cette partie en insistant sur l'importance de l'éclairage des experts. Une décision de suppression ou d'acceptation d'une question après analyse des rpbis nous entraîne dans des considérations liée à la validité de l'épreuve. Seuls ces experts (que nous avons rencontrés lors de deux réunions de débrieffing des épreuves) peuvent indiquer si la question se trouve bien à sa place dans le test. L'éclairage des experts permet donc de se prémunir contre d'éventuels effets pervers que pourraient entraîner des rectifications aveugles (sans contrôle des experts) sur la validité de l'épreuve. En effet, en supprimant par exemple une question aux faibles performances classiques et spectrales, nous pourrions peut-être augmenter la fidélité d'un test mais peut-être aussi en diminuer la validité dans la mesure où un objectif important pourrait ne plus être évalué.

Conclusions

Conclusions détaillées

A. Bilan

B. Perspectives

A. Bilan

En octobre 1999, dix épreuves de connaissances (voir liste, p. 286) ont été soumises à des milliers d'étudiants (entre 1.392 et 3.846 selon les tests) entrant en 1ère candidature dans 8 universités et facultés universitaires de la Communauté française de Belgique (projet « MOHICAN », voir p. 93).

Le format d'items utilisé pour ces épreuves standardisées était de type QCM. Lors des tests les étudiants furent invités à accompagner systématiquement le choix de chacune de leurs réponses d'un pourcentage de certitude. Voici l'échelle des pourcentages de certitude (%C) utilisée dans le contexte MOHICAN (détails, p. 274) :

```
DC = 0        1        2        3        4        5
     |--------|--------|--------|--------|--------|
%C = 0%      20%      40%      60%      80%     100%
```

Comparée à l'ancienne consigne FAPSE (voir p. 272), cette nouvelle échelle est plus simple : d'une part elle n'est plus dissymétrique et d'autre part, il ne s'agit plus pour l'étudiant de dire dans quel intervalle de certitude il se situe, mais de quel pourcentage de certitude proposé dans la consigne il estime qu'il est le plus proche.

Rappelons que les tests MOHICAN n'étaient pas cotés et que l'anonymat des étudiants était garanti. Le choix des pourcentages de certitude n'a donc pas été influencé par un barème de tarif de points ni même par l'octroi d'une cote finale qui aurait pu avoir une quelconque incidence sur leur parcours académique ultérieur.

L'observation des courbes de fréquences des scores de réalisme des étudiants lors des épreuves MOHICAN montre des taux d'erreurs dans les auto-estimations des participants superposables à ceux des 28 tests FAPSE (voir p. 303). L'échelle simplifiée et non tarifée de recueil des certitudes permet donc une utilisation normale des pourcentages de certitude par des étudiants entrant en 1ère candidature et qui sont pratiquement pas ou très peu entraînés à cette procédure d'auto-estimation de ses réponses.

Nous pensons que cette méthode non tarifée de recueil des pourcentages de certitude (il n'y a plus de tarifs de points associés à chaque degré de certitude comme c'était le cas dans l'ancienne consigne FAPSE, voir p. 69) constitue une voie d'avenir prometteuse même dans le contexte d'épreuves cotées. En effet, il est tout à fait possible d'imaginer un système de cotation où à l'issue d'une épreuve académique d'une part les étudiants recevraient une première cote pour l'exactitude des réponses fournies (le Taux d'Exactitude Observé, *TEA*) et d'autre part une seconde cote en rapport avec le nombre d'erreurs d'auto-estimations commises (l'indice de Réalisme du sujet, *Rs*). L'évaluateur pourrait dès lors construire la cote finale en déterminant la part liée au taux d'exactitude et la part liée au réalisme. Par exemple, 80% de la cote finale pourrait concerner l'exactitude des réponses et 20% le réalisme. Ces pourcentages devraient évidemment être adaptés en fonction du public, du contenu, des exigences de l'examinateur, …

1. L'intuition initiale

Dans le cadre de cette thèse nous avons saisi l'opportunité des milliers de données, réponses et certitudes, recueillies à l'occasion des dix tests MOHICAN[75] pour concevoir et étudier de nouveaux indices d'analyse de la qualité spectrale des épreuves standardisées.

Rappelons notre intuition de départ pour la construction de ces nouveaux indices spectraux : logiquement les étudiants qui répondent correctement à une question devraient fournir des pourcentages de certitude plus élevés que les étudiants qui répondent incorrectement. Ainsi, pour une question à choix multiple qui fonctionne logiquement d'un point de vue spectral nous devrions observer une tendance à fournir des pourcentages de certitude plus élevés chez les étudiants qui choisissent la proposition correcte, et, en parallèle, une tendance à fournir des certitudes moins élevées chez les étudiants qui choisissent un des distracteurs. Dès lors que cette situation ne se présente pas, par exemple lorsque les étudiants qui choisissent une proposition incorrecte accompagnent celle-ci de certitudes en moyenne plus élevées que les certitudes fournies par les étudiants qui répondent correctement, nous nous trouvons face à un problème d'incohérence spectrale du groupe dans l'utilisation des pourcentages de certitude.

Selon nous, trois types de facteurs peuvent intervenir dans l'explication d'un problème d'incohérence spectrale dans l'utilisation des certitudes.

Une première catégorie de facteurs concerne les caractéristiques de l'instrument de testing utilisé et plus particulièrement ceux qui sont liés à l'appareillage de la question (QCM mal formulée, distracteurs incohérents,…). Le cas le plus évident concerne l'erreur d'encodage de la réponse correcte dont pourrait être sujette une QCM et qui expliquerait l'incohérence spectrale dans l'utilisation des pourcentages de certitude. Nous avons pu observer un tel cas de figure dans le cadre des tests MOHICAN, il s'agit de la 20ème question du test d'Histoire ([H]q20). Nous y reviendrons plus loin dans ces conclusions.

Une seconde catégorie de facteurs serait focalisée sur les caractéristiques des examinés. Nous pourrions en effet imaginer que pour une série de raisons liées à « l'histoire » des étudiants, ces derniers soient amenés à choisir une proposition incorrecte en étant dans l'ensemble plus convaincus qu'elle est correcte que ne le sont les étudiants qui ont opté pour la solution correcte, ce qui entraînerait une utilisation incohérente des certitudes. Un tel cas de figure « conviction erronnée » a également été observé, il s'agit de la 3ème question du test d'Histoire ([H]q3) et nous y reviendrons aussi plus loin.

Enfin, une troisième catégorie de facteurs qui pourraient expliquer les incohérences spectrales dans l'utilisation des certitudes concerne les caractéristiques de l'enseignement. Il se pourrait en effet qu'une matière mal ou trop peu enseignée soit à l'origine de conceptions fausses chez les apprenants, ce qui provoquerait des utilisations incohérentes des pourcentages de certitude. C'est le cas de la 5ème question du test de Biologie ([B]q5).

A notre connaissance, jusqu'à présent, les informations fournies par les pourcentages de certitude n'avaient jamais été traitées en vue de déterminer les niveaux de qualité spectrale des épreuves standardisées, de leurs questions à choix multiple et des propositions au sein de celles-ci. C'est cette voie nouvelle de recherche édumétrique et docimologique que nous avons exploitée dans le cadre de cette thèse.

[75] Pour l'ensemble des 10 tests MOHICAN, le nombre de réponses et de certitudes s'élève à 941.576.

2. Une approche « *bottom-up* » dans la construction des nouveaux indices de la qualité spectrale des épreuves

Nous avons défini la « qualité spectrale » d'une solution donnée au sein d'une QCM comme étant la propension, lorsqu'il s'agit d'une réponse correcte, à obtenir des certitudes plus élevées que les certitudes qui accompagnent les réponses incorrectes. Lorsqu'il s'agit d'une proposition incorrecte, la qualité spectrale se définit comme étant la propension d'un distracteur à obtenir des certitudes moins élevées que celles qui ont accompagné la réponse correcte. Lorsque l'utilisation des pourcentages de certitude est conforme à cette définition de la qualité spectrale, nous parlons aussi de « cohérence spectrale ».

Selon nous, un des apports de cette thèse se situe au niveau de la mise au point de nouveaux indices spectraux qui permettent de quantifier la qualité spectrale des épreuves. Rappelons ici les principaux instruments que nous proposons pour l'analyse de la qualité spectrale des tests standardisés.

L'approche qui a présidé à la construction des indices d'analyse de la qualité spectrale des épreuves est de type « *bottom-up* ». C'est-à-dire une approche qui consiste à partir d'abord d'un premier niveau d'analyse qui s'appuie sur les informations que peuvent nous fournir les données relatives aux choix des propositions. A l'aide des informations sur la qualité spectrale des propositions, on peut ensuite passer à un second niveau d'analyse qui envisage la qualité spectrale de chaque QCM. Enfin, en globalisant l'ensemble des informations liées aux QCM on aboutit à un troisième niveau d'analyse de la qualité spectrale d'une épreuve entière.

a) Les indices du niveau « PROPOSITIONS »

Pour le niveau « PROPOSITION », le nouvel indice proposé est le coefficient de corrélation point bisérial Spectral auquel nous avons appliqué un traitement Contrasté. Dans la terminologie que nous proposons, il s'agit des *rpbis S*, *rpbis SC* et *rpbis SCT*.

Le *rpbis S* : nous avons choisi d'appeler ce nouvel indice « *rpbis S* » car son fonctionnement est proche de celui du coefficient de corrélation point bisérial classique (voir détails, p. 171). Dans le cas du *rpbis classique*, les choix ou les rejets (1 ou 0) d'une proposition d'une QCM sont corrélés avec les nombres de réponses correctes obtenues à l'ensemble des questions du test. Dès lors, on s'attend à ce que la réponse correcte soit corrélée avec des scores plus élevés que les scores liés aux choix des distracteurs. Dans ce cas, la question discrimine les étudiants en fonction du critère du nombre de réponses correctes récoltées à l'ensemble des questions du test.

Ce principe du *rpbis classique* est réutilisé dans le cas de l'analyse spectrale mais avec une différence de taille : le *rpbis S* n'utilise pas les nombres de réponses correctes obtenues pour l'ensemble du test mais les remplace par les pourcentages de certitude qui sont corrélés avec les choix ou les rejets de la proposition auxquels ils sont associés. La mesure critère n'est plus externe à la question (le nombre de réponses correctes à l'ensemble des questions du test, ce qui inclut les performances aux autres questions) mais interne à la question dans la mesure où les certitudes sont celles fournies dans le cadre de la QCM envisagée et uniquement dans ce cadre (la qualité des autres questions du test n'influence pas, comme c'est le cas pour le *rpbis classique*, la mesure de la qualité de la QCM envisagée). Nous rappellerons plus loin que cette propriété « *test free* » du *rpbis S* constitue un avantage qui ouvre de nouvelles perspectives en terme de suivi de la qualité des QCM dans la gestion des banques de question.

Le *rpbis SC* : parallèlement, nous avons distingué deux cas de figure : celui d'une proposition correcte et d'autre part celui des propositions incorrectes. En ce qui concerne le cas d'une proposition correcte, ce sont les données de tous les étudiants qui ont répondu à la question qui sont envisagées dans la corrélation. Pour ce qui est d'une proposition incorrecte, nous avons repris dans les données qui sont corrélées, d'une part les réponses et les certitudes des étudiants qui ont répondu correctement, et, d'autre

part, les réponses et les certitudes des étudiants qui ont choisi la proposition incorrecte envisagée, et uniquement celle là, sans prendre en compte les données des autres propositions incorrectes. C'est en cela que nous pouvons qualifier le traitement de « Contrasté », d'où le « *C* » dans l'appellation *rpbis SC* (voir détails p. 178). Notons qu'il serait tout à fait possible d'appliquer le principe du traitement contrasté aux *rpbis classiques* (ce que nous comptons faire dans le cadre de travaux de recherche ultérieurs).

Le *rpbis SCT* : enfin, nous avons aussi appliqué le principe de la « Turbo analyse » au nouvel indice de cohérence spectrale des propositions. L'idée sous-jacente est que nous pouvons accroître la confiance dans les informations liées aux pourcentages de certitude en ne prenant en compte que les données des étudiants qui commettent le moins d'erreurs dans leurs auto-estimations. Nous pouvons en effet mesurer cette qualité d'auto-estimation par l'intermédiaire de l'indice de Réalisme des sujets (*Rs*) qui nous donne une indication précise quant au taux d'Erreur Moyenne Absolue de Certitude (*EMAC*) de chaque étudiant (voir détails, p. 186 et p. 274). Dès lors, nous avons calculé les *rpbis SC* en sélectionnant les données des étudiants à des seuils *Rs* de plus en plus élevés. Le terme « Turbo analyse » utilisé pour qualifier ce type de traitement fait référence à la montée en puissance de l'instrument en terme de qualité d'information fournie au fur et à mesure que l'on progresse dans les « paliers Turbo », c'est-à-dire que l'on prend en compte les données des étudiants qui commettent de moins en moins d'erreurs dans leurs auto-estimations. Nous proposons l'appellation *rpbis SCT* suivie de l'indication du seuil *Rs* (un nombre entre 0 et 100) utilisé pour calculer le *rpbis SC* dans le cadre d'une Turbo analyse.

Ces nouveaux indices d'analyse de la qualité spectrale des propositions au sein d'une QCM sont particulièrement utiles lorsqu'on souhaite vérifier la qualité des questions auxquelles ont été soumis les examinés. Nous rappellerons plus loin dans le bilan des explorations des 173 questions des 10 épreuves, que les indices *rpbis SCT* permettent un nouvel éclairage sur la qualité des questions, celui de la cohérence spectrale, qui dans la plupart des cas converge avec celui de la discrimination classique basée sur les *rpbis classiques*, et, met en évidence les problèmes parfois de façon plus nette (surtout aux paliers de turbo analyse les plus élevés où les *rpbis SCT* sont plus marqués que les *rpbis classiques*).

b) Les instruments d'analyse de la qualité spectrale du niveau « QCM »

Les indices *rpbis SC* et *rpbis SCT* que nous venons de rappeler concernent les propositions au sein des QCM. Nous pensons qu'il est utile de globaliser ces informations afin de fournir une « image » de la qualité spectrale au niveau des questions. En effet, à l'aide des instruments qui permettent de dresser un bilan de la cohérence spectrale d'une question, nous pouvons localiser les problèmes éventuels sans devoir passer en revue, une par une, chacune des propositions de chacune des QCM. Nous proposons cinq types d'instruments d'analyse de la qualité spectrale des questions que nous rappelons ci-dessous et qui peuvent tous être calculés à différents paliers de turbo analyse. Nous utilisons le terme « instruments » dans la mesure où les indices proposés sont aussi accompagnés de représentations graphiques afin de visualiser les performances spectrales des questions et de faciliter l'interprétation des valeurs obtenues.

Le premier type d'instrument envisagé concerne les Niveaux de Cohérence Spectrale d'une question (*NCSq*). Les *NCSq* sont calculés à chaque niveau de turbo analyse en soustrayant la valeur de la moyenne pondérée des *rpbis SCT* des propositions incorrectes (p.189) à la valeur du *rpbis SCT* de la réponse correcte (voir détails, p. 231). Nous proposons deux types de représentations graphiques des valeurs obtenues aux indices *NCSq* et qui facilitent l'analyse de la cohérence spectrale des questions. Il s'agit d'une part des Brins Spectraux par question (*BSq*) qui figurent sur le graphique en Gerbe Spectrale d'un test (*GSt*) (voir détails, p. 228) et d'autre part des courbes des valeurs *NCSq* (voir p. 232). Nous avons également montré qu'il était possible de calculer un Niveau de cohérence Interne d'une question (*NCIq*) au départ des *rpbis classiques* des propositions (voir p. 233). L'avantage de la visualisation des *NCSq* est de permettre une identification rapide des questions dont les niveaux de cohérence spectrale sont faibles et qui dès lors doivent faire l'objet d'une analyse plus détaillée.

Le second type d'instrument proposé est intitulé Profil Spectral d'une question (*PSq*). Il s'agit des indices et graphiques liés aux calculs des répartitions de réponses correctes et incorrectes pour chaque degré de certitude. Nous distinguons ainsi pour chaque question, d'une part les Taux de Réponses

Correctes des différentes Certitudes (*TRC C0* à *C5*) et d'autre part les Taux de Réponses Incorrectes (*TRI C0* à *C5*). A partir de ces informations nous générons un graphique reprenant deux hémi-spectres par question, l'un concerne les *TRC* et l'autre les *TRI* (voir détails p. 235). Notre intuition est que les questions qui fonctionnent anormalement d'un point de vue spectral se caractériseront par des profils spectraux différents des questions dont la cohérence spectrale est bonne. Les profils spectraux ouvrent de nouvelles perspectives en matière d'étude des disfonctionnements de QCM que nous pourrions qualifier de spectralement incohérente ; l'analyse et l'exploitation des *PSq* des questions constitue un nouveau champ de recherches docimologiques que nous investiguerons dans nos prochains travaux.

L'indice de Réalisation des prédictions par question (*Rq*) est un troisième type d'outil d'analyse de la qualité spectrale des QCM. Il s'agit ici d'évaluer comment les Taux d'Exactitude Annoncés (TEA) par les étudiants à chaque fois qu'ils choisissent un pourcentage de certitude s'ajustent aux Taux d'Exactitude Observés (TEO). La formule de l'indice *Rq* est similaire à celle de l'indice *Rs* (Réalisme des sujets) mais s'en différencie cependant par la prise en compte des certitudes récoltées non plus pour chaque étudiant à toutes les questions, mais par chaque question à tous les étudiants (voir détails, p. 242). Il existe aussi une représentation graphique de l'indice *Rq*. Le graphe (voir exemples, p. 244) reprend en abscisse les pourcentages de certitude (ou taux d'exactitudes annoncés) et en ordonnée les taux d'exactitude associés à chaque certitude. Plus les points sont proches de la diagonale-repère où les TEO correspondent aux TEA, meilleures sont les concordances entre les prédictions (TEA ou pourcentages de certitude) et les réalisations de celles-ci (TEO). Les valeurs obtenues aux indices *Rq* et *NCSq* sont liées (voir p. 334), ce qui est logique : plus les erreurs d'auto-estimations (*Rq*) sont faibles, plus les pourcentages de certitude sont utilisés de façon cohérente (*NCSq*). Cependant, les corrélations positives entre les valeurs *Rq* et *NCSq* sont moyennement élevées (VOCABU : $r = 0{,}61$, MATHEM : $r = 0{,}52$, ARTACT : $r = 0{,}61$), ce qui renforce notre intuition selon laquelle ces indices spectraux sont complémentaires. Il s'agit de deux éclairages de la qualité spectrale d'une question, l'un, le *Rq*, nous donne une indication sur la quantité d'erreurs dans les auto-estimations et l'autre, le *NCSq*, nous montre dans quelle mesure il y a cohérence spectrale dans l'utilisation des pourcentages de certitude.

Le quatrième indice proposé est celui de la facilité introspective d'une question (*piq*, p. 251). Etant donné que le choix d'une solution est systématiquement accompagné d'un pourcentage de certitude, nous avons calculé pour chaque proposition la moyenne des pourcentages de certitude fournis par les étudiants (*CMp*, p. 252). Chaque *CMp* est ensuite multiplié par le nombre d'utilisations de la proposition envisagée et la somme des produits est effectuée. Enfin, nous divisons cette somme des produits par le nombre total de réponses afin d'obtenir la valeur de l'indice *piq*. D'une part l'indice *piq* nous renseigne à propos de la facilité de la question ressentie par les étudiants. D'autre part, comme nous allons le voir dans le paragraphe suivant, cet indice peut être confronté à la facilité objective d'une question, ce qui permet d'évaluer la quantité de sur ou sous-estimation dans les résultats.

Le cinquième type d'instrument d'analyse de la qualité spectrale des questions est lié à l'indice de Centration par question (*Cq*). *Cq* est obtenu en soustrayant la facilité objective d'une question (*poq*, ou le taux de réponses correctes) à la facilité introspective de la question (*piq*). Le signe de la valeur de l'indice *Cq* nous renseigne sur le niveau de sur (+) ou sous (-) estimation observée dans les résultats des étudiants (détails p. 254). Les performances d'une question à l'indice *Cq* peuvent être visualisées à l'aide d'un graphique en nuage de points où les valeurs des *piq* sont portées en abscisses et les valeurs des *poq* en ordonnées. L'indice *Cq* est complémentaire à l'indice *Rq*, l'information fournie par ce dernier nous donne une mesure précise de la quantité d'erreurs observées dans les auto-estimations qui peut être complétée par *Cq* qui nous informe sur la tendance à la sur ou sous-estimation dans ces erreurs d'auto-estimations.

c) Les instruments d'analyse spectrale du niveau « TEST »

Afin de permettre d'étudier la qualité spectrale au niveau d'analyse « TEST » nous avons globalisé les informations spectrales liées aux questions. L'intérêt des indices calculés au niveau « TEST » réside notamment dans la possibilité qui est ainsi offerte d'évaluer l'impact d'éventuelles rectifications opérées au niveau des questions sur les résultats spectraux de l'épreuve (voir l'étape « *6. Correction et discussion* » du cycle SMART, p. 82). On retrouvera donc à ce troisième niveau d'analyse les différents indices développés au niveau précédent « QCM » mais moyennés de façon à offrir une vue globale des performances spectrales de l'épreuve. Tous ces indices peuvent donc aussi être calculés pour les différents niveaux de turbo analyse envisagés.

En calculant la moyenne des *NCSq* d'une épreuve on obtient l'indice du Niveau de Cohérence Spectrale du test (*NCSt*). Comme nous pouvons aussi calculer un indice de Niveau de Cohérence Interne d'une Question (*NCIq*) au départ des *rpbis classiques*, nous avons mesuré le Niveau de Cohérence Interne d'un test (*NCIt*) à l'aide de la moyenne des *NCIq* (voir p. 261). Dès lors, nous sommes aussi en mesure de comparer le *NCSt* (mesure de cohérence spectrale) et le *NCIt* (mesure de cohérence interne) d'une épreuve (p. 263) sur la base des informations liées aux propositions des QCM, ce que nous avons fait pour les 10 épreuves (voir comparaison *NCSt* et *NCIt* pour les 10 tests MOHICAN, p. 317) et que nous rappellerons plus loin.

Nous proposons une seconde mesure de la qualité spectrale d'une épreuve sous la forme d'un indice de Réalisation des prédictions par test (*Rt*). Il s'agit de la moyenne des indices *Rq* d'une épreuve (voir p. 264).

L'indice de facilité introspective du test (*pit,* p. 268) est une troisième mesure de la qualité spectrale d'une épreuve qui reprend la moyenne des *piq*.

Le *pit* peut aussi être confronté à la facilité objective du test (*pot*). Dès lors, nous proposons un indice de Centration par test (*Ct*) qui est calculé en soustrayant le *pot* du *pit* (voir p. 270). Comme pour l'indice *Cq* le signe de la valeur obtenue à l'indice *Ct* nous renseigne sur une tendance à la sur ou sous-estimation dans les résultats, mais, dans le cas de *Ct*, en ce qui concerne l'ensemble des questions de l'épreuve. Les valeurs obtenues à l'indice *Ct* peuvent aussi être calculées aux différents paliers de turbo analyse et donc être représentées sur un graphique en nuage de points où les valeurs des *pit* sont portées en abscisses et celles des *pot* en ordonnées. Un graphique des *Ct* des tests MOHICAN tel que celui qui est présenté page 300 permet de visualiser deux types d'informations : (1) la tendance à la sous ou surestimation dans les résultats et (2) la facilité des épreuves, et ce, en prenant en compte les données des étudiants à chacun des 10 paliers de la turbo analyse (de T0 à T90 avec un pas de 10).

3. Une approche « *top-down* » dans l'exploration spectrale des épreuves

A l'inverse de l'approche « *bottom-up* » que nous avons utilisée dans la construction des instruments spectraux, nous préconisons une approche « *top-down* » dans l'exploration et l'analyse de la qualité spectrale des épreuves. C'est-à-dire une approche où nous analysons d'abord les résultats globaux du test, ensuite les performances classiques et spectrales de chacune des questions et enfin les indices liés aux propositions des questions dont la qualité spectrale est faible. L'avantage d'une telle approche réside dans le fait que nous évitons de devoir analyser un grand nombre de propositions pour finalement en identifier que quelques unes qui posent des problèmes. Nous pensons que les indices de performance des questions doivent nous permettent de mettre en évidence les QCM qui fonctionnent anormalement. Dès lors, une fois ces « questions à problèmes » identifiées, c'est sur elles seules que devraient porter les analyses approfondies des propositions. On éviterait ainsi une série d'analyses superflues.

Du point de vue de la qualité globale d'un test, il est utile de pouvoir comparer les informations sur l'état de qualité initiale du test. Ces informations pourront ensuite être comparées aux performances de l'épreuve après rectification de certaines questions qui posent problèmes. Nous sommes ainsi en mesure d'évaluer l'impact sur la qualité globale du test des rectifications opérées sur certaines QCM. Nous verrons plus loin dans la partie « perspectives » un exemple de mesure d'impact effectuée dans le cadre du test d'histoire.

Du point de vue de la cohérence interne, l'alpha de Cronbach (p. 137), le coefficient de bipartition (p. 130) avec correction de Spearman-Brown (p. 131), le coefficient de Guttman (p. 131) sont habituellement proposés dans les analyses de test classiques. Nous nous sommes aussi intéressé au coefficient d'allongement des tests en calculant le nombre de questions parallèles qu'il faudrait ajouter aux épreuves pour que leur alpha atteigne 0,8 (p. 139). Nous avons utilisé tous ces coefficients dans le cadre de cette thèse en vue de comparer les performances spectrales avec les performances en cohérence interne. Du point de vue de l'indice de difficulté globale, c'est l'indice de facilité objective du test (*pot*) que nous emploierons (p. 270). En ce qui concerne les nouveaux indices d'analyse de la qualité spectrale des épreuves, nous avons eu recours aux différents instruments rappelés plus haut (p. 320).

4. Faits saillants mis en lumière lors de l'exploration du niveau « TEST »

Pour calculer ces nouveaux indices spectraux sur les milliers de données des dix épreuves MOHICAN, nous avons été amené à programmer un logiciel de traitement spectral des résultats que nous avons intitulé « *SCANTEST 2.0* » (voir p. 195). Résumons ici les principaux faits saillants que nous avons mis en évidence lors de l'exploration du niveau « TEST ».

a) A propos de la qualité spectrale des dix épreuves MOHICAN

Du point de vue de la qualité spectrale des dix épreuves, nous avons d'abord calculé les **pourcentages d'effectifs d'étudiants aux différents paliers de turbo analyse** (p. 291). Ces pourcentages sont très proches pour 9 épreuves sur 10. Une épreuve se démarque des autres, il s'agit de l'épreuve de Compréhension qui obtient des pourcentages d'effectifs plus faibles de T40 à T80. Rappelons que ces effectifs observés aux différents paliers de turbo analyse nous donnent une indication sur les performances des étudiants à l'indice de réalisme (*Rs*). Par exemple, les pourcentages des effectifs à T40 représentent les pourcentages de sujets dont les performances à l'indice *Rs* sont plus élevées ou égales à 40 (*Rs* varie de 0 à 100). On peut dire que les performances en réalisme sont très proches d'une épreuve à l'autre sauf pour le test de Compréhension. Ceci est confirmé par le tableau des statistiques descriptives liées aux scores de réalisme des épreuves MOHICAN (voir p. 304), seule l'épreuve de Compréhension obtient un *Rg* inférieur à 70 (65,2), les neuf autres tests obtiennent des *Rg* compris entre 71,3 et 78,1. Les étudiants ont donc commis plus d'erreurs dans leurs auto-estimations au test de Compréhension que dans les autres tests où les performances moyennes étaient assez proches. Ces constats débouchent sur une autre question : les performances en réalisme des étudiants sont-elles corrélées d'un test à l'autre ? Cette question est actuellement étudiée par Leclercq & Detroz (2001) dans le cadre d'une recherche commanditée par le Conseil Inter Universitaire de la Communauté française de Belgique (CIUF).

Les effectifs commencent à véritablement infléchir à partir du palier T60, en effet, de T0 à T50 ils diminuent très peu en passant de 100% à 97% en moyenne (voir courbes, p. 291). Ce constat peut aussi être visualisé sur le graphique des fréquences en pourcentage des scores de réalisme des sujets (voir graphique, p. 303). On y observe en effet une distribution des fréquences moyennes des scores *Rs* très décentrée vers la droite. Selon nous, ce décentrage vers la droite des performances est lié à la façon dont l'indice *Rs* est calculé (voir pp. 184 et 274). En effet, pour qu'un étudiant obtienne un score *Rs* égal à zéro, il doit accompagner chaque réponse incorrecte de la certitude la plus élevée (100%) et chaque réponse correcte de la certitude la plus faible (0%). Un tel comportement totalement incohérent chez une personne amenée à auto-estimer ses réponses pourrait être qualifié de pathologique ou d'immature, or nous avons affaire à des étudiants entrant en 1ère candidature et, au vu des performances en réalisme observées, ceux-ci ont dans l'ensemble essayé de s'auto-estimer correctement en respectant la consigne. Selon nous, sous le seuil *Rs* de 50 on devrait plutôt parler d'« irréalisme » et réserver le terme « réalisme » aux performances *Rs* supérieures à 50. Ses constats et réflexions sont corroborés par les observations que nous avions pu faire lors d'une étude précédente portant sur 28 examens entre 1994 et 1996 (voir p. 303).

Nous constatons qu'à T80 il reste en moyenne 39% de sujets ce qui représente entre 584 et 1.806 étudiants selon les épreuves. Ce grand nombre de sujets très réalistes (à T80 le pourcentage d'erreurs dans les auto-estimations des sujets est $\leqslant$ 20%) disponibles à T80 nous a amené à privilégier ce palier de turbo analyse pour le calcul des indices spectraux.

En ce qui concerne le **niveau de cohérence spectrale des tests** (p. 295), nous observons une augmentation des valeurs *NCSt* des dix épreuves aux paliers de turbo analyse T0 à T90. Cette augmentation est logique dans la mesure où plus on monte dans les paliers turbo, moins les résultats à partir desquels les *NCSt* sont calculés contiennent des erreurs d'auto-estimations et donc, plus les étudiants ont tendance à

utiliser d'une part des pourcentages de certitude élevés pour accompagner leurs réponses correctes et d'autre part des pourcentages faibles pour les incorrectes.

Au palier de turbo analyse T80, la moyenne des *NCSt T80* des dix tests se situe à 0,8 (*NCSt* varie dans une plage de –2 à +2). Six épreuves obtiennent des *NCSt* ⩾0,8 : CHIMIE (0,96), COMPRE (0,96), GEOGRA (0,86), ARTACT (0,84), PHYSIQ (0,80) et BIOLOG (0,80) et quatre se situent sous cette moyenne : MATHEM (0,76), HISTOI (0,72), VOCABU (0,68) et SYNTAX (0,62).

Pour ce qui est de la **Réalisation des prédictions par test** (p. 296) mesurée à l'aide de l'indice *Rt* (moyenne des *Rq*), au palier de turbo analyse T80, cinq tests récoltent des *Rt T80* supérieurs à la moyenne (87,6) : CHIMIE (92,7), BIOLOG (90,2), MATHEM (89,6), PHYSIQ (89,4) et ARTACT (87,8). Les cinq autres tests dont l'indice *Rt T80* se situe sous la moyenne sont : SYNTAX (87,5), VOCABU (86,1), GEOGRA (85,1), HISTOI (84,8) et COMPRE (83).

Lorsque nous corrélons les valeurs de *NCSt T80* et *Rt T80* aux dix tests nous n'observons pas de corrélation (r = 0,09). Bien qu'utilisant les données liées aux pourcentages de certitude, les deux indices ne mesurent donc pas les mêmes phénomènes, ce qui est précieux du point de vue de l'analyse. Le *NCSt* permet au niveau d'une épreuve d'évaluer dans quelle mesure les résultats des questions sont cohérents du point de vue de l'utilisation des certitudes tandis que le *Rt* nous donne une indication sur la concordance moyenne entre les Taux d'Exactitude Annoncés (TEA) et les Taux d'Exactitude Observés (TEO) pour l'ensemble des questions de l'épreuve.

En ce qui concerne l'indice de **Centration par test** (p. 299), lorsqu'on prend en compte les données de tous les étudiants (palier turbo T0) deux observations semblent devoir être soulignées : d'une part la tendance à la surestimation plus élevée dans les résultats de COMPRE (*Ct T0* vaut 16 alors que les autres tests récoltent des valeurs qui varient entre –3,9 et 5,7) et d'autre part la tendance à la sous-estimation dans les résultats de PHYSIQ (*Ct T0* vaut -3,9 alors que dans les autres tests les valeurs sont positives ou très proches de zéro). A T80, lorsqu'on prend en compte que les données d'étudiants particulièrement réalistes, les différences s'amenuisent et ne varient plus qu'entre –0,1 (VOCABU) et 4,1 (GEOGRA) ce qui est logique dans la mesure où les indices *Ct T80* sont calculés à partir des résultats des sujets qui ne commettent pas plus de 20% d'erreurs dans leurs auto-estimations.

Le graphique des centrations par test calculé après turbo analyse de T50 à T90 avec un pas de 10 (voir p. 300) permet de visualiser les valeurs obtenues aux indices *Ct*. On y observe trois épreuves qui se détachent des autres par leurs *pit T90* et *pot T90* plus élevées (⩾0,75) : MATHEM, SYNTAX et PHYSIQ. L'agrandissement de la partie qui contient les sept autres épreuves (p. 301) permet de distinguer deux autres tests dont les *pit T90* et *pot T90* sont comprises entre 0,5 et 0,6 : CHIMIE et VOCABU. Pour ces cinq tests, plus les étudiants sont réalistes, plus ils ressentent les épreuves comme étant faciles et plus celles-ci le sont dans les faits (courbes montantes). En ce qui concerne les cinq épreuves qui restent et dont les *pit T90* et *pot T90* sont inférieures à 0,5 (BIOLOG, ARTACT, GEOGRA, HISTOI et COMPRE), plus les étudiants sont réalistes, plus ils les considèrent comme étant difficiles et plus elles le sont dans les faits (courbes descendantes).

L'indice *Ct* peut aussi être calculé à partir des valeurs absolues des indices de centration par question ($|Cq|$). Au niveau du test, nous avons appelé cet indice *Ct abs*, on perd alors l'information sur l'orientation des erreurs (la sur ou sous-estimation exprimée au travers du signe, + ou -) mais il devient par contre possible de comparer les valeurs de *Ct abs* avec celles obtenues aux indices *Rt* et *NCSt*. Rappelons qu'autrement (en gardant le signe), si par exemple nous corrélions *Ct* et *Rt*, nos résultats seraient biaisés par le fait qu'une erreur de sous-estimation de –10 n'est pas moins grande qu'une erreur de surestimation de +10, c'est en fait la même « quantité » d'erreur qui est en jeu (10%) mais dans le premier cas elle est orientée vers la sous-estimation (-) et dans le second cas vers la surestimation (+).

A l'aide des corrélations obtenues dans le cadre des dix épreuves MOHICAN, soulignons encore une fois la **complémentarité des trois indices *NCSt T80*, *Rt T80* et *Ct abs T80***. Le tableau ci-contre montre des corrélations non significatives proches de zéro lorsque nous corrélons les indices *Rt T80* et *Ct abs T80* avec l'indice *NCSt T80*. Par contre, nous observons une corrélation significative, négative et très élevée (-0,93) entre *Rt T80* et *Ct abs T80*. Cette dernière corrélation indique que plus la tendance à la sous ou surestimation est élevée dans les résultats, moins les prédictions se réalisent au niveau du test, ce qui est logique. On voit bien dès lors, tout l'intérêt qu'il y a à utiliser les deux indices *Rt T80* et *Ct T80* en complémentarité. En effet, *Rt T80* étant calculé au départ de la moyenne des erreurs absolues de prédictions observées pour chacun des six pourcentages de certitude de chaque question, il donne une idée précise du taux de réalisation de ces prédictions, et, *Ct T80*, étant un indice moyen plus grossier mais donnant une information supplémentaire quant à l'orientation globale des erreurs permet de compléter *Rt T80* par cette dernière information sur la tendance à la sous ou surestimation dans les résultats.

	NCSt T80	*Rt T80*
Rt T80	0,09 (ns)	
Ct abs T80	-0,02 (ns)	-0,93 (*p* ,001)

NCSt T80 non corrélé avec les deux indices précédents, bien qu'étant aussi calculé à partir des données liées à l'utilisation des pourcentages de certitude, fournit un autre type d'information lié à la cohérence d'utilisation des certitudes dans les résultats, c'est-à-dire la tendance à obtenir des certitudes élevées pour les propositions correctes et des certitudes faibles pour les incorrectes.

Enfin, au cours de l'exploration spectrale du niveau test, nous nous sommes aussi posé **la question d'une éventuelle relation entre les Taux d'Exactitude (*TE*) et les scores de Réalisme des sujets (*Rs*)**. Nous avons observé de faibles corrélations (p. 305) variant entre 0,13 et 0,25 pour sept épreuves (VOCABU, COMPRE, GEOGRA, ARTACT, BIOLOG, CHIMIE et PHYSIQ) voire pratiquement nulle (HISTOI) et deux corrélations un peu plus élevées pour les tests SYNTAX (0,33) et MATHEM (0,40). Dès lors, nous pouvons conclure d'une part qu'il existe un lien, mais qui est faible, voire très faible, pour huit épreuves sur dix et d'autre part que pour les deux tests où la liaison *TE* ⇔ *Rs* est plus forte, les corrélations demeurent peu marquées.

b) A propos de la cohérence interne des dix épreuves MOHICAN

Les performances en cohérence interne des tests ont été mesurées à l'aide d'une série de coefficients classiques : l'alpha de Cronbach (α), le coefficient de bipartition ($r_{xx'}$) avec correction de Spearman-Brown (*rS*), le coefficient de Guttman (*rG*) ainsi que le nombre de questions parallèles à ajouter à chaque test pour que l'alpha atteigne 0,8 ($kq[\alpha=0,8]$).

Ces coefficients, nous les avons calculés en utilisant deux types de matrices des résultats. D'une part les matrices binaires (*mb*), c'est-à-dire les données des réponses des étudiants sous forme de « 1 » (réponse correcte) ou de « 0 » (réponse incorrecte), donc sans l'information liée aux pourcentages de certitude qui ont accompagné ces réponses. D'autre part, les matrices spectrales (*ms*), c'est-à-dire les pourcentages de certitude en valeur négative lorsque la réponse était incorrecte et en valeur positive lorsque la réponse était correcte (p. 132).

A l'aide du tableau récapitulatif des valeurs obtenues à ces indices classiques de fidélité calculés à partir des matrices binaires et spectrales (p. 309), nous avons mis en évidence une majorité de valeurs plus élevées lorsque les matrices spectrales sont utilisées. Le tableau montre que sur 40 paires de mesures de la cohérence interne effectuées avec les matrices binaires et avec les matrices spectrales, dans 9 cas seulement les valeurs obtenues sont plus élevées avec une matrice binaire, dans 11 cas elles sont quasi identiques (les différences sont inférieures à 0,015) et dans 19 cas les valeurs obtenues sont plus élevées lorsqu'elles sont calculées à l'aide d'une matrice spectrale. On peut donc conclure que dans la majorité des tests

MOHICAN **la fidélité est améliorée par l'utilisation de matrices spectrales**, donc par le recours aux pourcentages de certitude qui sont à la base de ce type de matrice de résultats.

A l'aide des ingénogrammes de cohérence interne (p. 320), nous visualisons les disparités des performances en cohérence interne pour les dix épreuves. Trois tests obtiennent les valeurs les plus élevées : VOCABU, MATHEM et ARTACT. Trois autres tests récoltent les valeurs les moins élevées : COMPRE, CHIMIE et BIOLOG. Les quatre épreuves qui restent et dont la cohérence interne se situe de façon relativement intermédiaire entre les deux groupes précédents sont : HISTOI, PHYSIQ, SYNTAX et GEOGRA. Deux types de facteurs peuvent expliquer les différences de cohérence interne observées : d'une part le nombre de questions (les trois tests dont les performances sont les moins bonnes comportent peu de questions) et d'autre part la diversité des contenus abordés dans certaines épreuves (par exemple HISTOI qui comporte 25 questions mais qui est en fait composé de six sous-tests).

En ce qui concerne les performances en cohérence interne, nous nous sommes aussi demandé dans quelle mesure la qualité des auto-estimations des sujets pouvait influencer les valeurs obtenues au coefficient alpha de Cronbach (α) calculé d'une part sur matrice binaire (α_{mb}) et d'autre part sur matrice spectrale (α_{ms}). Dès lors, pour chacun des dix tests nous avons comparé les α_{mb} et α_{ms} obtenus aux 10 paliers de la turbo analyse (de T0 à T90 avec un pas de 10). Les courbes des graphiques (voir p. 310) montrent que les α_{ms} calculés à partir des données spectrales des sujets les plus réalistes ($Rs \geqslant 90$) sont plus élevés dans chaque épreuve. En ce qui concerne les matrices binaires (α_{mb}), nous constatons une augmentation pour toutes les épreuves sauf une (SYNTAX). Nous pouvons donc conclure que dans le cadre des épreuves MOHICAN, l'alpha s'améliore lorsqu'il est calculé sur la base des résultats des étudiants qui commettent moins de 10% d'erreurs dans leurs auto-estimations : dans 9 cas sur 10 lorsqu'on compare les α_{mb} et dans tous les cas en ce qui concerne les α_{ms}. Donc, sauf dans le seul cas de la matrice binaire du test SYNTAX, **lorsqu'on examine les données des étudiants les plus réalistes ($Rs \geqslant 90$), tous autres facteurs étant égaux, les résultats des QCM sont plus étroitement reliés entre eux que lorsqu'on envisage les données de l'ensemble des examinés**.

Enfin en ce qui concerne l'indice du Niveau de Cohérence Interne des tests (*NCIt*) qui est calculé à partir des *rpbis classiques* des propositions des questions (voir p. 262) nous remarquons que les valeurs obtenues ne dépassent jamais celles des *NCSt T80*. Comme nous le rappellerons plus loin, ceci est lié au fait que **les valeurs obtenues par les propositions des questions à l'aide des *rpbis SCT* calculés aux paliers turbo élevés sont plus marquées que celles obtenues par les *rpbis classiques***.

c) Comparaison de la cohérence interne et de la qualité spectrale des tests

Nous avons comparé les classements des 10 épreuves en fonction des scores qu'elles obtiennent en cohérence interne et en qualité spectrale (voir p. 325). Nous constatons que les classements obtenus avec les coefficients alpha calculés à l'aide des matrices binaires et des matrices spectrales sont assez stables. En effet, les deux épreuves qui obtiennent les meilleurs alpha sont toujours VOCABU et MATHEM et les classements changent peu pour les autres sauf pour COMPRE et SYNTAX qui dans le classement lié à l'indice $\alpha_{mb\ T80}$, échangent leur dernière et cinquième places.

La situation est différente en ce qui concerne les indices spectraux. En effet, nous observons des bouleversements importants dans les classements des épreuves élaborés à l'aide des indices *Rt T0*, *Rt T80*, *NCSt T0* et *NCSt T80.* L'utilisation des données des sujets les plus réalistes entraîne notamment des augmentations d'ampleurs différentes aux indices spectraux. La progression la plus spectaculaire étant celle du test COMPRE à l'indice *NCSt* (voir à ce sujet le graphique p. 295) qui passe de 0,32 à *NCSt T0* à 0,96 à *NCSt T80.*

Lorsque nous comparons les classements réalisés avec l'alpha et avec les indices spectraux, nous ne remarquons pas de similitudes sauf en ce qui concerne la dernière position attribuée au test COMPRE (3

fois sur 4 dans les deux types de classements). Ainsi, l'épreuve VOCABU peut être considérée comme étant de bonne qualité en ce qui concerne la cohérence interne et de qualité moyenne voire faible (comparativement aux autres épreuves) en ce qui concerne les indices spectraux. Nous pensons qu'il est normal d'observer des différences dans les classements dans la mesure où les nouveaux indices de qualité spectrale ne mesurent pas les mêmes phénomènes que les indices de cohérence interne classiques.

5. Faits saillants liés à l'exploration du niveau « QCM »

Nous avons procédé à une exploration de la qualité des QCM à l'aide de trois catégories d'indices. D'une part, les nouveaux indices spectraux : *NCSq* (p. 231), *Rq* (p. 242), *Cq* (p. 270) et *piq* (p. 251). D'autre part les indices classiques calculés à partir de matrices binaires : *poq mb* (p. 254), *NCIq* (p. 233), α-*q mb* (alpha calculé à partir d'une *m*atrice *b*inaire après suppression de la question *q* envisagée, voir p. 141) et $r_{qt\ mb}$ (corrélation *q*uestion-*t*otal calculée à partir d'une *m*atrice *b*inaire, p. 141). Enfin, deux indices classiques calculés à partir de matrices spectrales : α-*q ms* (voir p. 141) et $r_{qt\ ms}$ (p. 141).

a) Liaisons entre les indices classiques et spectraux d'évaluation de la qualité des QCM

Nous nous sommes d'abord demandé dans quelle mesure les indices classiques ont tendance à varier ensemble (voir p. 331) pour trois épreuves (VOCABU, MATHEM et ARTACT) dont la cohérence interne est supérieure aux sept autres (p. 309). Nous observons des liaisons très fortes entre les valeurs obtenues aux indices r_{qt} et α-*q*, qu'ils aient été calculés à l'aide de matrices binaires ou spectrales et un peu moins fortes entre ces derniers indices et les valeurs à *NCIq*. Les corrélations élevées, voire très élevées, obtenues par les *NCIq* avec r_{qt} et α-*q* calculés sur matrices binaires ou spectrales, montrent que l'indice *NCIq* fonctionne bien pour mesurer la cohérence interne des questions d'une épreuve (surtout lorsque *NCIq* est comparé avec r_{qt} et α-*q* calculés sur matrices binaires).

Nous avons ensuite comparé les valeurs obtenues aux indices spectraux (voir p. 334) : *NCSq*, *Rq*, *Cq* en valeur absolue (|*Cq*|) et *piq*. Ces indices ont été calculés en prenant en compte d'une part tous les sujets et d'autre part les étudiants dont le réalisme est supérieur ou égal à 80 (palier de turbo analyse *T80*). Pour les trois épreuves sélectionnées (VOCABU, MATHEM et ARTACT) nous remarquons des liaisons plus fortes entre l'indice *Rq T80* et |*Cq*| *T80* qu'entre ces deux indices et le *NCSq T80*. Les plus fortes liaisons entre *Rq T80* et |*Cq*| *T80* sont logiques dans la mesure où les sur ou sous-estimations élevées sont forcément accompagnées de taux d'erreurs d'estimation élevés eux aussi. Ces corrélations et celles, plus faibles, observées entre *NCSq T80* et *Rq T80* et non significatives observées entre *NCSq T80* et |*Cq*| *T80*, illustrent la complémentarité des trois indices spectraux. Le *NCSq T80* convient bien pour mesurer la cohérence d'utilisation des degrés de certitude. Le *Rq T80* est particulièrement indiqué pour évaluer la quantité d'erreurs d'auto-estimation. *Cq*, très lié à *Rq*, permet d'évaluer globalement le « sens » de ces erreurs d'auto-estimation, c'est-à-dire la tendance à la surestimation ou à la sous-estimation présente dans les données d'une question.

Enfin, nous avons comparé les valeurs obtenues aux indices classiques et aux indices spectraux par les QCM des trois épreuves (voir p. 337). Nous remarquons que des liaisons existent entre les indices spectraux et les indices de cohérence interne, mais les configurations de ces liaisons sont différentes d'une épreuve à l'autre. Pour le test VOCABU les corrélations les plus fortes sont observées entre les indices *NCSq*, *NCSq T80* et tous les autres indices de cohérence interne. Pour MATHEM on observe plus ces corrélations élevées mais c'est par contre *Rq*, *Rq T80*, |*Cq*|, |*Cq*| *T80* qui sont les plus corrélés avec les indices classiques de cohérence interne calculés à partir des matrices spectrales. Pour ARTACT, la configuration des corrélations est encore différente.

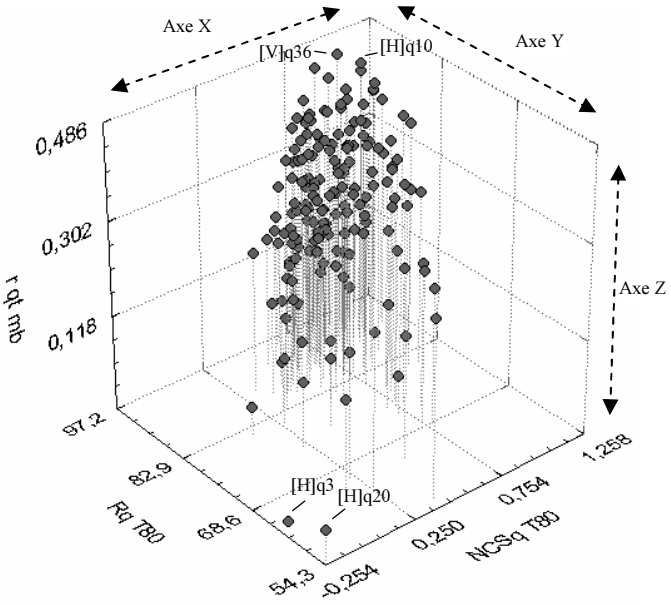
Il existe donc entre indices spectraux et indices classiques portant sur les questions des tendances à varier ensemble qui ne se présentent pas de la même manière d'une épreuve à l'autre. Dans certaines

épreuves des relations fortes existent entre une série d'indices spectraux et classiques qui n'apparaissent plus ou de façon beaucoup moins marquée dans les autres tests.

b) Identification des QCM présentant des performances faibles en qualité spectrale ou/et en cohérence interne

A l'aide de deux indices spectraux (*NCSq T80* et *Rq T80*) et de l'indice de corrélation question total calculée à partir de la matrice binaire ($r_{qt\ mb}$) nous avons analysé les 173 QCM des 10 épreuves MOHICAN. Nous proposons différentes méthodes et instruments pour mettre en évidence et visualiser les niveaux de performances des questions.

« 3D classico-spectral » des 173 QCM des 10 épreuves MOHICAN tracé en fonction des valeurs aux indices NCSq T80, Rq T80 et $r_{at\ mb}$

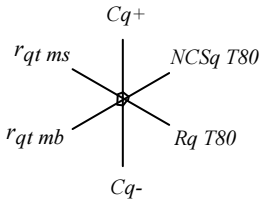


A l'aide du graphique « *3D classico-spectral* » (p. 344) nous visualisons les 173 questions dont les points forment une grappe allongée. Nous remarquons que les questions sont plus nombreuses pour les valeurs plus élevées en haut au fond du graphique que pour les valeurs plus faibles en bas vers l'avant.

Nous avons épinglé en haut du graphique les deux questions qui présentent les meilleures performances en qualité spectrale et en cohérence interne : la 36ème question du test VOCABU ([V]q36) et la 10ème du test HISTOI ([H]q10).

Les deux autres questions mises en évidence en bas du graphique ([H]q3 et [H]q20) sont celles dont les performances spectrales sont les moins élevées et dont la cohérence interne est faible.

Une autre méthode pour visualiser les niveaux de qualité spectrale et de cohérence interne des questions consiste à utiliser des ingénogrammes (graphiques à coordonnées polaires, voir p. 346).



Ici, deux indices de cohérence interne sont utilisés : la corrélation question-total calculée d'une part à l'aide d'une matrice spectrale ($r_{qt\ ms}$) et d'autre part à l'aide d'une matrice binaire ($r_{qt\ mb}$). En ce qui concerne la qualité spectrale nous avons retenu les indices *NCSq T80, Rq T80* et *Cq*. En ce qui concerne *Cq*, nous en avons extrait deux sous-indices : l'indice des surestimations ($Cq+ = Cq > 0$) et l'indice des sous-estimations ($Cq- = Cq < 0$). Dès lors, la moitié supérieure de l'axe vertical permet de mettre en évidence les questions qui présentent une tendance à la surestimation dans les résultats et la moitié inférieure celles qui présentent une tendance à la sous-estimation. Pour les indices *NCSq T80, Rq T80,* $r_{qt\ ms}$ et $r_{qt\ mb}$, plus les valeurs des indices sont élevées (donc plus les performances sont bonnes) plus les points représentant ces valeurs sur les axes sont proches du centre. Pour les indices *Cq+* et *Cq-*, plus les valeurs sont faibles (donc moins il y a de sur ou sous-estimations) plus le point sur l'axe vertical est proche du centre. L'exemple ci-dessus montre le cas d'une question dont les performances sont excellentes en cohérence interne et en qualité spectrale. Nous avons ainsi tracé les ingénogrammes des 173 questions (voir p. 347).

Nous avons aussi dressé un tableau des performances des QCM aux six indices utilisés pour la réalisation des ingénogrammes (p. 350). Pour chacun des indices trois niveaux de performances « + », « ≃ » et « - » ont été définis en fonction des plages de valeurs observées pour les 173 questions (p. 349).

Nous présentons ci-contre les ingénogrammes et le récapitulatif des niveaux de performances des huit questions qui obtiennent les valeurs les moins élevées en qualité spectrale et en cohérence interne (voir p. 356).

Tableau des performances des 8 QCM qui obtiennent les valeurs les moins élevées en qualité spectrale et en cohérence interne

		NCSq T80	*Rq T80*	*Cq+ T80*	*Cq- T80*	*$r_{qt\,mb}$*	*$r_{qt\,ms}$*
[H]q3		-	-	-		-	-
[H]q20		-	-	-		-	-
[V]q5		-	-	-		-	≃
[V]q12		-	-	-		≃	≃
[S]q1		-	-	-		≃	≃
[V]q27		-	≃		≃	-	≃
[A]q14		-	≃	+		-	≃
[B]q5		≃	≃	≃		-	-

Nous constatons d'une part que parmi ces huit questions, cinq ([H]q3, [H]q20, [V]q5, [V]q12, [S]q1) présentent une tendance prononcée à la surestimation (*Cq+ T80* = « - »). D'autre part, une seule ([V]q27) contient une tendance à la sous-estimation dans ses résultats et cette tendance est moyenne (*Cq- T80* = « ≃ »). Enfin, [A]q14 ne présente pas de tendance anormale à la surestimation (*Cq+ T80* = « + ») et pour [B]q5 la tendance à la surestimation est moyenne (*Cq+ T80* = « ≃ »).

Nous rappellerons plus loin les résultats des analyses détaillées des propositions de ces huit questions. Parmi elles, [H]q3 et [H]q20 sont les deux seules sur lesquelles les experts s'accordent à dire que de sérieuses anomalies dans les questions elles-mêmes expliquent les faibles performances spectrales et de cohérence interne. Notons qu'il s'agit donc des deux seules QCM pour lesquelles trois faisceaux d'arguments provenant des analyses de la cohérence interne, de la qualité spectrale et du contenu par des experts, convergent vers un diagnostic de disfonctionnement.

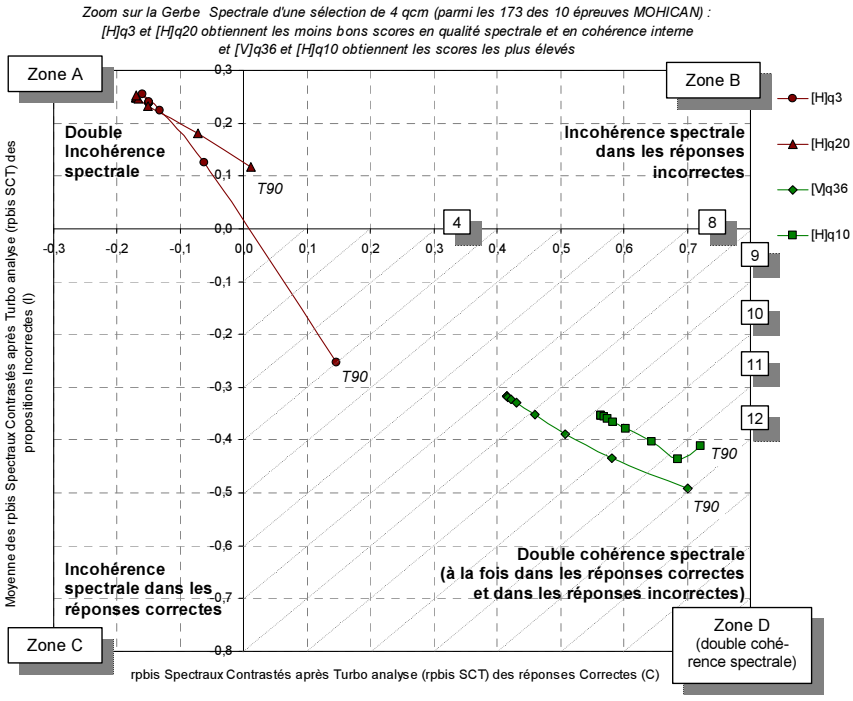
c) Lien entre qualité spectrale et cohérence interne peu élevées et surestimation prononcée

Bien qu'on observe une tendance à la surestimation dans cinq questions sur huit dont les performances en cohérence interne et spectrale sont faibles (notamment dans [H]q20 où 36% des étudiants ont fourni une réponse incorrecte avec le pourcentage de certitude le plus élevé), nous ne pouvons pas conclure qu'il y a forcément incohérence interne ou spectrale là où il y a surestimation. En effet, deux contre exemples : [V]q39 et [M]q17 (voir tableau p. 349), montrent qu'il est possible d'obtenir des valeurs moyennes aux indices de cohérence interne et à l'indice de cohérence spectrale tout en observant une tendance élevée à la surestimation (parmi les 173 questions, [M]q17 contient la plus forte tendance à la surestimation avec une *Cq+ T80* égale à 0,43, et, en ce qui concerne [V]q39, la *Cq+ T80* est élevée et vaut 0,33). De plus, lors du débriefing des épreuves, les experts n'ont pas relevé de problèmes particuliers liés aux contenus de [V]q39 et de [M]q17.

d) Brins spectraux contrastés en fonction des performances des questions

Basés sur les *NCSq* (p. 231), les Brins spectraux par question (*Bsq*) nous permettent de visualiser les performances en cohérence spectrale des QCM sur des graphiques cartésiens (p. 228).

Reprenons le graphique détaillé p. 357. Nous y avons tracé les *BSq* de deux couples de questions très contrastées du point de vue de leurs performances.



Les deux questions les moins performantes ([H]q3 et [H]q20) à la fois en cohérence spectrale et en cohérence interne, et, où les experts ont décelé des anomalies majeures, sont situées dans la partie supérieure gauche. Au palier de turbo analyse T80, le point qui précède « T90 » sur chaque courbe, nous remarquons que [H]q3 et [H]q20 se situent dans la zone de « double incohérence spectrale ». Lorsqu'une QCM se trouve dans cette zone de non qualité spectrale, cela signifie que les étudiants dont les données ont été sélectionnées (en l'occurrence ceux qui ont un Rs⩾80, donc ceux qui globalement commettent peu d'erreurs dans leurs auto-estimations) ont eu tendance à accompagner les propositions incorrectes de certitudes élevées et la réponse correcte de certitudes faibles. Après analyse des *BSq* des 171 autres questions des 10 épreuves MOHICAN (les graphiques sont présentés en annexe, voir pp. 511 à 521) nous pouvons conclure qu'il s'agit là des deux seuls cas de questions qui présentent une double incohérence spectrale à ce palier de turbo analyse T80. Au palier turbo T90, [H]q3 pénètre dans la zone de qualité spectrale tandis que [H]q20 pose encore des problèmes d'incohérence spectrale dans les réponses incorrectes. Nous rappellerons plus loin qu'une analyse détaillée des propositions éclairée par l'avis des experts permet d'expliquer l'origine du disfonctionnement spectral (corroboré par les faibles performances en cohérence interne) de ces deux questions.

Dans la partie inférieure droite du graphique nous trouvons deux autres questions : les plus performantes en cohérence spectrale parmi les 173, il s'agit de [V]q36 et [H]q10. Les traits pointillés obliques situés dans la zone de double cohérence spectrale nous donnent des repères quant aux niveaux de qualité. A T90, ces deux questions atteignent le niveau 12 de qualité en cohérence spectrale (sur une échelle de 1 à 20).

Nous avons également isolé sur une gerbe spectrale les *Bsq* de six autres questions présentées dans le tableau des performances des huit QCM globalement moins performantes (voir p. 356) : [V]q5 [V]q12, [S]q1, [V]q27, [A]q14 et [B]q5 (graphique p. 359). A T80 ces six questions se situent toutes dans la zone

de double cohérence spectrale. Donc, malgré de faibles performances observées en qualité spectrale et en cohérence interne (relativement aux autres questions des 10 épreuves), ces six QCM, contrairement à [H]q3 et [H]q20, présentent une double cohérence spectrale lorsqu'on prend en compte les données des étudiants qui commettent peu d'erreurs dans leurs auto-estimations. Les niveaux de qualité en cohérence spectrale n'atteignent cependant pas les valeurs observées pour les deux questions les plus performantes, en effet, à T90, [B]q5 atteint le niveau « 9 », [V]q27 et [S]q1 le niveau « 6 », [V]q12 et [A]q14 le niveau « 5 » et [V]q5 le niveau « 3 ». Nous reviendrons sur ces questions dans nos conclusions liées à l'exploration du niveau « proposition », notons cependant déjà que l'analyse du contenu de ces six QCM ne révèle pas de problème majeur alors que de sérieuses anomalies ont été détectées dans les deux questions globalement moins performantes et présentant une double incohérence spectrale à T80 : [H]q3 et [H]q20.

6. Conclusions liées à l'exploration du niveau « PROPOSITIONS »

Nous avons analysé les propositions des questions mises en évidence dans le cadre de l'exploration du niveau d'analyse « QCM » à l'aide des *rpbis classiques* et des rpbis Spectraux Contrastés calculés après Turbo analyse (*rpbis SCT*). Pour chaque proposition nous avons aussi envisagé les nombres et pourcentages de réponses ainsi que les certitudes moyennes, et ce, à chaque palier de turbo analyse (le plus souvent de T50 à T90 étant donné les faibles différences d'effectifs aux paliers les plus bas).

Contrairement aux *rpbis classiques* où les données de tous les sujets sont impliquées dans le calcul des corrélations, (voir p. 171), tous les couples de mesures ne sont pas utilisés dans le calcul des *rpbis SC* et *rpbis SCT* des propositions incorrectes (voir p. 178). Dès lors lorsque le nombre de couples de mesures impliquées dans le calcul d'une corrélation point bisériale spectrale est peu élevé (aux paliers de turbo analyse élevés où les effectifs des propositions sont les plus faibles), nous avons utilisé la formule de Fisher transformation du *r* en *t* de Student (voir p. 164) pour effectuer un test de signification $r \neq 0$.

Nous avons également confronté les valeurs des *rpbis classiques* et *rpbis SCT* de chaque proposition aux avis exprimés par les experts du contenu des questions consultés lors de deux réunions de débriefing des épreuves auxquelles nous étions présents.

Nous disposons dès lors de trois faisceaux d'éclairage pour l'analyse des propositions des questions : (1) la discrimination des propositions (*rpbis classiques*), (2) la cohérence spectrale d'utilisation des pourcentages de certitude associés aux choix des propositions (*rpbis SCT*) et (3) les avis des experts du contenu des questions.

Parmi les 173 questions des 10 épreuves MOHICAN, nous en avons relevé 16 qui nous paraissent devoir être analysées en profondeur au niveau d'exploration « PROPOSITIONS ». Il s'agit d'une part, d'une première série de 8 questions mises en évidence au niveau d'exploration « QCM » pour leurs faibles performances globales aux indices de cohérence interne ($r_{qt\ mb}$ et $r_{qt\ ms}$) ainsi qu'aux indices de qualité spectrale (*NCSq* et *Rq*). D'autre part, d'une seconde série de 8 questions épinglées pour les valeurs illogiques récoltées par les propositions aux *rpbis classiques* (dans les 8 cas) et les *rpbis spectraux* (dans 2 cas seulement). Notons que les propositions des 157 autres questions (173 - 16 = 157) des 10 épreuves obtiennent toutes des valeurs attendues aux *rpbis classiques* et *rpbis spectraux* (voir annexe, p. 544).

Lors des analyses de ces 16 questions « suspectes » (car elles qui obtiennent des *rpbis classiques* ou/et *rpbis spectraux* anormaux) nous avons remarqué des différences notables entre les valeurs récoltées aux trois faisceaux d'éclairage de la qualité des propositions. Les indices *rpbis classiques* et indices spectraux (*rpbis SC*, *rpbis SCT80* et *rpbis SCT90*) ne se comportent pas forcément de la même manière. Des différences apparaissent aussi au sein des rpbis spectraux, plus particulièrement entre les *rpbis SC* d'une part, et, d'autre part, les *rpbis SCT80*, *rpbis SCT90*. De plus, les avis des experts ne pointent pas toujours les propositions mises en évidence à l'aide des valeurs observées aux différents types de rpbis.

a) Récapitulatif des performances des propositions des 16 questions « suspectes »

Le tableau ci-dessous signale pour chacune des 16 questions qui récoltent des valeurs anormales aux *rpbis classiques* ou/et aux *rpbis spectraux*, les propositions qui posent problèmes selon que l'on se réfère à l'avis des experts ou à l'interprétation des *rpbis classiques*, des *rpbis SC*, des *rpbis SCT80* et des *rpbis SCT90*.

		Avis experts	*rpbis classique*	*rpbis SC*	*rpbis SCT80*	*rpbis SCT90*
8 **QCM** dont les indices de cohérence interne et de qualité spectrale sont les moins élevés	[H]q3	↗P1 ↗P6	↗P1 ↗P6	↗P1↗P5 **FA** ↗P6	↗P1 ↗P6	×
	[H]q20	↗P3 ↗P6	↗P3 ↗P6	↗P3 ↗P6	↗P3 ↗P6	↗P3 ↗P6
	[V]q5	ok	↗P5 **FA**	×	✓	✓
	[S]q1	ok	✓	↗P5 **FA** ↗P6 **FA**	✓	✓
	[V]q12	ok	✓	↗P6 **FA**	✓	✓
	[V]q27	ok	↗P1 **FA**	↗P1 **FA** ↗P3 **FA**	↗P1 **FA**	✓
	[A]q14	↗P3 ↗P4	×	↗P3 ↗P4	✓	✓
	[B]q5	ok	↗P3 **FA**	×	✓	✓
8 autres QCM épinglées après analyse des rpbis des **propositions**	[V]q43	ok	↗P3 **FA**	✓	✓	✓
	[B]q6	↗P2 ↗P3	↗P2	↗P2 ↗P3	✓	✓
	[G]q1	ok	↗P4 **FA**	✓	✓	✓
	[A]q24	ok	↗P1 **FA**	✓	✓	✓
	[H]q16	↗P1	↗P4 **FA**	↗P1 ↗P4 **FA**	✓	✓
	[B]q10	ok	↗P3 **FA**	✓	✓	✓
	[Ch]q1	ok	↗P4 **FA**	✓	✓	✓
	[M]q17	ok	↗P5 **FA**	✓	✓	✓

Légende :

« ✓ » : *globalement, le fonctionnement des propositions est cohérent ;*

« ✓ » : *mise en évidence d'un fonctionnement global cohérent, mais de façon peu marquée ;*

« ↗P… » : *(suivi d'un n°, nous soulignons en pointillés la proposition correcte) désigne une proposition qui fonctionne de manière incohérente ;*

« ↗P… » : *mise en évidence du fonctionnement incohérent d'une proposition, mais de façon peu marquée ;*

« × » : *les valeurs des indices des propositions sont proches de zéro;*

« **FA** » : *Fausse Alerte ;*

: *Non détection de propositions que les experts désignent comme étant problématiques.*

b) Convergence forte des analyses spectrales et classiques pour les propositions de deux questions particulièrement problématiques

Dans le tableau qui précède nous relevons deux questions qui se démarquent des autres du point de vue des valeurs obtenues aux indices de cohérence interne et de qualité spectrale, il s'agit de [H]q3 et [H]q20.

Les conclusions de l'analyse de [H]q3 (la question dont l'amorce est « *Quel est le secrétaire actuel de l'OTAN ?* », voir annexe, p. 494) indiquent qu'il existe un problème pour le distracteur P1 (« *Javier Solana* ») et la réponse correcte P6 (« *aucune* »[76]). En ce qui concerne P1 un fonctionnement incohérent est mis en évidence à l'aide des *rpbis classiques*, des *rpbis SC* et des *rpbis SCT80* (voir tableau, p. 396). Ce diagnostic recoupe celui des experts qui ont signalé que le contexte des évènements liés à l'intervention de

[76] Une Solution Générale Implicite (SGI), voir p. 6.

l'OTAN en Yougoslavie en 1998-1999, pendant l'année scolaire qui a précédé le test et qui ont été largement relayés par les médias, pouvait expliquer le fait que ce distracteur fonctionne comme une réponse correcte (voir détails, p. 375). En effet, la première proposition « *Javier Solana* » était correcte pendant l'année scolaire 1998-1999 mais incorrecte à partir du 4 août 1999 (pendant les vacances scolaires), date du remplacement de Javier Solana à la tête de l'OTAN, et *a fortiori* incorrecte lors de l'administration du test entre le 4 et le 8 octobre 1999. En ce qui concerne la réponse correcte P6 « *aucune* », seul le *rpbis SC* indique clairement (par une valeur négative significativement différente de zéro) un fonctionnement incohérent : les étudiants qui choisissent la réponse correcte P6 ont tendance à accompagner leurs choix de pourcentages de certitude moins élevés que ceux qui répondent incorrectement. Ceci dit, les *rpbis SC* désignent aussi une autre proposition, P5, mais dont l'analyse du contenu montre qu'elle ne contient pas d'anomalie particulière. En ce qui concerne les *rpbis classiques* et *rpbis SCT80* de P6, les valeurs sont très proches de zéro, mettant moins clairement en évidence le problème que le *rpbis SC*. Enfin, notons que les *rpbis SCT90* sont tous non significativement différents de zéro, donc guère informatifs pour cette question. Les *rpbis SCT90* ne parviennent pas ici à détecter les propositions jugées problématiques par les experts. On peut donc conclure dans le cas de [H]q3, que les informations fournies par les *rpbis classiques*, les *rpbis SC* et les *rpbis SCT80* se rejoignent et permettent de corroborer les conclusions de l'analyse du contenu des propositions P1 et P6. Les *rpbis SC* le font de façon plus nette en ce qui concerne P6, mais en déclenchant aussi une « fausse alerte » [77] en ce qui concerne P5 dont la vérification du contenu montre qu'elle ne pose pas de problème.

En ce qui concerne [H]q20 (« *Quelle chaîne de TV crée et diffuse la séquence 'No Comment' ?* », voir annexe, p. 496) les analyses des propositions effectuées à l'aide des *rpbis SC*, *rpbis SCT* et *rpbis classiques* convergent. Pour ces différents types d'indices, les propositions P3 (« *CNN* », réponse considérée initialement comme correcte) et P6 (« *aucune* ») fonctionnent de façon incohérente. L'analyse du contenu de la question (p. 380) montre qu'il y a eu erreur d'encodage en ce qui concerne la réponse correcte qui n'est pas P3 mais bien P6. Lorsqu'on compare les configurations des différents types de rpbis (voir tableau p. 397), on remarque que se sont les *rpbis SC* qui mettent le mieux en évidence le problème d'erreur d'encodage de la réponse correcte : P3 obtient une valeur négative significativement différente de zéro et P6 la valeur positive la plus élevée. Bien que le problème soit aussi mis en évidence à l'aide des *rpbis SC T80* et *rpbis SCT90*, ces rpbis spectraux sont moins marqués en ce qui concerne P3. Enfin, pour ce qui est des *rpbis classiques*, les valeurs observées sont peu marquées à la fois pour P3 et P6.

On voit donc que pour ces deux questions, [H]q3 et [H]q20, les plus problématiques des 173 QCM des 10 épreuves MOHICAN, les rpbis classiques et spectraux se rejoignent. Cependant, les *rpbis SC* et *rpbis SCT80* permettent une meilleure mise en évidence (plus contrastée) des propositions pointées par l'analyse du contenu que les *rpbis classiques*.

c) Convergence des avis des experts et des *rpbis SCT90* pour six questions dont les performances globales figurent parmi les moins élevées

Dans le tableau de synthèse des performances des propositions précédent on retrouve six autres questions qui ont été sélectionnées lors de l'exploration du niveau « QCM » pour leurs faibles performances en cohérence interne et en qualité spectrale. Il s'agit de [V]q5, [S]q1, [V]q12, [V]q27, [A]q14 et [B]q5.

Pour ces six questions, les situations se présentent assez différemment d'un cas à l'autre. En ce qui concerne [V]q5, le *rpbis classique* positif du distracteur P5 n'a pas trouvé d'explication lors de la réunion de débriefing de février 2000 (voir p. 383). Cette incohérence n'est pas non plus corroborée par les *rpbis spectraux* (voir tableau récapitulatif, p. 398) et dès lors, nous pensons qu'il s'agit là d'une « fausse

[77] C'est à dire la mise en évidence d'une valeur anormale récoltée par une proposition alors que les experts du contenu n'y décèlent pas d'anomalie particulière.

alerte ». Signalons à ce sujet que lors des corrections des épreuves traitées et analysées dans le contexte du Système Méthodologique d'Aide à la Réalisation de Tests (SMART, description p. 57), nous sommes souvent confronté à ce problème des « fausses alertes » provoquées par des valeurs *rpbis classiques* insatisfaisantes et qui après consultation des examinateurs et analyse du contenu des questions ne trouvent pas d'explications. Nous constatons que pour [V]q5, les *rpbis spectraux*, bien que peu marqués, ne montrent pas de problèmes majeurs et sont plus en phase avec les avis des experts que les *rpbis classiques*.

Dans le cas de [S]q1, ce sont les *rpbis SC* qui déclenchent deux fausses alertes pour le distracteur P5 (valeur positive) et pour la réponse correcte P6 (valeur légèrement négative) (voir p. 399). Les experts présents lors de la réunion de débriefing ont expliqué le choix massif de P5 par 49% des étudiants (seule proposition qui contienne une préposition devant le pronom relatif) et souligné la difficulté de cette question (liée au fait que l'amorce comporte plusieurs subordonnées) mais n'ont relevé aucune anomalie grave. Une incohérence spectrale moins marquée est aussi relevée au niveau du *rpbis SC* de P6 (réponse correcte). Par contre, il existe une convergence des éclairages fournis par les analyses du contenu et des valeurs des *rpbis classiques*, *rpbis SCT80* et *rpbis SCT90*.

En ce qui concerne [V]q12 (« *…à l'instar de…* »), l'analyse du contenu de la question ne met pas en évidence d'anomalie grave (p. 387) et nous ne trouvons pas d'explication au *rpbis SC* incohérent mais peu marqué de la réponse correcte P6 (p. 399). Les *rpbis SC* déclenchent donc une fausse alerte pour P6 tandis que les *rpbis SCT80* et *rpbis SCT90* corroborent les avis des experts, ainsi que les *rpbis classiques* mais en ce qui concerne ces derniers, de façon moins marquée.

Les experts du contenu n'ont pas trouvé d'anomalies graves dans la question [V]q27 (p. 388). Cependant, les *rpbis classiques*, *rpbis SC* et *rpbis SCT80* pointent des problèmes au niveau de la proposition incorrecte P1 (une fausse alerte) et aussi au niveau de P3 en ce qui concerne les *rpbis SC* (une seconde fausse alerte). Par contre, les *rpbis SCT90* sont ici plus en phase avec les avis des experts.

Pour ce qui est de la question [A]q14 (« *Lequel de ces 5 films traite de la 'décimation' ?* », voir annexe, p. 498), l'analyse du contenu de la question (p. 390) montre que la proposition correcte P3 (« *Les sentiers de la gloire* ») ne peut pas être interprétée comme étant une illustration du sens classique et premier du mot « décimation ». L'analyse du contenu nous permet aussi de comprendre le choix de P4 (« *Full Metal Jacket* ») par une proportion importante d'étudiants (23%). La comparaison des valeurs obtenues par les différents types de rpbis (voir tableau p. 400) montre que les *rpbis SC* mettent en évidence un problème au niveau de P3 et P4 tandis que les *rpbis SCT80* (peu marqués) et *rpbis SCT90* indiquent que la cohérence spectrale de la question est correcte. Les *rpbis classiques* ne discriminent pas les étudiants en fonction du critère du nombre de réponses correctes fournies à l'ensemble des questions du test. Les *rpbis SCT80*, *rpbis SCT90* et *rpbis classiques* ne détectent donc pas P3 et P4.

Enfin, en ce qui concerne [B]q5, les experts n'ont pas détecté d'anomalies dans les propositions ce qui est corroboré par les *rpbis SCT80* et *rpbis SCT90,* mais pas par les *rpbis SC* tous très proches de zéro, ni par les *rpbis classiques* qui mettent en évidence P3 (fausse alerte), mais de façon peu marquée.

Donc, pour ces six questions suspectes ([V]q5, [S]q1, [V]q12, [V]q27, [A]q14 et [B]q5), comme le tableau précédent (p. 434) permet de le visualiser, en ce qui concerne les *rpbis classiques* on observe un parallélisme avec les avis des experts pour 2 questions sur 6 ([S]q1 et [V]q12). En ce qui concerne les *rpbis SC* nous constatons qu'ils sont en phase avec les avis des experts pour une question ([A]q14). Les *rpbis SCT80* le sont pour quatre questions ([V]q5, [S]q1, [V]q12 et [B]q5). Enfin, les *rpbis SCT90* recoupent les analyses du contenu des questions dans 5 cas sur 6 ([V]q5, [S]q1, [V]q12, [V]q27 et [B]q5).

d) Meilleure détection des problémes par les *rpbis SC* pour les 8 questions épinglées après analyse des rpbis des propositions

Nous avons également examiné les propositions d'une seconde série de 8 questions épinglées parce qu'elles présentent des valeurs anormales aux *rpbis spectraux* et/ou aux *rpbis classiques* : [V]q43, [B]q6, [G]q1, [A]q24, [H]q16, [B]q10, [Ch]q1 et [M]q17.

Les experts désignent deux questions qui posent problèmes. Il s'agit de [B]q6 où deux propositions (P2 et P3) sont sujettes à controverse (voir p. 404) et [H]q16 dont le distracteur P1 est « acceptable » (voir p. 406). L'expertise du contenu des six autres questions n'a pas révélé la présence d'anomalies dans les propositions.

Les valeurs des *rpbis classiques* indiquent la présence d'une proposition problématique dans chacune des huit questions. Ainsi, les réponses correctes de sept questions obtiennent une valeur positive située sous le seuil repère et une question ([M]q17) obtient un *rpbis classique* positif pour un distracteur (P5). A chaque fois, il s'agit de fausses alertes sauf pour P2, une des deux propositions controversées de [B]q6.

En ce qui concerne les *rpbis spectraux*, nous remarquons qu'ils sont beaucoup plus en phase avec les avis des experts. Les *rpbis SC* indiquent des incohérences spectrales dans les deux propositions controversées de [B]q6. Les *rpbis SC* signalent aussi deux autres propositions problématiques dans la question ([H]q16) : le distracteur P1 accepté comme réponse correcte par les experts et la proposition P4 dont la mise en évidence constitue la seule fausse alerte des *rpbis SC* pour ce second lot de 8 questions. Les *rpbis SCT80* et *rpbis SCT90* ne pointent pas de propositions où les pourcentages de certitude auraient été utilisés de façon incohérente mais ils n'attirent pas non plus l'attention sur les propositions litigieuses de [B]q6 et [H]q16.

Donc, pour ce second lot de huit questions, seuls les *rpbis SC* mettent en évidence les trois propositions problématiques des deux questions signalées par les experts (ce qui se traduit par l'absence de cases sur fond jaune pour les *rpbis SC* dans le tableau p. 434).

e) Moins de « fausses alertes » dans le cas des *rpbis spectraux*

Envisageons maintenant l'ensemble des 16 questions suspectes (voir tableau, p. 434). Lorsque nous comparons les avis des experts avec les configurations des différents types de rpbis nous remarquons que les *rpbis spectraux* déclenchent moins de fausses alertes (signalée par « FA » en rouge dans le tableau) que les *rpbis classiques* :

- dix fausses alertes pour les *rpbis classiques* ;
- sept pour les *rpbis SC* ;
- une pour les *rpbis SCT80 ;*
- aucune pour les *rpbis SCT90.*

Les *rpbis SCT80* et *rpbis SCT90* déclenchent le moins de fausses alertes mais ils ne repèrent pas toutes les propositions désignées comme étant problématiques par les experts.

Lorsqu'on compare les *rpbis classiques* et *rpbis SC* des 16 questions du tableau, on remarque que les rpbis classiques déclenchent plus de fausses alertes (10 « FA » contre 7) et qu'ils ne détectent pas certaines propositions problématiques (3 cases sur fond jaune) alors qu'elles sont toutes signalées par les *rpbis SC* (aucune case sur fond jaune).

f) Implications au niveau des pratiques du Système Méthodologique d'Aide à la Réalisation de Tests

La pratique quotidienne d'un Système Méthodologique d'Aide à la Réalisation de Tests comme le SMART a encore plus besoin d'analyse de la qualité que les tests MOHICAN pour les raisons suivantes :

- Beaucoup plus de tests, questions et propositions sont traitées dans l'espace d'une année académique. Pour MOHICAN, c'était 10 tests, 173 questions et 1.119 propositions. Le SMART, lui, en 1999-2000 a traité 235 tests comptant en moyenne une trentaine de questions à 5 propositions, soit 35.250 propositions, environ 30 fois plus ;

- Beaucoup moins d'étudiants par test, en moyenne 100, ce qui rend les analyses turbo souvent impossibles au seuil T90 et ce qui rend les fausses alertes plus fréquentes étant donné la plus grande « instabilité » de la population (les réponses d'un petit nombre d'étudiants ayant un impact plus grand sur les indices).

Dès lors, les qualités de « moins de fausses alertes » et de « meilleure détection » prennent beaucoup d'importance pour la pratique. Heureusement, il est possible de combiner les approches et de compenser les faiblesses de l'une par les forces de l'autre.

Jusqu'à présent, les analyses d'items se faisaient en utilisant deux filtres : celui des indices *rpbis classiques* et celui des *avis des experts*. Ces filtres pouvant dans le meilleur des cas détecter les mêmes questions suspectes, ou, dans certains cas, détecter des questions que l'autre filtre ne détecte pas, ou encore, dans le pire des cas, tous les deux « laisser passer » des questions cependant problématiques. Désormais les *filtres spectraux* viennent s'ajouter aux deux précédents et diminuent les risques de non-détection d'items suspects tout en offrant l'avantage par rapport aux *rpbis classiques* de déclencher moins de fausses alertes.

Ces nouveaux filtres spectraux sont totalement originaux car ils introduisent une autre expertise que celle des enseignants : celle des apprenants. Nous disons expertise car ils sont les seuls à avoir cet accès privilégié à leur degré de doute, et ce, pour chacune des réponses qu'ils fournissent.

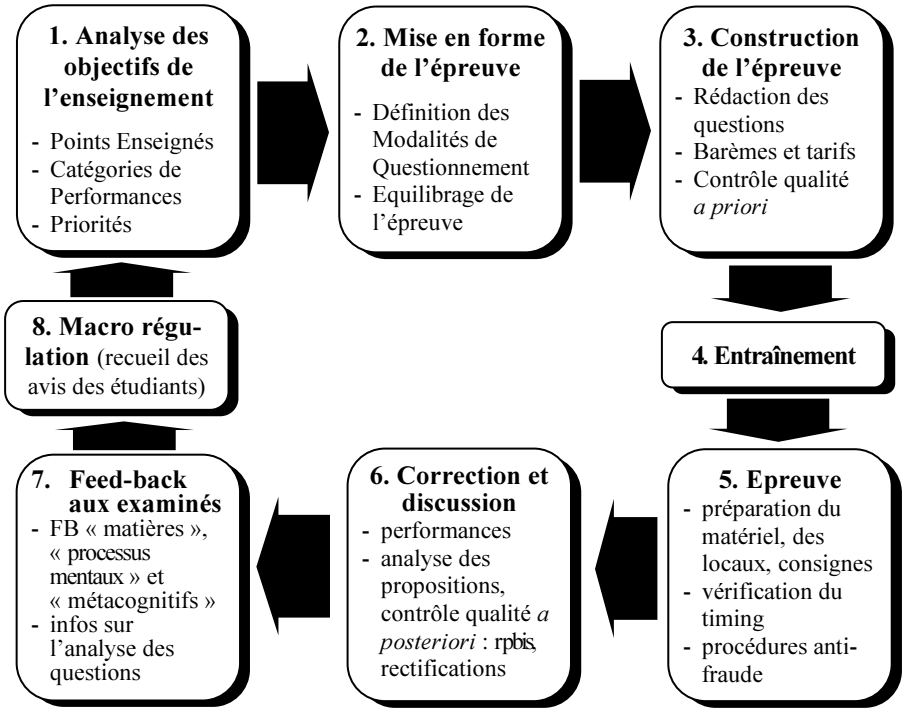
B. Perspectives

Quelles pourraient être les retombées pratiques de l'analyse spectrale appliquée aux tests standardisés d'évaluation des acquis des étudiants universitaires ?

Nous l'avons montré précédemment, l'analyse spectrale d'une épreuve ayant recours aux QCM apporte un nouvel éclairage sur la qualité des tests, des questions et des propositions au sein de celles-ci. Comment les enseignants d'une institution d'enseignement supérieur comme l'Université de Liège (ULg) pourraient-ils exploiter les apports de l'analyse spectrale des épreuves standardisées ? Dans quel contexte d'utilisation ? Avec quels bénéfices ?

1. Vers l'autorégulation des trois niveaux d'exploration « TEST - QCM - PROPOSITIONS » de la qualité classique et spectrale des épreuves

Dans la première partie de cette thèse nous avons décrit le Système Méthodologique d'Aide à la Réalisation de Tests (SMART, voir p. 57), une cellule d'aide aux enseignants de l'ULg qui souhaitent obtenir un soutien logistique et méthodologique dans la réalisation des évaluations des étudiants. Dans le cadre du SMART nous proposons un modèle de construction en « *spirale de qualité* » des examens standardisés qui comporte 8 étapes (voir explications détaillées, pp. 74-92).



C'est dans le contexte de la 6ème étape de « correction et discussion » de ce cycle de réalisation que doivent s'insérer les instruments d'analyse spectrale d'une épreuve ayant recours aux pourcentages de certitude.

Actuellement, lors de la correction, un contrôle *a posteriori*[78] de la qualité des questions est effectué à l'aide des seuls *rpbis classiques*. Ceux-ci sont calculés puis commentés par l'équipe du SMART et ensuite envoyés à l'enseignant. Ce dernier sur base des informations et avis transmis prend les décisions de rectification de l'épreuve (par exemple la suppression d'une question, la valorisation d'un distracteur, …). A ce stade, des discussions sont fréquentes entre enseignants et scientifiques du SMART spécialisés dans les problèmes de testing standardisés à propos des mesures à prendre. Lorsque les décisions de rectification sont prises, le SMART exécute les demandes et communique une seconde version des résultats. Cette

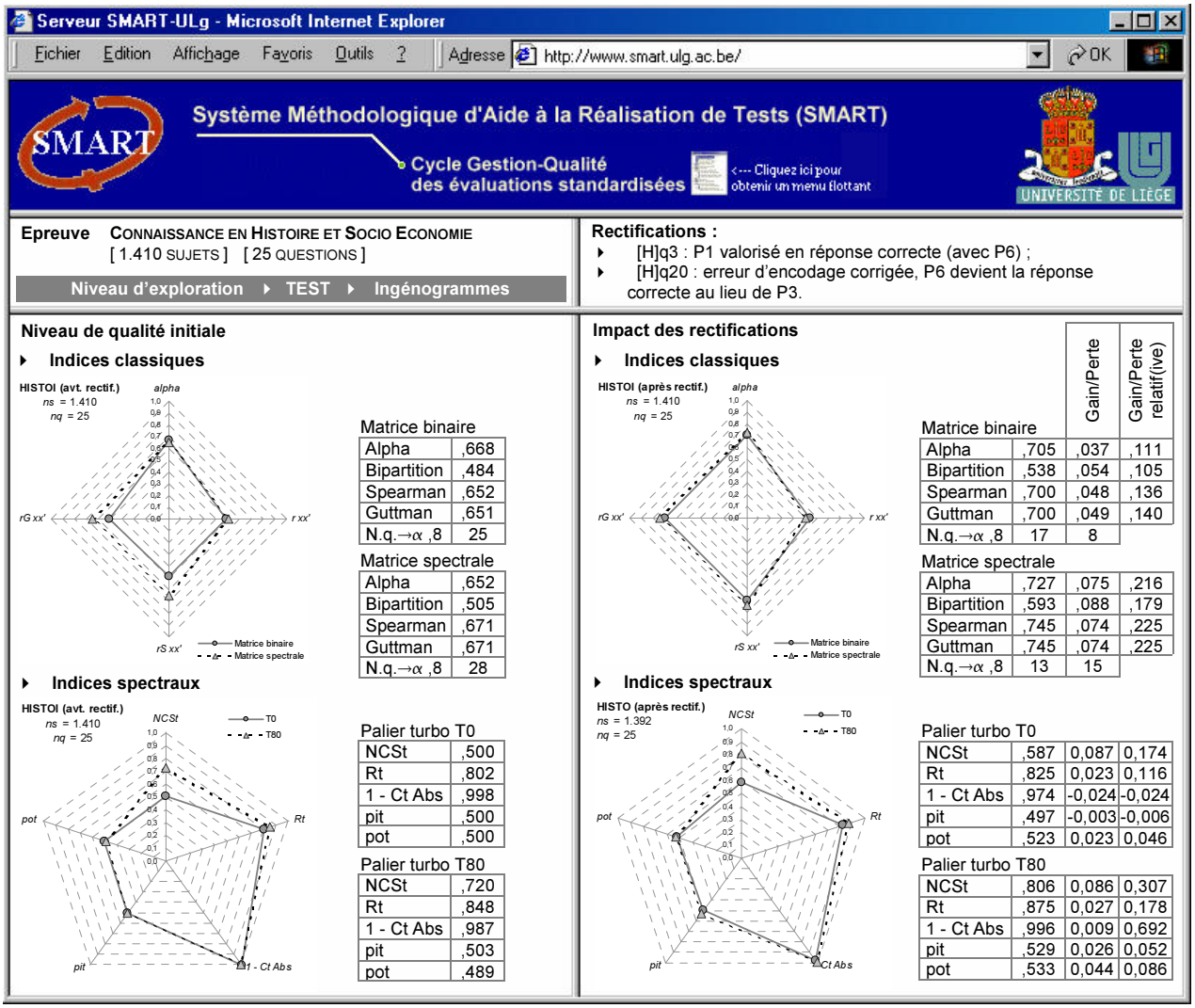
[78] C'est à dire <u>après</u> que l'épreuve ait eu lieu, par opposition aux contrôles qualités *a priori* qui ont lieu avant la mise en œuvre de l'épreuve (*cf.* étape « 3. Construction de l'épreuve »).

version peut encore être rediscutée et améliorée, jusqu'à ce que l'enseignant juge la qualité *a posteriori* de son épreuve suffisante.

Depuis l'année académique 1999-2000, nous travaillons à la mise en place d'une série de procédures informatisées qui a terme permettront (1) de procéder à l'analyse classique ou/et spectrale de la qualité d'une épreuve en envisageant les trois niveaux d'exploration « TEST – QCM – PROPOSITIONS » et non plus seulement du seul niveau « PROPOSITIONS », (2) de rectifier un test par la **suppression de questions**, la **valorisation de distracteurs** ou le **changement de réponses correctes** (voir à ce propos l'exemple de régulation d'un test formatif ayant eu lieu en novembre 2000, p. 47) et (3) de mesurer l'impact de ces rectifications sur la qualité de l'épreuve (nous détaillerons ce point plus loin). Ces outils en cours de réalisation sont conçus pour pouvoir être utilisés **à distance** soit par l'équipe du SMART, soit par les enseignants qui souhaitent procéder eux-mêmes à la régulation de la qualité de leurs tests (d'où le terme « *autorégulation* » dans le titre de cette section).

2. Vers de nouvelles interfaces de gestion de la qualité des épreuves

Comment présenter les indices édumétriques d'évaluation de la qualité classique et spectrale en vue d'améliorer les épreuves ? Pour répondre à cette question, reprenons le cas du test de Connaissance en Histoire et Socio Economie. Voici une façon d'exposer les résultats des analyses classiques et spectrales du niveau « TEST » dans la perspective d'une interface de consultation et de régulation à distance de la qualité des épreuves. L'écran ci-dessous présente une comparaison à l'aide d'ingénogrammes de la qualité obtenue avant (à gauche) et après (à droite) rectification du test.



La partie gauche montre donc les ingénogrammes et les valeurs obtenues aux indices classiques et spectraux avant la mise en œuvre des modifications (« *Niveau de qualité initiale* ») tandis que la partie droite concerne les changements observés après rectifications de deux questions du test (« *Impact des rectifications* »). Un résumé des améliorations apportées aux questions figure en haut de la partie droite. Ces modifications ont été réalisées conformément aux suggestions proposées suite à l'analyse des indices et du contenu de ces questions. Pour [H]q3 (voir p. 375), P1 est valorisé en réponse correcte (avec P6) et pour [H]q20 (p. 380), l'erreur d'encodage est corrigée, P6 devient la réponse correcte au lieu de P3.

3. Vers des mesures d'impact classique et spectral des régulations

Sur l'écran précédent nous avons placé près des ingénogrammes[79] (voir détails et comparaisons, pp. 320 et 323) les valeurs récoltées par les indices qui ont permis leur construction. Deux colonnes supplémentaires sur fond grisé qui figurent dans la partie droite permettent de chiffrer l'impact de ces rectifications. Les formules utilisées reprennent les principes de celle de Cox et Vargas (1966) pour le calcul des gains/pertes brut(e)s (voir p. 145), de celle de Mac Guigan (1967) pour le gain relatif (p. 145) et de celle de D'Hainaut (1975) pour le calcul des pertes relatives (p. 146).

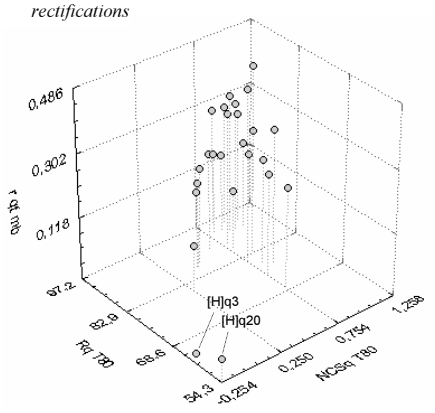
Remarquons sur l'écran précédent l'impact des rectifications sur les indices classiques du niveau « TEST », en particulier sur l'alpha. Lorsque ce dernier est calculé au départ de la matrice binaire (p. 132), il passe de 0,668 à 0,705, soit une progression en gain brut de 0,037 (≃4%) et en gain relatif de 0,111 (11%). Calculé à partir de la matrice spectrale, l'alpha passe de 0,652 à 0,727, soit un gain brut de 0,075 (7,5%) ou un gain relatif de 0,216 (≃22%). Après rectification des deux questions, la valeur de l'alpha calculé à l'aide d'une matrice spectrale dépasse la valeur obtenue à l'aide de la matrice binaire.

En ce qui concerne les indices spectraux *NCSt* et *Rt*, nous observons également une amélioration de la situation tant au palier de turbo analyse T0 que T80. Après rectification des questions [H]q3 et [H]q20, le *NCSt* passe de 0,5 à 0,587 (gain brut de 0,087, gain relatif[80] de 0,058) et le *Rt* de 0,802 à 0,825 (gain brut de 0,023, gain relatif de 0,116).

Dans nos prochaines recherches et développements, nous prévoyons aussi d'enrichir les interfaces de gestion de la qualité des épreuves par les indices et graphiques qui ont été mis au point pour le niveau « QCM ». Voici ce que donnerait une évaluation de l'impact envisagée au niveau des « 3D classico-spectraux » calculés avant et après rectification des QCM [H]q3 et [H]q20.

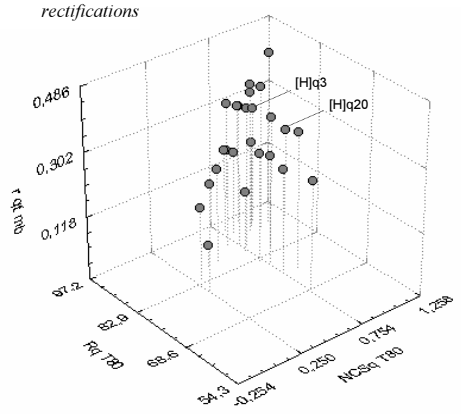
Niveau de qualité initiale

« 3D classico-spectral » des 25 QCM avant rectifications



Impact des rectifications

« 3D classico-spectral » des 25 QCM après rectifications



[79] Graphiques polygonaux à coordonnées polaires.

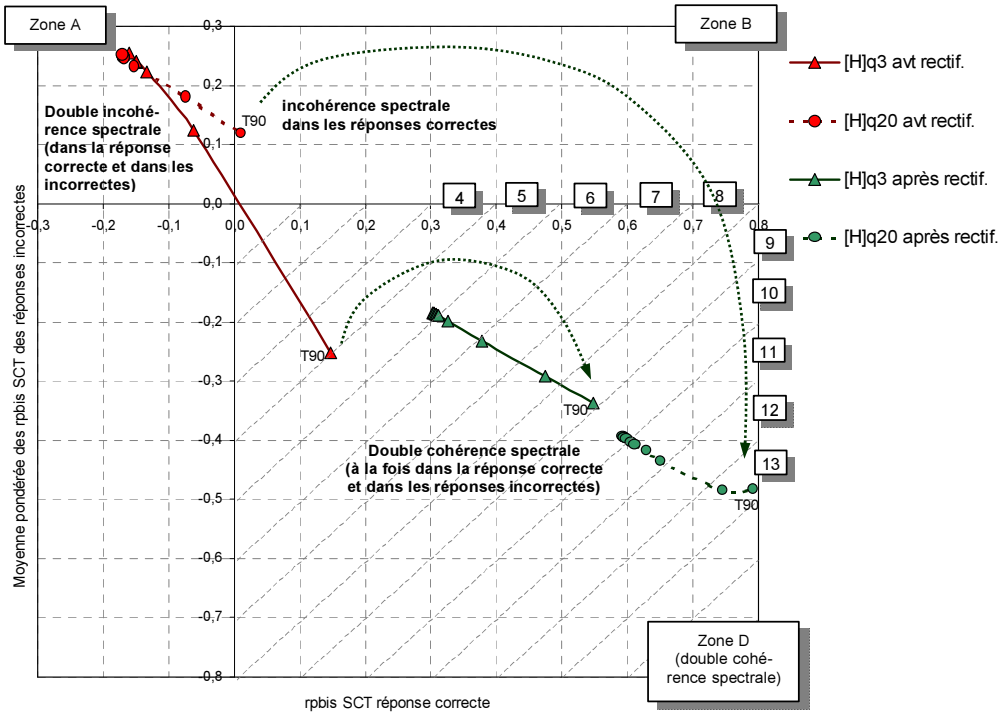
[80] Notons qu'en ce qui concerne le calcul du gain relatif lié au *NCSt*, le maximum possible vaut 2 (voir détails sur le calcul du *NCSt*, pp. 6 et 6).

Lorsqu'on compare les deux graphiques, on observe une nette amélioration des performances de [H]q3 et [H]q20 (remontée vers le coin supérieur droit des points représentant les deux questions). Le tableau ci-dessous chiffre cette progression.

	NCSq T80		*Gain*		*Rq t80*		*Gain*		*R qt mb*		*Gain*	
	avant	*après*	*brut*	*relat.*	*avant*	*après*	*brut*	*relat.*	*avant*	*après*	*brut*	*relat.*
[H]q3	-,187	,766	,953	,436	64,0	89,1	25,1	,697	-,066	,339	,405	,380
[H]q20	-,254	1,232	1,486	,659	54,3	91,3	37	,586	,001	,194	,193	,193

Comme le montrent les gains bruts et relatifs, l'impact spectral des rectifications est énorme : en cohérence d'utilisation des pourcentages de certitude, [H]q3 progresse de 43,6% et [H]q20 de 65,9% en gains relatifs. En ce qui concerne la réalisation des prédictions au niveau de ces deux questions, la progression en gain relatif pour [H]q3 est de 69,7% et pour [H]q20 de 58,6%. La cohérence interne augmente aussi considérablement : le gain relatif [H]q3 est de 38% et celui de [H]q20 de 19,3%.

Les Brins Spectraux par question (*BSq*) permettent aussi une visualisation de l'impact des rectifications au niveau de la cohérence d'utilisation des pourcentages de certitude. Dans notre esprit, des représentations graphiques des *BSq* avant et après rectifications, comme ci-dessous, devraient également faire partie des instruments d'analyse spectrale dans les futures interfaces de gestion de la qualité des épreuves.



On remarque sur le graphique ci-contre qu'au palier de turbo analyse T90 après rectification la QCM [H]q3 passe du niveau de qualité 4 au niveau de qualité 9. Soit une progression de 5 niveaux vers le coin inférieur droit.

En ce qui concerne [H]q20, la progression est encore plus forte, avant rectification, à T90, la QCM se trouve dans la zone B d'incohérence d'utilisation des pourcentages de certitude au niveau des réponses incorrectes, et, après améliorations, elle atteint le niveau de qualité 13 dans la zone D de double cohérence spectrale.

La gerbe spectrale de l'ensemble des questions du test après rectification de [H]q3 et [H]q20 figure en annexe (p. 522).

Enfin, les mesures d'impact classique et spectral des régulations peuvent aussi être réalisées au niveau « PROPOSITIONS ». Voici comment se présentent ces mesures pour **la question [H]q3** (le protocole détaillé des rpbis figure en annexe, p. 541).

Le premier tableau ci-contre montre les configurations des rpbis de [H]q3 **avant** rectification. On remarque notamment les valeurs positives élevées récoltées par la proposition P1 dont le statut d'erreur a été remis en cause après analyse (voir p. 375).

Le second tableau ci-contre montre la situation **après** rectification de la question. Les hachures dans la colonne P1 s'expliquent par le transfert des réponses fournies à cette proposition dans le groupe des réponses correctes P6.

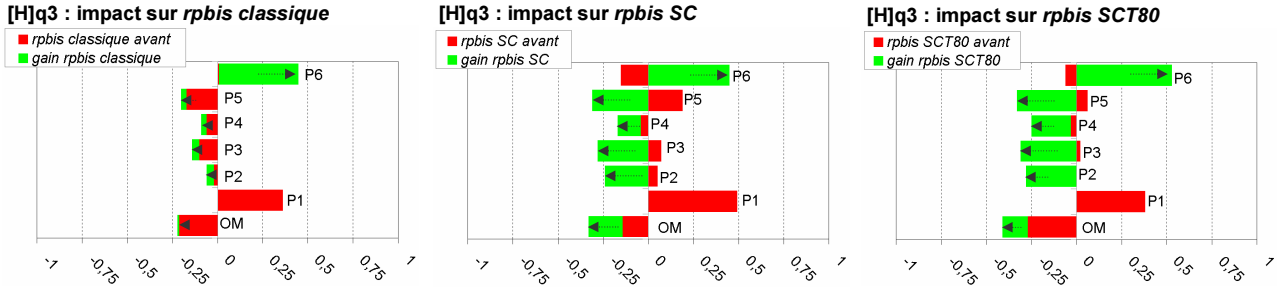
Le troisième tableau présente **l'impact** des améliorations apportées. On observe pour les différents types de rpbis des valeurs plus négatives en ce qui concerne les distracteurs et plus positives et élevées pour ce qui est de la réponse correcte.

[H]q3 (avant)	OM	P1	P2	P3	P4	P5	**P6**
N Rép.	86	501	132	197	104	258	**128**
% Rép.	6%	36%	9%	14%	7%	18%	**9%**
rpbis classique	-,21	,36	-,02	-,10	-,06	-,17	**,01**
rpbis SC	-,14	,49	,05	,07	-,04	,19	**-,15**
N Rép. T80	55	218	67	97	52	111	**54**
% Rép. T80	8%	33%	10%	15%	8%	17%	**8%**
rpbis SCT80	-,27	,38	,00	,02	-,03	,06	**-,06**

[H]q3 (après)	OM	P1	P2	P3	P4	P5	**P6 & P1**
N Rép.	86		132	197	104	258	**629**
% Rép.	6%		9%	14%	7%	18%	**45%**
rpbis classique	-,22		-,06	-,14	-,09	-,20	**,45**
rpbis SC	-,33		-,19	-,21	-,17	-,12	**,30**
N Rép.T80	41		69	99	54	119	**337**
% Rép.T80	1%		10%	14%	7%	17%	**47%**
rpbis SCT80	-,41		-,28	-,29	-,25	-,27	**,47**

[H]q3 (impact)	OM	P1	P2	P3	P4	P5	**P6 & P1**
N Rép.	-		-	-	-	-	**+501**
% Rép.	-		-	-	-	-	**+36%**
rpbis classique	-,01		-,04	-,04	-,03	-,03	**+,44**
rpbis SC	-,19		-,24	-,28	-,13	-,31	**+,45**
N Rép.T80	-14		+2	+2	+2	+8	**+283**
% Rép.T80	-2%		0%	-1%	-1%	0%	**+39%**
rpbis SCT80	-,14		-,28	-,31	-,22	-,33	**+,53**

Les graphiques ci-dessous permettent de visualiser l'ampleur de l'impact des rectifications sur les valeurs récoltées par les propositions aux différents types de rpbis. Pour chaque proposition ainsi que pour l'omission, la partie coloriée en rouge du bâtonnet d'une solution représente les valeurs initiales (avant rectification). La partie coloriée en vert montre les gains obtenus après rectification des résultats de la question. Lorsqu'une question fonctionne bien, les distracteurs doivent obtenir des valeurs négatives, donc sur ces graphiques, leurs bâtonnets doivent être orientés vers la gauche. En ce qui concerne la réponse correcte (ici P6) les valeurs des rpbis doivent être positives et les bâtonnets orientés vers la droite.



En ce qui concerne les *rpbis classiques*, on remarque que l'impact de la régulation de la question est énorme en ce qui concerne la réponse correcte (P6) et peu élevé pour les distracteurs, sauf pour P1 dont les réponses ont été assimilées à P6 (ce qui explique l'absence d'une partie colorée en vert pour le bâtonnet de P1). En ce qui concerne les *rpbis SC*, l'impact sur les valeurs récoltées par les distracteurs est plus élevé. Enfin, pour ce qui est des *rpbis SCT80*, l'impact sur les valeurs récoltées est encore plus élevé à la fois pour les distracteurs et pour la réponse correcte.

On visualise à l'aide de ces graphiques des configurations de rpbis qui redeviennent normales après rectification de la question. Pour les distracteurs, les parties vertes des bâtonnets sont orientés vers la gauche (valeurs négatives) et pour la réponse correcte la partie verte est à chaque fois orientée vers la droite (valeur positive).

Les tableaux des valeurs récoltées par les propositions de **la question [H]q20** avant et après rectification ainsi que les graphiques de visualisation de l'impact sur les rpbis figurent ci-dessous (protocole détaillé des rpbis en annexe, p. 542).

On observe sur le premier tableau ci-contre des valeurs initiales positives pour les différents types de rpbis de la proposition P6, ce qui est anormal pour un distracteur. Par ailleurs, la réponse correcte récolte une valeur positive peu élevée pour le *rpbis classique* et des valeurs négatives pour les rpbis spectraux.

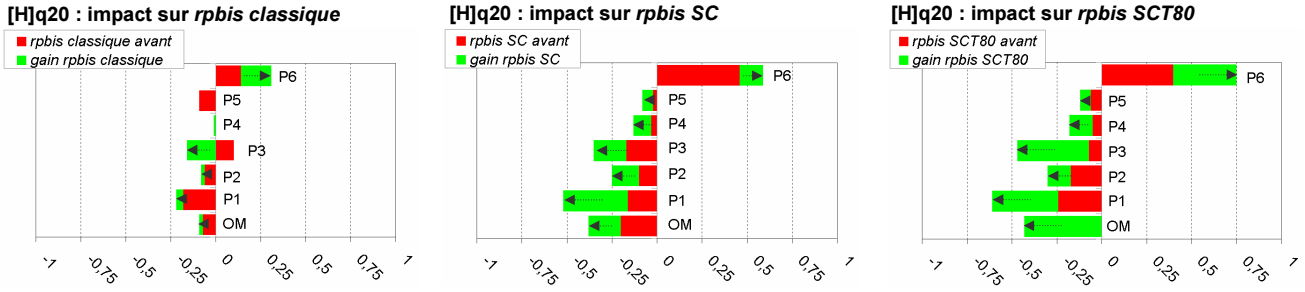
Le second tableau montre les valeurs obtenues après rectification de la question (changement de la réponse correcte qui devient P6). Les configurations des différents types de rpbis redeviennent normales.

Le troisième tableau montre l'impact des améliorations sur les valeurs obtenues aux rpbis. Comme le montrent les graphiques ci-dessous, après rectification, les configurations des rpbis redeviennent normales : la proposition correcte (P6 et non plus P3 comme initialement) récolte des valeurs encore plus positives et les distracteurs de valeurs encore plus négatives, sauf pour le *rpbis classique* de P5 où on assiste à un *statu quo* avant et après amélioration.

[H]q20 (avant)	OM	P1	P2	**P3**	P4	P5	P6
N Rép.	58	263	39	**256**	20	7	744
% Rép.	4%	19%	3%	**18%**	1%	0%	53%
rpbis classique	-,07	-,18	-,06	**,10**	,00	-,09	,14
rpbis SC	-,20	-,16	-,10	**-,17**	-,03	-,02	,46
N Rép. T80	38	137	18	**113**	16	3	320
% Rép. T80	6%	21%	3%	**17%**	2%	0%	49%
rpbis SCT80	xxxx	-,24	-,17	**-,07**	-,05	-,06	,40

[H]q20 (après)	OM	P1	P2	P3	P4	P5	**P6**
N Rép.	58	263	39	256	20	7	**744**
% Rép.	4%	19%	3%	18%	1%	0%	**53%**
rpbis classique	-,09	-,22	-,08	-,06	-,01	-,09	**,31**
rpbis SC	-,38	-,52	-,25	-,35	-,13	-,08	**,59**
N Rép. T80	27	131	19	114	14	3	**404**
% Rép. T80	4%	18%	3%	16%	2%	0%	**56%**
rpbis SCT80	-,43	-,61	-,30	-,47	-,18	-,12	**,75**

[H]q20 (impact)	OM	P1	P2	P3	P4	P5	**P6**
N Rép.	-	-	-	-	-	-	**-**
% Rép.	-	-	-	-	-	-	**-**
rpbis classique	-,02	-,04	-,02	-,16	-,01	,00	**+,17**
rpbis SC	-,18	-,36	-,15	-,18	-,10	-,06	**+,13**
N Rép. T80	-11	-6	-1	-1	-2	0	**+84**
% Rép. T80	-2%	-3%	0%	-1%	0%	0%	**+7%**
rpbis SCT80	-0,43	-,37	-,13	-,40	-,13	-,06	**+,35**



De telles mesures d'impact et leurs visualisations graphiques permettent à l'évaluateur de vérifier la pertinence des décisions qu'il est amené à prendre en vue d'améliorer la qualité des épreuves.

Enfin, terminons en montrant l'impact de ces régulations sur la fidélité de l'épreuve à l'aide de l'indice alpha de Cronbach calculé sur base des données des matrices binaires (α_{mb}) ou spectrales (α_{ms}).

La fidélité de l'épreuve s'améliore après régulation. Remarquons que le gain calculé à l'aide des données de la matrice spectrale est quasi le double du gain observé à l'aide des données de la matrice biniaire.

	HISTOI AVANT	HISTOI APRES	GAIN	GAIN RELATIF
α_{mb}	0,668	0,706	+ 0,038	11,4%
α_{ms}	0,652	0,727	+ 0,075	21,6%

Nous avons aussi calculé le coefficient d'allongement du test (p. 139). Avant régulation, il fallait doubler la taille du test (ajouter 25 questions parallèles aux 25 questions du test) pour obtenir un alpha de 0,80 lorsque le coefficient était calculé sur base de la matrice binaire ($kq[\alpha=0,8]_{mb}$). Lorsque le coefficient d'allongement était calculé sur la base des résultats de la matrice spectrale ($kq[\alpha=0,8]_{ms}$), il fallait en ajouter 28.

Après régulation, pour obtenir un alpha de 0,80 nous pouvons ajouter moins de questions : 17 lorsque le coefficient d'allongement est calculé sur base de la matrice binaire ($kq[\alpha=0,8]_{mb}$) et 13 questions lorsqu'il est calculé à l'aide de la matrice spectrale ($kq[\alpha=0,8]_{ms}$).

	HISTOI AVANT	HISTOI APRES
$kq[\alpha=0,8]_{mb}$	25	17
$kq[\alpha=0,8]_{ms}$	28	13

On voit donc que globalement la fidélité de l'épreuve a été améliorée suite aux prises de décisions de rectifications effectuées après contrôle de la qualité des propositions des questions du test. Rappelons que dans le cas de cette épreuve HISTOI le contrôle qualité a été opéré en fonction de trois éclairages : l'analyse classique, l'analyse spectrale et les avis des experts.

Ce dernier éclairage des experts nous paraît particulièrement crucial car il permet de se prémunir contre d'éventuels effets pervers que pourraient entraîner des rectifications aveugles (sans contrôle des experts) sur la validité de l'épreuve. En effet, en supprimant par exemple une question aux faibles performances classiques et spectrales, nous pourrions peut-être augmenter la fidélité d'un test mais peut-être aussi en diminuer la validité dans la mesure où un objectif important pourrait ne plus être évalué au travers d'une question supprimée dans l'épreuve.

Des indices spectraux tels que les *rpbis SC* peuvent aussi conforter les experts dans leurs analyses lorsque par exemple ces derniers ne détectent rien d'anormal dans une question et que les *rpbis SC* n'indiquent rien d'anormal non plus alors qu'au contraire les *rpbis classiques* montrent eux des anomalies qui sont en fait de « fausses alertes ».

4. Vers une gestion des banques de questions à l'aide d'indices « *test free* »

Nous l'avons déjà souligné précédemment, le *rpbis classique* n'est pas un indice « *test free* » (pp. 177 et 420). Pour ce premier type de rpbis la mesure critère est externe à la question, il s'agit des nombres de réponses correctes obtenus par les étudiants pour l'ensemble des questions du test (voir explications détaillées sur le principe du *rpbis classique*, p. 171). Dès lors, lorsqu'on corrèle les réussites/échecs de tous les sujets à une question avec leurs niveaux de performances pour l'ensemble de l'épreuve (leurs nombres de réponses correctes) ces dernières mesures critères peuvent différer d'un test à l'autre en fonction des questions (différentes) retenues pour construire les épreuves. Ainsi, la valeur du *rpbis classique* d'une même question pourra varier d'une épreuve à l'autre notamment parce que les questions qui l'accompagnent dans l'autre test sont différentes. Notons aussi que la valeur du *rpbis classique* d'une question sera influencée par le nombre de questions présentes dans l'épreuve, c'est ce qui nous amène à calculer une valeur repère pour contrer le problème de la corrélation automatique (voir p. 176). Cette dépendance au contexte des autres questions rend difficile l'exploitation des *rpbis classiques* pour renseigner le constructeur d'une épreuve sur la qualité d'une QCM qu'il compte insérer dans une nouvelle évaluation qui contiendra d'autres questions que celles qui ont permis le calcul du *rpbis classique* initial de cette QCM.

Dans le cas du *rpbis Spectral Contrasté* (voir détails, p. 178) à la différence du *rpbis classique*, l'influence des autres questions contenues dans le test ne joue plus, les choix ou non choix d'une proposition d'une QCM n'étant plus corrélés avec les scores totaux à l'ensemble de l'épreuve mais avec les degrés de certitude qui ont accompagné les réponses des sujets à la question envisagée. Ceci dit, comme pour le *rpbis classique*, dans le contexte d'examens universitaires, les étudiants sont évidemment différents d'une année académique à l'autre et on ne peut pas être certain que la façon d'utiliser les degrés de certitude soit similaire dans deux groupes de répondants différents. Néanmoins, la mesure critère n'est plus externe mais interne à la question dans la mesure où les certitudes sont celles fournies dans le cadre de la QCM envisagée et uniquement dans ce cadre (la qualité des autres questions du test n'influence pas, comme c'est le cas pour le *rpbis classique*, la mesure de la qualité de la QCM envisagée).

Notons à ce propos qu'une turbo analyse (p. 186) enlève cette propriété « *test free* » aux *rpbis spectraux* dans la mesure où une sélection de données est alors opérée sur la base du niveau de réalisme des sujets (*Rs*, pp. 184 et 277) exigé à un palier turbo donné (par exemple, au palier turbo T80 on ne prendra en compte que les données des étudiants dont *Rs* est supérieur ou égal à 80). Or, le réalisme est calculé au départ des réponses et certitudes provenant de toutes les questions d'une épreuve et, dès lors, on retombe dans le problème de la dépendance au contexte des autres questions dont la qualité spectrale peut influencer le réalisme des sujets. Au contraire, le calcul du *rpbis SC* ne nécessite pas de sélection de sujets en fonction du réalisme, ce sont en effet les données de tous les étudiants qui sont prises en compte dans le calcul de cet indice.

Le caractère « *test free* » du *rpbis SC* constitue un avantage indéniable qui ouvre de nouvelles perspectives en terme de suivi de la qualité des QCM dans la gestion des banques de questions. En d'autres termes, les informations livrées par les *rpbis SC* peuvent être réutilisées dans le cadre de l'étape « *3. Construction de l'épreuve* » du cycle de réalisation en « *spirale de qualité* » des examens standardisés (voir p. 78). A cette étape l'évaluateur peut en effet être amené à consulter des banques de questions en vue d'y puiser les items de son épreuve et on pourrait imaginer d'inclure dans ces banques des informations relatives à la qualité spectrale des questions posées lors de tests où le recours aux pourcentages de certitude aurait été autorisé. Ceci permettrait d'obtenir une information sur la qualité spectrale *a priori* des questions, et ce, non seulement du point de vue du niveau « PROPOSITIONS » grâce aux *rpbis SC*, mais aussi du point de vue du niveau « QCM » à l'aide des indices *NCSq* (p. 231) et *Rq* (p. 242).

Conclusions générales

Dans les préliminaires de cette thèse nous avons rappelé que les examens standardisés ayant recours aux questions à choix multiple (QCM) offrent potentiellement de nombreux avantages : étudiants tous traités de la même façon, correction automatisée, large éventail de la matière évalué, utilisation systématique des pourcentages de certitude, rétroactions rapides à l'aide de feedbacks détaillés automatisés, …

Nous avons aussi souligné une série de points faibles liés aux QCM. Par exemple, certaines performances complexes ne peuvent être évaluées par ce type de questions, notamment la capacité à s'exprimer par écrit. Dès lors, lorsque les objectifs de l'évaluateur l'exigent, nous préconisons une utilisation combinée de l'approche standardisée et de l'approche plus traditionnelle d'évaluation à l'aide de questions ouvertes[81].

Un autre inconvénient lié aux QCM concerne le caractère « binaire » de l'évaluation de la performance de l'étudiant : la proposition choisie est soit correcte, soit incorrecte. Il est heureux qu'une technique ancienne[82], celle des pourcentages de certitude, y apporte une solution, à condition de veiller à respecter une série de règles méthodologiques que Shufford & al. (1966) appellent « *admissible probability measurement procedures* ».

En invitant l'étudiant à accompagner son choix d'une proposition du pourcentage de chances qu'il lui attribue d'être correct, nous permettons plus de nuances dans l'analyse de ses performances. En effet, le choix d'un distracteur accompagné du pourcentage de certitude maximum (100%) présente la pire des situations, celle où l'étudiant fournit une réponse erronée en estimant qu'elle a un maximum de chances d'être correcte. A l'opposé, l'étudiant qui répond correctement avec une certitude maximale fait preuve d'une connaissance assurée. Entre ces deux extrêmes, s'ouvre tout l'espace d'une analyse « spectrale » (et non plus « binaire ») des performances, espace invisible lorsque les pourcentages de certitude ne sont pas utilisés. Ainsi, dans le cas d'une réponse correcte, Jans & Leclercq (1999) proposent une terminologie *ad hoc* pour distinguer une « *ignorance* » (réponse correcte et certitude faible), d'une « *connaissance partielle* » (réponse correcte et certitude moyenne), d'une « *connaissance parfaite* » (réponse correcte et certitude élevée). De telles nuances spectrales ont aussi été envisagées par ces auteurs dans le cas d'une réponse incorrecte (« *méprise* » et « *connaissance dangereuse* »).

Par ailleurs, nous savons que depuis plusieurs décennies la plupart des institutions universitaires européennes sont confrontées à une forte augmentation de leurs effectifs d'étudiants alors que les budgets alloués n'augmentent pas en proportion (Gibbs & Jenkins, 1992). Les universités de la Communauté Française de Belgique n'échappent pas à cette tendance lourde : par rapport aux chiffres de 1972, le nombre d'inscrits est passé à 150% et, en francs constants, les subsides sont restés les mêmes (Debry & al., 1998). Cette situation entraîne dans les sections des premiers cycles d'études où les étudiants sont les plus nombreux, un recours massif aux examens standardisés avec QCM dont la correction peut être automatisée, ce qui permet d'évaluer dans des délais raisonnables et avec une procédure « qualité » de grands groupes d'étudiants (parfois plus de 600 !).

Etant donné les avantages potentiels des épreuves ayant recours aux QCM et leur caractère actuellement incontournable lorsqu'il s'agit d'évaluer de grands groupes d'étudiants, nous pensons qu'il est indispensable que la recherche édumétrique innove et progresse dans la mise en œuvre de concepts et de principes qui contribuent à améliorer la qualité des examens standardisés universitaires.

Nous espérons avoir fait progresser cette problématique de recherche en proposant dans le cadre de notre thèse une nouvelle voie d'analyse, celle de la qualité spectrale des épreuves standardisées. L'idée

[81] Questions ouvertes dont la qualité peut aussi être améliorée, notamment à l'aide des « *Echelles descriptives en évaluation* » (De Bal, De Landsheere et Beckers, 1977).

[82] A notre connaissance, les premières recherches sur les pourcentages de certitude remontent à Henmon (1911).

étant d'exploiter les informations liées aux pourcentages de certitude pour créer de nouveaux indicateurs de la qualité des tests.

Nous avons vu qu'habituellement les pourcentages de certitude qui accompagnent les réponses aux QCM sont utilisés pour livrer des informations nuancées, spectrales (et non plus binaires), sur la qualité des performances des étudiants. L'aspect novateur de notre démarche réside dans le fait que nous avons exploité les certitudes fournies par les étudiants pour livrer cette fois des informations spectrales sur la qualité des questions (différentes des informations sur la qualité des performances des étudiants).

Notre recherche a ainsi débouché sur l'élaboration d'une série d'indices originaux d'analyse de la qualité spectrale des épreuves. Ces indices spectraux sont destinés à être utilisés lors de la phase de correction, lorsqu'il s'agit de mettre en évidence les QCM problématiques et, au sein de celles-ci, les propositions qui contiennent des anomalies.

Nous avons mis ces instruments de détection des QCM suspectes à l'épreuve des données en calculant les indices spectraux au départ de plusieurs milliers [83] de réponses et certitudes récoltées lors des dix tests standardisés du projet MOHICAN (Leclercq & al., 2001). Il s'agissait de dix tests de connaissance des principales matières de fin du secondaire qui ont été soumis à des groupes d'étudiants entrant en première année dans huit des neuf institutions universitaires de la Communauté Française de Belgique.

L'intuition de départ pour la construction des indices spectraux était la suivante : logiquement les étudiants qui répondent correctement à une question doivent fournir des pourcentages de certitude plus élevés que les étudiants qui répondent incorrectement. Dans un tel cas de figure, il y a cohérence dans l'utilisation des pourcentages, nous disons « cohérence spectrale ». Les nouveaux indices d'analyse spectrale des propositions des QCM, les « *rpbis spectraux* », mesurent cette « cohérence spectrale ». Permettent-ils de détecter les questions dont les propositions contiennent des anomalies ?

Les dix épreuves MOHICAN comptaient au total 173 QCM et pour deux d'entre elles, la 3ème et la 20ème question du test de Connaissance en Histoire et Socio Economie, les valeurs obtenues aux *rpbis spectraux* indiquent des situations « d'incohérence spectrale », les étudiants ayant tendance à fournir des certitudes moins élevées pour la réponse considérée comme correcte et plus élevées pour les propositions incorrectes. Ces deux questions présentent donc des problèmes « d'incohérence spectrale ».

Nous avons aussi analysé les performances des propositions de ces deux questions à la lumière d'indices qui sont habituellement utilisés (et qui ne sont pas spectraux) : les *rpbis classiques*. Rappelons que dans le cas des *rpbis classiques*, les choix ou les rejets (1 ou 0) des propositions d'une QCM par les étudiants sont corrélés avec les nombres de réponses correctes à l'ensemble des questions du test. Logiquement, lorsque la question fonctionne correctement, on s'attend à ce que les choix de la réponse correcte soient corrélés avec des scores plus élevés au total du test et les choix des distracteurs avec des scores moins élevés. Dans ce cas, la question discrimine les étudiants en fonction du critère du nombre de réponses correctes fournies à l'ensemble des questions du test. Les sujets qui récoltent un nombre élevé de réponses correctes ayant tendance à choisir la proposition correcte et les sujets qui récoltent un nombre moins élevé ayant eux tendance à choisir une proposition incorrecte.

Lorsque nous étudions les propositions des deux QCM problématiques à l'aide des indices *rpbis classiques*, nous remarquons qu'elles ne fonctionnent pas correctement du point de vue de la discrimination classique. Lorsque nous demandons l'avis des experts du contenu, ces derniers confirment que ces QCM posent problèmes. Pour une des questions un distracteur pourrait aussi être considéré comme étant correct et pour l'autre, il y a erreur dans l'encodage de la réponse correcte. Pour ces deux questions il y a donc convergence de trois éclairages différents : (1) celui des *rpbis classiques*, (2) celui des experts et, (3) celui de la cohérence spectrale mesurée à l'aide des *rpbis spectraux*. Dans le cas des épreuves MOHICAN, l'analyse spectrale permet donc de mettre en évidence deux questions qu'une analyse plus

[83] Rappelons que le nombre d'étudiants interrogés variait entre 1.392 et 3.846 selon les épreuves.

qualitative (les avis des experts) ainsi qu'une analyse de discrimination classique (les *rpbis classiques*) désignent aussi comme questions à problèmes.

L'analyse spectrale permet-elle de faire mieux que l'analyse de discrimination classique lorsqu'il s'agit de repérer les questions problématiques et en leur sein les propositions qui contiennent des anomalies ?

La réponse doit être nuancée. Mais avant, revenons sur les particularités des traitements que nous proposons dans le cadre des nouveaux indices spectraux. Rappelons d'abord deux principes que nous avons appliqués aux *rpbis Spectraux*.

D'une part nous proposons des *rpbis Spectraux Contrastés* (*rpbis SC*) pour les propositions incorrectes. Ce qui consiste à mettre les données des étudiants qui ont choisi un distracteur *en contraste* avec les seules données des étudiants qui ont choisi la réponse correcte. L'avantage réside dans l'élimination des données des sujets ayant opté pour les autres propositions incorrectes, ce qui évite d'introduire dans la mesure de la cohérence spectrale du distracteur envisagé, le « bruit » qu'engendreraient les données des autres propositions incorrectes.

D'autre part, nous appliquons le principe de la « turbo analyse » aux indices spectraux. Il s'agit alors d'opérer une sélection dans les données utilisées de manière à ne garder pour le calcul des *rpbis Spectraux Contrastés* que les résultats des étudiants qui utilisent les certitudes de façon judicieuse, c'est-à-dire en commettant peu d'erreurs dans leurs auto-estimations. Nous obtenons alors des *rpbis Spectraux Contrastés Turbo* (*rpbis SCT*) plus valides car fondés sur les informations fournies par les sujets dont les données sont plus fiables. Dans le cadre de nos analyses, il s'agit des *rpbis SCT80* et *rpbis SCT90* (ces derniers étant les plus valides).

Pour détecter les incohérences spectrales dans les propositions des questions nous avons donc utilisé trois types de *rpbis spectraux* : les *rpbis SC*, les *rpbis SCT80* et les *rpbis SCT90*.

Dans le cadre de cette thèse nous avons analysé les 173 QCM des 10 tests MOHICAN en utilisant ces trois indices spectraux ainsi que les indices *rpbis classiques*. Nous avons également passé en revue les commentaires effectués par les experts du contenu à propos de chaque question. De ces analyses il ressort qu'en plus des deux QCM déjà épinglées précédemment, 14 autres questions présentent des valeurs anormales soit aux *rpbis classiques* ou/et aux *rpbis SC* et, parmi elles, une seule obtient un *rpbis SCT80* anormal. En ce qui concerne les *rpbis SCT90* de ces 14 QCM, nous n'observons pas de valeurs anormales.

Que disent les experts ? Parmi ces 14 questions, ils en pointent trois qui posent problèmes : la 14ème QCM du test de Connaissances artistiques, la 6ème du test de Biologie et la 16ème du test d'Histoire.

Seuls les *rpbis SC* signalent les propositions problématiques de ces trois questions épinglées par les experts. Quant aux *rpbis SCT80* et *rpbis SCT90*, ils ne les mettent pas en évidence. Les *rpbis classiques*, eux, ne signalent qu'une des deux propositions problématiques pour une seule des trois questions. Donc, du point de vue de la « détection », les *rpbis SC* on été plus efficaces pour mettre en évidence les problèmes relevés par les experts.

Ceci dit, les différents types de rpbis déclenchent aussi ce que nous avons appelé des « fausses alertes »[84]. De ce point de vue, les *rpbis SC*, avec 7 fausses alertes, sont moins efficaces que les *rpbis SCT80* qui en provoquent une seule et que les *rpbis SCT90* qui en déclenchent aucune (mais les *rpbis SCT80* et *rpbis SCT90* ne détectent pas les trois questions pointées par les experts). Par contre les *rpbis SC* déclenchent moins de fausses alertes que les *rpbis classiques* qui en ont 10 à leur actif.

[84] C'est à dire la mise en évidence d'une valeur anormale récoltée par une proposition alors que les experts du contenu n'y décèlent pas d'anomalie particulière.

Ces qualités de meilleure « détection » et de moins de « fausses alertes » sont cruciales lorsqu'il s'agit de mettre en évidence les QCM problématiques. Notamment dans le cadre de l'étape de correction des examens du cycle « en spirale de qualité » de réalisation des épreuves standardisées qui est proposé aux enseignants de l'Université de Liège.

D'autres indices spectraux ont aussi été créés en vue d'offrir une image de la qualité spectrale à un second niveau plus global, celui des questions (et non plus uniquement au niveau des propositions au sein de celles-ci). A l'aide de l'indice du Niveau de Cohérence Spectrale d'une question (*NCSq*) nous pouvons détecter les questions où l'utilisation des pourcentages de certitude est particulièrement incohérente. Cet indice permet d'épingler directement les QCM dont les propositions récoltent des valeurs anormales sans devoir passer en revue tous les *rpbis spectraux* de toutes les propositions de toutes les questions d'une épreuve.

Mais à ce niveau QCM deux autres indices spectraux peuvent aussi être calculés. Nous avons proposé d'évaluer dans quelle mesure les pourcentages de certitudes choisis par les étudiants (leurs prédictions) s'ajustent aux taux d'exactitude observés (la réalisation de ces prédictions). Cet indice de Réalisation des prédictions par question (*Rq*) permet de mesurer l'ampleur des erreurs de certitude contenues dans les résultats d'une QCM. Un autre indice spectral vient compléter cette mesure, il s'agit de la Centration par question (*Cq*) qui permet d'évaluer, lorsque *Rq* est faible, si les erreurs d'ajustement des certitudes sont liées à une tendance à la surestimation ou au contraire à une tendance à la sous-estimation dans les résultats. Comme précédemment pour les *rpbis spectraux*, ces trois indices *NCSq*, *Rq* et *Cq* peuvent aussi être calculés à différents niveaux de turbo analyse, ce qui permet d'en améliorer la fiabilité.

Lorsque nous corrigeons les anomalies contenues dans certaines propositions au sein des questions, nous pouvons désormais non seulement évaluer l'impact des rectifications sur ces propositions, mais aussi sur la question entière en comparant les valeurs obtenues aux indices *NCSq*, *Rq* et *Cq* avant et après les changements opérés. Nous l'avons fait pour les deux questions problématiques du test de Connaissances en Histoire et Socio Economie et avons ainsi pu observer une amélioration sensible de la cohérence spectrale ainsi que de la réalisation des prédictions.

En effectuant la moyenne des valeurs obtenues aux indices spectraux des QCM d'une épreuve, nous obtenons une image de la qualité spectrale moyenne du test. Cette information là est aussi utile lorsqu'il s'agit d'évaluer l'impact au niveau de l'épreuve des rectifications effectuées sur les propositions des QCM.

Ainsi, à l'aide des indices spectraux développés dans le cadre de notre thèse et utilisables à trois niveaux d'analyse : « PROPOSITIONS », « QCM » et « TEST », nous ouvrons une nouvelle voie pour l'analyse de la qualité des épreuves standardisées et leur régulation. Désormais, nous sommes en effet en mesure : d'évaluer la qualité spectrale des épreuves standardisées universitaires ayant recours aux pourcentages de certitude ; de mettre en évidence d'éventuelles anomalies dans les questions ; et, après rectifications, d'évaluer l'impact spectral des améliorations. C'est là notre contribution à l'amélioration des procédures visant à produire des tests de qualité et, par extension, à l'amélioration de la fiabilité des notes, ce qui, *in fine*, constitue l'enjeu de nos préoccupations édumétriques.

Bibliographie

Adams, J.-K. & Adams, P.-A., (1961). Realism of confidence judgments, *Psychological Review*, 68, 33-45.

Agazzi, A. (1967). *Les aspects pédagogiques des examens*, Strasbourg : Conseil de l'Europe,C.C.C.

Ahlgren, A. (1967). *Confidence on Achievment Tests and the Prediction of Retention*. Harvard University : Thèse de doctorat.

Albert, M. & Raiffa, H. (1982). A progress report on the training of probability assessors. In Kahneman, D., Slovic, P. & A. Tversky A. (Eds), *Judgment under uncertainty : Heuristics and biases* (294-305), Cambridge : University Press, 1982.

Baker J.-D. (1969). The uncertain student and the understanding computer, *La Recherche en enseignement programmé*, Tendances actuelles, Paris : Dunod, 303-319.

Barnabé, C. (1995). *Introduction à la qualité totale en éducation*. Québec : Presses Inter Universitaires.

Barrows, H.-S. & Tamblyn, R.-M. (1977). The portable patient problem pack (P4), a problem based learning unit, *Journal of Medical Education*, 52, 1002-1004.

Barrows, H.-S. & Tamblyn, R.-M. (1980). *Problem based learning, an approach to medical education*. New York : Springer Publishing.

Blais, J.-G., Laurier, M., Van der Maren, J.-M., Gervais, C., Lévesque & M., Pelletier, G. (1997). *L'évaluation des apprentissages à l'Université de Montréal et dans ses écoles affiliées*. Montréal : Université de Montréal, Faculté des sciences de l'éducation, Groupe de Recherche Interdisciplinaire en Pédagogie Universitaire (GRIPU).

Bloom, B. (1969). *Taxonomie des objectifs pédagogiques - I. Domaine cognitif*, traduit par M. Lavallée. Montréal : Education Nouvelle.

Bloom, B., Engelhart, M.-D., Forst, E.-J., Hill, W.-H. & Krathwohl, D.-R. (1956). *Taxonomy of educational objectives : handbook I, cognitive domaine*. New York : D. Mac Kay.

Bonniol, J.-J. (1972). *Les comportements d'estimation dans une tâche d'évaluation d'épreuves scolaires - Etude de quelques-uns de leurs déterminants*. Aix-En-Provence : Université de Provence.

Boucher, S. *et al.* (1997). *L'évaluation de la qualité de l'enseignement dans les institutions universitaires de la Communauté française de Belgique*, groupe de travail du CreF (Conseil des recteurs Francophones).

Boxus , E. (1981). Une tentative d'insertion de l'évaluation formative dans la pratique quotidienne : l'opération PREDIC, in *Revue de l'Organisation des Etudes*, Bruxelles, 9, 9-39.

Boxus, E. & al. (1991). *Principes communs pour évaluer les résultats cognitifs de la formation*. Bruxelles : Commissions des Communautés européennes, programme Eurotecnet.

Brennan, R.-L. (1972). A generalized upper-lower item discrimination index. *Educational and psychological measurement*, 32, 289-303.

Brier, G.W. (1950).Verification of forecasts expressed in terms of probability, *Montly Weather Review*, 75, 1-3.

Brisbois R. (1967). Les corrélations en pédagogie et en psychologie. Montréal, Canada : Les éditions Lidec.

Bruno J. (1990). Confidence contour tests item analysis with information referenced testing. *Proceedings of the Seventh International Conference on Technology and Education*, Brussels.

Bruno, J. & Baxter, J. (1989). An application of information reference testing, *Proceedings of the Sixth International Conference on Technology and Education*, Orlando, vol. 2, 191-192.

Bruno, J. (1986). Assessing the Knowledge base of students : An information theoric approach to testing. *Measurement and evaluation in counseling and development*, 19(3), 116-130.

Bruno, J. (1987). Admissible probability measurement ininstructional management. Journal of computer based instruction, 14(1), 23-30.

Cardinet, J. , Tourneur, Y. et Allal, L. (1975). Extension of generalisability theory and its applications in educational measurement, *Journal of Educational Measurement*, 18, 183-204.

Carver, R.-P. (1974). Two dimensions of tests : Psychometric and Edumetric, *American Psychologist*, 29, 512-518.

Castaigne, J.-L. & Gilles, J.-L. (2000). Application du cycle gestion qualité SMART des tests pédagogiques dans le cours d'Obstétrique et de Pathologie de la Reproduction des ruminants, équidés et porcins, Communication présentée lors des journées Form@sup, Liège : Université de Liège, Service de Technologie de l'Education.

Castaigne, J.-L., Gilles, J.-L. & Hansen, C. (2001). Application du cycle gestion qualité SMART des tests pédagogiques au cours d'Obstétrique et de Pathologie de la Reproduction des ruminants, équidés et porcins, Communication au 18ème Congrès de l'Association Internationale de Pédagogie Universitaire (AIPU), Les stratégies de réussite dans l'enseignement supérieur, Dakar, 5-7 avril 2001.

Chicago Board of Education (1999). *Standardized Test Preparation – Preparating Your High School Students to take Standardized Tests*, Chicago : Chicago Public Schools.

Cox, R.C. & Vargas, J.S. (1966). *A comparison of item selection techniques for norm-referenced and criterion-referenced tests*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16.

Cronbach, L., Gleser, G.C. & Rajaratnam, N. (1963). Theory of generalizability. A liberalization of reliability theory. *British journal of mathematical and statistical psychology*, 16, 137-173.

Cross, L. & Frary (1977). An empirical test of Lord's theorical results regarding formula scoring of multiple choice tests, *Journal of Educational Measurement*, vol 14, 313-321.

Dagnélie, P. (1992). Statistique Théorique et Appliquée – Tome I. Gembloux : Les presses agronomiques de Gembloux.

D'Hainaut, L. (1973). Etude d'une nouvelle variable pour l'analyse statistique des expériences pédagogiques, *Bulletin de psychologie*, 305, vol. 26, 622-630.

D'Hainaut, L. (1975). *Arrondir et estimer, cours programmé*. Paris : Hachette.

De Bal, R., De Landsheere, G., & Beckers, J. (1977). *Construire des échelles d'évaluation descriptives*. Bruxelles : Ministère de l'Education, Organisation des Etudes.

De Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item, *British Journal of Mathematical and Statistical Psychology*, 18, 87-123.

De Landsheere (1980). *Evaluation continue et examens - Précis de docimologie*. Bruxelles : Ed. Labor, Eucation 2000.

De Landsheere, G. (1976). *La formation des enseignants demain*, Collection E3, Orientation, Ed. Casterman.

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

De Landsheere, V. (1986). *Faire réussir, faire échouer. La compétence minimale et son évaluation*. Paris : PUF.

Debry, M., Deltour, J.-J., Demeuse, M., Tanguy, E., Gilles, J.-L., Leclercq, D., Malherbe, E., Perée, F., & Poncin, P. (1999). *Observations sur les Etudes Universitaires à la Faculté (ŒUF) – Rapport de synthèse de la Commission Premier Cycle*. Liège : Université de Liège, Faculté de Psychologie et des Sciences de l'Education.

Debry, M., Leclercq, D. & Boxus, E. (1998). De nouveaux défis pour la pédagogie universitaire. In D. Leclercq (Ed), *Pour une pédagogie universitaire de qualité*. Liège Sprimont : Mardaga.

Deming, W.E. (1992). *The New Economics for Industry, Government, Education*. Cambridge MA : MIT Center for Advanced Engineering Study.

Descartes, R. (1628). *Règles pour la conduite de l'esprit*.

Descartes, R. (1636). *Discours de la méthode pour bien conduire sa raison et chercher la vérité dans les sciences*.

Didier, P., Faivre, M., Fauroux, R., Grandpierre, A., Millord, D. & Rousselet, M. (1966). *Le bouton du mandarin. L'école face à notre avenir*. Casterman, Centre d'Etudes Pédagogiques.

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

D'Ivernois J.-F., Gagnayre R. (1995). *Apprendre à éduquer le patient*. Paris : Vigot.

Droesbeke, J.-J., (2001). *Eléments de statistique*. Bruxelles : Editions de l'Université de Bruxelles.

Dupont, P. & Ossandon, M. (1994). *La pédagogie universitaire*. Paris : Presses Universitaires de France.

Ebel, R.L. (1965). Confidence–Weighting and Test Reliability, *Journal of Educational Measurement*, 2, 49-57 B.

Ebel, R.L. (1968). Review of Valid confidence testing demonstration kit*, Journal of Educational Measurement*, 5, 353-354.

Ebel, R.L. (1979). Using tests to improve learning, *Arithmetic-Teacher*, v27, n3, p10-12.

Edwards, W. (1967). Probabilistic information processing by men and man-machine systems, in *La simulation du comportement humain*. Paris : Dunod, p. 187.

Findley, W.G. (1956). A rationale for evaluation of item discrimination statistics. *Educational and Psychological Measurement*, 16, 175-180.

Georges, F., Gilles, J.-L., Pirson, M., Simon, F. & Leclercq, D. (2001). Les feedbacks aux étudiants et aux enseignants du projet MOHICAN, in Leclercq, D. (Ed.), *Le premier des MOHICANs – Une recherche-action de Monitoring Historique des CANdidatures*, Rapport relatif au contrat 798363 entre le Conseil Inter Universitaire Francophone (CIUF) et l'Université de Liège. Bruxelles : CIUF (à paraître).

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

Gilles, J.-L. (1995). Entraînement à l'autoévaluation : une comparaison filles/garçons à l'université, *Actes du Colloque de l'AIPU « Enseignement supérieur : stratégies d'enseignement appropriées »*, Hull : Université du Québec à Hull, pp. 159-166.

Gilles, J.-L. (1996a). Utilisation des degrés de certitude et normes de réalisme en situation d'examen et d'auto-estimation à FA.P.S.E. - ULG, Communication présentée au *Colloque de l'Association pour le Développement des Méthodologies d'Evaluation en Education (ADMEE-EUROPE) "Dix années de travaux de recherche en évaluation"* (18, 19 et 20 septembre 1996), Grenoble : Université Pierre Mendes France.

Gilles, J.-L. (1996b). Profils d'auto-évaluation et d'états de connaissance partielle chez les étudiants de première candidature (1995-1996) de la Faculté de Psychologie et des Sciences de l'Education de l'Université de Liège, communication présentée lors du *Colloque de l'Association Internationale de Pédagogie Universitaire (A.I.P.U.)*. Tunis : 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, in Boxus, E., Gilles, J.-L., Jans, V. & Leclercq, D. (Eds), *Actes du 15ème Colloque de l'Association Internationale de Pédagogie Universitaire (A.I.P.U.)*. Liège : Affaires Académiques de l'Université de Liège, pp. 311-326.

Gilles, J.-L. (1998a). Apports des mesures métacognitives lors d'un test de compréhension d'un article scientifique, in Depover & Noël (Eds), *Approches plurielles de l'évaluation des processus cognitifs*. Mons : Université de Mons-Hainaut, pp. 19-30.

Gilles, J.-L. (1998b). Mise en œuvre de tests formatifs à l'aide de l'Internet, in Depover, C. & Noël, B.(Eds), *Approches plurielles de l'évaluation des processus cognitifs*. Mons : Université de Mons-Hainaut, pp. 193-204.

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

Gilles, J.-L. & Melon, S. (2000). Comparaison de trois modalités de « testing » des compétences en français chez les étudiants médecins lors de leur première candidature à l'ULg, in Defays et al. (Eds), *La maîtrise du français du niveau secondaire au niveau supérieur*. Bruxelles : Editions De Boeck, pp. 161-178.

Gilles, J.-L., Bourguignon, J.-P. et Detroz, P., (2000). Les questionnaires à choix multiple : utilisation pour l'enseignement en groupe avec boîtiers électroniques. *Revue Médicale de Liège*, 55 : 12, pp. 1047-1050.

Gilles, J.-L., Collet, M., Debry, M., Denis, B., Etienne, A.-M., Geuzaine, C., Jans, V., Leclercq, D., Lejeune, M. & Paheau, C. (1998). *Evaluation des enseignements en 1ère et 2ème candidatures, année académique 1997-1998 - Rapport de synthèse pour le Conseil de Faculté du 10 novembre 1998*. Liège : Université de Liège, Faculté de Psychologie et des Sciences de l'Education.

Gilles, J.-L., Poncin, P., Ruwet, J.-C. et Leclercq, D. (1999). Les travaux dirigés virtuels d'Anthropologie biologique – Bilan d'une première utilisation, in J.-P. Béchard et D. Gregoire (Eds), *Apprendre et enseigner autrement*, Montréal : Ecole des Hautes Etudes Commerciales, Vol. 1, pp. 294-307.

Glasser, W. (1992). *The Quality School*. New York : Harper and Row.

Guliksen, H. (1950). *Theory of mental tests*. New York : Wiley.

Hardy, J.-L. (1981). Adjustment of choice tests correlation in multiple choice item analysis, *Série LPE*, 81-07-001.

Hardy, J.-L. (1983). Plusieurs solutions au problème du recouvrement dans la corrélation entre un test et un item dichotomique, *Scientia Paedagogica Experimentalis*, 49-71.

Hartog, P. & Rhodes, E.C. (1936). *An Examination of Examinations*. London : Mac Millan.

Henmon, V. (1911). The relation of the time of a judgment to its accuracy*, Psychological Review*, 18, 186-201.

Hevner, K.-A. (1932). A method of correcting for guessing in true-false tests and empirical evidence in support of it, *Journal of Social Psychology*, 3, 359-362.

Holingworth, R.-L. (1913). Experimental Studies in judgment, *Archives of Psychology*, 29, 1-119.

Hoteling, H. & Pabst, M.R. (1936). Rank correlation and tests significance involving no assumption of normality. *Ann. Math. Statist.*, 7, 29-43.

Howell, D.C., (1998). *Méthodes statistiques en sciences humaines*. Bruxelles : De Boeck Université.

Hunt, D. (1977). *The human self assessment process. Study II : The effects of the number of self-assessment categories on acquisition*. Interim Report from U.S. Army Research Institute for the Behavioral and Social Sciences Grant #DAHC19-76-G-002, New Mexico State University, Las Cruces, NM.

Jacobs S.-S. (1968). *An empirical investigation of the relationship between selected aspects ofpersonality and confidence-weighting behaviors*, Doctoral dissertation, University of Maryland, University Micro-films, 68, 16, 676.

Jacobs S.-S. (1971). Correlates of unwarranted confidence in response to objective test items, *Journal of Educational Measurement*, 8, 1.

Jans & Leclercq (1999). Mesurer l'effet de l'apprentissage à l'aide de l'analyse spectrale des performances, in C. Depover & B. Noel (Ed.), *Evaluation des compétences et des processus cognitifs.* Bruxelles : De Boeck, pp. 303-317.

Jans (2000). Confrontations Instrumentées et Dialectiques des Jugements Auto- et Alloévaluatifs, thèse de doctorat en Sciences de l'Education. Liège : Université de Liège, Faculté de Psychologie et des Sciences de l'Education.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64, 425-461.

Karraker R.J. (1967). Knowledge of results and incorrect recall of plausible multiple choice alternatives, Journal of Educational Psychology, 58, 11-14.

Kauffman, Ch., Dupont, P ., Philippart, A. (1995). *Projet pilote européen pour l'évaluation de la qualité dans l'enseignement supérieur*. Bruxelles : Ministère de l'Education, de la Recherche et de la Formation.

Kaufman, R. (1992). The challenge of total quality management in education. *International Journal of Educational Reform*, 1(2), 149-165.

Kehoe, J. (1995). *Basic item Analysis for Multiple-Choice Tests*. ERIC/AE Digest. Washington : ERIC Clearinghouse on Assessment and Evaluation, The Catholic University of America, Department of Education, O'Boyle Hall.

Kelley, T.L. (1939). Selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30, 17-24.

Kurst, A.K. & Mayo, S.T. (1979). *Statistical methods in education and psychology*. New York : Springer-Verlag.

Landercy, A. (1981). *Elements de statistiques*. Mons : Université de Mons, Département d'Edition des Cours.

Landercy, A. (1983). *Initiation statistique*. Bruxelles : Editions De Boeck.

Laveault, D. & Gregoire, J. (1997). *Introduction aux théories des tests en sciences humaines*. Bruxelles : De Boeck.

Laveault, D. (1999). *Autoévaluation et régulation des apprentissages*. in C. Depover & B. Noel (Ed.), *Evaluation des compétences et des processus cognitifs.* Bruxelles : De Boeck, pp. 57-79.

Leclercq, D. (1973). Critiques de méthodes d'application, de correction et de cotation des questions à choix multiple. *Scientia Paedagogica Experimentalis*, X, 1, 46-57.

Leclercq, D. (1975). *L'évaluation subjective de la probabilité d'exactitude des réponses en situation pédagogique*. Thèse de doctorat en Sciences de l'Education. Liège : Université de Liège, Institut de Psychologie et des Sciences de l'Education.

Leclercq, D. (1982). Confidence marking, its use in testing, in B. Choppin & N. Postlethwaite (eds.), *Evaluation in Education : International Review Series*, Oxford : Pergamon, vol. 6, n°2, pp. 161-287.

Leclercq, D. (1986). *La conception des questions à choix multiple*. Bruxelles : Labor.

Leclercq, D. (1987). *Qualité des questions et signification des scores avec application aux QCM.* Bruxelles : Labor.

Leclercq, D. (1993). Validity, Reliability and Acuity of Self-Assessment in Educational Testing, in Leclercq, D. et Bruno, J. (Eds), *Item Banking : Interactive Testing and Self-Assessment*, NATO ASI Series. Heidelberg : Springer Verlag, pp. 113-131.

Leclercq, D. (1998). Approche Technologique de l'Education et de la Formation. Liège : Université de Liège, Service de Technologie de l'Education, 4ème édition.

Leclercq, D. & Detroz, P. (2001). Les résultats globaux du projet MOHICAN, in Leclercq, D. (Ed.), *Le premier des MOHICANs – Une recherche-action de Monitoring Historique des CANdidatures*, Rapport relatif au contrat 798363 entre le Conseil Inter Universitaire Francophone (CIUF) et l'Université de Liège. Bruxelles : CIUF (à paraître).

Leclercq, D. & Gilles, J.-L. (1993). Hypermedia : Teaching Through Assessment. In D. Leclercq et J. Bruno (Eds), NATO ASI Series, *Item Banking: Interactive Testing and Self Assessment*. Berlin: Springer Verlag, Vol. 112, pp. 31-47.

Leclercq, D. & Gilles, J.-L. (1994). GUESS, un logiciel pour entraîner à l'auto-estimation de sa compétence cognitive, in A. Dumont et J. Weber (Eds), *Actes du 3ème colloque international ESIEE* - Marne-La-Vallée *« QCM et questionnaires fermés »*. Paris : Université Paris 7 – Denis Diderot, Laboratoire d'ingénierie didactique, pp.137-158.

Leclercq, D. & Gilles J.-L. (1995). Le kaléidoscope des techniques de questionnement, *Colloque National de l'Association Internationale de Pédagogie Universitaire (A.I.P.U.),* Colonster-Liège, *22 septembre 1995*.

Leclercq, D. & Gilles, J.-L. (2001). Techniques de mesure dans l'autoévaluation (dix tecniques d'auto-estimation de la qualité de ses réponses), in G. Figari & M. Achouche (Eds), *L'activité évaluative réinterrogée – Regards scolaires et socioprofessionnels*. Bruxelles : Editions De Boeck, pp. 134-142.

Leclercq, D., Conti, C., De Ketele, J.-M., Delhaxhe, M., Dupont, P., Lambert, J.-P., Lambotte, J.-P., Noël, B., Romainville, M. & Wolfs, J.-L. (2001). Le projet MOHICAN-check-up : son origine, ses intentions, ses méthodes, in Leclercq, D. (Ed.), *Le premier des MOHICANs – Une recherche-action de Monitoring Historique des CANdidatures*, Rapport relatif au contrat 798363 entre le Conseil Inter Universitaire Francophone (CIUF) et l'Université de Liège. Bruxelles : CIUF (à paraître).

Leclercq, D., Denis, B., Boxus, E. & Gilles, J.-L. (1994). DOUBLE CHECK - Un concept d'évaluation interactive permettant de distinguer les capacités de compréhension et d'analyse, in *Colloque itinérant sur la Pédagogie Universitaire entre les Universités du Québec et les Universités de la Communauté française de Belgique*. Liège : Université de Liège – STE.

Leclercq, D., Denis, B., Jans, V., Poumay, M. & Gilles, J.-L. (1998). L'amphithéâtre électronique – Une application : le LQRT - SAFE. In D. Leclercq (Ed), *Pour une pédagogie universitaire de qualité*. Liège Sprimont : Mardaga, pp. 161-186.

Leclercq, D., Georges, F., Gilles, J.-L., Reggers, T., and Rommes, O. (1998b). Interactive Multimedia Programmed Biographies (IMPB) : a new method for clinical training, *proceedings of the BITE (Bringing Information Technology fot Education) conference (25-27 March 1998)*. Maastricht : pp. 406-417.

Leclercq, D., Jans, V., Georges, F. & Gilles, J.-L. (2000). Objective assessment of subjectivity : applying confidence marking to partial knowledge, Communication présentée à la conférence EARLI SIG Assessment, Maastricht, septembre 2000.

Leclercq, D., Peeters, R., Reggers, T., Charlier, J.E., De Ketele, J.M., Delhaxhe, M., Dupont, P., Lambert, J.P., Lambotte, J.P., Loeckx, E., Pilatte, A., Romainville, M. & Wolfs, J.-L. (1997). *Franchir le cap des candis – Inscriptions et réussites en candidatures universitaires dans la Communauté Française de Belgique de 1988 à 1995*. Bruxelles : Conseil Interuniversitaire de la Communauté Française.

Leclercq, D., Rommes, O., Georges, F. et Gilles J.-L. (1998). Une préparation multimédias à affronter les conflits professeur-élèves en classe, *Education formation*, pp. 90-93.

Leclercq, D., Willain, J.-C. , Denis, B., Poumay, M., Gilles, J.-L., Orban, M. and Jans, V. (1999). Votes en amphithéâtre électronique pour animer de grands auditoires universitaires selon six paradigmes d'apprentissage/enseignement, in J.-P. Béchard et D. Gregoire (Eds), *Apprendre et enseigner autrement*. Montréal : Ecole des Hautes Etudes Commerciales, Vol 2, pp. 567-578.

Lichtenstein, S., Fischhoff, B. & Philips, L. (1977). Calibration of probabilities : The state of the art, in Jungermann & Dezeeuw (Eds), *Decision making and change in human affairs*. *Proceedings of the 5th SPUDM Conference*. Darmstadt : Reidel Publishing Company, 275-324.

Lion, Y. (2000). Tests chez les ingénieurs. Communication présentée lors des journées Form@sup, Liège : Université de Liège, Service de Technologie de l'Education.

Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Reading, Mass. : Addison-Wesley.

Mac Guigan, F.J. (1967). The G Statistics, An Index of Amount Learned. *N.S.P. Journal*, 69, pp. 14-16.

Magnuson (1967). *Test theory*. Boston : Addison-Wesley.

Messick (1988). The once and the future issues of validity : Assessing the meaning and consequences of measurement. In H. Wainer & H.I. Braun (Eds), *Test Validity*. Hillsdale, NJ : Lawrence Erlbaum.

Michael, J.-J. (1968). The reliability of a multiple-choice examination under various test-making instructions. *Journal of Educational Measurement*, 5, 307-314.

Miller, G.-A. (1956). The magical number of seven, plus minus two, *Psychological Review*, Vol 63, 81-97.

Murphy A.H. & Winkler R.L. (1974). Subjective probability forecasting experiments in meteorology : some preliminary results. *Bulletin of the American Meteorological Society*, 55, 1206-1216, 1974.

Nightingale, P. & O'Neil, M. (1994). *Achieving quality learning in higher education*. London : Kogan Page.

Nightingale, P. & O'Neil, M. (1994). *Achieving quality learning in higher education*. London : Kogan Page.

Nitko, A. (1996). *Educational Assessment of Students*. Englewood Cliffs : Merill, second edition.

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction, *Psychological Monographs*, 76.

Passeron, J.C. (1970). Sociologie des examens. *Education et Gestion*, 2, 6-16.

Piaget, J. (1969). *Psychologie et pédagogie*. Paris : Denoël.

Pieron, H., Reuchlin, M. & Bacher, F. (1962). Une recherche expérimentale de docimologie sur des examens oraux de physique au niveau du baccalauréat de mathématiques, *Biotypologie*, 23, 48-73.

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

Pitz, G.F. (1974). Subjective probability distributions for imerfectly known quantities. In Gregg, L.W. (Ed.) *Knowledge and Cognition*. New York : Wiley, pp. 29-41.

Preston, R.C. (1965). Multiple-choice test as an instrument in perpetuating false concepts, Educational and Psychological Measurement, 34, 499-509.

Ramsden, P. (1991). A performance indicator of teaching quality in higher education : the course experience questionnaire, *Studies in Higher Education*, 16, 129-150.

Rasch, G. (1966). An item analysis which takes individual differences into account, *British Journal of Mathematical and Statistical Psychology*, 19(1), 49-57.

Rhodes, L.A. (1992). *On the road to quality*. Educational leadership, 49(6), 76-80.

Romainville, M. & Boxus, E. (1998). La qualité en pédagogie universitaire. In D. Leclercq (Ed), *Pour une pédagogie universitaire de qualité*. Liège Sprimont : Mardaga, pp. 13-32.

Sandbergen (1971). Guessing and Confidence in testing educational achievment, In Choppin, B., *A/106 IEA Memorandum*.

Segers, M. & Dochy, F. (1996). Quality assurance in higher education : theorical considerations and empirical evidences, *Studies in Educational Evaluation*, 1996, 22 (2), 115-137.

Shufford, A. (1993). In Pursuit of the Fallacy : Resurrecting the Penalty, in D. Leclercq & J. Bruno (Eds), *Item Banking : Interactive Testing and Self-Assessment*, NATO ASI Series, Berlin : Springer Verlag, pp. 76-98.

Shufford, A., Albert, A. & Massengil, N.E. (1966). Admissable probability measurement procedures, *Psychometrika*, 31, 125-145.

Skinner, B.F. (1961). Why we need teaching machines, *Harvard Educational Revue*, 31, 377-398.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology*, 18, 161-169.

Swineford, F. (1938). The measurement of a personality trait., *Journ. of Educ. Psych*., 29, 289-292.

Tourneur, Y. (1988). *La psychométrie et la construction des instruments d'évaluation - Recueil de textes*. Mons : Université de Mons, Département d'Edition des Cours.

Trow W.-C. (1923). The psychology of confidence, an experimental inquiry, *Archives of Psychology*, 67, 1-47.

Van Der Vleuten, C. & Wijnen, W. (1990). *Problem-Based learning : Perspective from the Maastricht experience*. Amsterdam : Thesis.

Van Der Vleuten, C. (1996). *Beyond intuition, inaugural lecture*. Maastricht : Universitaire Press Maastricht.

Van Lenthe, J. (1993). The Development and Evaluation of ELI, an Interactive Elicitation Technique for Subjective Probability Distributions, in Leclercq, D. et Bruno, J. (Eds), *Item Banking : Interactive Testing and Self-Assessment*, NATO ASI Series. Heidelberg : Springer Verlag.

Van Naerssen, R.-F. & Van Beaumont, E. (1965). Ervaringen met een Zekerheidsaanduiding bij objektieve Tentamens, *Nederlands Tijdschrift Psychologie*, 20, 308-315.

Van Naerssen, R.F. (1962). A scale for measurement of subjective probability, *Acta Psychologica*, 20, 2, 159-166.

Vandevelde, L. (1971). *Réflexions* sur les QCM. Bruxelles : ULB, document ronéotyp.

Wiersma, W. & Jurs, S.G. (1990). Educational Measurement and Testing. Boston : Allyn and Bacon.

Wiley, L.-N. & Trimble, D.-C. (1936). The ordinary objective test as a possible criterion of certain personality, *School and Society*, 43, 446-448.

Wood, R. (1977). Multiple-choice : A state of the art report, in Choppin and Postlethwaite (Eds), *Evaluation in Education International Progress*. Oxford : Pergamon.

Wright, B. & Stone (1979). *Best test Design*, Chicago IL : MESA Press.

Zink, K.J. & Schmidt, A. (1995). Measuring universities against the european quality award criteria. *Total quality management*, 6 (5-6), 547-561.

Glossaires

A. Glossaire des principaux indices et instruments d'analyse spectrale utilisés dans cette recherche

BSq & GSt
Brin Spectral par question (*BSq*) et Gerbe Spectrale d'un test (*GSt*)

Les Niveaux de Cohérence Spectrale des questions (*NCSq*) peuvent se visualiser sur un graphique en nuage de points divisé en quatre quadrants. Pour chaque QCM, la valeur placée en abscisse correspond à celle du *rpbis SC* de la réponse correcte et la valeur placée en ordonnée correspond à la moyenne des *rpbis SC* des propositions incorrectes. La zone de « *cohérence spectrale* » est définie par le 4ème quadrant, là où la valeur de la moyenne des *rpbis SC* des propositions incorrectes peut varier entre 0 et –1 et où la valeur du *rpbis SC* de la réponse correcte peut quant à elle varier entre 0 et 1. Plus un point représentant une QCM se rapproche du coin inférieur droit du graphique, plus la cohérence d'utilisation des pourcentages de certitude est élevée. Ce type de graphique permet aussi de visualiser les *NCSq* obtenus aux différents paliers de turbo analyse. Pour chaque QCM on obtient alors une série de points représentant les valeurs du *NCSq* aux différents paliers de la turbo analyse. Ces points peuvent être reliés, ils forment alors un Brin Spectral par question (*BSq*) représentant la cohérence d'utilisation des pourcentages de certitude aux différents seuils de réalisme exigés par la turbo analyse. On s'attend à ce que plus le palier *T* de turbo analyse est élevé, plus le point représentant la valeur du *NCSq* soit proche du coin inférieur droit du graphique. Lorsqu'on visualise ainsi toutes les QCM d'une épreuve on obtient un ensemble de brins formant une Gerbe Spectrale du test (*GSt*) où nous pouvons identifier les niveaux de cohérence d'utilisation des pourcentages de certitude pour chaque question.

CM
Certitudes moyennes

Le module « *Traitements basiques* » de *SCANTEST 2.0* permet d'obtenir une matrice de résultats contenant une série d'informations liées aux rpbis Spectraux Contrastés. Cette matrice de résultats offre pour chaque proposition d'une QCM la moyenne des pourcentages de certitude qui ont accompagné le choix de la proposition (voir p. 216). La certitude moyenne par proposition permet de chiffrer la conviction avec laquelle les répondants ont opté pour une proposition. On s'attend à ce que la certitude moyenne soit plus élevée lorsqu'elle accompagne la réponse correcte que lorsqu'elle accompagne une proposition incorrecte.

Cq
L'indice de Centration par question

Nous pouvons confronter pour chaque question l'indice spectral de facilité/difficulté introspective par question (*piq*), avec l'indice classique de facilité/difficulté objective par question (*poq*). On obtient alors l'indice de Centration par question (*Cq*) qui se calcule : $Cq = piq - poq$.

Le signe de la valeur obtenue à l'indice *Cq* nous informe sur la tendance à la sous-estimation (-) ou à la surestimation (+) dans les résultats des sujets.

Le principe de la turbo analyse peut aussi être appliqué au calcul des *Cq*. On obtient alors aux paliers de turbo analyse les plus élevés, des indices *Cq* calculés sur la base des données des étudiants les plus réalistes.

Il est possible de représenter graphiquement les indices de Centration par question (*Cq*) obtenus aux différents paliers de turbo analyse sur un graphique en nuage de points. Pour chaque question, un point sur le graphique représente la valeur du taux d'exactitude (en ordonnée) et la valeur de la certitude moyenne (en abscisse) obtenues à un palier *T* de turbo analyse. On observe ainsi une série de points par question. Ces points peuvent être reliés et former un « brin » pour chaque question. Chaque brin tend, en principe, à se rapprocher, en fonction des paliers *t* de plus en

plus élevés, de la diagonale idéale qui est l'expression des situations où les certitudes moyennes correspondent aux taux d'exactitude, autrement dit, où la valeur de l'indice de facilité/difficulté introspective par question (*piq*) est égale à la valeur de la facilité/difficulté objective par question (*poq*).

NCSq
Niveau de Cohérence Spectrale d'une question

Dans le cas d'une QCM, le Niveau de Cohérence Spectrale de la question (*NCSq*) est obtenu au départ des rpbis Spectraux Contrastés (*rpbis SC*) récoltés pour chaque proposition. Le calcul du *NCSq* consiste alors à soustraire du *rpbis SC* de la réponse correcte la moyenne pondérée des *rpbis SC* des propositions incorrectes.

Les *rpbis SC* calculés pour les propositions d'une QCM nous donnent une information sur la cohérence avec laquelle les répondants ont utilisé les pourcentages de certitude. Lorsqu'une question fonctionne bien, logiquement les sujets qui ont choisi la réponse correcte accompagnent leur choix d'un degré de certitude en moyenne plus élevé que ceux qui ont choisi une proposition incorrecte. A l'aide du *NCSq* nous globalisons au niveau d'une QCM les informations obtenues au départ des différentes propositions.

Ayant repéré les QCM qui pourraient poser des problèmes de cohérence spectrale à l'aide des *NCSq*, nous pouvons ensuite affiner notre diagnostic en analysant les *rpbis SC* des propositions.

Le principe de la turbo analyse peut aussi être appliqué au calcul des *NCSq*. Lorsqu'on calcule les *NCSq* à partir des données des sujets les plus réalistes (le réalisme des sujets est mesuré à l'aide de l'indice *Rs*), par exemple au palier de turbo analyse T80, on obtient alors des informations sur la cohérence d'utilisation des pourcentages de certitude plus valides car basées sur les sujets dont *Rs* est plus grand ou égal à 80, donc qui commettent moins de 20% d'erreurs dans leurs auto-estimations.

N Rep.
Nombre de répondants

La ventilation des répondants dans les différentes propositions nous informe sur l'attractivité de celles-ci (voir p. 216).

N Rep. T
Nombre de répondants pour chaque proposition à un palier *t* de turbo analyse

L'utilisation de la turbo analyse dans le cadre du calcul des *rpbis SCT* entraîne parmi le groupe initial de répondants des sélections d'étudiants en moyenne de plus en plus réalistes aux niveaux élevés des paliers *T*.

NCIq
L'indice de Niveau de Cohérence Interne d'une question

Il est possible de calculer le Niveau de Cohérence Interne d'une question (*NCIq*) en se basant sur les valeurs des *rpbis classiques* des propositions. L'indice *NCIq* se calcule sur le même principe que le Niveau de Cohérence Spectrale d'une question (*NCSq*) mais en utilisant cette fois les *rpbis classiques*. On soustrait la moyenne des *rpbis classiques* des propositions incorrectes à la valeur du *rpbis classiques* de la réponse correcte. L'indice varie entre –2 et +2. Plus le *NCIq* est proche de la valeur maximale, plus la question participe à la cohérence interne de l'épreuve.

Les Niveaux de Cohérence Interne des Questions (*NCIq*) peuvent se visualiser sur un graphique en nuage de points divisé en quatre quadrants. Pour chaque QCM, la valeur placée en abscisse correspond à celle du *rpbis classiques* de la réponse correcte et la valeur placée en ordonnée à celle de la moyenne des *rpbis classiques* des propositions incorrectes. Une zone de « *cohérence interne* » est définie dans le 4ème quadrant, là où la valeur de la moyenne des *rpbis classiques* des propositions incorrectes peut varier entre 0 et –1 et où la valeur du *rpbis classiques* de la réponse correcte peut varier entre 0 et 1. Plus un point représentant une QCM se rapproche du coin inférieur droit du graphique, plus cette question participe à la cohérence interne de l'épreuve.

% Rep.
Pourcentage de répondants

La matrice des résultats des rpbis Spectraux Contrastés du module « *Traitements basiques* » du logiciel *SCANTEST 2.0* nous donne pour chaque proposition le pourcentage de sujets ayant choisi celle-ci (voir p. 216). Ce pourcentage nous informe sur l'attractivité de chaque proposition, et, lorsqu'on considère la proposition correspondant à la réponse correcte, sur la facilité de la question.

% Rep. T
Pourcentage de répondants pour chaque proposition à un palier *t* de turbo analyse

Le pourcentage de répondants à un palier *T* de turbo analyse correspond à la valeur du *N Rep. T* traduite en pourcentage. Cette information nous permet de comparer les répartitions des sujets dans les propositions aux différents paliers de turbo analyse.

piq
Indice de facilité/difficulté introspective par question

L'utilisation des pourcentages de certitude permet de calculer la facilité/difficulté introspective de chaque question via le pourcentage moyen des certitudes qui ont accompagné le choix de la réponse correcte. Le principe de turbo analyse peut aussi être appliqué aux indices *piq*, dans ce cas, aux paliers de turbo analyse les plus élevés, c'est à partir des résultats des étudiants les plus réalistes que l'indice est calculé.

poq
L'indice de facilité/difficulté objective par question

L'indice de difficulté *p* d'un item est la proportion de sujets qui fournissent la réponse correcte à la question.

PSq
Profil Spectral d'une question

Les pourcentages de réponses correctes et incorrectes associées à chaque pourcentage de certitude permettent de tracer le Profil Spectral d'une question (*PSq*). Ce type de profil se compose de deux hémi-spectres : un premier pour les certitudes associées à la réponse correcte et un second pour les certitudes associées aux réponses incorrectes.

L'observation de la répartition des répondants dans les différentes certitudes accompagnant d'une part la réponse correcte et d'autre part les propositions incorrectes nous permet d'identifier les questions qui posent problème en ce qui concerne la cohérence d'utilisation des pourcentages de certitude. Par exemple, pour l'hémi-spectre des réponses incorrecte d'une question qui fonctionne bien du point de vue de la cohérence avec laquelle les répondants choisissent les pourcentages de certitude, on ne s'attend pas à ce que une majorité de sujets ayant répondu incorrectement opte pour la certitude 100%. De même pour l'hémi-spectre de la réponse correcte : on ne s'attend pas à ce qu'une majorité de sujets choisisse la certitude 0%.

Le principe de la turbo analyse peut aussi être appliqué aux tracés des *PSq*. Dès lors, la validité des profils obtenus augmente avec le seuil de réalisme exigé pour la turbo analyse. En effet aux paliers turbo les plus élevés, les profils sont construits à partir des données des étudiants les plus réalistes.

rpbis classique
Corrélation point bisériale

Dans le cas du *rpbis classique*, les choix ou les rejets (1 ou 0) d'une proposition d'une QCM sont corrélés avec les nombres de réponses correctes obtenues à l'ensemble des questions du test. Dès lors, on s'attend à ce que la proposition correcte soit corrélée avec des nombres de réponses correctes plus élevés et les distracteurs avec des nombres de réponses correctes moins élevés. Dans ce cas, la question discrimine les étudiants en fonction du critère du nombre de réponses correctes fournies à l'ensemble des questions du test.

Le module « *Traitements basiques* » du logiciel *SCANTEST 2.0* permet de générer une matrice de résultat « *rpbis classiques* » (voir pp. 214 et 216). Le principe et la procédure de calcul des rpbis classiques ont été expliqués en détail p. 171.

rpbis S
Corrélation point bisériale Spectrale

Le coefficient de corrélation point bisériale spectrale (*rpbis S*) est la corrélation entre les choix ou les rejets (0 ou 1) de la réponse correcte avec les degrés de certitude qui ont accompagné cette réponse correcte. Il s'agit d'un indice de cohérence spectrale, il permet d'évaluer dans quelle mesure les étudiants qui ont répondu correctement ont accompagné leurs réponses de pourcentages de certitude plus élevés et ceux qui ont répondu incorrectement de pourcentages de certitude moins élevés. Contrairement au *rpbis classique*, le *rpbis spectral* est calculé pour chaque question indépendamment des autres questions de l'épreuve. Avec le *rpbis spectral* il n'est donc pas nécessaire de calculer une valeur repère. Plus la réponse correcte d'une question « fonctionne bien » du point de vue de la cohérence spectrale, plus son *rpbis spectral* sera proche de 1. En ce qui concerne les distracteurs, on s'attend à des *rpbis spectraux* négatifs lorsque la QCM fonctionne normalement.

rpbis SC
Corrélation point bisériale Spectrale Contrastée

Le principe du *rpbis classique* est réutilisé dans le cas de l'analyse spectrale mais avec une différence de taille : le *rpbis SC* n'utilise pas les nombres de réponses correctes obtenues pour l'ensemble du test mais les remplace par les pourcentages de certitude qui sont corrélés avec les choix ou les rejets de la proposition auxquels ils sont associés. La mesure critère n'est plus externe à la question (le nombre de réponses correctes à l'ensemble des questions du test, ce qui inclut les performances aux autres questions) mais interne à la question dans la mesure où les certitudes sont celles fournies dans le cadre de la QCM envisagée et uniquement dans ce cadre (la qualité des autres questions du test n'influence pas, comme c'est le cas pour le *rpbis classique*, la mesure de la qualité de la QCM envisagée).

Le principe et la procédure de calcul du rpbis Spectral Contrasté (*rpbis SC*) ont été expliqués en détail, p. 178. La matrice de résultat d'un traitement « *Calcul des rpbis Spectraux Contrastés* » générée par le logiciel *SCANTEST 2.0* permet d'obtenir la valeur du *rpbis SC* de chaque proposition (voir pp. 213 et 216). La valeur du *rpbis SC* varie entre –1 et +1. On s'attend généralement à ce que le *rpbis SC* de la réponse correcte soit positif et le *rpbis SC* d'une proposition incorrecte soit négatif. Lorsqu'il est calculé pour la réponse correcte, le *rpbis SC* est identique au *rpbis S*.

rpbis SCT
Corrélation point bisériale Spectrale Contrastée calculée dans le cadre d'une Turbo analyse

Nous avons calculé les *rpbis SC* en sélectionnant les données des étudiants à des seuils de réalisme (*Rs*) de plus en plus élevés. Le terme « Turbo analyse » utilisé pour qualifier ce type de traitement fait référence à la montée en puissance de l'instrument en terme de qualité d'information fournie au fur et à mesure que l'on progresse dans les « paliers Turbo », c'est-à-dire que l'on prend en compte les données des étudiants qui commettent de moins en moins d'erreurs dans leurs auto-estimations. Nous proposons l'appellation *rpbis SCT* suivie de l'indication du seuil *Rs* (un nombre entre 0 et 100) utilisé pour calculer le *rpbis SC* dans le cadre d'une Turbo analyse.

La turbo analyse peut être appliquée au calcul du rpbis Spectral Contrasté (voir pp. 214 et 217). On obtient alors aux paliers *T* élevés de turbo analyse des *rpbis SCT* plus valides car calculés à partir des données des étudiants les plus réalistes (qui commettent moins d'erreurs d'auto-estimations).

Indice de Réalisation des prédictions par question (*Rq*)

On calcule l'Indice de Réalisation des prédictions par question (*Rq*) en établissant la moyenne pondérée des écarts entre chaque pourcentage de certitude et son taux d'exactitude, on soustrait ensuite la moyenne pondérée des écarts à 100 pour obtenir *Rq*.

Si les écarts sont faibles entre pourcentages de certitude (les prédictions) et les taux d'exactitude (les réalisations de ces prédictions), les résultats de la question enferment de bonnes auto-estimations chez les répondants. Lorsque ces écarts sont grands, on

peut s'attendre à des problèmes de surestimation ou de sous-estimation dans les réponses fournies par les sujets à la question (la surestimation ou la sous-estimation par question est mesurée à l'aide de l'indice de Centration par question, *Cq*).

Le principe de la turbo analyse peut aussi être appliqué au calcul des *Rq*. Comme pour les autres indices de cohérence d'utilisation des pourcentages de certitude on obtient alors aux paliers élevés de turbo analyse, des informations plus valides car basées sur les données des étudiants les plus réalistes.

Il est possible de visualiser sur un graphique en nuage de points la concordance des prédictions par rapport à leur degré de réalisation. L'axe des abscisses (X) du graphique reprend les différents pourcentages de certitude utilisés. L'axe des ordonnées (Y) reprend les taux d'exactitude observés pour chaque pourcentage de certitude. Dans le cadre des épreuves MOHICAN, on obtient 6 points, représentant les taux d'exactitudes des 6 pourcentages de certitude utilisés (0%, 20%, 40%, 60%, 80% et 100%). Chacun des six points tracés sur le graphique représente le taux d'exactitude (en ordonnée) du pourcentage de certitude (en abscisse) envisagé. La situation idéale de concordance parfaite des prédictions aux taux d'exactitude peut être visualisée sur le graphique par une diagonale dont une extrémité se situe au coin inférieur gauche et l'autre au coin supérieur droit. L'erreur de certitude moyenne est la moyenne pondérée des distances entre chaque points tracé et le point qui lui correspond sur la diagonale idéale.

$r_{qt\ mb}$
La corrélation « *question-total* » à partir d'une matrice binaire des résultats

La corrélation entre les résultats à une question *q* et les résultats totaux pour l'ensemble des autres questions du test sans cette question *q* nous donne une indication de la tendance des scores *q* et totaux à varier ensemble. Si la corrélation est négative, on observe une tendance à ce que les scores élevés à *q* correspondent à des scores totaux faibles et inversement, donc, *q* participe peu à la cohérence interne de l'épreuve.

$r_{qt\ ms}$
La corrélation « *question-total* » à partir d'une matrice spectrale des résultats

Il est possible de calculer la corrélation question-total à partir des résultats spectraux qui reprennent les pourcentages de certitude tout en tenant compte de l'aspect incorrect ou correct de la réponse (le signe – ou +). Comme pour la corrélation item-total calculée à partir des résultats binaires, si la corrélation est négative entre une question *q* et les scores totaux alors les scores élevés à *q* on tendance à correspondre à des scores totaux faibles et inversement. Dès lors, *q* participe peu à la cohérence interne de l'épreuve.

Turbo analyse

L'idée sous-jacente de la « Turbo analyse » est que nous pouvons accroître la confiance dans les informations liées aux indices spectraux si nous ne prenons en compte que les données des étudiants qui commettent le moins d'erreurs dans leurs auto-estimations. Nous pouvons en effet mesurer cette qualité d'auto-estimation par l'intermédiaire de l'indice de Réalisme des sujets (*Rs*) qui nous donne une indication précise quant au taux d'Erreur Moyenne Absolue de Certitude (*EMAC*) de chaque étudiant (voir détails, p. 186 et p. 274).

B. Glossaire des principaux symboles et abréviations

RC	Abréviation de « Réponse Correcte »
RI	Abréviation de « Réponse Incorrecte »
C0	Notation de la Certitude 0%.
C1	Notation de la Certitude 20%.
C2	Notation de la Certitude 40%.
C3	Notation de la Certitude 60%.
C4	Notation de la Certitude 80%.
C5	Notation de la Certitude 100%.
t	Notation accompagnant un indice et indiquant que ce dernier s'applique au test.
q	Notation accompagnant un indice et indiquant que ce dernier s'applique à une question.
nq	Le nombre de questions.
s	Notation accompagnant un indice et indiquant que ce dernier s'applique aux sujets, les étudiants.
ns	Le nombre de sujets.
Rs	Notation désignant le score de Réalisme des sujets.
Cs	Notation désignant le score de Centration des sujets.
T	Abréviation de Turbo analyse signifiant qu'elle a été appliquée au calcul des indices qui l'accompagnent. Généralement cette abréviation est suivie d'un chiffre entre 0 et 100 qui indique le seuil de réalisme (ou palier de turbo analyse) à partir duquel les données des étudiants ont été prises en compte pour le calcul des indices.
P	Indice de facilité/difficulté.
po	Indice de facilité/difficulté objective
pi	Indice de facilité/difficulté introspective.
poq	Indice de facilité/difficulté objective d'une question.
piq	Indice de facilité/difficulté introspective d'une question.
pot	Indice de facilité/difficulté objective d'un test.
pit	Indice de facilité/difficulté introspective d'une question.
r	Coefficient de corrélation Bravais-Pearson.
$r_{xx'}$	Coefficient de bipartition (« *Split half test coefficient* »).
rS	Coefficient de bipartition après correction de Spearman-Brown.
rG	Coefficient de fiabilité de Guttman.
α	Coefficient alpha de Cronbach.
$kq_{\alpha=0,8}$	Nombre de questions à ajouter au test pour obtenir un alpha de 0,8.
rpbis	Indice du coefficient de corrélation bisériale de point classique.
$rpbis^{c}$	Indice du coefficient de corrélation bisériale de point classique de la réponse Correcte d'une question.
$\overline{rpbis}^{i}$	Indice de la moyenne pondérée des coefficients de corrélation bisériale de point classique des réponses Incorrectes d'une question.
NCIq	Indice du Niveau de Cohérence Interne d'une question.
rpbis SC	Indice du coefficient de corrélation bisériale de point Spectral Contrasté.

rpbis SC ^{c}	Indice du coefficient de corrélation bisériale de point Spectral Contrasté d'une réponse Correcte.
$\overline{rpbis\ SC}^{i}$	Indice de la moyenne des coefficients de corrélation bisériale de point Spectral Contrasté des réponses Incorrectes.
rpbis SCT	Indice du coefficient de corrélation bisériale de point Spectral Contrasté après Turbo analyse.
rpbis SCT ^{c}	Indice du coefficient de corrélation bisériale de point Spectral Contrasté d'une réponse Correcte après Turbo analyse.
$\overline{rpbis\ SCT}^{i}$	Indice de la moyenne des coefficients de corrélation bisériale de point Spectral Contrasté des réponses Incorrectes après Turbo analyse.
NCSq	Indice du Niveau de Cohérence Spectrale d'une question (généralement le numéro de la question suit l'indice).
NCSt	Indice du Niveau de Cohérence Spectrale d'un test.
GSt	Abréviation de Gerbe Spectrale d'un test.
BSq	Abréviation de Brin Spectral d'une question.
PSq	Abréviation de Profil Spectral d'une question.
TRC	Abréviation de Taux de Réponses Correctes.
TRI	Abréviation de Taux de Réponses Incorrectes.
Rq	Indice de Réalisation des prédictions par question.
Cq	Indice de Centration des questions.
Ct	Indice de Centration d'un test.
TE	Taux d'Exactitude des étudiants, équivalent à l'indice de facilité/difficulté objective multiplié par 100 (*po* x 100), nous préférons utiliser les appellations *poq* ou *pot* lorsqu'il est question des propriétés des QCM ou des tests plutôt que des performances des étudiants.
CM	Certitude Moyenne d'un sujet, équivalent à l'indice de facilité/difficulté introspective multiplié par 100 (*pi* x 100), nous préférons utiliser les appellations *piq* ou *pit* lorsque qu'il est question des propriétés des QCM ou des tests plutôt que des performances des étudiants.
CMp	Certitude Moyenne obtenues par une proposition (la moyenne des certitudes fournies par un groupe de sujets ayant choisi une proposition *p*).

Index des formules

Annexes

Annexes

A. Formulom d'évaluation des examens (version FAPSE-ULg, 1996-1997 et 1997-1998)

B. Questionnaires des épreuves MOHICAN

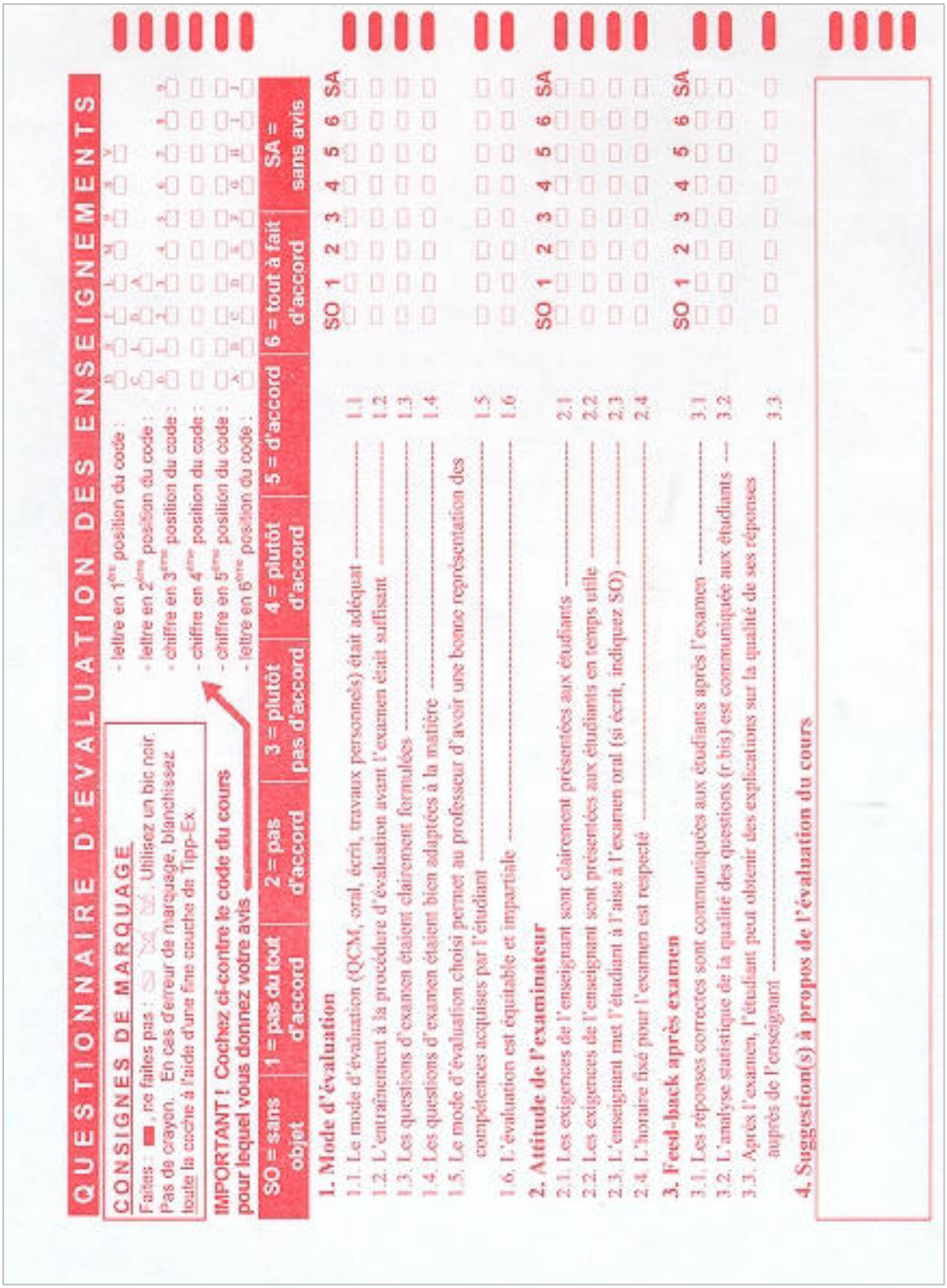
C. Gerbes Spectrales des tests MOHICAN

D. Protocoles SCANTEST 2.0 d'analyse des propositions

E. Table du T de Student

F. Tableaux des valeurs obtenues par les 173 QCM aux rpbis classiques, rpbis SC, rpbis SCT80 et rpbis SCT90

A. Formulom d'évaluation des examens (version FAPSE-ULg, 1997 à 1999)



QUESTIONNAIRE D'EVALUATION DES ENSEIGNEMENTS

CONSIGNES DE MARQUAGE

Faites : ■, ne faites pas : ⊘ ☒ ✓. Utilisez un bic noir. Pas de crayon. En cas d'erreur de marquage, blanchissez toute la case à l'aide d'une fine couche de Tipp-Ex.

IMPORTANT ! Cochez ci-contre le code du cours pour lequel vous donnez votre avis

- lettre en 1ère position du code :
- lettre en 2ème position du code :
- chiffre en 3ème position du code :
- chiffre en 4ème position du code :
- chiffre en 5ème position du code :
- lettre en 6ème position du code :

SO = sans objet	1 = pas du tout d'accord	2 = pas d'accord	3 = plutôt pas d'accord	4 = plutôt d'accord	5 = d'accord	6 = tout à fait d'accord	SA = sans avis

1. Mode d'évaluation

1.1. Le mode d'évaluation (QCM, oral, écrit, travaux personnels) était adéquat ---------- 1.1
1.2. L'entraînement à la procédure d'évaluation avant l'examen était suffisant ---------- 1.2
1.3. Les questions d'examen étaient clairement formulées ---------- 1.3
1.4. Les questions d'examen étaient bien adaptées à la matière ---------- 1.4
1.5. Le mode d'évaluation choisi permet au professeur d'avoir une bonne représentation des compétences acquises par l'étudiant ---------- 1.5
1.6. L'évaluation est équitable et impartiale ---------- 1.6

2. Attitude de l'examinateur

2.1. Les exigences de l'enseignant sont clairement présentées aux étudiants ---------- 2.1
2.2. Les exigences de l'enseignant sont présentées aux étudiants en temps utile ---------- 2.2
2.3. L'enseignant met l'étudiant à l'aise à l'examen oral (si écrit, indiquez SO) ---------- 2.3
2.4. L'horaire fixé pour l'examen est respecté ---------- 2.4

3. Feed-back après examen

3.1. Les réponses correctes sont communiquées aux étudiants après l'examen ---------- 3.1
3.2. L'analyse statistique de la qualité des questions (r bis) est communiquée aux étudiants --- 3.2
3.3. Après l'examen, l'étudiant peut obtenir des explications sur la qualité de ses réponses auprès de l'enseignant ---------- 3.3

4. Suggestion(s) à propos de l'évaluation du cours

B. Questionnaires des épreuves MOHICAN

Epreuve de vocabulaire (VOCABU)

NOM :	PRENOM :	N° Check up : …… N° (univ./Fac./Sec.) : ……..

Vocabulaire

Check up (Projet Mohican – CIUF) Auteur : M. MONBALLIN - FUNDP

Forme de questionnaire 1

Consignes : Questions à choix multiple avec degrés de certitude
(une feuille de consignes détaillées est disponible en annexe)

- **Répondez à l'aide du formulaire de réponses destiné à la lecture optique. Gardez le présent Questionnaire.**

A. Parmi les cinq mots proposés à la suite de chaque énoncé, choisissez CELUI QUI REPREND LE PLUS EXACTEMENT POSSIBLE le sens que le mot souligné a dans la phrase donnée.

Q1. Ce nouvel élément corrobore les thèses que nous avons développées précédemment.
1. complète
2. confirme
3. nuance
4. contredit
5. détaille
6. Aucune
7. Toutes

Q2. Quand il fut nommé, il fit largement valoir ses prérogatives.
1. droits
2. relations
3. antécédants
4. revendications
5. qualités
6. Aucune
7. Toutes

Q3. Le responsable des forces de l'ordre a stigmatisé l'attitude de la police lors des manifestations d'étudiants.
1. approuvé
2. exagéré
3. condamné
4. surveillé
5. commenté
6. Aucune
7. Toutes

Q4. Il a présenté son projet sans en mentionner les difficultés sous-jacentes.
1. majeures
2. inévitables
3. annexes
4. cachées
5. plausibles
6. Aucune
7. Toutes

Q5. Des haines <u>latentes</u> opposaient les deux hommes.
1. profondes
2. anciennes
3. farouches
4. douloureuses
5. patentes
6. Aucune
7. Toutes

Q6. Leurs interprétations <u>divergent</u> sur ce point.
1. se renforcent
2. s'embrouillent
3. se contredisent
4. s'écartent
5. sont fausses
6. Aucune
7. Toutes

Q7. Personne n'a réagi à ses <u>dénégations</u>.
1. mensonges
2. interdictions
3. démentis
4. aveux
5. dénonciations
6. Aucune
7. Toutes

Q8. Les événements du mois passé ont sans doute largement contribué au <u>discrédit</u> du gouvernement actuel.
1. à la déconsidération
2. à la condamnation
3. à l'impopularité
4. à la dénonciation
5. à la pénurie
6. Aucune
7. Toutes

B. La phrase incomplète qui suit chaque énoncé en reformule l'idée. Parmi les cinq fins de phrase proposées, choisissez CELLE QUI CONSERVE LE PLUS EXACTEMENT le sens que le mot souligné a dans l'énoncé de départ.

Q9. De la lecture du curriculum vitae de la candidate, le directeur <u>a inféré</u> qu'elle était apte à remplir les fonctions du poste à pourvoir.

Suite à la lecture du curriculum vitae de la candidate, le directeur ..... qu'elle était apte à remplir les fonctions du poste
1. a annoncé
2. a décidé
3. a admis
4. a conclu
5. a fait savoir
6. Aucune
7. Toutes

Q 10. Le conférencier a par ailleurs relevé certaines caractéristiques <u>inhérentes</u> à ce type de régime politique.

Les caractéristiques relevées par le conférencier.... ce type de régime.
1. sont indispensables au bon fonctionnement de
2. sont à la source de
3. sont incompatibles avec
4. s'observent occasionnellement dans
5. sont inséparablement liées à
6. Aucune
7. Toutes

Q11. Ces mesures accentueront les <u>clivages</u> sociaux.

Suite à ces mesures, ...
1. les négociations vont s'intensifier
2. les conflits vont redoubler
3. les dérapages vont se multiplier entre les classes sociales
4. les échanges vont augmenter
5. les distances vont s'accroître
6. Aucune
7. Toutes

Q12. Le valet de don Juan agit <u>à l'instar de</u> son maître, ce qui crée un décalage comique.

Le valet de don Juan...
1. désobéit effrontément à son maître
2. fait tout le contraire de ce que fait son maître
3. seconde et aide maladroitement son maître
4. cache à son maître tout ce qu'il fait
5. exagère ce que fait son maître
6. Aucune
7. Toutes

C. Parmi les cinq mots proposés à la suite de chaque énoncé, choisissez CELUI QUI EXPRIME LE PLUS EXACTEMENT POSSIBLE LE CONTRAIRE du sens qu'a le MOT souligné dans la phrase donnée.

Q13. Dans l'interview à laquelle elle s'est prêtée, la chanteuse a éludé les Questions sur sa vie privée.
LE CONTRAIRE de « éludé » =
1. Autorisé
2. Apprécié
3. joué avec
4. trié
5. affronté
6. Aucune
7. Toutes

Q14. La remarque qu'il a faite à son collègue était tout à fait judicieuse.
LE CONTRAIRE de « judicieuse » =
1. injuste
2. absurde
3. blessante
4. naïve
5. pertinente
6. Aucune
7. Toutes

Q15. Dans son nouvel ouvrage, cet auteur réfute les thèses de ses prédécesseurs.
LE CONTRAIRE de « réfute » =
1. modifie
2. reformule
3. commente
4. confirme
5. affine
6. Aucune
7. Toutes

Q16. Tous les chiffres de cette étude mettent en évidence la précarité des emplois de beaucoup de jeunes.
LE CONTRAIRE de « précarité » =
1. qualité
2. diversité
3. stabilité
4. rentabilité
5. facilité
6. Aucune
7. Toutes

Q17. Dans sa plaidoirie, l'avocat a accumulé les présomptions.
LE CONTRAIRE de « présomptions » =
1. témoignages
2. contre-vérités
3. éloges
4. mensonges
5. détails
6. Aucune
7. Toutes

Q18. L'histoire est la plus partiale des sciences.
LE CONTRAIRE de « partiale » =
1. objective
2. complète
3. générale
4. ancienne
5. positive
6. Aucune
7. Toutes

D. La phrase incomplète qui suit chaque énoncé en formule l'idée contraire.
Parmi les quatre fins de phrase proposées, choisissez CELLE QUI EXPRIME LE PLUS EXACTEMENT LE CONTRAIRE du sens qu'a le mot souligné dans LA PHRASE donnée.

Q19. Sa secrétaire était habilitée à délivrer certains documents officiels.

Au contraire, sa secrétaire... délivrer certains documents officiels.
1. n'avait pas le droit de
2. était lente à
3. n'avait pas coutume de
4. faisait des difficultés pour
5. s'était peu à peu habituée à
6. Aucune
7. Toutes

Q20. A travers son oeuvre, l'auteur a dressé un tableau exhaustif de la France du XIX^{e} siècle.

Au contraire, dans le tableau qu'il a dressé, l'auteur...la France du XIX^{e} siècle.
1. a établi un constat très critique de
2. a fait l'éloge de
3. a accentué les traits de
4. a posé un regard pessimiste sur
5. a ironisé sur la situation de
6. Aucune
7. Toutes

Q21. Il a formule des griefs à l'égard de certains de ses collaborateurs.

Au contraire, il... ses collaborateurs.
1. est resté très poli en s'adressant à
2. fait des suggestions à
3. a félicité certains de
4. a adressé des voeux de réussite à
5. a laissé toute initiative à
6. Aucune
7. Toutes

E. Comparez un à un les mots des listes qui suivent avec le mot en gras auquel ils sont associés.
Pour chacune des paires ainsi constituées, déterminez le rapport de sens en écrivant

- ***1. si, dans certains contextes, les deux mots peuvent avoir à peu près le MEME sens.***
- ***2. si, dans certains contextes, les deux mots peuvent avoir des sens à peu près CONTRAIRES.***
- ***3. si, dans aucun contexte, les deux mots n'entretiennent l'un des deux rapports précédents.***

Assujettir				
Q22.	- émanciper	1	2	3
Q23.	- subordonner	1	2	3
Q24.	- asservir	1	2	3
Q 25.	- affranchir	1	2	3

Injonction				
Q 26.	- sommation	1	2	3
Q 27.	- invective	1	2	3
Q 28.	- ordre	1	2	3
Q 29.	- addition	1	2	3

Prohibé				
Q 30.	- permis	1	2	3
Q 31.	- inhibé	1	2	3
Q 32.	- proscrit	1	2	3
Q 33.	- légal	1	2	3

Différer				
Q 34.	- retarder	1	2	3
Q 35.	- contester	1	2	3
Q 36.	- ajourner	1	2	3
Q 37.	- anticiper	1	2	3

F. Choisissez, pour chacun des mots en majuscules, LA DEFINITION QUI LUI CORRESPOND.

Q38. PROBITE
1. désir immodéré de posséder une chose
2. observation rigoureuse des devoirs de la justice et de la morale
3. penchant effréné ou irrésistible pour la luxure
4. sentiment violent qui pousse à vouloir du mal à quelqu'un
5. période où l'on doit faire ses preuves
6. Aucune
7. Toutes

Q39. RÉACTIONNAIRE
1. qui s'oppose au progrès et vise à rétablir un État antérieur
2. qui conteste l'ordre établi et est partisan de changements radicaux et rapides
3. qui possède des titres de propriété dans une entreprise
4. qui exerce une force de propulsion
5. qui est prompt à s'irriter, à s'emporter
6. Aucune
7. Toutes

Q40. EXHORTER
1. empêcher quelqu'un de réaliser son projet
2. arracher quelque chose à quelqu'un
3. s'efforcer de persuader quelqu'un de faire quelque chose
4. élever très haut par ses discours, par ses enseignements
5. émettre une odeur particulière
6. Aucune
7. Toutes

Q41. ARBITRAIRE
1. qui est prononcé par un ou plusieurs arbitres
2. qui dépend du bon plaisir, du caprice de quelqu'un
3. qui est le résultat d'une concertation
4. qui augmente progressivement
5. qui ne peut être annulé
6. Aucune
7. Toutes

Q42. EMPIRIQUE
1. qui va de plus en plus mal
2. qui relève du style baroque
3. qui détient l'autorité absolue
4. qui s'appuie sur l'expérience
5. qui est incertain
6. Aucune
7. Toutes

Q43. DÉPERDITION
1. État de ce qui se déteriore, se délabre, tombe en ruine
2. État d'un navire en danger de naufrage
3. perte qui se fait graduellement
4. perte des principes moraux
5. acte de vandalisme
6. Aucune
7. Toutes

Q44. ALLEGATION
1. affirmation quelconque
2. obligation de fidélité et d'obéissance à une nation
3. fait de rendre moins pesant, moins pénible
4. fait de donner à quelqu'un le droit d'agir au nom d'un autre
5. fait de minimiser un propos
6. Aucune
7. Toutes

Q45.SUBVERSION
1. aide que l'État, qu'une association accorde à un groupement, à une personne
2. moyen habile et détourné pour échapper à une situation
3. action de dérober avec adresse
4. bouleversement des idées et des valeurs reçues, renversement de l'ordre établi
5. action de revenir à la surface
6. Aucune
7. Toutes

Epreuve de Syntaxe et articulation logique (SYNTAX)

NOM :	PRENOM :	N° Check up : …… N° (univ./Fac./Sec.) : ……

Syntaxe et articulation logique

Check up (Projet Mohican – CIUF) Auteur : Département de français – ISLV - ULg

Forme de questionnaire 1

Consignes : Questions à choix multiple avec degrés de certitude
(une feuille de consignes détaillées est disponible en annexe)

- **Répondez à l'aide du formulaire de réponses destiné à la lecture optique. Gardez le présent questionnaire.**

Le pronom relatif

Q 1. Cet étudiant fait des efforts considérables pour suivre la conférence du professeur ———— on voit qu'il ne comprend que quelques bribes.

1. que l'
2. qu'
3. duquel
4. laquelle
5. pour laquelle
6. Aucune
7. Toutes

Q2. Une dernière chose ———— nécessaire que tu saches : les inscriptions doivent être renvoyées avant la fin de ce mois.

1. qu'il est
2. qui est
3. qui y est
4. qui l'est
5. qui t'est
6. Aucune
7. Toutes

Conjonction (ou locution conjonctive)

Q3. Il ne remettra jamais les pieds ici ———— tu ne lui présentes tes excuses.

1. à moins que
2. d'autant plus que
3. jusqu'à ce que
4. alors que
5. pour que
6. Aucune
7. Toutes

Q4. La conférence était ———— les participants à la formation n'ont pas vu le temps passer.

1. passionnante **à tel point que**
2. passionnante **de sorte que**
3. passionnante **au point que**
4. **tellement** passionnante **que**
5. **si** passionnante **que**
6. Aucune
7. Toutes

Les pronoms

Q5. Que représente le pronom <u>les</u> ?

Les solutions sont le lot des politiciens qui, prisonniers du court terme, en deviennent les otages. Leurs électeurs exigent au moins des promesses de solutions rapides. Ils ne se privent pas d'en distribuer. On se garderait bien de **<u>les</u>** en exempter. [V. FORRESTER, *L'Horreur économique*]

1. les otages
2. les solutions
3. les électeurs
4. les politiciens
5. les promesses
6. Aucune
7. Toutes

Q6. Que représente le pronom <u>y</u> ?

On met de plus en plus ces résultats sur le compte de défauts inhérents aux tests ou à l'éducation reçue par ceux qui passent les tests. On n'<u>y</u> voit plus une preuve de l'existence de différences innées entre les races.

1. dans les défauts
2. dans l'éducation
3. dans les résultats
4. dans les tests
5. dans les races
6. Aucune
7. Toutes

Inférence

Q7. Déterminez quelle(s) conclusion(s) on peut tirer de la phrase suivante :

Personne n'est sans savoir que Nicolas n'est pas peu lésé dans cette affaire.

1. Personne ne sait que Nicolas est largement lésé
2. Personne ne sait que Nicolas n'est pas lésé du tout
3. Tout le monde sait que Nicolas est largement lésé
4. Tout le monde sait que Nicolas n'est pas lésé du tout
5. Certaines personnes bien informées savent que Nicolas est largement lésé
6. Aucune
7. Toutes

Q8. Quelle(s) fin(s) cohérente(s) pourrait-on donner à la phrase suivante :

La commission vient de décider de suspendre ses travaux, son président estimant…

1. que plusieurs éléments neufs au dossier justifiaient qu'elle les reprenne
2. que les circonstances ne l'empêchaient désormais plus de les poursuivre
3. que la situation politique les rendait actuellement impossibles
4. qu'il était urgent d'arriver à dégager des conclusions fiables
5. qu'il lui fallait les mener le plus rapidement possible à leur terme
6. Aucune
7. Toutes

Ponctuation

Comment peut-on interpréter les phrases suivantes ?

Q9. Les questions du deuxième syllabus qui ont été traitées lors des travaux pratiques ne devront pas être révisées pour l'examen.

1. Aucune question du deuxième syllabus ne doit être révisée pour l'examen
2. Il reste dans le deuxième syllabus des questions qui devront être révisées pour l'examen
3. Toutes les questions du deuxième syllabus ont été traitées lors des travaux pratiques
4. Aucune des questions du deuxième syllabus n'a été à la fois traitée lors des travaux pratiques et révisée pour les examens
5. Toutes les questions des travaux pratiques devront être révisées pour l'examen
6. Aucune
7. Toutes

Q10.

Pierre aime Marie; Claude, Dominique.

1. Pierre aime Claude
2. Pierre aime Marie, Claude et Dominique
3. Claude aime Dominique
4. Pierre aime Dominique
5. Marie aime Claude
6. Aucune
7. Toutes

Articulation logique

Dans le texte suivant, certains mots ont été supprimés et remplacés par des lettres (a, b). Choisissez la combinaison qui convient pour obtenir un texte cohérent.

Q11. Nous absorbons de l'eau sans penser, le plus souvent, que nous sommes en train de satisfaire un besoin physiologique; fréquemment, nous buvons [..a..] par habitude [..b..] par soif parce que nous sommes soumis à un véritable conditionnement familial et social.

1. [..a..] toujours [..b..] jamais
2. [..a..] comme [..b..] si
3. [..a..] rarement [..b..] et
4. [..a..] plutôt [..b..] que
5. [..a..] alors [..b..] que
6. Aucune
7. Toutes

Q12. Lorsque nous transpirons, nous perdons de l'eau et du sel. Or, nous ressentons le besoin d'eau et non le besoin de sel; logiquement, nous ne devrions boire [..a..] de l'eau pure [..b..] un liquide sucré, mais de l'eau légèrement salée.

1. [..a..] jamais [..b..] plutôt
2. [..a..] ni [..b..] ni
3. [..a..] ni [..b..] plutôt
4. [..a..] rarement [..b..] ou
5. [..a..] que [..b..] ou
6. Aucune
7. Toutes

Epreuve de Compréhension (COMPRE)

NOM :	PRENOM :	N° Check up : …… N° (univ./Fac./Sec.) : ……..

Compréhension

Check up (Projet Mohican – CIUF) Auteur : Ph. HOUGARDY - ULB

Forme de questionnaire 1

Consignes : Questions à choix multiple avec degrés de certitude
(une feuille de consignes détaillées est disponible en annexe)

- **Répondez à l'aide du formulaire de réponses destiné à la lecture optique. Gardez le présent questionnaire.**

A lire : extraits du texte : « *Les Neurones* », Olivier BLOND, La Recherche, 316, janvier 1999, lignes 1 à 115 et 143 à 181.

Fournir un titre

Q1. Choisissez le ou les titre(s) le(s) qui pourrai(en)t convenir pour le premier paragraphe (ligne 1 à 60) :
1. Quelle est la composition des cellules nerveuses ?
2. Qui est l'inventeur de l'expression « cellule nerveuse » ?
3. Quelle est l'origine du microscope électronique ?
4. A quand remonte la découverte des premiers neurones ?
5. Le cerveau : une substance résistante
6. Aucune
7. Toutes

Interprétation

Q2. Le cerveau a résisté longtemps à l'observation parce que, d'après cet article :
1. le microscope n'est pas un instrument permettant d'observer de fine partie du cerveau
2. les autorités religieuses de l'époque s'y opposaient fermement
3. on craignait l'extension de maladies infectieuses particulièrement redoutées
4. une technique de rigidification et de coloration des tissus était nécessaire pour l'observation
5. le cerveau forme un réseau continu
6. Aucune
7. Toutes

Q3. Que dit l'article sur la fonction du neurone ?
1. la composition asymétrique du neurone correspond rigoureusement à une séparation de ses fonctions
2. la fonction d'un neurone est d'interrompre un signal passant dans un réseau
3. les neurones se présentent tous sous la même morphologie
4. l'arbre dendritique des neurones est toujours extrêmement ramifié et étendu
5. les cellules de la rétine forment un petit réseau longiligne
6. Aucune
7. Toutes

Caractéristiques(s) de concept

Q4. Quelle est la nature du différend qui séparait Camillo Golgi de son confrère Santiago Ramon y Cajal ?
1. Les deux souhaitaient être reconnus comme « le » précurseur
2. une rivalité d'ordre personnel
3. une polémique d'ordre scientifique
4. l'opinion sur la forme (continue ou non) du réseau neuronal
5. la comparabilité du réseau neuronal à celui de la circulation sanguine
6. Aucune
7. Toutes

Q5. La communication entre neurones a la caractéristique suivante :
1. Dans certains cas très limités, la transmission entre neurones est électrique
2. Dans la plupart des cas, le signal est émis sous forme électrique et transformé pour sa réception sous forme chimique
3. Le signal chimique réceptionné est converti en signal électrique
4. Elle se fait à l'aide d'éléments chimiques « voyageants » dans l'espace synaptique
5. De façon unidirectionnelle
6. Aucune
7. Toutes

Définition de mots

Q6. Le mot *synapse* dans cet article désigne :
1. la région de quasi-contact entre deux neurones
2. les excroissances autour du soma
3. l'arbre des neurones, ramifié et étendu
4. une cellule bipolaire de la rétine
5. le bout de l'axone qui reçoit le message de l'autre cellule
6. Aucune
7. Toutes

Epreuve de Lecture de document et géographie (GEOGRA)

NOM :	PRENOM :	N° Check up : …… N° (univ./Fac./Sec.) : ……

Lecture de documents et géographie

Check up (Projet Mohican – CIUF) Auteur : F. ORBAN - FUNDP

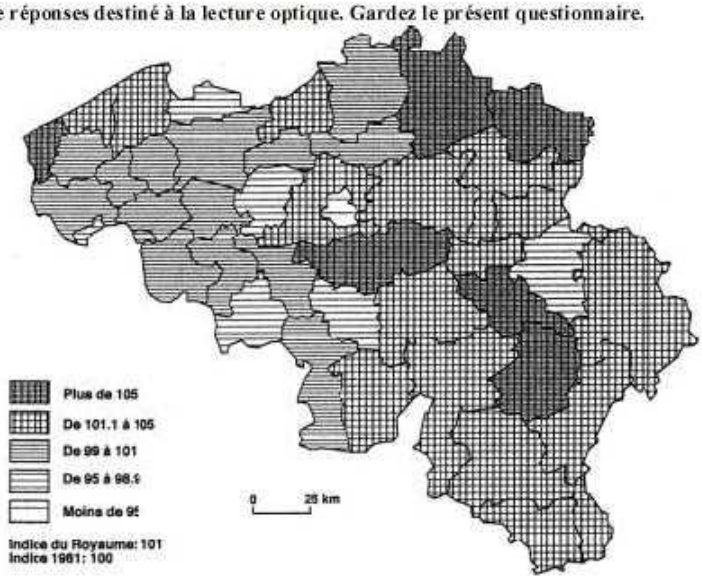
Forme de questionnaire 1

Consignes : Questions à choix multiple avec degrés de certitude
(une feuille de consignes détaillées est disponible en annexe)

- **Répondez à l'aide du formulaire de réponses destiné à la lecture optique. Gardez le présent questionnaire.**

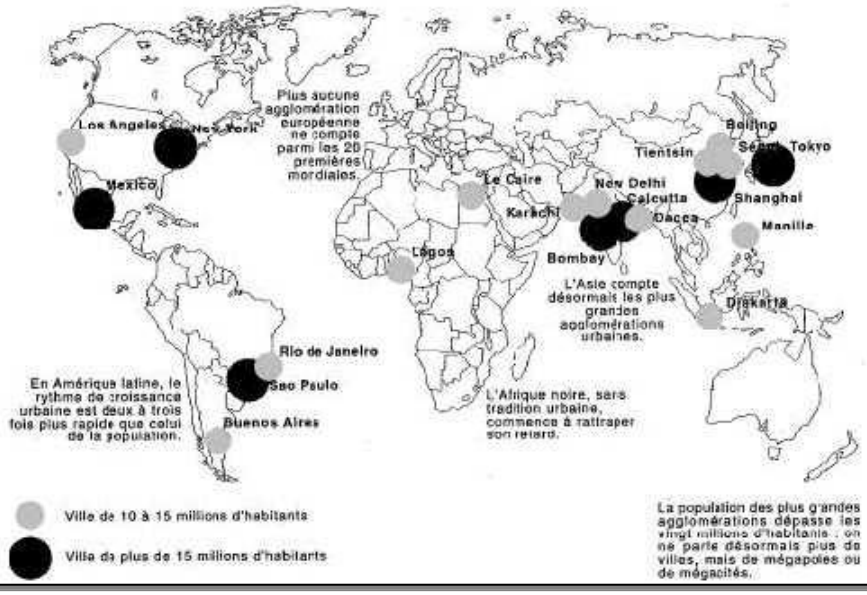
Q1. Selon la carte ci-contre:

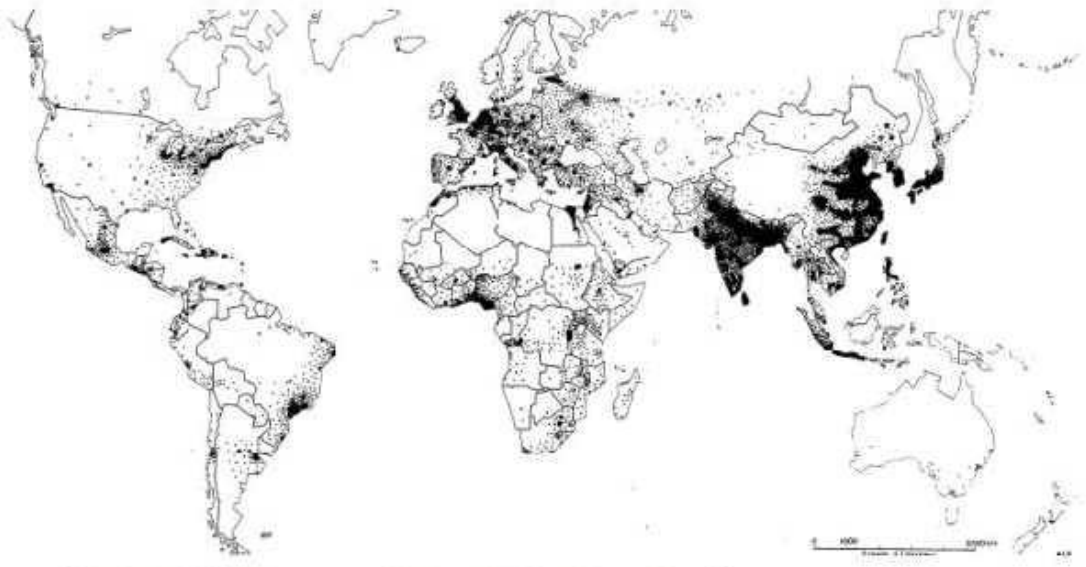
1. L'arrondissement de Bruxelles voit sa population diminuer de moins de 5 points
2. La Flandre voit sa population augmenter tandis qu'en Wallonie, elle diminue
3. La Campine est une région où la population est en déclin depuis 1981
4. Les grandes agglomérations wallonnes sont en déclin démographique
5. Les zones rurales wallonnes sont en décroissance démographique
6. Aucune
7. Toutes



Q2. Quel titre donner à la carte ci-dessous ?

1. La revanche des villes du Nord
2. Les grandes agglomérations les plus riches du monde
3. Villes du monde
4. Les 20 plus grandes agglomérations du monde
5. La croissance des agglomérations urbaines de 1980 à 2000
6. Aucune
7. Toutes





1 point représente 500.000 habitants
Population mondiale: 4219 million

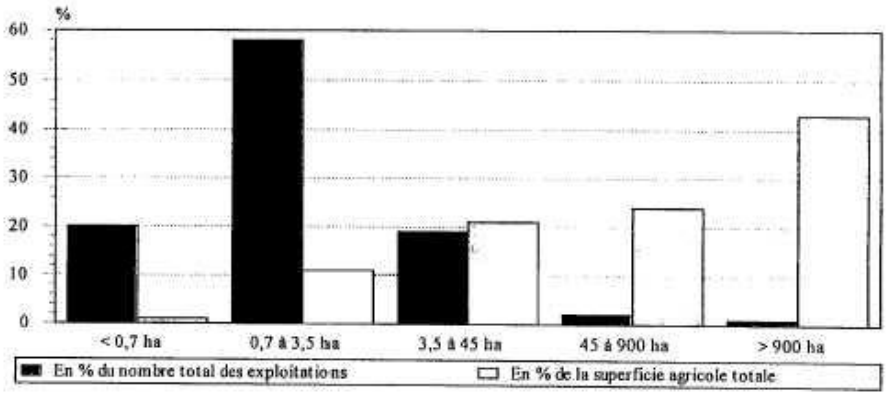
Distribution spatiale de la population mondiale en 1978
(Chiffres utilisés : World Popul. Data sheet).

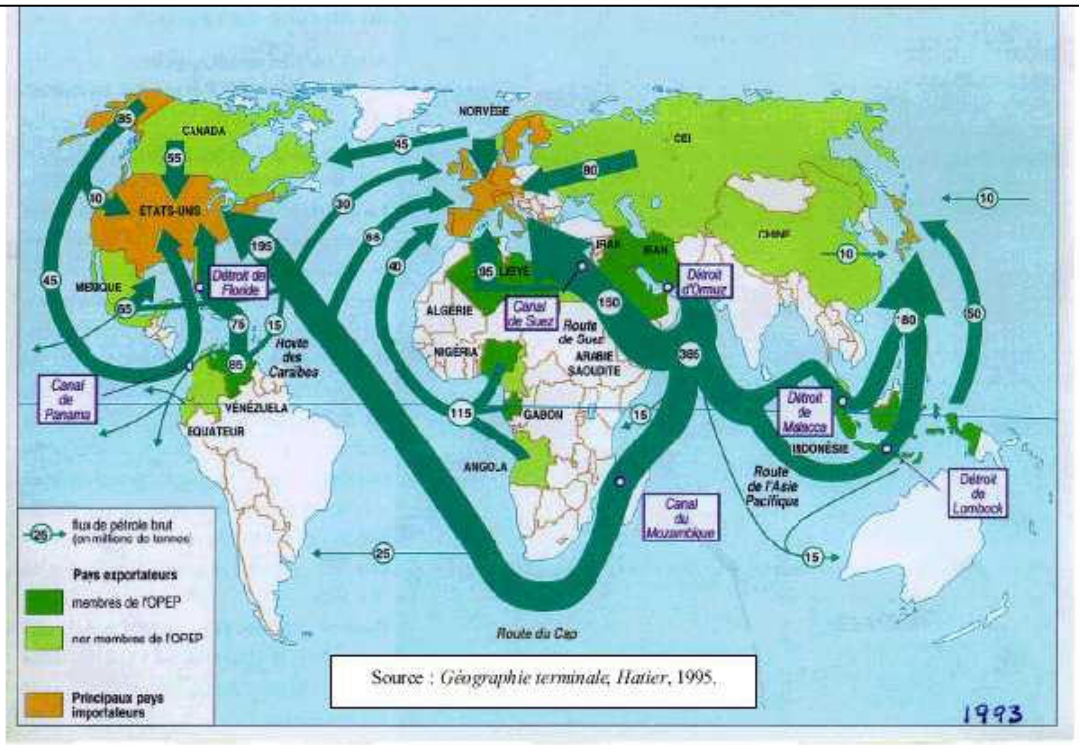
Q3. Selon la carte ci-dessus :

1. La population d'Amérique du Sud observe un modèle d'opposition Centre / Périphérie
2. Tous les grands foyers de population sont dans la zone intertropicale
3. Les zones très froides et très sèches attirent généralement les populations
4. Les zones chaudes et humides sont toujours densément peuplées
5. La population européenne est répartie de manière moins uniforme que la population d'Amérique du Nord
6. Aucune
7. Toutes

Q4. Selon le graphique ci-dessous :

1. Les petites exploitations (< à 3,5 ha) représentent plus de 70% de la superficie agricole totale
2. Les grands domaines agricoles (> à 900 ha) occupent plus de la moitié de la surface agricole
3. Les petites exploitations (< à 3,5 ha) représentent près de 80% des exploitations agricoles
4. Les exploitations agricoles supérieures à 45 ha représentent 25% de la surface agricole totale
5. Les grandes exploitations agricoles (> 45 ha) correspondent à plus de 10% du nombre total des exploitations
6. Aucune
7. Toutes



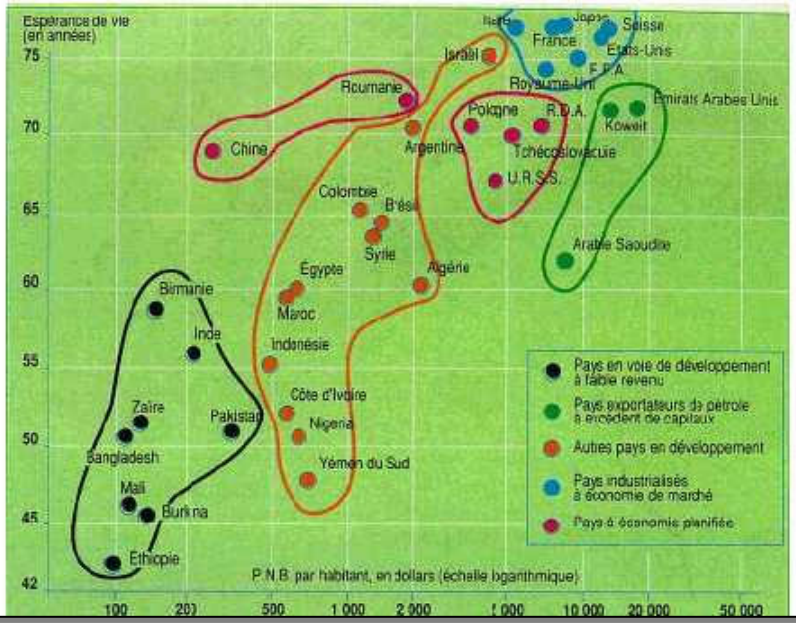


Q5. Selon la carte ci-dessus

1. Le plus grand importateur de pétrole est le Japon
2. Les Etats-Unis sont exportateurs de pétrole
3. Le Venezuela exporte principalement vers les Etats-Unis
4. L'OPEP regroupe exclusivement des pays Africains et Asiatiques
5. L'Europe ne produit pas de pétrole
6. Aucune
7. Toutes

Q6. Selon le graphique ci-dessous

1. Lorsque le Produit National Brut (PNB) augmente, l'espérance de vie diminue
2. La Pologne se situe en avance de développement socio-économique par rapport à l'Italie
3. Le Pakistan possède une production économique par habitant moins élevée que celle de la Birmanie
4. Si un pays a un PNB plus élevé qu'un autre, il a toujours une espérance de vie supérieure
5. Les pays à économie planifiée ont une espérance de vie inférieure à celle des pays industrialisés à économie de marché
6. Aucune
7. Toutes



Un monde inégalement développé, « Nord » et Sud

Source : *Le système Monde en question*, Magnard, 1989.

Le développement inégal : P.N.B. et espérance de vie à la naissance

Réflexion géographique

Q7. Sachant que les coordonnées géographiques de villes sélectionnées ci-dessous sont :

Canberra : 36°S, 149°E
Madrid : 40°N, 40°O
Mexico : 19°N, 99°O
New York : 41°N, 74°O
Pretoria : 25°S, 28°E
Vorkouta : 67° N, 64°E

sachant que :

Tropique du Cancer : 23°27' N
Tropique du Capricorne : 23°27' S
Cercle Polaire Arctique : 66°33' N
Cercle Polaire Antarctique : 66°33' S

Cochez la(les) affirmation(s) exacte(s).
Note : Il peut être utile que vous visualisiez les situations en effectuant quelque croquis positionnant la terre par rapport au soleil au moment précisé.

1. Le 1 décembre, le jour est moins long que la nuit à Pretoria.
2. Le 1 juillet, au midi solaire local, le soleil culmine plus haut à Madrid qu'à New York
3. Le 1 février, le soleil se lève exactement à l'est de Mexico
4. En avril, on peut voir le soleil de minuit à Vorkouta
5. Pour bien fonctionner, des panneaux solaires doivent être orientés vers le Sud à Canberra
6. Aucune
7. Toutes

Q8. La Cordillère des Andes s'est formée:

1. Suite à la séparation de la plaque d'Amérique du Sud et de l'Australie
2. Par la rencontre de la plaque Pacifique et de la plaque d'Amérique du Sud
3. Par la rencontre de la plaque Atlantique et de la plaque Amérique du Sud
4. Par la rencontre de la plaque Amérique du Nord et de la plaque Amérique du Sud
5. Par la rencontre de la plaque Antarctique et de la plaque d'Amérique du sud
6. Aucune
7. Toutes

Définitions

Q9.

Cochez la(les) définition(s) qui recouvre(nt) le concept de « Mégalopole »

1. Plus grande ville d'un pays
2. Réunion de plusieurs grandes agglomérations
3. Toujours un pôle de développement industriel
6. Aucune
7. Toutes

Q10.

Cochez la(les) définition(s) qui recouvre(nt) le concept de « Délocalisation »

1. Transfert des activités de production ou de services d'une entreprise d'un pays industrialisé vers un autre pays
2. Transfert des bénéfices d'une entreprise nationale vers un autre pays
3. Transfert du centre de décision d'une entreprise d'un pays vers un autre
6. Aucune
7. Toutes

Epreuve de Connaissances en Histoire et Socio Economie (HISTOI)

NOM :	PRENOM :	N° Check up : …… N° (univ./Fac./Sec.) : ……..

Connaissances en Histoire et Socio Economie

Check up (Projet Mohican – CIUF) Auteur : B. NOEL (FUCAM) – D. LECLERCQ (Ulg)

Forme de questionnaire 1

Consignes : Questions à choix multiple avec degrés de certitude
(une feuille de consignes détaillées est disponible en annexe)

- **Répondez à l'aide du formulaire de réponses destiné à la lecture optique. Gardez le présent questionnaire.**

Institutions internationales

Q 1. A l'heure actuelle, quels sont les 5 membres permanents de l'ONU ?
1. Japon – Allemagne – USA – Russie – France
2. République populaire de Chine – France – Grande-Bretagne – USA – Russie
3. Brésil – Inde – République populaire de Chine – USA – URSS
4. République populaire de Chine – Grande-Bretagne – France – USA – URSS
5. Italie – France – USA – URSS – Royaume-Uni
6. Aucune

Q2. Quel est le Secrétaire général actuel de l'ONU ?
1. Javier Solana
2. Boutros Boutros Ghali
3. Jacques Santer
4. Antonio Samaranch
5. Kofi Anan
6. Aucune

Q3. Quel est le Secrétaire général actuel de l'OTAN ?
1. Javier Solana
2. Boutros Boutros Ghali
3. Jacques Santer
4. Antonio Samaranch
5. Kofi Anan
6. Aucune

Q4. Dans l'abréviation G.A.T.T. (General Agreement …) que signifie un des deux T ?
1. Tourisme
2. Total
3. Tolerance
4. Tax
5. Trade
6. Aucune

Q5. Dans l'abréviation O.M.S. que signifie le S ?
1. Sport
2. Santé
3. Sida
4. Sécurité
5. Société
6. Aucune

Q6. Dans l'abréviation U.N.E.S.C.O. que signifie le E ?
1. Enfants
2. Elèves
3. Education
4. Environnement
5. Economie
6. Aucune

Q7. Dans l'abréviation B.I.R.D. que signifie le B ?
1. Banque
2. Bombe
3. Bataillon
4. Bilatéral
5. Bureau
6. Aucune

Q8. Dans l'abréviation U.N.I.C.E.F. que signifie le C ?
1. Children
2. Community
3. Christian
4. Country
5. Communist
6. Aucune

L'union Européenne

Q9. Voici 5 dates de la Construction Européenne. Laquelle correspond à l'entrée en vigueur de l'Euro ?
1. 1944
2. 1951
3. 1972
4. 1981
5. 1999
6. Aucune

Q10. Quel est le nom du nouveau Président de la Commission européenne (depuis trois mois) ?
1. SANTERRE
2. PRODI
3. SOLANA
4. FISCHLER
5. LAFONTAINE
6. Aucune

Repères historiques

Q11. En quelle année Hiroshima a-t-elle été détruite par la première bombe atomique ?

1. 1940
2. 1943
3. 1946
4. 1950
5. 1955
6. Aucune

Q12. En quelle année le Congo a-t-il cessé d'être une colonie belge ?

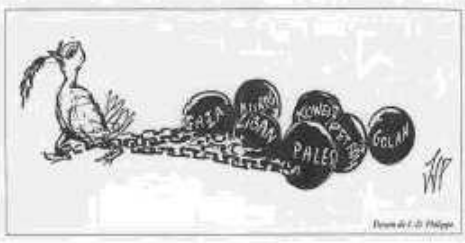
1. 1940
2. 1950
3. 1960
4. 1970
5. 1980
6. Aucune

Q13. En quelle année a eu lieu la Révolution française ?

1. 1432
2. 1524
3. 1618
4. 1789
5. 1815
6. Aucune

Géopolitique

Q14. Quelle est la région à laquelle ce document fait allusion ?



1. L'Amérique du Sud
2. Les Balkans
3. La Méditérannée
4. L'Asie du Sud-Est
5. Le Proche et le Moyen-Orient
6. Aucune

Vocabulaire français à référence historique

Q15. L'adjectif « Précolombien » est attribué :

1. A une civilisation, un objet, un art… de l'Amérique latine existant avant la découverte des Amériques par Christophe Colomb au 14ème Siècle
2. A un animal qui existait avant l'existence de la Colombe (Crétacé)
3. Aux institutions, régime politique… qui régissaient l'Amérique latine avant l'indépendance de ces pays et la création de la Colombie au 19ème Siècle
4. A une civilisation, un objet, ou un art… de l'Amérique latine existant avant la découverte des Amériques au 15ème Siècle
5. Aux civilisations dans l'état où elles existaient avant leur colonisation par l'Occident
6. Aucune

Q16. Iconoclaste fait référence à :

1. Quelqu'un qui détruit les images pieuses, les statues… à l'époque de la Réforme protestante au 16ème Siècle
2. Un adorateur d'Icônes dans le monde de la religion orthodoxe depuis le 11ème Siècle
3. Un Vandale, appartenant à ce peuple, lors des grandes invasions qui détruisirent l'Empire Romain d'Occident au 5ème Siècle et qui détruisait tout sur son passage
4. Un partisan d'une idéologie née dans l'empire byzantin au 9ème Siècle et qui refuse l'adoration des images pieuses, allant jusqu'à les détruire
5. Un destructeur des peintures religieuses dans les églises russes durant la période communiste (XXème Siècle)
6. Aucune

Q17. Un Kolkhoze est :

1. Une institution d'Etat en URSS au 19ème Siècle
2. Une grande ferme collective en URSS au 20ème Siècle
3. Un membre des jeunesses communistes, en URSS, au 20ème Siècle
4. Une usine faisant partie d'un Combinat dans le système de l'économie planifiée mise en place en URSS au 20ème Siècle
5. Une communauté agricole en Israël au XXème Siècle
6. Aucune

Q18. L'Hégire est :

1. Un phénomène météorologique qui permit au prophète Mohamed de conquérir l'Arabie
2. En 622, la fuite de Mahomet de Médine, début du calendrier islamique
3. Evénement historique qui marque la fuite de Mahomet de la Mecque au 5ème Siècle et qui date la création de l'Islam
4. Un phénomène historique qui marque, en 742, la naissance de l'Empire musulman
5. Une cérémonie religieuse musulmane célébrée lors des mariages
6. Aucune

Q19. Dans l'expression « Une vie spartiate », le terme spartiate :
1. Désigne la façon de vivre d'un habitant de la ville grecque de Sparte au 20ème Siècle
2. Fait référence aux habitudes des habitants de Sparte, dans la Grèce Antique, qui sacrifiaient confort et richesse… aux devoirs militaires et à la puissance de leur ville
3. Désigne le sort d'un esclave romain qui avait suivi Spartacus lors de la révolte des esclaves qui embrasa la Péninsule Italienne au 1er siècle avant Jésus-Christ
4. Etait le nom d'une exigente combinaison (triple) d'épreuves des Jeux Olympiques de l'Antiquité, qui réunissait le marathon, le lancement du javelot et la lutte
5. Désigne une vie où tous les habitants d'une même ville sont tous traités égalitairement comme à Sparte
6. Aucune

Médias

Q20. Quelle chaîne de TV crée et diffuse la séquence « No Comment » ?
1. La BBC
2. FR3
3. CNN
4. RTL
5. La RAI
6. Aucune

Q21. Quel journal appartient au groupe de presse Rossel ?
1. Le Monde
2. Le New York Times
3. Le matin
4. Le Soir
5. La Libre Belgique
6. Aucune

Economie

Q22. Quel était, approximativement, le taux d'inflation annuel de la Belgique en 1998 ?
1. 1%
2. 5%
3. 10%
4. 15%
5. 20%
6. Aucune

Q23. Combien a-t-on de FB (environ) pour un Euro ?
1. Dix
2. Vingt
3. Trente
4. Quarante
5. Cinquante
6. Aucune

Q24. Qui délivre le minimex ?
1. Le FOREM
2. L'ONEM
3. Le CPAS
4. La mutuelle
5. Le syndicat
6. Aucune

Q25. Le taux de chômage était, en Wallonie, en 1998, aux alentours de :
1. 5%
2. 10%
3. 15%
4. 20%
5. 25%
6. Aucune

Epreuve de Connaissances Artistiques (ARTACT)

NOM :	PRENOM :	N° Check up : ..... N° (univ./Fac./Sec.) : .....

Connaissances artistiques

Check up (Projet Mohican – CIUF) B. NOEL (FUCAM) D. LECLERCQ (Ulg)[1]

Forme de questionnaire 1

Consignes : Questions à choix multiple avec degrés de certitude
(une feuille de consignes détaillées est disponible en annexe)

- **Répondez à l'aide du formulaire de réponses destiné à la lecture optique. Gardez le présent questionnaire.**

Architecture

Q1. La cathédrale « Notre Dame de Paris » relève de l'architecture

1. baroque
2. romane
3. gothique
4. art nouveau
5. jésuite
6. Aucune
7. Toutes

Q2. Et celle d'Orléans que voici ? (mêmes propositions de solutions)

Q3. Et le Parthénon, sur l'Acropole ? (mêmes propositions de solutions).

Q4. Dans quelle ville est situé le Golden Gate bridge ?

1. New York
2. Los Angeles
3. Londres
4. Pittsburgh
5. Dallas
6. Aucune
7. Toutes

Peinture / sculpture

Q5. Parmi les paires d'artistes constituées ci-après, laquelle (paire) peut-on associer à l'impressionnisme en peinture ?

1. Rembrandt – David
2. Monet – Sisley
3. Magritte – Alechinsky
4. Bruegel – Bosch
5. Picasso – Permeke
6. Aucune
7. Toutes

Q6. Laquelle de ces 5 mêmes paires regroupe des peintres belges qui ont vécu au XX° siècle ?

[1] Avec la collaboration de A. LAMBLIN

Q7. La parabole des aveugles (ci-dessous) relève de la peinture

1. cubiste
2. roman
3. gothique
4. baroque
5. surréaliste
6. Aucun
7. Tous

Q8. Et cette sculpture de Jacques Lipchitz ?
(mêmes propositions de solutions que Q7)

Cinéma

Q9. Quelle est l'œuvre cinématographique qui a été inspirée par le drame du SIDA ?

1. Philadelphia
2. La ligne rouge
3. Love Story
4. Basic instinct
5. Rain Man
6. Aucune
7. Toutes

Q10. Et laquelle de ces 5 mêmes œuvres a été inspirée par le problème de l'autisme ?

Q11. Quel film a été réalisé par Steven Spielberg ?

1. Il faut sauver le soldat Ryan
2. Jurassic Park
3. La couleur pourpre
4. Rencontre du troisième type
5. La liste de Schindler
6. Aucun
7. Tous

Q12. Et laquelle de ces 5 mêmes œuvres a été inspirée par la Shoah (extermination du peuple juif) ?

Q13. Quel film a été réalisé par Stanley Kubrik ?

1. Orange mécanique
2. 2001 Odyssée de l'espace
3. Les sentiers de la gloire
4. Full Metal Jacket
5. Barry Lindon
6. Aucun
7. Tous

Q14. Lequel de ces mêmes 5 films traite de la « décimation» ?

Théâtre

Q15. Quel spectacle a été conçu par Bertold Brecht ?

1. L'opéra de quatre sous
2. La résistible ascension d'Arturo Ui
3. Mère courage
4. Le Brave soldat Schweick
5. Le cercle de craie caucasien
6. Aucun
7. Tous

Q16. Voici un extrait de pièce de théâtre :

Le docteur : Mais vous n'avez jamais exercé.
Knock : Autre erreur.
Le docteur : Comment ? Ne m'avez-vous pas dit que vous veniez de passer votre thèse l'été dernier ?
Knock : Oui, trente-deux pages in-octavo : *Sur les prétendus états de santé*, avec cette épigraphe, que j'ai attribuée à Claude Bernard : « les gens bien portants sont des malades qui s'ignorent. »

Qui a écrit cette pièce ?

1. Molière
2. Racine
3. Corneille
4. Marivaux
5. Rostand
6. Aucun

Littérature

Q17. Voici une strophe du poème « Napoléon II » :

Oui, l'aigle, un soir, planait aux voûtes éternelles,
Lorsqu'un grand coup de vent lui cassa les deux ailes ;
Sa chute fit dans l'air un foudroyant sillon ;
Tous alors sur son nid fondirent pleins de joie ;
Chacun selon ses dents se partagea la proie ;
L'angleterre prit l'aigle, et l'Autriche l'aiglon.

Quel est l'auteur de ce poème ?

1. Rimbaud
2. Hugo
3. Baudelaire
4. Hemingway
5. Balzac
6. Aucun
7. Tous

Q18 Voici le début du poème « L'invitation au voyage » :

Mon enfant, ma sœur,
Songe à la douceur
D'aller là-bas vivre ensemble !
Aimer à loisir,
Aimer et mourir
Au pays qui te ressemble !

Quel est l'auteur de ce poème ?
(mêmes propositions de solutions que Q17)

Q19. Quel ouvrage Kant a-t-il écrit ?
1. Le discours de la méthode
2. Mein Kampf
3. La critique de la raison pure
4. Le capital
5. Le meilleur des mondes
6. Aucun
7. Tous

Q20. Lequel a été écrit par Aldous Huxley ?
(mêmes propositions de solutions que Q19)

Q21. Lequel est antisémite ?
(mêmes propositions de solutions que Q19)

Musique et danse

Q22. La chanson « We are the champions, my friend » (souvent entonnée sur les stades) a été lancée par quel interprète et quel groupe ?
1. Lou et les Hollywood Bananas
2. Bono et U2
3. Mick Jagger et les Rolling Stones
4. Paul Mc Cartney et les Beatles
5. Freddy Mercury et le groupe Queen
6. Aucun
7. Tous

Q23. L'hymne « européen » aussi appelé l'Hymne à la joie, est l'œuvre de
1. Mozart
2. Berlioz
3. Debussy
4. Bach
Dvorak
6. Aucun
7. Tous

Q24. Dans le film « Les uns et les autres », le danseur, Jorge Donn (voir photo ci-dessous à côté de Tania Bari), qui y interprète le « Boléro » a fait partie
1. Du ballet du XX° siècle
2. De la troupe de danse de Serge Lifar
3. De la troupe de danse de Rudolf Noureev
4. De la troupe de danse de Balanchine
5. De la troupe de danse de Patrick Dupont
6. Aucun
7. Tous

Q25. De quel musicien est cet air « Le boléro » (dansé dans le film « Les Uns et les Autres ») par le danseur Jorge Donn?
1. Stravinsky
2. Mozart
3. Rodrigo
4. Paco de Lucia
5. Tchaikowski
6. Aucun

Epreuve de Mathématiques (MATHEM)

NOM :	PRENOM :	N° Check up : …… N° (univ./Fac./Sec.) :……

Mathématique

Check up (Projet Mohican – CIUF) Auteur : M. LEBRUN & J. LEGA – UCL

Forme de questionnaire 1

Consignes : Questions à choix multiple avec degrés de certitude
(une feuille de consignes détaillées est disponible en annexe)

- **Répondez à l'aide du formulaire de réponses destiné à la lecture optique. Gardez le présent questionnaire.**

CE. Transcrire mathématiquement un énoncé

Pierre a actuellement 6 ans de plus que Paul. Il y a 10 ans, Pierre avait un âge double de celui de Paul.
En supposant que Y représente l'âge de Pierre et X l'âge de Paul.

Q1. Choisissez les équations représentatives du problème.

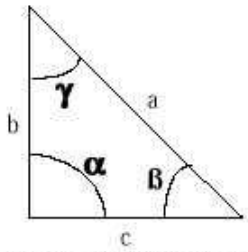
1. Y = X+6 et 2Y = X
2. Y = X - 6 et 2Y – 10 = X – 10
3. Y = X+6 et Y = 2X – 10
4. Y = X + 6 et (Y – 10) = 2(X - 10)
5. Y = X – 6 et (X – 10) = 2(Y - 10)
6. Aucune
7. Toutes

Q2. Donnez les solutions de ce problème.

1. Y = 12 ans et X = 6 ans
2. Y = 22 ans et X = 16 ans
3. Y = 16 ans et X = 22 ans
4. Y = 18 ans et X = 12 ans
5. Y = 26 ans et X = 20 ans
6. Aucune
7. Toutes

GE. Déduire une formule mathématique à partir d'un graphique

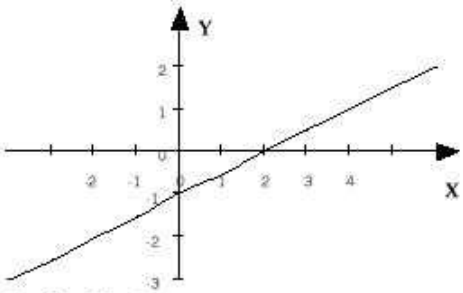
La figure ci-dessous représente un triangle dont l'angle α vaut 90°.



Q3. Exprimez la valeur du côté c à l'aide des variables indiquées sur le dessin.

1. $c = \sqrt{a^2 + b^2}$
2. $c = \sqrt{a^2 - b^2}$
3. $c = a^2 + b^2$
4. $c = \sqrt{b^2 - a^2}$
5. $c = a^2 - b^2$
6. Aucune
7. Toutes

Q4. Ecrivez l'équation de la droite représentée sur le graphique ci-dessous :



1. Y = 2X – 1
2. Y = 2X – 2
3. Y = (X / 2) –2
4. Y = X – 1
5. Y = (X / 2) – 1
6. Aucune

EC. Préciser le sens d'une formule

Q5. A l'aide des symboles de la figure ci-dessus , précisez la signification concrète de la formule de tan ß. Elle représente :

1. La pente du segment de droite a
2. La longueur du coté b en face de l'angle ß
3. La longueur du coté a qui touche l'angle ß
4. La valeur de l'angle ß en degrés
5. La grandeur de l'hypothénuse
6. Aucune
7. Toutes

Soit la relation $Y = \log(X)$

Q6. De combien doit évoluer X pour que Y augmente d'une unité ?

1. Il doit être multiplié par 2
2. Il doit être multiplié par 10
3. Il doit être diminué de 10
4. Il doit être divisé par 10
5. Il doit être augmenté de 10
6. Aucune
7. Toutes

EE. Déduire une relation d'autres relations

Un satellite de masse m tourne autour de la terre de masse M à une distance r du centre de celle-ci et avec une vitesse v. En admettant :
- l'hypothèse du mouvement circulaire, on trouve que la force centripète est :

$$F = m\frac{v^2}{r}$$

- la loi de gravitation universelle, on trouve que cette force est :

$$F = \frac{GMm}{r^2}$$

Etablissez en fonction de M, m, r et G :

Q7. la formule de la vitesse (v) du satellite

1. $V = \sqrt{\frac{GM}{r^2}}$
2. $V = \sqrt{\frac{GM}{r}}$
3. $V = \sqrt{\frac{F}{GM}}$
4. $V = \sqrt{\frac{Fr}{m}}$
5. $V = \sqrt{\frac{GMm}{r}}$
6. Aucune
7. Toutes

Q8. Calculez la fonction suivante :

$$\int x dx =$$

1. $\ln(x) + C$
2. $2x + \frac{2}{X}$
3. $x^2 + C$
4. $\frac{X^2}{2} + C$
5. $2x + C$
6. Aucune
7. Toutes

NC. Décrire le comportement d'une suite de nombres

Dans le tableau ci-dessous, on présente l'évolution d'un ensemble de données Y en fonction d'autres données X :

X :	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Y :	**2**	**4**	**8**	**16**	**32**	**64**	**128**

Q9. A quel type de série numérique, la progression des nombres de la série Y appartient-elle ?

1. une série alternée
2. une série logarithmique
3. une série géométrique
4. une série arithmétique
5. une série cubique
6. Aucune
7. Toutes

NE. Transformer un tableau de données en une équation

Dans le tableau ci-dessous, on présente l'évolution d'un ensemble de données Y en fonction d'autres données X :

X :	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Y :	**2**	**4**	**8**	**16**	**32**	**64**	**128**

Q10. Laquelle des fonction suivantes correspond le mieux à la relation entre les données X et Y :

1. $Y = X^2$
2. $X = 2Y$
3. $Y = \log(X)$
4. $Y = 2X$
5. $Y = 2^X$
6. Aucune
7. Toutes

NN. Réduire ou développer une relation donnée :

Q11. Réduisez ou développez les expressions suivantes :

$$(X^5)^3$$

1. X^8
2. X^{15}
3. $3 \log(X^5)$
4. $\log(X^5) + \log(X^3)$
5. $2X^4$
6. Aucune
7. Toutes

Q12. Que vaut 1 M/S :

1. 3,6 Km/H
2. 1 (Km/H) / 3,6
3. 10 Km/H
4. 0,36 Km/H
5. 1 (Km/H) / 0,36
6. Aucune
7. Toutes

CC. Définir un concept

Q13. Parmi les affirmations ci-dessous, indique celle qui définit le mieux la notion de dérivée ?

1. La limite de y(x) quand x tend vers 0
2. La dérivée représente la pente de la droite tangente à la fonction en l'origine
3. La dérivée donne la pente de la fonction limite pour tout point de la fonction initiale
4. La dérivée permet de calculer l'aire d'une surface délimitée par la fonction
5. C'est la fonction qui donne la pente de la tangente à la fonction en tout point
6. Aucune
7. Toutes

Q14. Parmi les affirmations ci-dessous, indique celle qui définit le mieux la notion d'intégrale définie ?

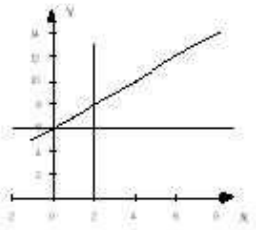
1. La surface totale comprise sous la courbe
2. L'intégrale définie représente la pente de la droite tangente à la fonction en l'origine
3. L'intégrale définie permet de calculer l'aire de la surface entre deux abscisses, l'axe des X et la fonction
4. L'intégrale est la fonction inverse de la dérivée
5. C'est la somme des surfaces élémentaires définies par la fonction d'origine
6. Aucune
7. Toutes

NG. Se représenter l'objet décrit par l'équation

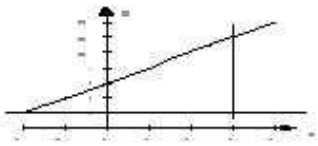
Q15. Identifiez le graphique qui représente les trois fonctions suivantes:

$Y = X + 6$
$Y = 6$
$X = 2$

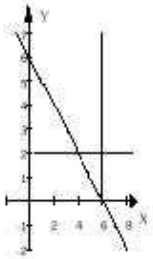
1.



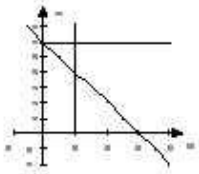
2.



3.



4.



6. Aucune
7. Toutes

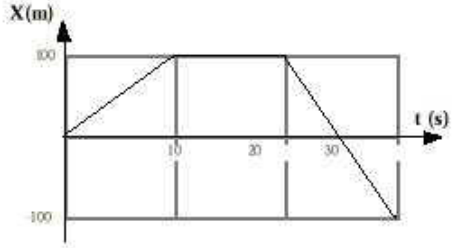
GN. Interpréter un graphique en nombres

Q16. Quelle est la valeur de l'aire de la surface délimitée par les trois droites données ci-dessus (en unité de surface)

1. 2
2. 4
3. 8
4. 16
5. 50
6. Aucune
7. Toutes

GN. Interpréter un graphique en nombres

La position X (exprimée en mètres) d'un mobile est donnée en fonction du temps t (exprimé en secondes) sur le graphique suivant :



Q17. Que vaut la vitesse du mobile au temps t = 25 s ?

1. -20 m/s
2. 0 m/s
3. 5 m/s
4. 10 m/s
5. 20 m/s
6. Aucune
7. Toutes

EN. Analyser numériquement une fonction

Q18 Le tableau ci-dessous fournit quelques points des deux paraboles

$Y_1 = -X^2 + 3X - 12$ **et** $Y_2 = X^2 + X - 36$

X	-4	-3	-2	-1	0	1	2	3	4	5
Y1	-40	-30	-22	-16	-12	-10	-10	-12	-16	-22
Y2	-24	-30	-34	-36	-36	-34	-30	-24	-16	-6

Identifiez la ou les valeur(s) de X correspondant à d'éventuels points d'intersection des deux courbes.

1. X = -3
2. X = -3 et X = 2
3. X = 2 et X = 4
4. X = 2
5. X = -3 et X = 4
6. Aucune
7. Toutes

Q19. Trouvez directement une solution du système

$$\begin{cases} Y = X^2 - X - 6 \\ Y = X - 3 \end{cases}$$

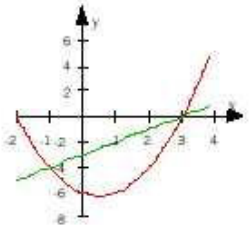
1. X = 0 et Y = 3
2. X = 2,2 et Y = -0,8
3. X = 0 et Y = -3
4. X = 1,8 et Y = 1,2
5. X = -1 et Y = -4
6. Aucune
7. Toutes

EG. Analyser graphiquement une fonction

Q20. Identifiez les graphes des deux fonctions suivantes

$Y_1 = X^2 - X - 6$ et $Y_2 = X - 3$

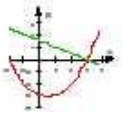
1.



2.

3.

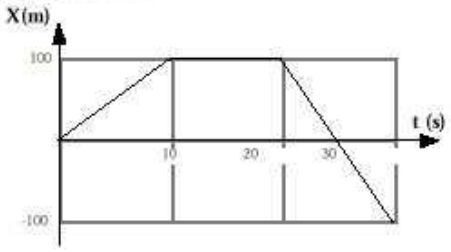
4.



6. Aucune
7. Toutes

GC. Exprimer un grahique en mots

La position X (exprimée en mètres) d'un mobile est donnée en fonction du temps t (exprimé en secondes) sur le graphique suivant :

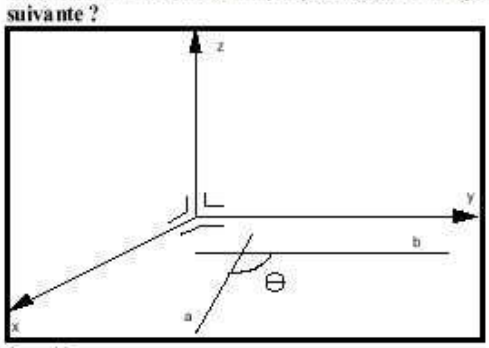


Pour décrire le mieux possible le mouvement de ce mobile, on demande de compléter les phrases ci-dessous.

Q21. Entre 10 et 20 secondes, le mobile …

1. avance le long de l'axe des X positifs
2. avance le long de l'axe des X positifs à vitesse constante
3. avance le long de l'axe des X positifs en accélérant
4. recule le long de l'axe des X positifs à vitesse constante
5. est immobile
6. Aucune
7. Toutes

Q22. Comment peut-on qualifier l'angle θ formé par les droites a et b situées dans le plan (x, y) de la figure suivante ?



1. Aigu
2. Droit
3. Obtu
4. Alterne
5. Interne
6. Aucune
7. Toutes

Epreuve de Biologie (BIOLOG)

NOM :	PRENOM :	N° Check up : …… N° (univ./Fac./Sec.) : ……

Biologie

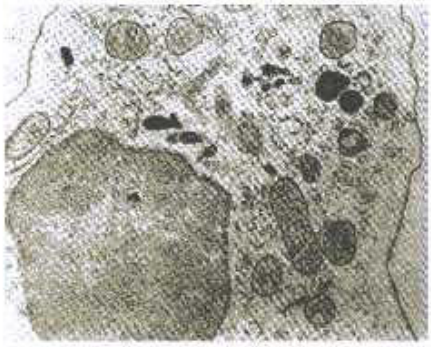
Check up (Projet Mohican – CIUF) Auteur : J.-C. VERHAEGHE - ULB

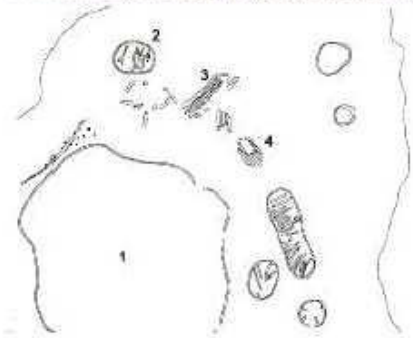
FORME DE QUESTIONNAIRE 1

Consignes : Questions à choix multiple avec degrés de certitude
(une feuille de consignes détaillées est disponible en annexe)

- **Répondez à l'aide du formulaire de réponses destiné à la lecture optique. Gardez le présent questionnaire.**

Cellule





Q1. Sur cette photo et le schéma qui l'accompagne, donne le nom de l'organite indiqués par le numéro 2.

1. Appareil de Golgi
2. Centriole
3. Vacuole
4. Rybosome
5. Mitochondrie
6. Aucune
7. Toutes

Hérédité

Q2. La production d'une molécule d'ARN à partir d'un segment d'ADN est appelée :

1. transcription
2. traduction
3. épissage de l'ARN
4. réplication
5. recombinaison
6. Aucune
7. Toutes

Embryologie / génétique

Q3. Apparemment, nos cellules musculaires diffèrent de nos cellules nerveuses surtout parce qu'elles :

1. expriment des gènes différents
2. contiennent des gènes différents
3. utilisent des codes génétiques différents
4. possèdent des ribosomes qui leur sont propres
5. possèdent des chromosomes différents
6. Aucune
7. Toutes

Reproduction humaine

Q4. La fécondation de l'ovule chez l'Humain se produit le plus souvent dans :

1. le vagin
2. l'ovaire
3. l'utérus
4. la trompe utérine
5. le conduit déférent
6. Aucune
7. Toutes

Ecologie

Q5. La *niche écologique* est

1. la place de la population dans un ensemble de variables physiques, chimiques et climatiques
2. l'ensemble des adaptations morphologiques de la population à son environnement
3. la place de la population dans une communauté, c'est-à-dire l'ensemble des relations biologiques
4. la place de chaque individu au sein de sa population
6. Aucune
7. Toutes

Evolution

Q6. Lorsque l'on verse un antibiotique dans un flacon contenant un très grand nombre de bactéries, cette population peut devenir résistante à cet antibiotique.

Ceci est dû au fait :

1. que l'antibiotique a provoqué des mutations donnant la résistance à cet antibiotique chez toutes les bactéries
2. que l'antibiotique a provoqué des mutations donnant la résistance à cet antibiotique chez certaines bactéries; celles-ci se sont alors multipliées aux dépens des bactéries normales après l'ajout de l'antibiotique
3. que des bactéries résistantes aux antibiotiques existaient avant que l'on verse cet antibiotique et se sont multipliées aux dépens des bactéries normales après cette opération
4. que toutes les bactéries du flacon étaient résistantes aux antibiotiques
6. Aucune
7. Toutes

Biochimie - écologie

Q7. Laquelle de ces formules représente le bilan chimique de la photosynthèse :

1. $C_6H_{12}O_6$ ---> $2\ CH_3\_CH_2OH + 2\ CO_2$
2. $C_6H_{12}O_6$ ---> $2\ CH_3\_CHOH\text{-}COOH$
3. $6\ CO_2 + 6\ H_2O$ ---> $C_6H_{12}O_6 + 6\ O_2$
4. $C_6H_{12}O_6 + 6\ O_2$ ---> $6\ CO_2 + 6\ H_2O$
6. Aucune
7. Toutes

Q8. Laquelle de ces substances est une protéine ?

1. la cellulose
2. l'hémoglobine
3. l'ADN
4. le cholestérol
5. l'adrénaline
6. Aucune
7. Toutes

Q9. Laquelle de ces propositions est exacte :

L'insuline :

1. est une hormone stéroïde
2. est sécrétée par le pancréas
3. fait augmenter la concentration de sucre dans le sang (glycémie)
4. est une maladie grave qui concerne la régulation de la concentration du glucose dans le sang
5. est un enzyme digestif
6. Aucune
7. Toutes

Q10. Actuellement, le meilleur moyen de diminuer les risques d'être infecté par le SIDA par relation sexuelle est :

1. de se faire vacciner contre le SIDA
2. de faire préventivement une cure du médicament AZT
3. d'utiliser un préservatif (condom)
4. de se limiter aux relations hétérosexuelles (homme-femme)
5. de se faire prescrire des antibiotiques la plus vite possible après toute relation suspecte
6. Aucune
7. Toutes

Epreuve de Chimie (CHIMIE)

NOM :	PRENOM :	N° Check up : …… N° (univ./Fac./Sec.) : ……

Chimie

Check up (Projet Mohican – CIUF) Auteur : A. CORNELIS - ULg

FORME DE QUESTIONNAIRE 1

Consignes : Questions à choix multiple avec degrés de certitude
(une feuille de consignes détaillées est disponible en annexe)

- **Répondez à l'aide du formulaire de réponses destiné à la lecture optique. Gardez le présent questionnaire.**

Formules : Compréhension de la notation chimique

Q1. Un composé gazeux a pour formule moléculaire $(CH_3)_2CCH_2$.
Vous en concluez que chaque molécule de ce composé comporte :

1. Cinq atomes d'hydrogène
2. Six atomes d'hydrogène
3. Sept atomes d'hydrogène
4. Huit atomes d'hydrogène
5. Neuf atomes d'hydrogène
6. Aucune
7. Toutes

Stoechiométrie

Q2. Parmi les équations chimiques suivantes, quelle est celle ou le (les) coefficient(s) stœchiométrique x a (ont) pour valeur 3

1. $H_2 + Cl_2 \rightarrow x\ HCl$
2. $N_2 + x\ H_2 \rightarrow 2\ NH_3$
3. $x\ O_3 \rightarrow 3\ O_2$
4. $x\ NO + Cl_2 \rightarrow x\ NOCl$
5. $x\ SO_2 + O_2 \rightarrow x\ SO_3$
6. Aucune
7. Toutes

Loi de Le Châtelier

Q3. La réaction de décomposition de N_2O_4 gazeux en NO_2 gazeux donne lieu à l'établissement d'un équilibre chimique. La réaction est endothermique dans le sens de la décomposition.
Ce caractère endothermique est mis en évidence par le comportement suivant :

1. la proportion de NO_2 dans le mélange à l'équilibre augmente quand la température augmente
2. deux moles de NO_2 sont produites par mole de N_2O_4 consommée
3. la valeur de la constante d'équilibre de cette réaction est toujours positive à toute température
4. NO_2 est plus intensément coloré en brun que N_2O_4
5. la masse molaire de N_2O_4 est double de celle de NO_2
6. Aucune
7. Toutes

Dissociation ionique

Q4. Une solution aqueuse doit contenir les ions ci-dessous aux concentrations indiquées :

Ion	Concentration (mol/L)
Mg^{2+}	0,020
K^+	0,010
Na^+	0,030
Cl^-	0,040
$(SO_4)^{2-}$	0,005
$(NO_3)^-$	0,030

La façon la plus simple de réaliser cette solution est de dissoudre dans l'eau les quantités adéquates de :

1. $Mg(NO_3)_2$, K_2SO_4 et NaCl
2. $MgSO_4$, KCl et $NaNO_3$
3. $Mg(NO_3)_2$, KCl et Na_2SO_4
4. $MgCl_2$, K_2SO_4 et $NaNO_3$
5. $MgCl_2$, KNO_3 et Na_2SO_4
6. Aucune
7. Toutes

Le PH

Q5. On dispose d'un ensemble de solutions aqueuses contenant chacune une et une seule des substances ci-après, toujours à la concentration de 0,010 mol/L :

a) NH_3 (ammoniac)
b) CH_3-CO_2H (acide acétique)
c) HCl (chlorure d'hydrogène)
d) KBr (bromure de potassium)
e) NaOH (hydroxyde de sodium)

En ce qui concerne les valeurs de leur pH, ces solutions se classent dans l'ordre :

1. a < c < e < b < d
2. e < c < a < b < d
3. d < c < a < e < b
4. c < b < d < a < e
5. b < c < a < e < d
6. Aucune
7. Toutes

Mélanges et corps purs

Q6. Une émulsion est une dispersion de particules très fines d'une substance dans une autre, liquide.
Le « Petit Robert », édition 1993, définit la margarine comme une « émulsion de corps gras alimentaires (surtout de graisses végétales) et d'eau ».
De la définition ci-dessus, il ressort que la margarine est

1. un corps pur
2. une solution aqueuse
3. un mélange homogène
4. un corps simple
5. un mélange hétérogène
6. Aucune
7. Toutes

Nombres d'oxydation

Q7. Parmi les composés ci-dessous, quel est celui dont un des éléments constitutifs est dans un état caractérisé par un nombre d'oxydation (+III) ?

1. $MgSO_4$
2. Na_2CO_3
3. $(NH_4)_3PO_4$
4. $Al_2(SO_4)_3$
5. NH_3
6. Aucune
7. Toutes

Masse atomique

Q8. La matière grasse constituant la margarine est le plus souvent obtenue par une modification chimique des huiles végétales, auxquelles on fait subir une réaction appelée « hydrogénation catalytique ». Cette réaction est représentée par l'équation générale :
$R\text{-}CH{=}CH\text{-}R' + H_2 \rightarrow R\text{-}CH_2\text{-}CH_2\text{-}R'$
où R et R' représentent des chaînons et fonctions organiques divers.
De l'équation chimique ci-dessus, on peut conclure que la réaction d'hydrogénation des huiles végétales est un exemple de réaction :

1. d'hydrolyse
2. d'addition
3. de polymérisation
4. d'élimination
5. de substitution
6. Aucune
7. Toutes

Epreuve de Physique (PHYSIQ)

NOM :	PRENOM :	N° Check up : …… N° (univ./Fac./Sec.) : ……..

Physique

Check up (Projet Mohican – CIUF) Auteur : P. CHAPELLE – FPMs

Forme de questionnaire 1

Consignes : Questions à choix multiple avec degrés de certitude
(une feuille de consignes détaillées est disponible en annexe)

- **Répondez à l'aide du formulaire de réponses destiné à la lecture optique. Gardez le présent questionnaire.**

Gravitation

Q 1. La Terre exerce une force d'attraction sur la Lune. Pourquoi, dès lors, la Lune ne tombe-t-elle pas sur la Terre ?

1. parce qu'elle tourne autour de la terre
2. parce qu'elle est trop légère
3. parce que l'attraction solaire compense l'attraction terrestre
4. parce qu'elle n'est pas freinée par l'atmosphère terrestre
5. parce qu'elle est repoussée par la masse de la Terre
6. Aucune
7. Toutes

Accélération

Q2. Un mobile se déplace sur une trajectoire rectiligne. On relève la vitesse v à différents instants t :

t(s) :	0	2	4	6	8	10
V(m/s)	0	5	10	15	20	25

Calculez l'accélération en m/s²

1. – 2,5 m/s²
2. 0 m/s²
3. 2,5 m/s²
4. 5 m/s²
5. 7,5 m/s²
6. Aucune
7. Toutes

Masse

Q3. On considère un dl (décilitre) d'eau dans des conditions normales de température et de pression. Calculez la masse.

1. 100,0 g
2. 100,0 N
3. 10,0 g
4. 1,0 Kg
5. 100,0 Pa
6. Aucune
7. Toutes

Propriétés des corps

Q 4. Quelle est la cause majeure d'éclatement des canalisations lorsque l'eau se congèle ?

1. Le froid fragilise les canalisations
2. En se congelant l'eau se dilate
3. Le frottement de la glace fragilise les canalisations
4. En se congelant, l'eau se contracte
5. Le froid dilate les canalisations
6. Aucune
7. Toutes

Optique

Q5. Un objet est très éloigné d'une lentille mince convergente de 10 cm de distance focale. A quelle distance de la lentille se forme l'image de l'objet ?

1. 0 cm
2. 10 cm
3. 100 cm
4. à l'infini
5. ça dépend de la taille de l'objet
6. Aucun
7. Toutes

Forces

Q6. Une personne se trouve dans un ascenseur qui est accéléré vers le haut.
La force exercée par la personne sur le plancher est :

1. nulle
2. inférieure au poids de la personne
3. égale au poids de la personne
4. supérieure au poids de la personne
5. sans rapport avec le poids de la personne
6. Aucune
7. Toutes

Vitesse

Q7. On élève une masse m à une hauteur h du sol et on l'abandonne librement.
Déterminer l'expression littérale de la vitesse v lorsque la masse retombe sur le sol.

1. $v=\frac{1}{2}mv^2$
2. $v=\sqrt{2gh}$
3. $v=mgh$
4. $V=mc^2$
5. $V=\frac{(mh)}{g}$
6. Aucune
7. Toutes

Electricité

Q8. Une résistance R de 10 ohms est reliée à une batterie de 12 volts.
Calculer le courant dans la résistance.

1. 1,2 A
2. 1,2 w
3. 12 A
4. 1,2 W
5. 1,2 C
6. Aucune
7. Toutes

Q9. Lors d'un orage, pourquoi percevons-nous le bruit du tonnerre après avoir vu l'éclair ?

1. notre oreille est moins sensible que notre œil
2. la lumière de l'éclair se propage plus vite que le tonnerre
3. le bruit (tonnerre) se produit toujours après l'éclair
4. la pluie ralentit davantage le son que la lumière
5. la pluie accélère la transmission lumineuse
6. Aucune
7. Toutes

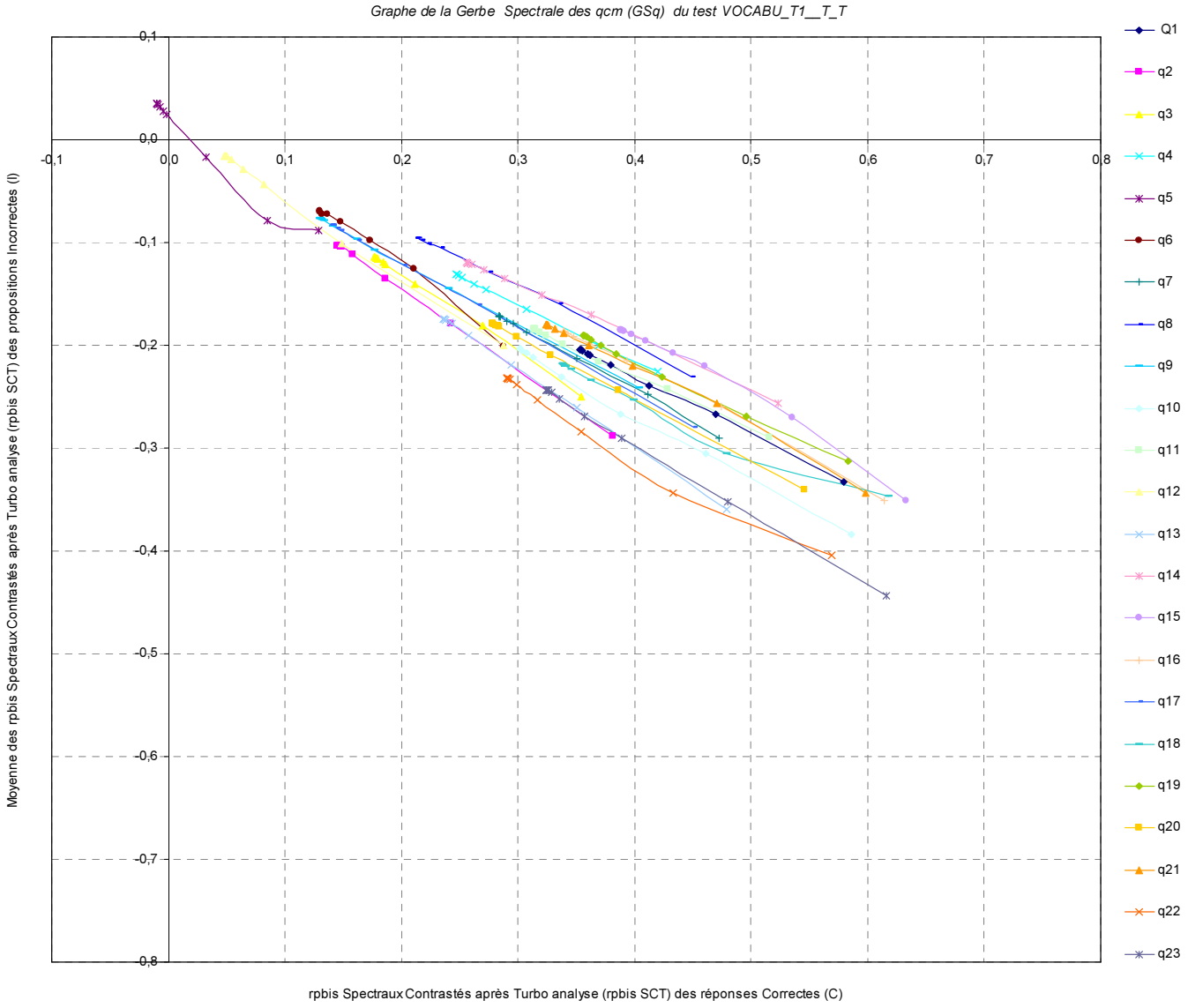
Q10. Pourquoi une bobine parcourue par un courant attire-t-elle une aiguille aimantée ?

1. la bobine crée une champ magnétique
2. la bobine chauffe
3. la bobine crée un appel d'air
4. parce que les corps chauds et les corps froids s'attirent
5. par interaction électrostatique
6. Aucune
7. Toutes

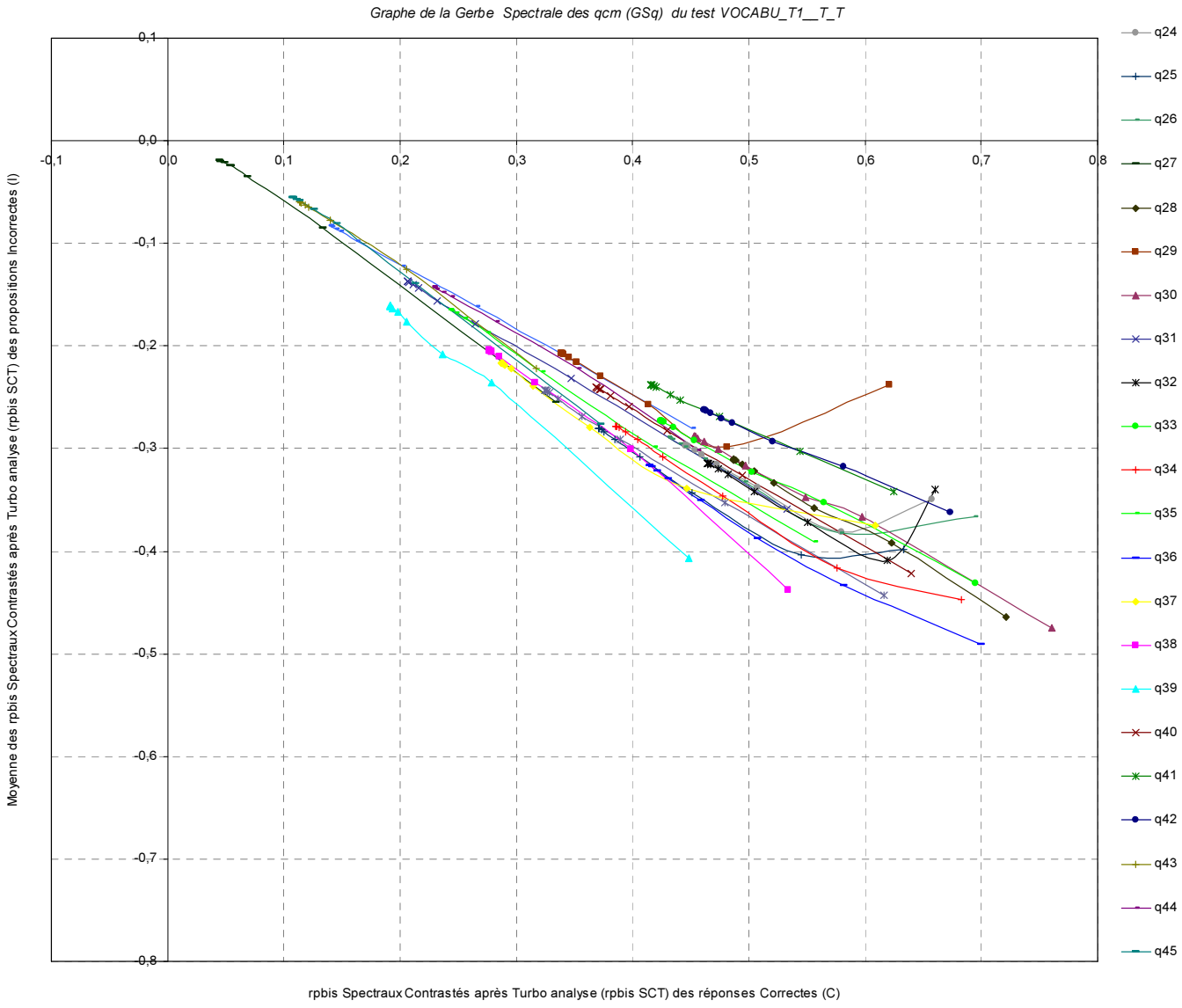
C. Gerbes Spectrales des tests MOHICAN

1. Epreuve de vocabulaire (VOCABU)

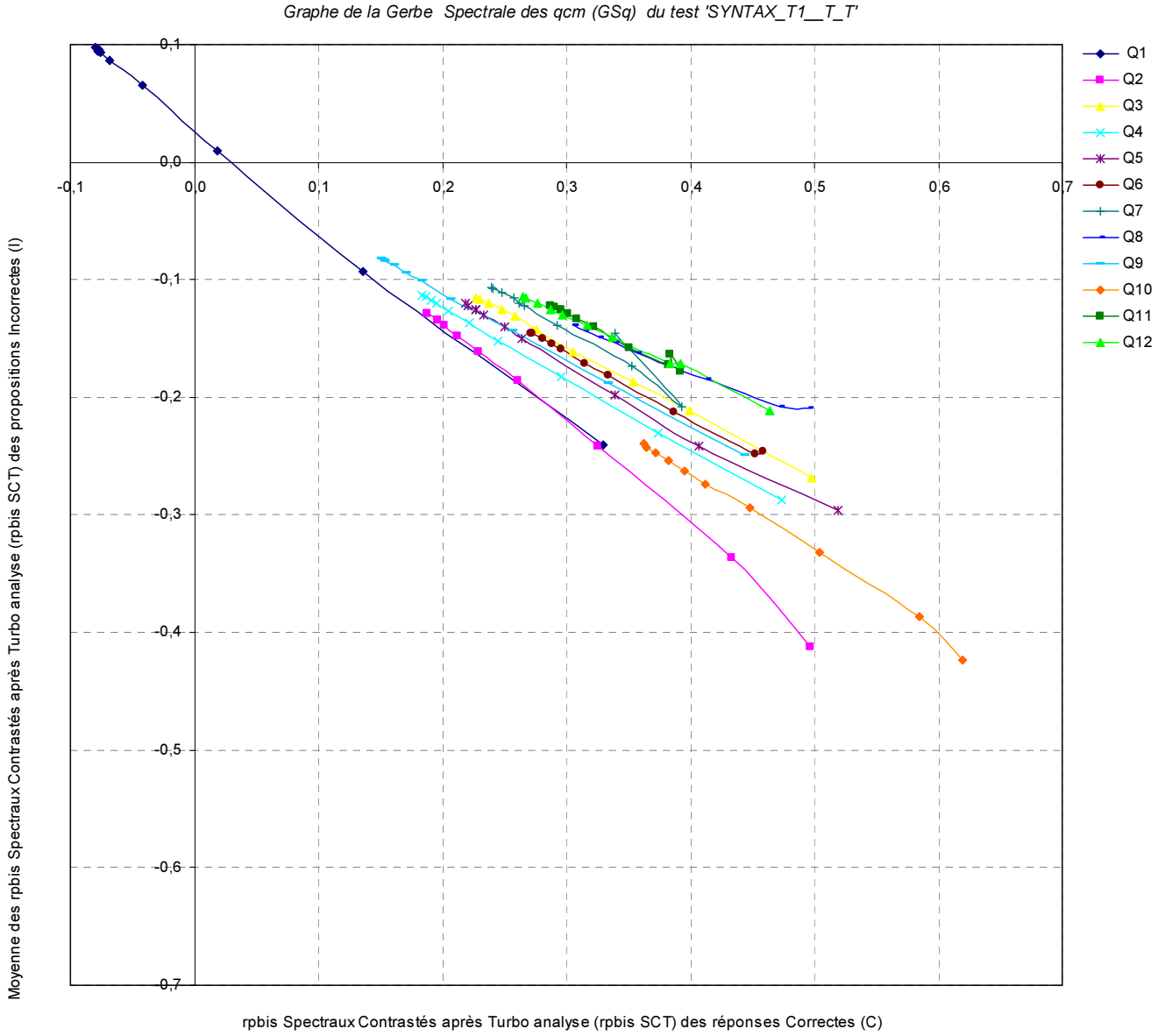
a) Questions 1 à 23



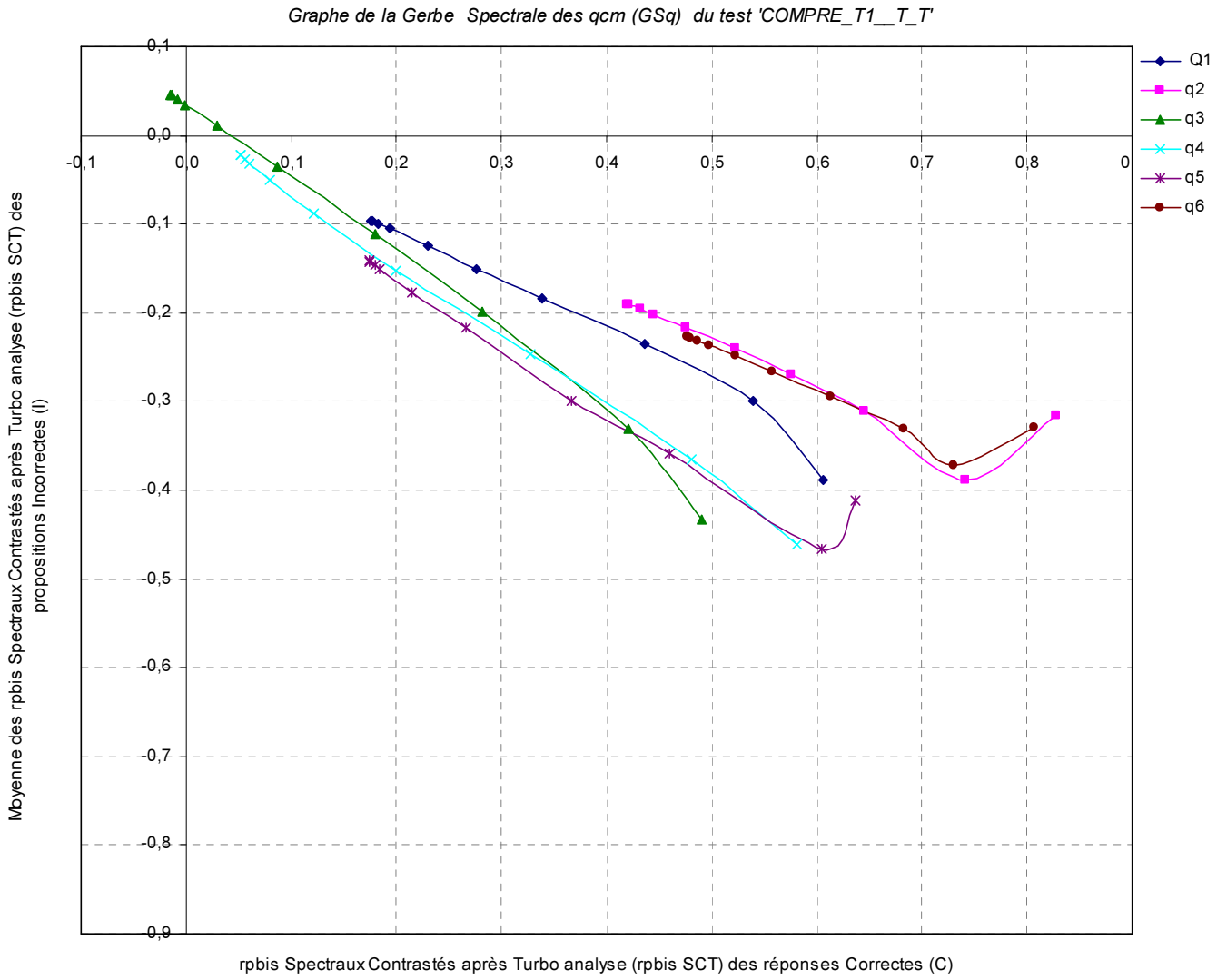
b) Questions 24 à 45



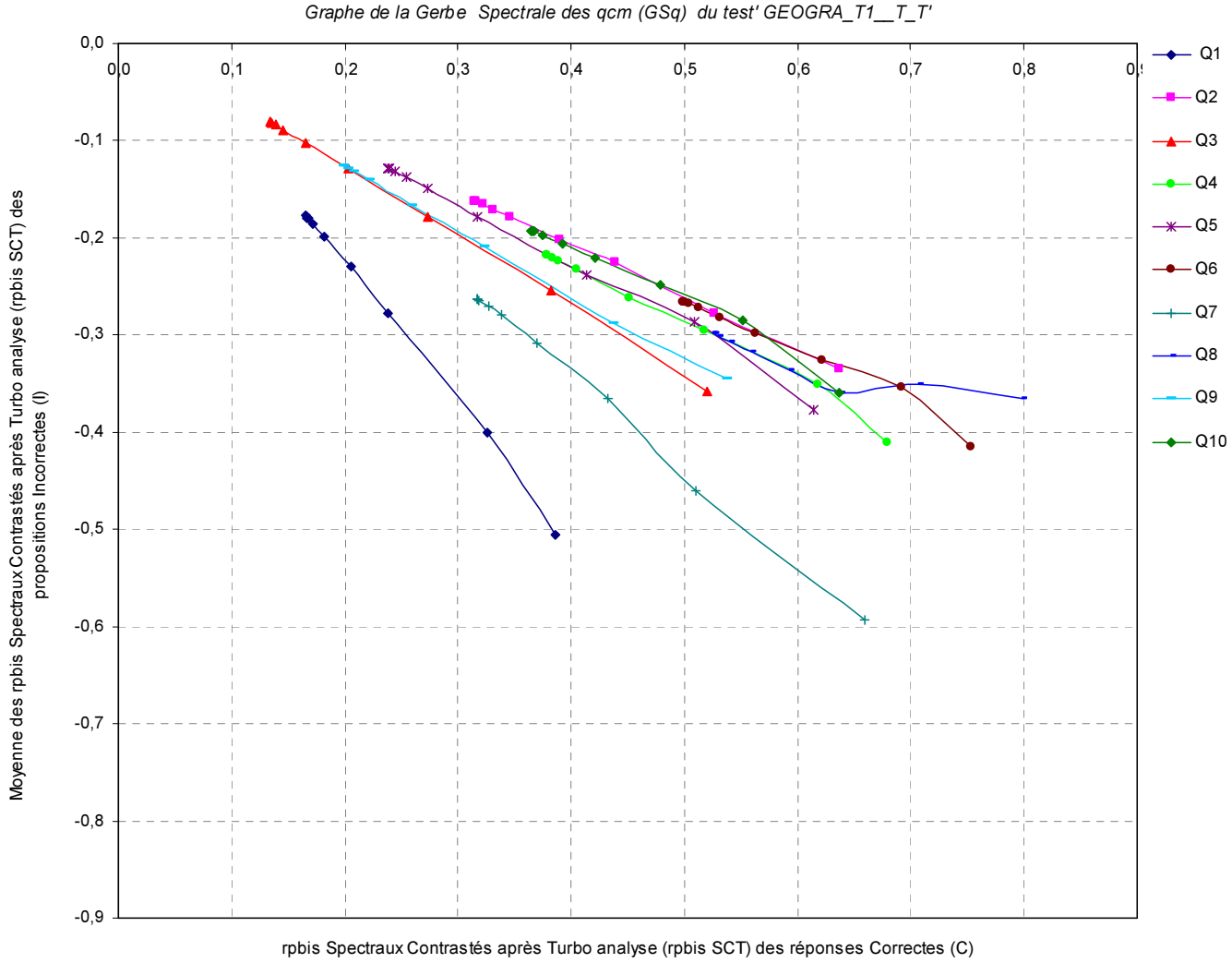
2. Epreuve de Syntaxe et articulation logique (SYNTAX)



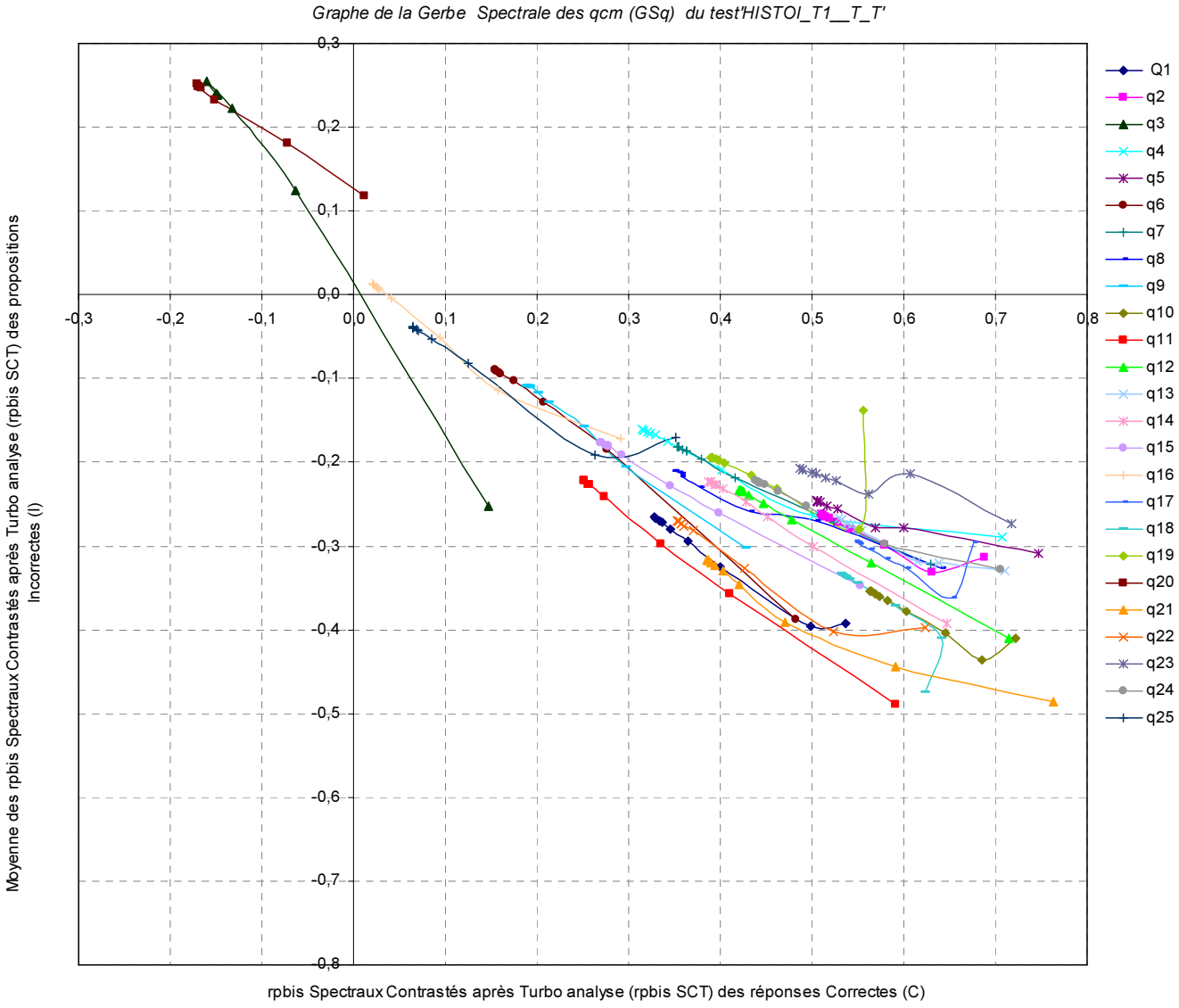
3. Epreuve de Compréhension (COMPRE)



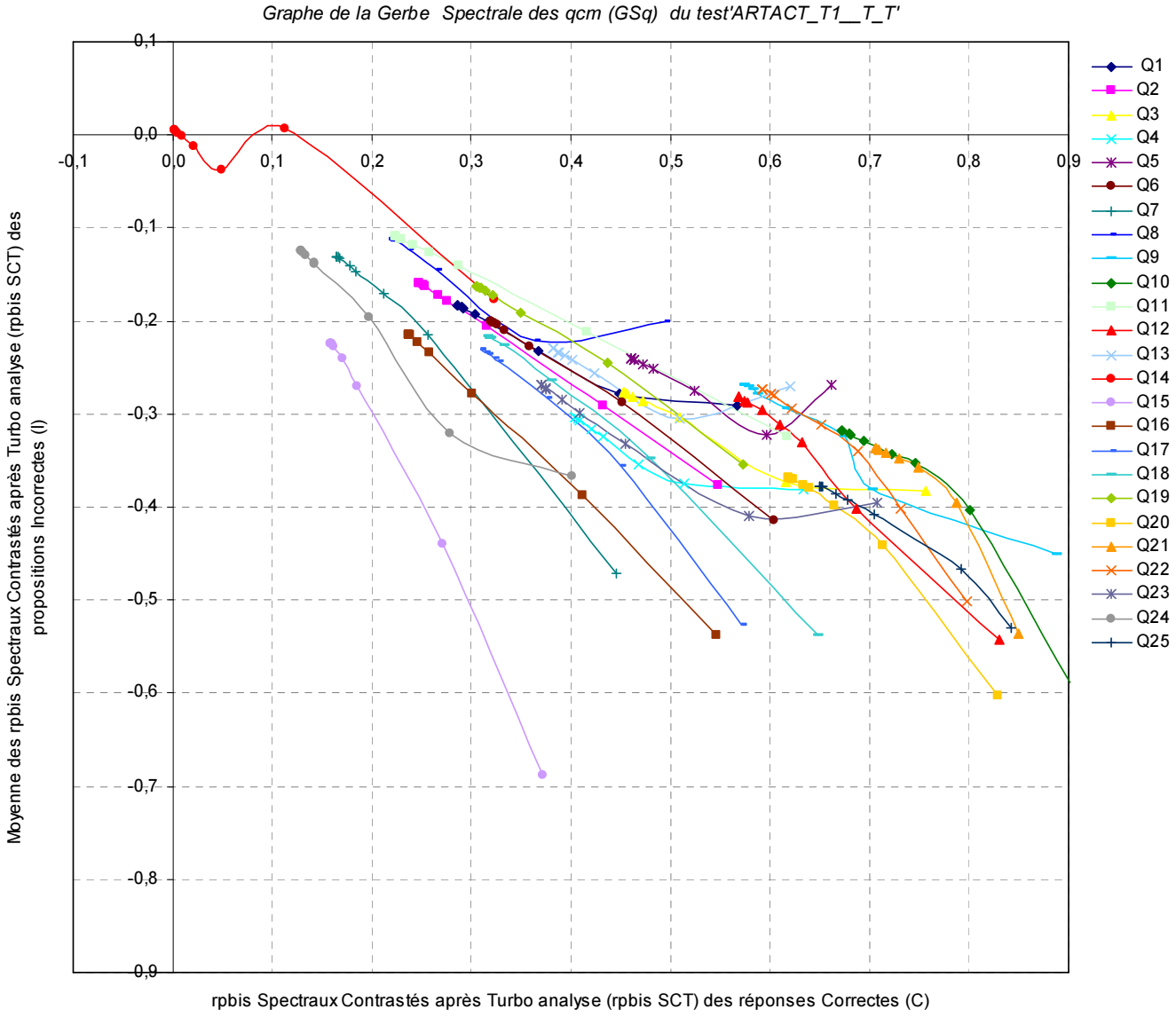
4. Epreuve de Lecture de document et géographie (GEOGRA)



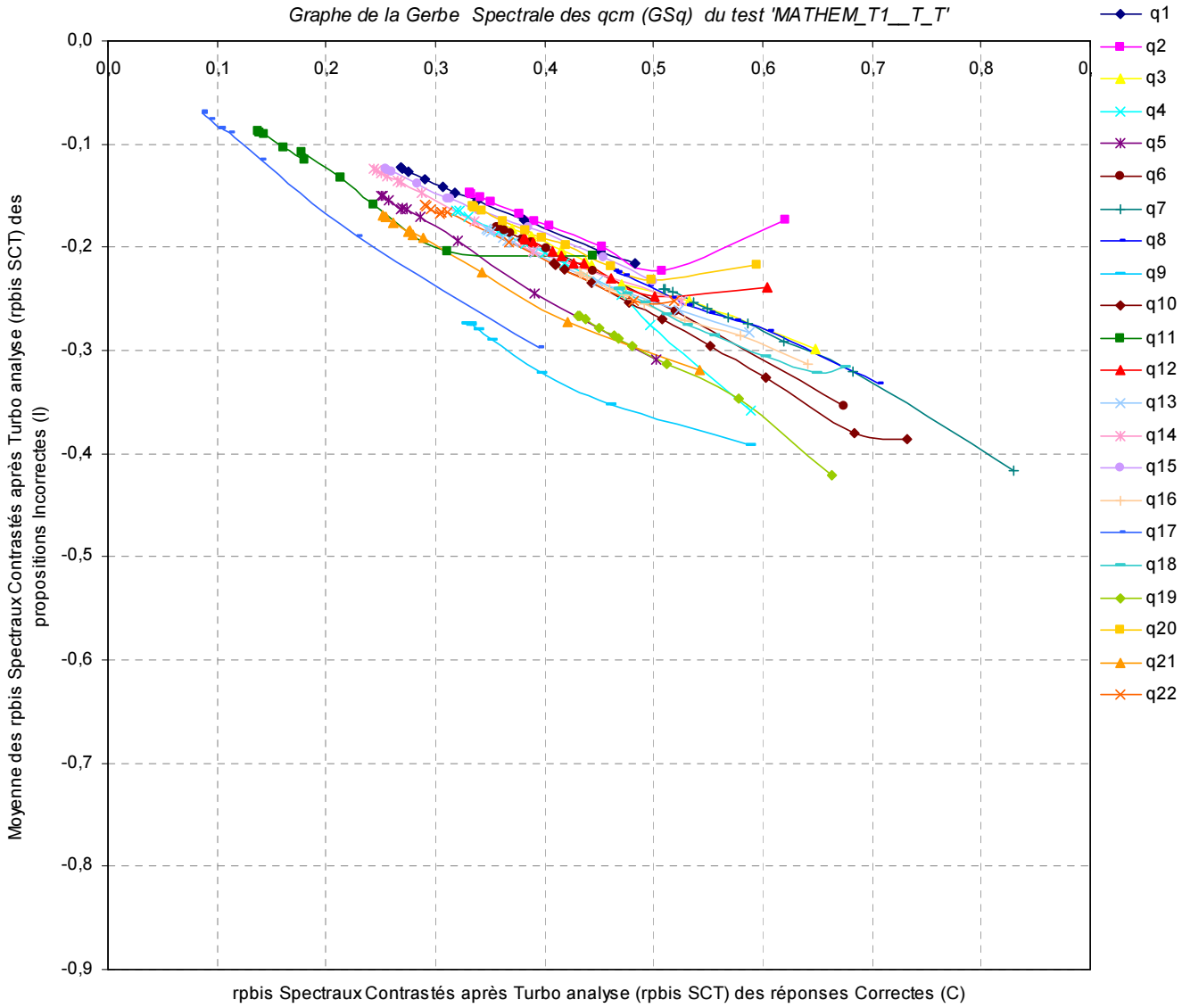
5. Epreuve de Connaissances en Histoire et Socio Economie (HISTOI)



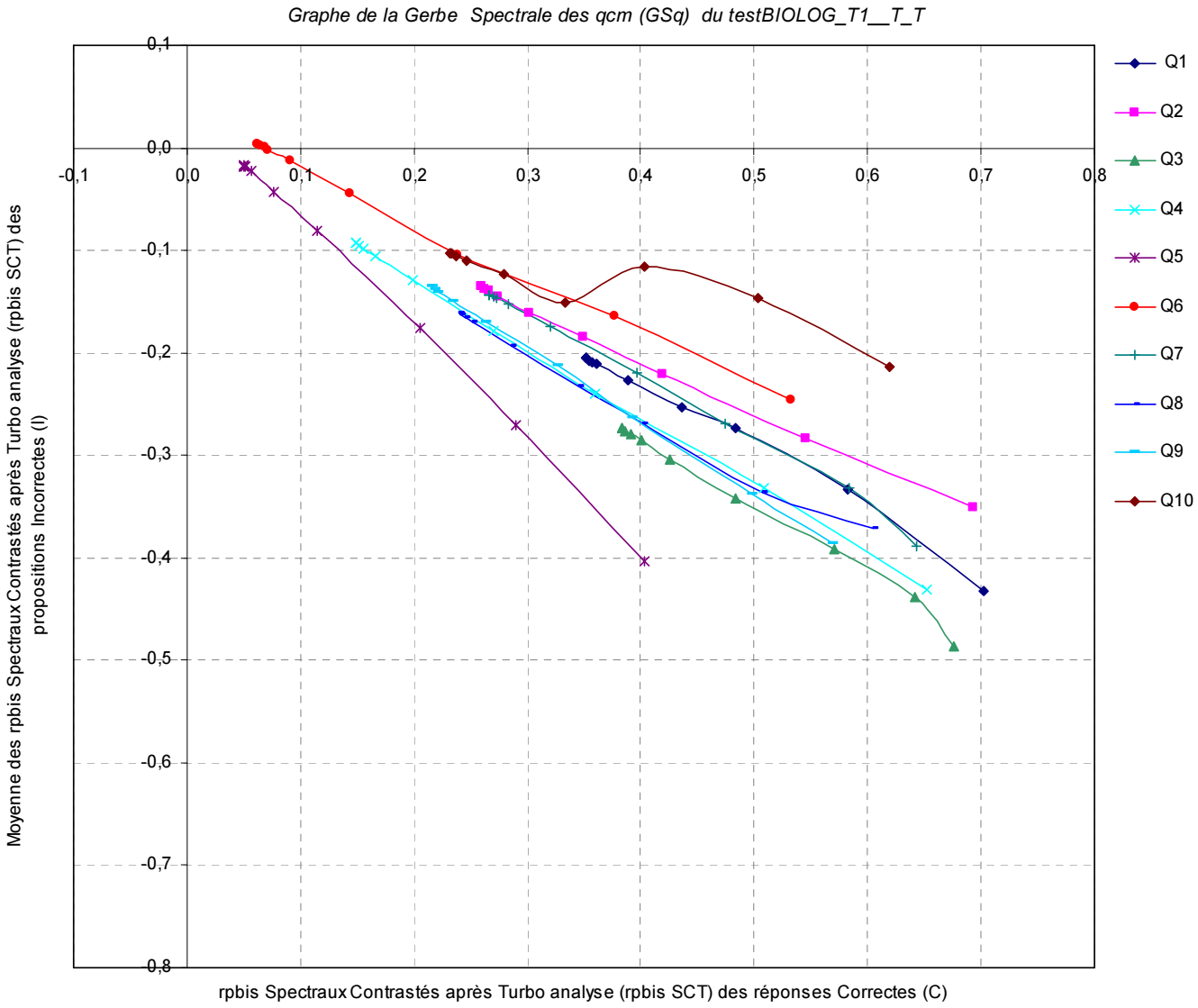
6. Epreuve de Connaissances Artistiques (ARTACT)



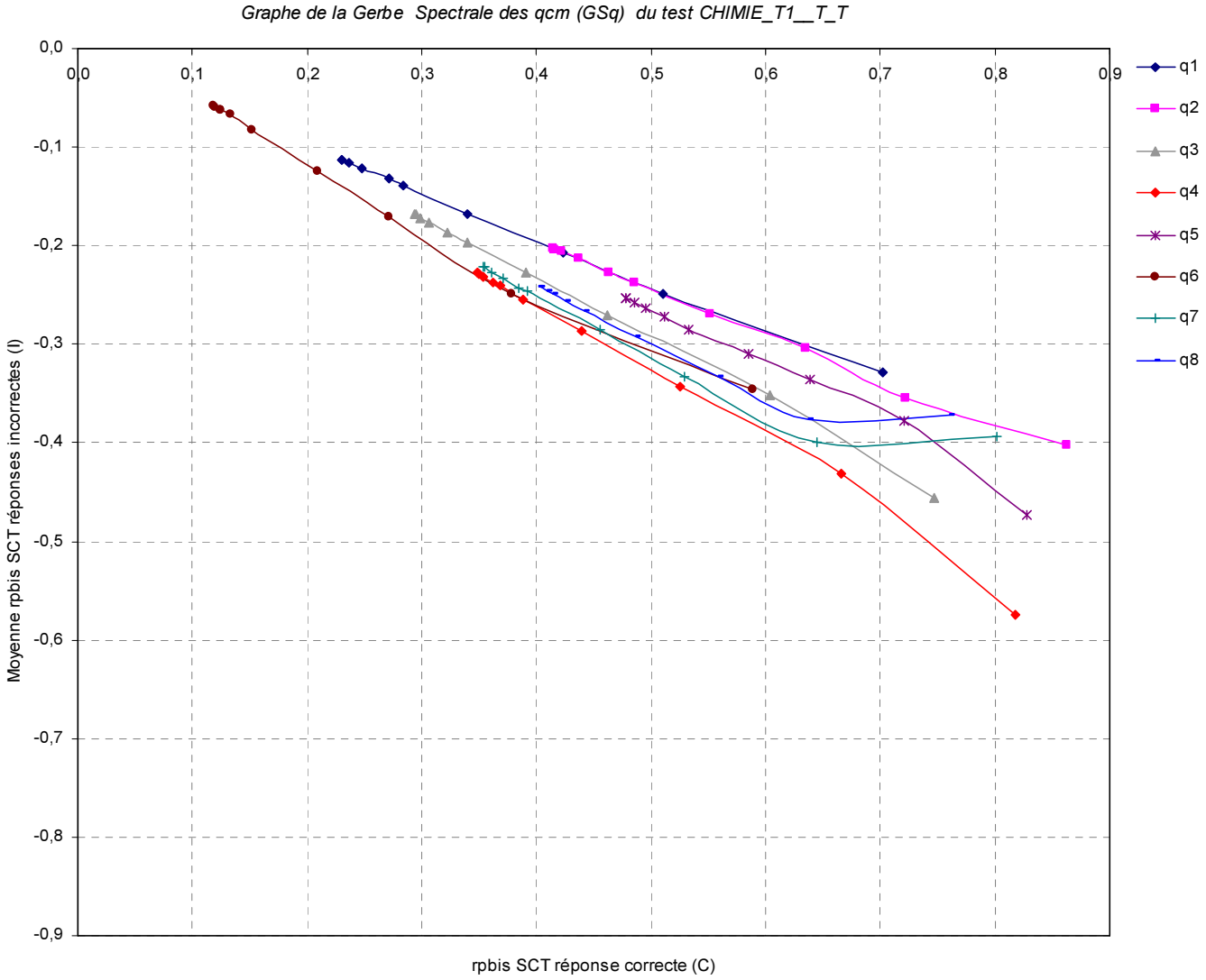
7. Epreuve de Mathématiques (MATHEM)



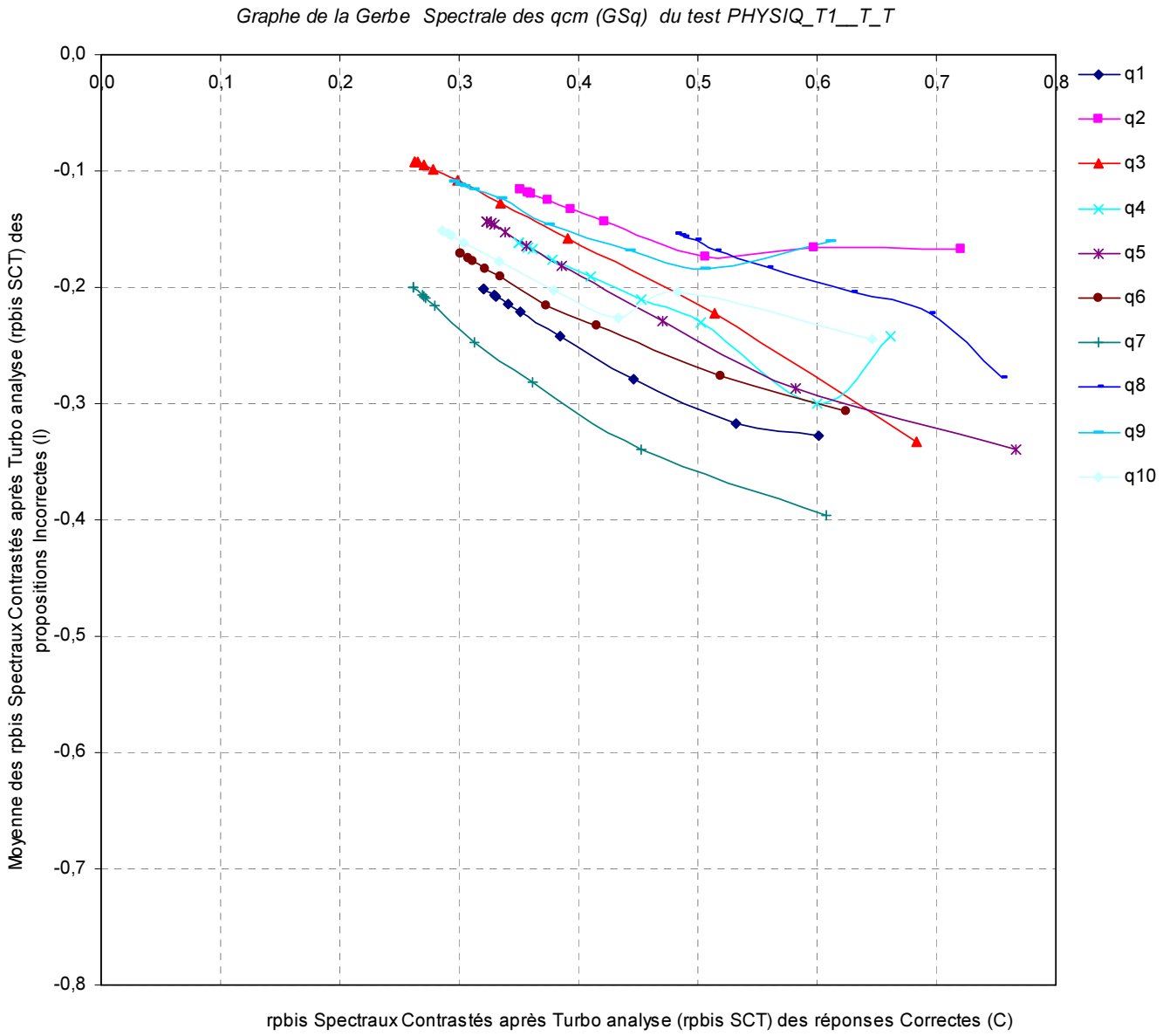
8. Epreuve de Biologie (BIOLOG)



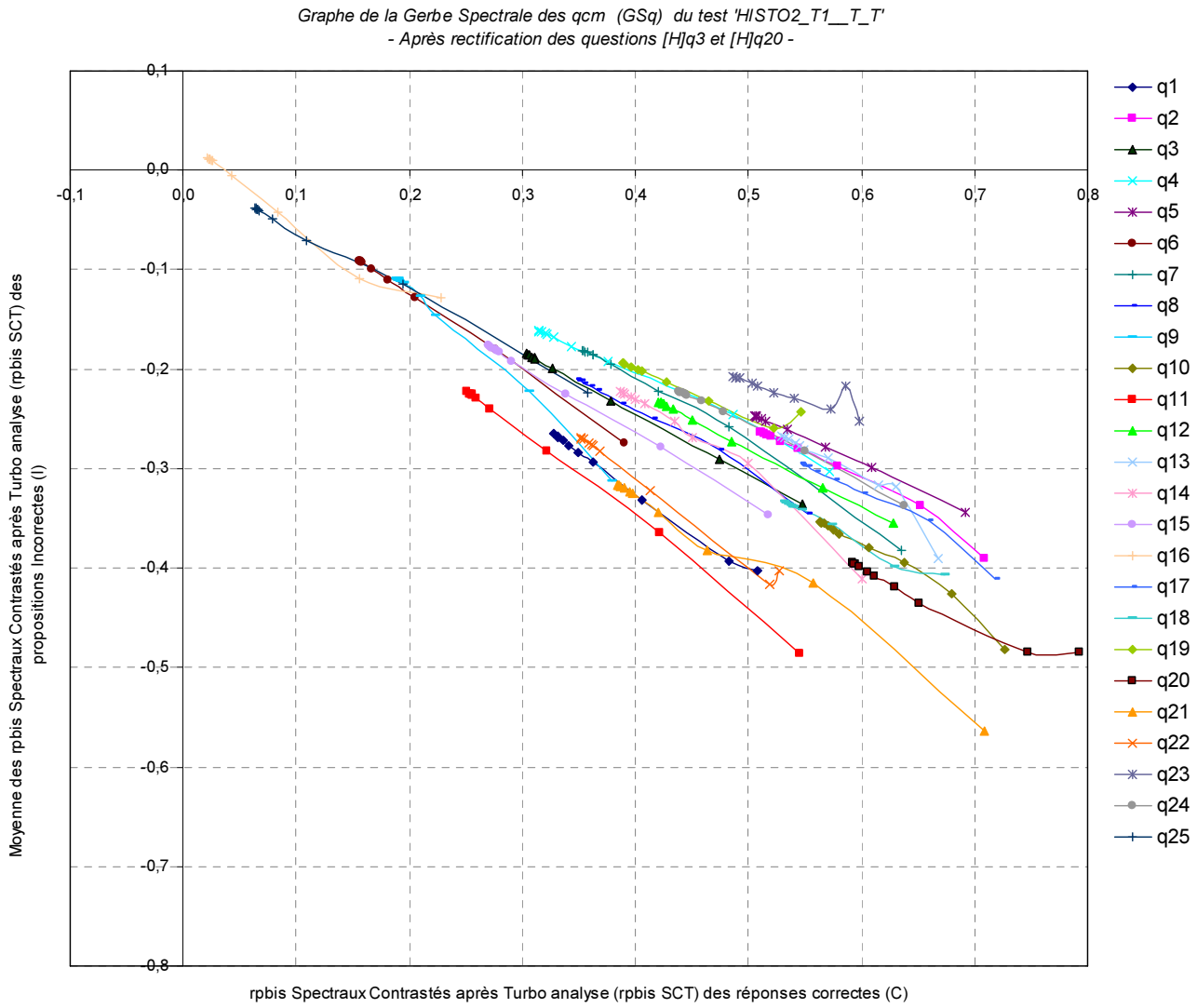
9. Epreuve de Chimie (CHIMIE)



10. Epreuve de Physique (PHYSIQ)



11. Epreuve de Connaissance en Histoire et Socio Economie après rectification des questions [H]q3 et [H]q20 (HISTO2)



D. Protocoles SCANTEST 2.0 d'analyse des propositions

1. Question 36 de l'épreuve de Vocabulaire - [V]q36

a) Statistiques classiques

	OM	**P1**	P2	P3
1) N Rép.	157	**2080**	323	1252
2) % Rép.	4%	**54%**	8%	33%
3) rpbis	-0,27	**0,51**	-0,16	-0,30

b) Statistiques spectrales

	OM	**P1**	P2	P3
1) C Moy.	9%	**68%**	42%	44%
2) rpbis SC	-0,35	**0,41**	-0,22	-0,35

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	**P1**	P2	P3
1.1 N Rép. T10	157	**2080**	323	1252
1.2 % Rép. T10	4%	**54%**	8%	33%
1.3 C. Moy. T10	9%	**68%**	42%	44%
1.4 rpbis SC T10	-0,35	**0,41**	-0,22	-0,35

2. Palier de Turbo analyse : T20

	OM	**P1**	P2	P3
2.1 N Rép. T20	157	**2080**	323	1252
2.2 % Rép. T20	4%	**54%**	8%	33%
2.3 C. Moy. T20	9%	**68%**	42%	44%
2.4 rpbis SC T20	-0,35	**0,41**	-0,22	-0,35

3. Palier de Turbo analyse : T30

	OM	**P1**	P2	P3
3.1 N Rép. T30	157	**2078**	323	1252
3.2 % Rép. T30	4%	**54%**	8%	33%
3.3 C. Moy. T30	9%	**68%**	42%	44%
3.4 rpbis SC T30	-0,35	**0,42**	-0,22	-0,35

4. Palier de Turbo analyse : T40

	OM	**P1**	P2	P3
4.1 N Rép. T40	157	**2073**	321	1250
4.2 % Rép. T40	4%	**54%**	8%	33%
4.3 C. Moy. T40	9%	**69%**	41%	44%
4.4 rpbis SC T40	-0,35	**0,42**	-0,23	-0,35

5. Palier de Turbo analyse : T50

	OM	**P1**	P2	P3
5.1 N Rép. T50	156	**2054**	318	1242
5.2 % Rép. T50	4%	**54%**	8%	33%
5.3 C. Moy. T50	9%	**69%**	41%	44%
5.4 rpbis SC T50	-0,36	**0,43**	-0,23	-0,36

6. Palier de Turbo analyse : T60

	OM	**P1**	P2	P3
6.1 N Rép. T60	152	**2003**	307	1194
6.2 % Rép. T60	4%	**54%**	8%	32%
6.3 C. Moy. T60	7%	**70%**	41%	42%
6.4 rpbis SC T60	-0,37	**0,46**	-0,24	-0,38

7. Palier de Turbo analyse : T70

	OM	**P1**	P2	P3
7.1 N Rép. T70	133	**1789**	252	1033
7.2 % Rép. T70	4%	**56%**	8%	32%
7.3 C. Moy. T70	6%	**71%**	39%	41%
7.4 rpbis SC T70	-0,40	**0,51**	-0,26	-0,42

8. Palier de Turbo analyse : T80

	OM	**P1**	P2	P3
8.1 N Rép. T80	103	**1199**	135	612
8.2 % Rép. T80	5%	**58%**	7%	30%
8.3 C. Moy. T80	3%	**72%**	37%	40%
8.4 rpbis SC T80	-0,48	**0,58**	-0,28	-0,47

9. Palier de Turbo analyse : T90

	OM	**P1**	P2	P3
9.1 N Rép. T90	41	**274**	24	113
9.2 % Rép. T90	9%	**60%**	5%	25%
9.3 C. Moy. T90	2%	**78%**	37%	39%
9.4 rpbis SC T90	-0,67	**0,70**	-0,28	-0,49

2. Question 3 de l'épreuve de Connaissances en Histoire et Socio Eco. - [H]q3

a) Statistiques classiques

	OM	P1	P2	P3	P4	P5	**P6**
1) N Rép.	86	501	132	197	104	258	**128**
2) % Rép.	6%	36%	9%	14%	7%	18%	**9%**
3) rpbis	-0,21	0,36	-0,02	-0,10	-0,06	-0,17	**0,01**

b) Statistiques spectrales

	OM	P1	P2	P3	P4	P5	**P6**
1) C Moy.	4%	60%	30%	31%	30%	42%	**24%**
2) rpbis SC	-0,14	0,49	0,05	0,07	0,04	0,19	**-0,15**

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	P3	P4	P5	**P6**
1.1 N Rép. T10	86	501	132	196	104	258	**128**
1.2 % Rép. T10	6%	36%	9%	14%	7%	18%	**9%**
1.3 C. Moy. T10	4%	60%	30%	31%	30%	42%	**24%**
1.4 rpbis SC T10	-0,14	0,49	0,05	0,07	0,04	0,19	**-0,15**

2. Palier de Turbo analyse : T20

	OM	P1	P2	P3	P4	P5	**P6**
2.1 N Rép. T20	86	501	132	196	104	258	**128**
2.2 % Rép. T20	6%	36%	9%	14%	7%	18%	**9%**
2.3 C. Moy. T20	4%	60%	30%	31%	30%	42%	**24%**
2.4 rpbis SC T20	-0,14	0,49	0,05	0,07	0,04	0,19	**-0,15**

3. Palier de Turbo analyse : T30

	OM	P1	P2	P3	P4	P5	**P6**
3.1 N Rép. T30	86	499	132	196	104	257	**128**
3.2 % Rép. T30	6%	35%	9%	14%	7%	18%	**9%**
3.3 C. Moy. T30	4%	60%	30%	31%	30%	42%	**24%**
3.4 rpbis SC T30	-0,14	0,49	0,05	0,07	0,04	0,19	**-0,15**

4. Palier de Turbo analyse : T40

	OM	P1	P2	P3	P4	P5	**P6**
4.1 N Rép. T40	85	498	132	196	103	257	**128**
4.2 % Rép. T40	6%	35%	9%	14%	7%	18%	**9%**
4.3 C. Moy. T40	3%	60%	30%	31%	30%	42%	**24%**
4.4 rpbis SC T40	-0,14	0,49	0,05	0,07	0,04	0,19	**-0,15**

5. Palier de Turbo analyse : T50

	OM	P1	P2	P3	P4	P5	**P6**
5.1 N Rép. T50	85	495	130	195	103	254	**128**
5.2 % Rép. T50	6%	36%	9%	14%	7%	18%	**9%**
5.3 C. Moy. T50	3%	60%	30%	31%	30%	42%	**24%**
5.4 rpbis SC T50	-0,15	0,49	0,05	0,07	0,04	0,19	**-0,15**

6. Palier de Turbo analyse : T60

	OM	P1	P2	P3	P4	P5	**P6**
6.1 N Rép. T60	84	483	124	186	100	247	**124**
6.2 % Rép. T60	6%	36%	9%	14%	7%	18%	**9%**
6.3 C. Moy. T60	3%	61%	31%	31%	28%	42%	**23%**
6.4 rpbis SC T60	-0,14	0,52	0,06	0,07	0,04	0,20	**-0,16**

7. Palier de Turbo analyse : T70

	OM	P1	P2	P3	P4	P5	**P6**
7.1 N Rép. T70	79	416	114	164	88	209	**103**
7.2 % Rép. T70	7%	35%	10%	14%	7%	18%	**9%**
7.3 C. Moy. T70	2%	61%	30%	31%	28%	40%	**26%**
7.4 rpbis SC T70	-0,17	0,49	0,04	0,05	0,01	0,16	**-0,13**

8. Palier de Turbo analyse : T80

	OM	P1	P2	P3	P4	P5	**P6**
8.1 N Rép. T80	55	218	67	97	52	111	**54**
8.2 % Rép. T80	8%	33%	10%	15%	8%	17%	**8%**
8.3 C. Moy. T80	0%	58%	31%	34%	30%	36%	**33%**
8.4 rpbis SC T80	-0,27	0,38	0,00	0,02	-0,03	0,06	**-0,06**

9. Palier de Turbo analyse : T90

	OM	P1	P2	P3	P4	P5	**P6**
9.1 N Rép. T90	25	36	12	13	14	17	**6**
9.2 % Rép. T90	20%	29%	10%	11%	11%	14%	**5%**
9.3 C. Moy. T90	1%	52%	29%	40%	42%	28%	**53%**
9.4 rpbis SC T90	-0,70	-0,03	-0,18	-0,17	-0,14	-0,27	**0,15**

3. Question 20 de l'épreuve de Connaissances en Histoire et Socio Eco. - [H]q20

a) Statistiques classiques

	OM	P1	P2	**P3**	P4	P5	P6
1) N Rép.	58	263	39	**256**	20	7	744
2) % Rép.	4%	19%	3%	**18%**	1%	0%	53%
3) rpbis	-0,07	-0,18	-0,06	**0,10**	0,00	-0,09	0,14

b) Statistiques spectrales

	OM	P1	P2	**P3**	P4	P5	P6
1) C Moy.	7%	30%	24%	**46%**	38%	37%	82%
2) rpbis SC	-0,20	-0,16	-0,10	**-0,17**	-0,03	-0,02	0,46

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	**P3**	P4	P5	P6
1.1 N Rép. T10	58	263	39	**256**	20	7	743
1.2 % Rép. T10	4%	19%	3%	**18%**	1%	0%	53%
1.3 C. Moy. T10	7%	30%	24%	**46%**	38%	37%	82%
1.4 rpbis SC T10	-0,20	-0,16	-0,10	**-0,17**	-0,03	-0,02	0,46

2. Palier de Turbo analyse : T20

	OM	P1	P2	**P3**	P4	P5	P6
2.1 N Rép. T20	58	263	39	**256**	20	7	743
2.2 % Rép. T20	4%	19%	3%	**18%**	1%	0%	53%
2.3 C. Moy. T20	7%	30%	24%	**46%**	38%	37%	82%
2.4 rpbis SC T20	-0,20	-0,16	-0,10	**-0,17**	-0,03	-0,02	0,46

3. Palier de Turbo analyse : T30

	OM	P1	P2	**P3**	P4	P5	P6
3.1 N Rép. T30	58	262	39	**256**	20	6	742
3.2 % Rép. T30	4%	19%	3%	**18%**	1%	0%	53%
3.3 C. Moy. T30	7%	30%	24%	**46%**	38%	27%	83%
3.4 rpbis SC T30	-0,20	-0,16	-0,10	**-0,17**	-0,03	-0,03	0,46

4. Palier de Turbo analyse : T40

	OM	P1	P2	**P3**	P4	P5	P6
4.1 N Rép. T40	58	261	39	**256**	20	6	740
4.2 % Rép. T40	4%	19%	3%	**18%**	1%	0%	53%
4.3 C. Moy. T40	7%	30%	24%	**47%**	38%	27%	83%
4.4 rpbis SC T40	-0,20	-0,17	-0,10	**-0,17**	-0,03	-0,03	0,46

5. Palier de Turbo analyse : T50

	OM	P1	P2	**P3**	P4	P5	P6
5.1 N Rép. T50	58	261	38	**254**	20	6	734
5.2 % Rép. T50	4%	19%	3%	**18%**	1%	0%	53%
5.3 C. Moy. T50	7%	30%	22%	**47%**	38%	27%	83%
5.4 rpbis SC T50	-0,20	-0,17	-0,11	**-0,17**	-0,03	-0,03	0,46

6. Palier de Turbo analyse : T60

	OM	P1	P2	**P3**	P4	P5	P6
6.1 N Rép. T60	58	250	36	**247**	19	6	715
6.2 % Rép. T60	4%	18%	3%	**18%**	1%	0%	53%
6.3 C. Moy. T60	7%	29%	20%	**47%**	39%	27%	83%
6.4 rpbis SC T60	-0,21	-0,17	-0,11	**-0,17**	-0,02	-0,03	0,47

7. Palier de Turbo analyse : T70

	OM	P1	P2	**P3**	P4	P5	P6
7.1 N Rép. T70	54	226	34	**209**	18	4	611
7.2 % Rép. T70	5%	19%	3%	**18%**	2%	0%	52%
7.3 C. Moy. T70	4%	30%	20%	**48%**	44%	20%	83%
7.4 rpbis SC T70	-0,24	-0,18	-0,12	**-0,15**	-0,02	-0,04	0,46

8. Palier de Turbo analyse : T80

	OM	P1	P2	**P3**	P4	P5	P6
8.1 N Rép. T80	38	137	18	**113**	16	3	320
8.2 % Rép. T80	6%	21%	3%	**17%**	2%	0%	49%
8.3 C. Moy. T80	0%	31%	13%	**53%**	40%	20%	82%
8.4 rpbis SC T80	xxxx	-0,24	-0,17	**-0,07**	-0,05	-0,06	0,40

9. Palier de Turbo analyse : T90

	OM	P1	P2	**P3**	P4	P5	P6
9.1 N Rép. T90	14	30	3	**16**	5	1	53
9.2 % Rép. T90	11%	24%	2%	**13%**	4%	1%	43%
9.3 C. Moy. T90	0%	29%	7%	**47%**	20%	60%	68%
9.4 rpbis SC T90	xxxx	-0,22	-0,18	**0,01**	-0,13	0,03	0,35

4. Question 5 de l'épreuve de Vocabulaire - [V]q5

a) Statistiques classiques

	OM	P1	P2	P3	P4	P5	**P6**	P7
1) N Rép.	135	761	1999	78	101	229	**455**	88
2) % Rép.	4%	20%	52%	2%	3%	6%	**12%**	2%
3) rpbis	-0,15	-0,07	0,01	-0,07	-0,10	0,12	**0,15**	0,02

b) Statistiques spectrales

	OM	P1	P2	P3	P4	P5	**P6**	P7
1) C Moy.	8%	44%	51%	34%	37%	51%	**46%**	50%
2) rpbis SC	-0,23	-0,03	0,09	-0,06	-0,05	0,04	**-0,01**	0,02

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	P3	P4	P5	**P6**	P7
1.1 N Rép. T10	135	761	1999	78	101	229	**455**	88
1.2 % Rép. T10	4%	20%	52%	2%	3%	6%	**12%**	2%
1.3 C. Moy. T10	8%	44%	51%	34%	37%	51%	**46%**	50%
1.4 rpbis SC T10	-0,23	-0,03	0,09	-0,06	-0,05	0,04	**-0,01**	0,02

2. Palier de Turbo analyse : T20

	OM	P1	P2	P3	P4	P5	**P6**	P7
2.1 N Rép. T20	135	761	1999	77	101	229	**455**	88
2.2 % Rép. T20	4%	20%	52%	2%	3%	6%	**12%**	2%
2.3 C. Moy. T20	8%	44%	51%	34%	37%	51%	**46%**	50%
2.4 rpbis SC T20	-0,23	-0,03	0,09	-0,06	-0,05	0,04	**-0,01**	0,02

3. Palier de Turbo analyse : T30

	OM	P1	P2	P3	P4	P5	**P6**	P7
3.1 N Rép. T30	135	760	1999	76	101	229	**454**	88
3.2 % Rép. T30	4%	20%	52%	2%	3%	6%	**12%**	2%
3.3 C. Moy. T30	8%	44%	51%	33%	37%	51%	**46%**	50%
3.4 rpbis SC T30	-0,23	-0,03	0,08	-0,06	-0,05	0,04	**-0,01**	0,02

4. Palier de Turbo analyse : T40

	OM	P1	P2	P3	P4	P5	**P6**	P7
4.1 N Rép. T40	135	753	1995	75	101	229	**453**	88
4.2 % Rép. T40	4%	20%	52%	2%	3%	6%	**12%**	2%
4.3 C. Moy. T40	8%	43%	51%	32%	37%	52%	**46%**	50%
4.4 rpbis SC T40	-0,24	-0,04	0,08	-0,07	-0,05	0,04	**-0,01**	0,02

5. Palier de Turbo analyse : T50

	OM	P1	P2	P3	P4	P5	**P6**	P7
5.1 N Rép. T50	134	745	1983	75	98	225	**449**	88
5.2 % Rép. T50	4%	20%	52%	2%	3%	6%	**12%**	2%
5.3 C. Moy. T50	8%	43%	51%	32%	38%	52%	**47%**	50%
5.4 rpbis SC T50	-0,24	-0,05	0,08	-0,07	-0,05	0,04	**0,00**	0,02

6. Palier de Turbo analyse : T60

	OM	P1	P2	P3	P4	P5	**P6**	P7
6.1 N Rép. T60	131	721	1921	72	93	215	**439**	87
6.2 % Rép. T60	4%	20%	52%	2%	3%	6%	**12%**	2%
6.3 C. Moy. T60	8%	43%	51%	31%	36%	52%	**47%**	50%
6.4 rpbis SC T60	-0,25	-0,05	0,08	-0,07	-0,06	0,05	**0,00**	0,02

7. Palier de Turbo analyse : T70

	OM	P1	P2	P3	P4	P5	**P6**	P7
7.1 N Rép. T70	119	627	1676	63	73	200	**387**	78
7.2 % Rép. T70	4%	19%	52%	2%	2%	6%	**12%**	2%
7.3 C. Moy. T70	6%	42%	50%	30%	32%	51%	**49%**	49%
7.4 rpbis SC T70	-0,29	-0,09	0,03	-0,09	-0,09	0,03	**0,03**	0,00

8. Palier de Turbo analyse : T80

	OM	P1	P2	P3	P4	P5	**P6**	P7
8.1 N Rép. T80	92	403	1032	37	40	137	**262**	53
8.2 % Rép. T80	4%	20%	50%	2%	2%	7%	**13%**	3%
8.3 C. Moy. T80	4%	42%	50%	30%	30%	53%	**52%**	48%
8.4 rpbis SC T80	-0,37	-0,15	-0,04	-0,11	-0,12	0,01	**0,08**	-0,02

9. Palier de Turbo analyse : T90

	OM	P1	P2	P3	P4	P5	**P6**	P7
9.1 N Rép. T90	32	84	225	10	8	33	**58**	7
9.2 % Rép. T90	7%	18%	49%	2%	2%	7%	**13%**	2%
9.3 C. Moy. T90	0%	41%	49%	43%	20%	50%	**53%**	54%
9.4 rpbis SC T90	xxxx	-0,18	-0,06	-0,09	-0,15	-0,01	**0,13**	0,01

5. Question 12 de l'épreuve de Vocabulaire - [V]q12

a) Statistiques classiques

	OM	P1	P2	P3	P4	P5	**P6**	P7
1) N Rép.	100	246	799	349	1335	280	**712**	25
2) % Rép.	3%	6%	21%	9%	35%	7%	**19%**	1%
3) rpbis	-0,19	-0,13	-0,08	-0,06	-0,01	0,03	**0,29**	-0,05

b) Statistiques spectrales

	OM	P1	P2	P3	P4	P5	**P6**	P7
1) C Moy.	11%	46%	55%	49%	67%	49%	**61%**	49%
2) rpbis SC	-0,25	-0,11	-0,07	-0,10	0,09	-0,10	**0,05**	-0,03

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	P3	P4	P5	**P6**	P7
1.1 N Rép. T10	100	246	799	349	1335	280	**712**	25
1.2 % Rép. T10	3%	6%	21%	9%	35%	7%	**19%**	1%
1.3 C. Moy. T10	11%	46%	55%	49%	67%	49%	**61%**	49%
1.4 rpbis SC T10	-0,25	-0,11	-0,07	-0,10	0,09	-0,10	**0,05**	-0,03

2. Palier de Turbo analyse : T20

	OM	P1	P2	P3	P4	P5	**P6**	P7
2.1 N Rép. T20	100	246	799	348	1335	280	**712**	25
2.2 % Rép. T20	3%	6%	21%	9%	35%	7%	**19%**	1%
2.3 C. Moy. T20	11%	46%	55%	49%	67%	49%	**61%**	49%
2.4 rpbis SC T20	-0,25	-0,11	-0,07	-0,11	0,09	-0,10	**0,05**	-0,03

3. Palier de Turbo analyse : T30

	OM	P1	P2	P3	P4	P5	**P6**	P7
3.1 N Rép. T30	100	245	799	348	1333	280	**712**	25
3.2 % Rép. T30	3%	6%	21%	9%	35%	7%	**19%**	1%
3.3 C. Moy. T30	11%	46%	55%	49%	67%	49%	**61%**	49%
3.4 rpbis SC T30	-0,25	-0,12	-0,07	-0,11	0,10	-0,10	**0,05**	-0,03

4. Palier de Turbo analyse : T40

	OM	P1	P2	P3	P4	P5	**P6**	P7
4.1 N Rép. T40	100	243	793	346	1332	280	**710**	25
4.2 % Rép. T40	3%	6%	21%	9%	35%	7%	**19%**	1%
4.3 C. Moy. T40	11%	45%	55%	49%	67%	49%	**61%**	49%
4.4 rpbis SC T40	-0,25	-0,12	-0,08	-0,11	0,10	-0,10	**0,05**	-0,03

5. Palier de Turbo analyse : T50

	OM	P1	P2	P3	P4	P5	**P6**	P7
5.1 N Rép. T50	98	242	791	343	1317	279	**702**	25
5.2 % Rép. T50	3%	6%	21%	9%	35%	7%	**18%**	1%
5.3 C. Moy. T50	11%	45%	55%	49%	68%	49%	**61%**	49%
5.4 rpbis SC T50	-0,25	-0,13	-0,08	-0,11	0,10	-0,10	**0,05**	-0,03

6. Palier de Turbo analyse : T60

	OM	P1	P2	P3	P4	P5	**P6**	P7
6.1 N Rép. T60	95	233	760	327	1283	273	**684**	24
6.2 % Rép. T60	3%	6%	21%	9%	35%	7%	**19%**	1%
6.3 C. Moy. T60	8%	45%	54%	48%	68%	49%	**62%**	47%
6.4 rpbis SC T60	-0,27	-0,13	-0,10	-0,12	0,09	-0,11	**0,06**	-0,04

7. Palier de Turbo analyse : T70

	OM	P1	P2	P3	P4	P5	**P6**	P7
7.1 N Rép. T70	88	204	663	279	1112	241	**616**	20
7.2 % Rép. T70	3%	6%	21%	9%	35%	7%	**19%**	1%
7.3 C. Moy. T70	7%	45%	54%	48%	67%	48%	**63%**	47%
7.4 rpbis SC T70	-0,30	-0,14	-0,12	-0,13	0,08	-0,12	**0,08**	-0,04

8. Palier de Turbo analyse : T80

	OM	P1	P2	P3	P4	P5	**P6**	P7
8.1 N Rép. T80	69	116	410	167	702	164	**416**	12
8.2 % Rép. T80	3%	6%	20%	8%	34%	8%	**20%**	1%
8.3 C. Moy. T80	4%	42%	53%	46%	66%	49%	**66%**	45%
8.4 rpbis SC T80	-0,38	-0,18	-0,17	-0,18	0,01	-0,16	**0,15**	-0,06

9. Palier de Turbo analyse : T90

	OM	P1	P2	P3	P4	P5	**P6**	P7
9.1 N Rép. T90	24	19	104	42	123	38	**102**	5
9.2 % Rép. T90	5%	4%	23%	9%	27%	8%	**22%**	1%
9.3 C. Moy. T90	0%	48%	48%	41%	61%	48%	**69%**	52%
9.4 rpbis SC T90	xxxx	-0,15	-0,27	-0,29	-0,12	-0,19	**0,29**	-0,06

6. Question 1 de l'épreuve de Syntaxe et articulation logique - [S]q1

a) Statistiques classiques

	OM	P1	P2	P3	P4	P5	**P6**	P7
1) N Rép.	54	59	39	453	82	1848	**1194**	10
2) % Rép.	1%	2%	1%	12%	2%	49%	**32%**	0%
3) rpbis	-0,07	-0,15	-0,11	-0,15	-0,06	-0,16	**0,38**	-0,03

b) Statistiques spectrales

	OM	P1	P2	P3	P4	P5	**P6**	P7
1) C Moy.	19%	53%	42%	63%	58%	66%	**58%**	64%
2) rpbis SC	-0,17	-0,03	-0,06	0,06	0,00	0,13	**-0,08**	0,01

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	P3	P4	P5	**P6**	P7
1.1 N Rép. T10	53	59	39	453	82	1848	**1190**	10
1.2 % Rép. T10	1%	2%	1%	12%	2%	49%	**32%**	0%
1.3 C. Moy. T10	19%	53%	42%	63%	58%	66%	**59%**	64%
1.4 rpbis SC T10	-0,17	-0,03	-0,06	0,06	0,00	0,12	**-0,08**	0,01

2. Palier de Turbo analyse : T20

	OM	P1	P2	P3	P4	P5	**P6**	P7
2.1 N Rép. T20	52	58	38	452	81	1844	**1186**	10
2.2 % Rép. T20	1%	2%	1%	12%	2%	50%	**32%**	0%
2.3 C. Moy. T20	17%	52%	41%	63%	57%	66%	**59%**	64%
2.4 rpbis SC T20	-0,17	-0,03	-0,07	0,05	-0,01	0,12	**-0,08**	0,01

3. Palier de Turbo analyse : T30

	OM	P1	P2	P3	P4	P5	**P6**	P7
3.1 N Rép. T30	52	58	38	444	79	1835	**1183**	10
3.2 % Rép. T30	1%	2%	1%	12%	2%	50%	**32%**	0%
3.3 C. Moy. T30	17%	52%	41%	63%	57%	66%	**59%**	64%
3.4 rpbis SC T30	-0,18	-0,03	-0,07	0,05	-0,01	0,13	**-0,08**	0,01

4. Palier de Turbo analyse : T40

	OM	P1	P2	P3	P4	P5	**P6**	P7
4.1 N Rép. T40	51	57	35	434	76	1820	**1177**	10
4.2 % Rép. T40	1%	2%	1%	12%	2%	50%	**32%**	0%
4.3 C. Moy. T40	18%	53%	42%	63%	60%	66%	**59%**	64%
4.4 rpbis SC T40	-0,18	-0,03	-0,06	0,04	0,00	0,13	**-0,08**	0,01

5. Palier de Turbo analyse : T50

	OM	P1	P2	P3	P4	P5	**P6**	P7
5.1 N Rép. T50	47	54	31	417	72	1780	**1152**	8
5.2 % Rép. T50	1%	2%	1%	12%	2%	50%	**32%**	0%
5.3 C. Moy. T50	19%	51%	45%	63%	59%	66%	**60%**	68%
5.4 rpbis SC T50	-0,17	-0,04	-0,05	0,03	0,00	0,12	**-0,07**	0,01

6. Palier de Turbo analyse : T60

	OM	P1	P2	P3	P4	P5	**P6**	P7
6.1 N Rép. T60	43	47	25	384	70	1679	**1080**	7
6.2 % Rép. T60	1%	1%	1%	12%	2%	50%	**32%**	0%
6.3 C. Moy. T60	17%	51%	41%	63%	58%	67%	**62%**	63%
6.4 rpbis SC T60	-0,19	-0,05	-0,08	0,01	-0,02	0,09	**-0,04**	0,00

7. Palier de Turbo analyse : T70

	OM	P1	P2	P3	P4	P5	**P6**	P7
7.1 N Rép. T70	36	40	20	320	58	1393	**900**	7
7.2 % Rép. T70	1%	1%	1%	12%	2%	50%	**32%**	0%
7.3 C. Moy. T70	17%	51%	41%	64%	57%	67%	**65%**	57%
7.4 rpbis SC T70	-0,22	-0,07	-0,08	-0,02	-0,05	0,03	**0,02**	0,00

8. Palier de Turbo analyse : T80

	OM	P1	P2	P3	P4	P5	**P6**	P7
8.1 N Rép. T80	24	27	13	191	37	847	**594**	3
8.2 % Rép. T80	1%	2%	1%	11%	2%	49%	**34%**	0%
8.3 C. Moy. T80	13%	47%	40%	64%	60%	66%	**71%**	33%
8.4 rpbis SC T80	-0,28	-0,13	-0,11	-0,09	-0,06	-0,09	**0,14**	-0,06

9. Palier de Turbo analyse : T90

	OM	P1	P2	P3	P4	P5	**P6**	P7
9.1 N Rép. T90	8	8	4	47	17	234	**214**	1
9.2 % Rép. T90	2%	2%	1%	9%	3%	44%	**40%**	0%
9.3 C. Moy. T90	0%	45%	50%	67%	65%	67%	**79%**	20%
9.4 rpbis SC T90	xxxx	-0,19	-0,12	-0,14	-0,12	-0,27	**0,33**	-0,12

7. Question 27 de l'épreuve de Vocabulaire - [V]q27

a) Statistiques classiques

	OM	P1	P2	**P3**
1) N Rép.	321	718	874	**1892**
2) % Rép.	8%	19%	23%	**49%**
3) rpbis	-0,24	0,13	-0,09	**0,13**

b) Statistiques spectrales

	OM	P1	P2	**P3**
1) C Moy.	3%	42%	28%	**32%**
2) rpbis SC	-0,27	0,14	-0,06	**0,04**

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	**P3**
1.1 N Rép. T10	321	718	874	**1892**
1.2 % Rép. T10	8%	19%	23%	**49%**
1.3 C. Moy. T10	3%	42%	28%	**32%**
1.4 rpbis SC T10	-0,27	0,14	-0,06	**0,04**

2. Palier de Turbo analyse : T20

	OM	P1	P2	**P3**
2.1 N Rép. T20	321	718	874	**1892**
2.2 % Rép. T20	8%	19%	23%	**49%**
2.3 C. Moy. T20	3%	42%	28%	**32%**
2.4 rpbis SC T20	-0,27	0,14	-0,06	**0,04**

3. Palier de Turbo analyse : T30

	OM	P1	P2	**P3**
3.1 N Rép. T30	321	717	874	**1891**
3.2 % Rép. T30	8%	19%	23%	**49%**
3.3 C. Moy. T30	3%	42%	28%	**32%**
3.4 rpbis SC T30	-0,27	0,14	-0,06	**0,05**

4. Palier de Turbo analyse : T40

	OM	P1	P2	**P3**
4.1 N Rép. T40	321	716	873	**1884**
4.2 % Rép. T40	8%	19%	23%	**49%**
4.3 C. Moy. T40	3%	42%	28%	**32%**
4.4 rpbis SC T40	-0,27	0,14	-0,06	**0,05**

5. Palier de Turbo analyse : T50

	OM	P1	P2	**P3**
5.1 N Rép. T50	320	704	867	**1873**
5.2 % Rép. T50	8%	19%	23%	**49%**
5.3 C. Moy. T50	3%	42%	28%	**32%**
5.4 rpbis SC T50	-0,27	0,13	-0,06	**0,05**

6. Palier de Turbo analyse : T60

	OM	P1	P2	**P3**
6.1 N Rép. T60	310	683	846	**1811**
6.2 % Rép. T60	8%	19%	23%	**49%**
6.3 C. Moy. T60	2%	43%	27%	**32%**
6.4 rpbis SC T60	-0,27	0,14	-0,07	**0,05**

7. Palier de Turbo analyse : T70

	OM	P1	P2	**P3**
7.1 N Rép. T70	275	611	741	**1572**
7.2 % Rép. T70	9%	19%	23%	**49%**
7.3 C. Moy. T70	2%	43%	27%	**33%**

8. Palier de Turbo analyse : T80

	OM	P1	P2	**P3**
8.1 N Rép. T80	207	413	488	**938**
8.2 % Rép. T80	10%	20%	24%	**46%**
8.3 C. Moy. T80	1%	45%	28%	**36%**
8.4 rpbis SC T80	-0,37	0,12	-0,13	**0,13**

9. Palier de Turbo analyse : T90

	OM	P1	P2	**P3**
9.1 N Rép. T90	75	99	92	**189**
9.2 % Rép. T90	16%	22%	20%	**41%**
9.3 C. Moy. T90	0%	43%	34%	**49%**
9.4 rpbis SC T90	-0,61	-0,03	-0,21	**0,33**

8. Question 14 de Connaissances artistiques - [A]q14

a) Statistiques classiques

	OM	P1	P2	**P3**	P4	P5	P6	P7
1) N Rép.	210	162	113	**296**	316	130	152	13
2) % Rép.	15%	12%	8%	**21%**	23%	9%	11%	1%
3) rpbis	-0,18	0,03	-0,01	**0,09**	0,08	0,01	-0,06	0,02

b) Statistiques spectrales

	OM	P1	P2	**P3**	P4	P5	P6	P7
1) C Moy.	0%	29%	27%	**25%**	39%	25%	21%	35%
2) rpbis SC	-0,30	0,04	0,03	**0,00**	0,20	0,01	-0,04	0,04

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	**P3**	P4	P5	P6	P7
1.1 N Rép. T10	210	162	113	**296**	316	130	152	13
1.2 % Rép. T10	15%	12%	8%	**21%**	23%	9%	11%	1%
1.3 C. Moy. T10	0%	29%	27%	**25%**	39%	25%	21%	35%
1.4 rpbis SC T10	-0,30	0,04	0,03	**0,00**	0,20	0,01	-0,04	0,04

2. Palier de Turbo analyse : T20

	OM	P1	P2	**P3**	P4	P5	P6	P7
2.1 N Rép. T20	210	162	113	**296**	316	130	152	13
2.2 % Rép. T20	15%	12%	8%	**21%**	23%	9%	11%	1%
2.3 C. Moy. T20	0%	29%	27%	**25%**	39%	25%	21%	35%
2.4 rpbis SC T20	-0,30	0,04	0,03	**0,00**	0,20	0,01	-0,04	0,04

3. Palier de Turbo analyse : T30

	OM	P1	P2	**P3**	P4	P5	P6	P7
3.1 N Rép. T30	210	162	113	**294**	316	129	152	13
3.2 % Rép. T30	15%	12%	8%	**21%**	23%	9%	11%	1%
3.3 C. Moy. T30	0%	29%	27%	**25%**	39%	25%	21%	35%
3.4 rpbis SC T30	-0,30	0,04	0,02	**0,00**	0,20	0,00	-0,05	0,03

4. Palier de Turbo analyse : T40

	OM	P1	P2	**P3**	P4	P5	P6	P7
4.1 N Rép. T40	210	162	113	**294**	316	128	152	13
4.2 % Rép. T40	15%	12%	8%	**21%**	23%	9%	11%	1%
4.3 C. Moy. T40	0%	29%	27%	**25%**	39%	25%	21%	35%
4.4 rpbis SC T40	-0,30	0,04	0,02	**0,01**	0,20	0,00	-0,05	0,03

5. Palier de Turbo analyse : T50

	OM	P1	P2	**P3**	P4	P5	P6	P7
5.1 N Rép. T50	207	160	112	**292**	311	128	150	13
5.2 % Rép. T50	15%	12%	8%	**21%**	23%	9%	11%	1%
5.3 C. Moy. T50	0%	28%	28%	**25%**	38%	25%	19%	35%
5.4 rpbis SC T50	-0,30	0,04	0,03	**0,01**	0,19	0,00	-0,06	0,04

6. Palier de Turbo analyse : T60

	OM	P1	P2	**P3**	P4	P5	P6	P7
6.1 N Rép. T60	205	155	109	**287**	303	124	147	13
6.2 % Rép. T60	15%	12%	8%	**21%**	23%	9%	11%	1%
6.3 C. Moy. T60	0%	27%	28%	**25%**	37%	23%	19%	35%
6.4 rpbis SC T60	-0,31	0,03	0,04	**0,02**	0,18	-0,02	-0,07	0,04

7. Palier de Turbo analyse : T70

	OM	P1	P2	**P3**	P4	P5	P6	P7
7.1 N Rép. T70	196	145	98	**256**	274	110	124	12
7.2 % Rép. T70	16%	12%	8%	**21%**	23%	9%	10%	1%
7.3 C. Moy. T70	0%	27%	29%	**26%**	36%	22%	18%	32%
7.4 rpbis SC T70	-0,34	0,01	0,03	**0,05**	0,16	-0,05	-0,09	0,02

8. Palier de Turbo analyse : T80

	OM	P1	P2	**P3**	P4	P5	P6	P7
8.1 N Rép. T80	156	89	62	**147**	174	67	74	9
8.2 % Rép. T80	20%	11%	8%	**19%**	22%	9%	10%	1%
8.3 C. Moy. T80	0%	26%	30%	**30%**	37%	24%	20%	27%
8.4 rpbis SC T80	xxxx	-0,04	-0,01	**0,11**	0,10	-0,05	-0,10	-0,01

9. Palier de Turbo analyse : T90

	OM	P1	P2	**P3**	P4	P5	P6	P7
9.1 N Rép. T90	54	20	10	**29**	47	22	17	2
9.2 % Rép. T90	27%	10%	5%	**14%**	23%	11%	8%	1%
9.3 C. Moy. T90	0%	28%	36%	**47%**	37%	24%	28%	60%
9.4 rpbis SC T90	xxxx	-0,20	-0,08	**0,32**	-0,13	-0,28	-0,23	-0,06

9. Question 5 de l'épreuve de Biologie - [B]q5

a) Statistiques classiques

	OM	P1	P2	**P3**	P4		P6	P7
1) N Rép.	142	1175	277	**533**	213		139	20
2) % Rép.	6%	47%	11%	**21%**	8%		6%	1%
3) rpbis	-0,13	0,02	-0,10	**0,24**	-0,12		-0,05	0,00

b) Statistiques spectrales

	OM	P1	P2	**P3**	P4		P6	P7
1) C Moy.	4%	41%	30%	**38%**	29%		27%	32%
2) rpbis SC	-0,26	0,05	-0,08	**0,05**	-0,08		-0,09	-0,02

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	**P3**	P4		P6	P7
1.1 N Rép. T10	142	1175	277	**533**	213		139	20
1.2 % Rép. T10	6%	47%	11%	**21%**	8%		6%	1%
1.3 C. Moy. T10	4%	41%	30%	**38%**	29%		27%	32%
1.4 rpbis SC T10	-0,26	0,05	-0,08	**0,05**	-0,08		-0,09	-0,02

2. Palier de Turbo analyse : T20

	OM	P1	P2	**P3**	P4		P6	P7
2.1 N Rép. T20	142	1172	277	**533**	212		139	20
2.2 % Rép. T20	6%	47%	11%	**21%**	8%		6%	1%
2.3 C. Moy. T20	4%	41%	30%	**38%**	29%		27%	32%
2.4 rpbis SC T20	-0,26	0,05	-0,08	**0,05**	-0,08		-0,09	-0,02

3. Palier de Turbo analyse : T30

	OM	P1	P2	**P3**	P4		P6	P7
3.1 N Rép. T30	142	1169	277	**531**	211		139	20
3.2 % Rép. T30	6%	47%	11%	**21%**	8%		6%	1%
3.3 C. Moy. T30	4%	41%	30%	**38%**	29%		27%	32%
3.4 rpbis SC T30	-0,26	0,05	-0,08	**0,05**	-0,08		-0,09	-0,02

4. Palier de Turbo analyse : T40

	OM	P1	P2	**P3**	P4		P6	P7
4.1 N Rép. T40	140	1159	274	**527**	209		139	20
4.2 % Rép. T40	6%	47%	11%	**21%**	8%		6%	1%
4.3 C. Moy. T40	4%	41%	30%	**39%**	29%		27%	32%
4.4 rpbis SC T40	-0,26	0,04	-0,09	**0,06**	-0,08		-0,09	-0,02

5. Palier de Turbo analyse : T50

	OM	P1	P2	**P3**	P4		P6	P7
5.1 N Rép. T50	135	1115	263	**498**	202		137	18
5.2 % Rép. T50	6%	47%	11%	**21%**	9%		6%	1%
5.3 C. Moy. T50	3%	40%	29%	**39%**	29%		26%	36%
5.4 rpbis SC T50	-0,27	0,02	-0,11	**0,08**	-0,10		-0,11	-0,01

6. Palier de Turbo analyse : T60

	OM	P1	P2	**P3**	P4		P6	P7
6.1 N Rép. T60	124	993	235	**435**	172		124	16
6.2 % Rép. T60	6%	47%	11%	**21%**	8%		6%	1%
6.3 C. Moy. T60	3%	40%	28%	**42%**	29%		27%	33%
6.4 rpbis SC T60	-0,30	-0,03	-0,14	**0,11**	-0,12		-0,12	-0,03

7. Palier de Turbo analyse : T70

	OM	P1	P2	**P3**	P4		P6	P7
7.1 N Rép. T70	103	777	178	**302**	120		108	12
7.2 % Rép. T70	6%	48%	11%	**19%**	7%		7%	1%
7.3 C. Moy. T70	1%	39%	26%	**46%**	25%		25%	25%
7.4 rpbis SC T70	-0,38	-0,13	-0,22	**0,21**	-0,20		-0,18	-0,06

8. Palier de Turbo analyse : T80

	OM	P1	P2	**P3**	P4		P6	P7
8.1 N Rép. T80	67	410	96	**151**	64		53	10
8.2 % Rép. T80	8%	48%	11%	**18%**	7%		6%	1%
8.3 C. Moy. T80	0%	37%	23%	**52%**	22%		25%	24%
8.4 rpbis SC T80	-0,47	-0,24	-0,32	**0,29**	-0,26		-0,21	-0,10

9. Palier de Turbo analyse : T90

	OM	P1	P2	**P3**	P4		P6	P7
9.1 N Rép. T90	25	95	24	**27**	18		13	2
9.2 % Rép. T90	12%	46%	12%	**13%**	9%		6%	1%
9.3 C. Moy. T90	0%	36%	18%	**58%**	29%		20%	0%
9.4 rpbis SC T90	xxxx	-0,42	-0,47	**0,40**	-0,33		-0,35	-0,07

10. Question 7 de l'épreuve de Physique - [P]q7

a) Statistiques classiques

	OM	P1	**P2**	P3	P4	P5	P6	P7
1) N Rép.	86	188	**602**	1200	57	135	225	4
2) % Rép.	3%	8%	**24%**	48%	2%	5%	9%	0%
3) rpbis	-0,14	-0,08	**0,43**	-0,20	-0,09	-0,12	0,01	-0,06

b) Statistiques spectrales

	OM	P1	**P2**	P3	P4	P5	P6	P7
1) C Moy.	11%	55%	**71%**	55%	40%	29%	43%	20%
2) rpbis SC	-0,31	-0,12	**0,27**	-0,23	-0,13	-0,27	-0,23	-0,06

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	**P2**	P3	P4	P5	P6	P7
1.1 N Rép. T10	85	188	**600**	1200	57	135	225	4
1.2 % Rép. T10	3%	8%	**24%**	48%	2%	5%	9%	0%
1.3 C. Moy. T10	12%	55%	**71%**	55%	40%	29%	43%	20%
1.4 rpbis SC T10	-0,31	-0,13	**0,27**	-0,23	-0,14	-0,27	-0,23	-0,06

2. Palier de Turbo analyse : T20

	OM	P1	**P2**	P3	P4	P5	P6	P7
2.1 N Rép. T20	80	188	**594**	1198	57	134	224	4
2.2 % Rép. T20	3%	8%	**24%**	48%	2%	5%	9%	0%
2.3 C. Moy. T20	12%	55%	**72%**	55%	40%	30%	43%	20%
2.4 rpbis SC T20	-0,30	-0,13	**0,27**	-0,24	-0,14	-0,28	-0,24	-0,06

3. Palier de Turbo analyse : T30

	OM	P1	**P2**	P3	P4	P5	P6	P7
3.1 N Rép. T30	78	185	**591**	1189	57	130	224	4
3.2 % Rép. T30	3%	8%	**24%**	48%	2%	5%	9%	0%
3.3 C. Moy. T30	13%	55%	**72%**	56%	40%	31%	43%	20%
3.4 rpbis SC T30	-0,30	-0,13	**0,28**	-0,24	-0,14	-0,27	-0,25	-0,06

4. Palier de Turbo analyse : T40

	OM	P1	**P2**	P3	P4	P5	P6	P7
4.1 N Rép. T40	75	180	**587**	1169	55	128	223	4
4.2 % Rép. T40	3%	7%	**24%**	48%	2%	5%	9%	0%
4.3 C. Moy. T40	13%	56%	**73%**	57%	40%	32%	43%	20%
4.4 rpbis SC T40	-0,30	-0,13	**0,28**	-0,24	-0,14	-0,27	-0,25	-0,06

5. Palier de Turbo analyse : T50

	OM	P1	**P2**	P3	P4	P5	P6	P7
5.1 N Rép. T50	68	176	**573**	1131	55	121	220	3
5.2 % Rép. T50	3%	7%	**24%**	48%	2%	5%	9%	0%
5.3 C. Moy. T50	13%	57%	**74%**	57%	40%	33%	44%	0%
5.4 rpbis SC T50	-0,30	-0,13	**0,28**	-0,25	-0,15	-0,27	-0,26	xxxx

6. Palier de Turbo analyse : T60

	OM	P1	**P2**	P3	P4	P5	P6	P7
6.1 N Rép. T60	62	163	**530**	1041	51	106	203	3
6.2 % Rép. T60	3%	8%	**25%**	48%	2%	5%	9%	0%
6.3 C. Moy. T60	12%	58%	**77%**	58%	40%	34%	44%	0%
6.4 rpbis SC T60	-0,33	-0,15	**0,32**	-0,28	-0,17	-0,28	-0,29	xxxx

7. Palier de Turbo analyse : T70

	OM	P1	**P2**	P3	P4	P5	P6	P7
7.1 N Rép. T70	46	126	**469**	838	35	75	167	2
7.2 % Rép. T70	3%	7%	**27%**	48%	2%	4%	9%	0%
7.3 C. Moy. T70	14%	58%	**79%**	58%	45%	37%	43%	0%
7.4 rpbis SC T70	-0,32	-0,18	**0,36**	-0,32	-0,16	-0,26	-0,33	xxxx

8. Palier de Turbo analyse : T80

	OM	P1	**P2**	P3	P4	P5	P6	P7
8.1 N Rép. T80	25	65	**330**	470	19	36	108	1
8.2 % Rép. T80	2%	6%	**31%**	45%	2%	3%	10%	0%
8.3 C. Moy. T80	19%	58%	**83%**	59%	43%	39%	44%	0%
8.4 rpbis SC T80	-0,31	-0,20	**0,45**	-0,40	-0,17	-0,26	-0,38	xxxx

9. Palier de Turbo analyse : T90

	OM	P1	**P2**	P3	P4	P5	P6	P7
9.1 N Rép. T90	15	17	**142**	96	2	11	29	0
9.2 % Rép. T90	5%	5%	**46%**	31%	1%	4%	9%	0%
9.3 C. Moy. T90	15%	52%	**89%**	59%	80%	38%	41%	xxxx
9.4 rpbis SC T90	-0,51	-0,25	**0,60**	-0,42	-0,12	-0,30	-0,41	xxxx

11. Question 43 de l'épreuve de Vocabulaire - [V]q43

a) Statistiques classiques

	OM	P1	P2	**P3**	P4	P5	P6	P7
1) N Rép.	249	1038	250	**1195**	540	80	243	251
2) % Rép.	6%	27%	7%	**31%**	14%	2%	6%	7%
3) rpbis	-0,21	0,06	0,03	**0,09**	-0,03	-0,05	-0,05	0,02

b) Statistiques spectrales

	OM	P1	P2	**P3**	P4	P5	P6	P7
1) C Moy.	7%	48%	45%	**45%**	35%	28%	22%	45%
2) rpbis SC	-0,31	0,04	0,00	**0,11**	-0,12	-0,08	-0,19	0,00

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	**P3**	P4	P5	P6	P7
1.1 N Rép. T10	249	1038	250	**1195**	540	80	243	251
1.2 % Rép. T10	6%	27%	7%	**31%**	14%	2%	6%	7%
1.3 C. Moy. T10	7%	48%	45%	**45%**	35%	28%	22%	45%
1.4 rpbis SC T10	-0,31	0,04	0,00	**0,11**	-0,12	-0,08	-0,19	0,00

2. Palier de Turbo analyse : T20

	OM	P1	P2	**P3**	P4	P5	P6	P7
2.1 N Rép. T20	249	1038	249	**1195**	540	80	243	251
2.2 % Rép. T20	6%	27%	6%	**31%**	14%	2%	6%	7%
2.3 C. Moy. T20	7%	48%	45%	**45%**	35%	28%	22%	45%
2.4 rpbis SC T20	-0,31	0,04	0,00	**0,11**	-0,12	-0,08	-0,19	0,00

3. Palier de Turbo analyse : T30

	OM	P1	P2	**P3**	P4	P5	P6	P7
3.1 N Rép. T30	249	1037	249	**1194**	539	80	243	251
3.2 % Rép. T30	6%	27%	6%	**31%**	14%	2%	6%	7%
3.3 C. Moy. T30	7%	48%	45%	**45%**	35%	28%	22%	45%
3.4 rpbis SC T30	-0,31	0,03	0,00	**0,11**	-0,12	-0,08	-0,19	0,00

4. Palier de Turbo analyse : T40

	OM	P1	P2	**P3**	P4	P5	P6	P7
4.1 N Rép. T40	249	1031	245	**1192**	538	80	243	251
4.2 % Rép. T40	7%	27%	6%	**31%**	14%	2%	6%	7%
4.3 C. Moy. T40	7%	48%	45%	**45%**	35%	28%	22%	45%
4.4 rpbis SC T40	-0,31	0,03	0,00	**0,11**	-0,12	-0,08	-0,19	0,00

5. Palier de Turbo analyse : T50

	OM	P1	P2	**P3**	P4	P5	P6	P7
5.1 N Rép. T50	246	1023	243	**1181**	535	78	242	249
5.2 % Rép. T50	6%	27%	6%	**31%**	14%	2%	6%	7%
5.3 C. Moy. T50	6%	48%	45%	**45%**	35%	28%	22%	45%
5.4 rpbis SC T50	-0,32	0,03	0,00	**0,12**	-0,12	-0,08	-0,19	0,00

6. Palier de Turbo analyse : T60

	OM	P1	P2	**P3**	P4	P5	P6	P7
6.1 N Rép. T60	239	987	238	**1149**	514	77	235	240
6.2 % Rép. T60	6%	27%	6%	**31%**	14%	2%	6%	7%
6.3 C. Moy. T60	7%	47%	44%	**45%**	34%	26%	22%	46%
6.4 rpbis SC T60	-0,32	0,03	-0,01	**0,12**	-0,13	-0,09	-0,19	0,00

7. Palier de Turbo analyse : T70

	OM	P1	P2	**P3**	P4	P5	P6	P7
7.1 N Rép. T70	212	871	206	**1010**	429	67	210	218
7.2 % Rép. T70	7%	27%	6%	**31%**	13%	2%	7%	7%
7.3 C. Moy. T70	6%	47%	45%	**46%**	34%	27%	21%	45%
7.4 rpbis SC T70	-0,34	0,02	-0,01	**0,14**	-0,14	-0,09	-0,21	0,00

8. Palier de Turbo analyse : T80

	OM	P1	P2	**P3**	P4	P5	P6	P7
8.1 N Rép. T80	150	549	137	**634**	251	47	132	156
8.2 % Rép. T80	7%	27%	7%	**31%**	12%	2%	6%	8%
8.3 C. Moy. T80	3%	47%	44%	**48%**	32%	28%	20%	44%
8.4 rpbis SC T80	-0,42	-0,04	-0,02	**0,21**	-0,19	-0,11	-0,24	-0,04

9. Palier de Turbo analyse : T90

	OM	P1	P2	**P3**	P4	P5	P6	P7
9.1 N Rép. T90	53	118	26	**125**	64	9	32	30
9.2 % Rép. T90	12%	26%	6%	**27%**	14%	2%	7%	7%
9.3 C. Moy. T90	1%	49%	46%	**55%**	33%	28%	21%	46%
9.4 rpbis SC T90	-0,61	-0,07	-0,10	**0,32**	-0,27	-0,13	-0,29	-0,09

12. Question 6 de l'épreuve de Biologie - [B]q6

a) Statistiques classiques

	OM	P1	**P2**	P3	P4	P5	P6	P7
1) N Rép.	87	236	**1301**	703	83	5	37	55
2) % Rép.	3%	9%	**52%**	28%	3%	0%	1%	2%
3) rpbis	-0,18	-0,16	**0,29**	-0,05	-0,11	-0,06	-0,07	-0,07

b) Statistiques spectrales

	OM	P1	**P2**	P3	P4	P5	P6	P7
1) C Moy.	8%	46%	**54%**	61%	32%	48%	39%	39%
2) rpbis SC	-0,27	-0,07	**0,06**	0,09	-0,12	-0,01	-0,06	-0,07

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	**P2**	P3	P4	P5	P6	P7
1.1 N Rép. T10	87	236	**1301**	703	83	5	37	55
1.2 % Rép. T10	3%	9%	**52%**	28%	3%	0%	1%	2%
1.3 C. Moy. T10	8%	46%	**54%**	61%	32%	48%	39%	39%
1.4 rpbis SC T10	-0,27	-0,07	**0,06**	0,09	-0,12	-0,01	-0,06	-0,07

2. Palier de Turbo analyse : T20

	OM	P1	**P2**	P3	P4	P5	P6	P7
2.1 N Rép. T20	87	235	**1300**	702	82	5	37	55
2.2 % Rép. T20	3%	9%	**52%**	28%	3%	0%	1%	2%
2.3 C. Moy. T20	8%	46%	**54%**	61%	31%	48%	39%	39%
2.4 rpbis SC T20	-0,27	-0,08	**0,07**	0,09	-0,13	-0,01	-0,06	-0,07

3. Palier de Turbo analyse : T30

	OM	P1	**P2**	P3	P4	P5	P6	P7
3.1 N Rép. T30	87	234	**1298**	700	82	4	37	55
3.2 % Rép. T30	3%	9%	**52%**	28%	3%	0%	1%	2%
3.3 C. Moy. T30	8%	46%	**54%**	60%	31%	40%	39%	39%
3.4 rpbis SC T30	-0,27	-0,08	**0,07**	0,08	-0,13	-0,02	-0,06	-0,07

4. Palier de Turbo analyse : T40

	OM	P1	**P2**	P3	P4	P5	P6	P7
4.1 N Rép. T40	86	231	**1286**	695	81	4	37	55
4.2 % Rép. T40	3%	9%	**52%**	28%	3%	0%	1%	2%
4.3 C. Moy. T40	8%	46%	**55%**	61%	32%	40%	39%	39%
4.4 rpbis SC T40	-0,27	-0,08	**0,07**	0,08	-0,13	-0,02	-0,06	-0,07

5. Palier de Turbo analyse : T50

	OM	P1	**P2**	P3	P4	P5	P6	P7
5.1 N Rép. T50	81	218	**1233**	676	75	3	36	53
5.2 % Rép. T50	3%	9%	**52%**	28%	3%	0%	2%	2%
5.3 C. Moy. T50	5%	44%	**55%**	60%	30%	20%	41%	39%
5.4 rpbis SC T50	-0,29	-0,10	**0,09**	0,07	-0,14	-0,04	-0,06	-0,08

6. Palier de Turbo analyse : T60

	OM	P1	**P2**	P3	P4	P5	P6	P7
6.1 N Rép. T60	72	183	**1108**	593	68	3	31	47
6.2 % Rép. T60	3%	9%	**53%**	28%	3%	0%	1%	2%
6.3 C. Moy. T60	2%	42%	**58%**	60%	27%	20%	39%	38%
6.4 rpbis SC T60	-0,32	-0,14	**0,14**	0,04	-0,18	-0,05	-0,07	-0,09

7. Palier de Turbo analyse : T70

	OM	P1	**P2**	P3	P4	P5	P6	P7
7.1 N Rép. T70	67	129	**879**	412	53	2	26	38
7.2 % Rép. T70	4%	8%	**55%**	26%	3%	0%	2%	2%
7.3 C. Moy. T70	1%	39%	**59%**	58%	26%	0%	37%	38%
7.4 rpbis SC T70	-0,39	-0,19	**0,24**	-0,02	-0,21	xxxx	-0,09	-0,11

8. Palier de Turbo analyse : T80

	OM	P1	**P2**	P3	P4	P5	P6	P7
8.1 N Rép. T80	44	74	**476**	196	29	2	14	19
8.2 % Rép. T80	5%	9%	**56%**	23%	3%	0%	2%	2%
8.3 C. Moy. T80	0%	38%	**62%**	52%	22%	0%	20%	40%
8.4 rpbis SC T80	xxxx	-0,22	**0,38**	-0,13	-0,24	xxxx	-0,17	-0,11

9. Palier de Turbo analyse : T90

	OM	P1	**P2**	P3	P4	P5	P6	P7
9.1 N Rép. T90	14	19	**109**	41	11	1	6	5
9.2 % Rép. T90	7%	9%	**53%**	20%	5%	0%	3%	2%
9.3 C. Moy. T90	0%	35%	**63%**	51%	28%	0%	12%	25%
9.4 rpbis SC T90	xxxx	-0,28	**0,53**	-0,21	-0,31	xxxx	-0,30	-0,20

13. Question 1 de l'épreuve de Lecture de documents et géographie - [G]q1

a) Statistiques classiques

	OM	**P1**	P2	P3	P4	P5	P6	P7
1) N Rép.	286	**2561**	75	310	15	13	240	188
2) % Rép.	8%	**69%**	2%	8%	0%	0%	7%	5%
3) rpbis	-0,27	**0,39**	-0,08	-0,16	-0,07	-0,07	-0,13	-0,05

b) Statistiques spectrales

	OM	**P1**	P2	P3	P4	P5	P6	P7
1) C Moy.	24%	**64%**	33%	46%	23%	22%	39%	52%
2) rpbis SC	-0,33	**0,37**	-0,14	-0,16	-0,08	-0,08	-0,19	-0,08

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	**P1**	P2	P3	P4	P5	P6	P7
1.1 N Rép. T10	286	**2561**	75	310	15	13	240	188
1.2 % Rép. T10	8%	**69%**	2%	8%	0%	0%	7%	5%
1.3 C. Moy. T10	24%	**64%**	33%	46%	23%	22%	39%	52%
1.4 rpbis SC T10	-0,33	**0,37**	-0,14	-0,16	-0,08	-0,08	-0,19	-0,08

2. Palier de Turbo analyse : T20

	OM	**P1**	P2	P3	P4	P5	P6	P7
2.1 N Rép. T20	286	**2559**	75	310	15	13	240	188
2.2 % Rép. T20	8%	**69%**	2%	8%	0%	0%	7%	5%
2.3 C. Moy. T20	24%	**64%**	33%	46%	23%	22%	39%	52%
2.4 rpbis SC T20	-0,33	**0,37**	-0,14	-0,16	-0,08	-0,08	-0,19	-0,09

3. Palier de Turbo analyse : T30

	OM	**P1**	P2	P3	P4	P5	P6	P7
3.1 N Rép. T30	283	**2553**	74	310	15	13	239	188
3.2 % Rép. T30	8%	**69%**	2%	8%	0%	0%	7%	5%
3.3 C. Moy. T30	25%	**65%**	32%	46%	23%	22%	39%	52%
3.4 rpbis SC T30	-0,33	**0,37**	-0,14	-0,16	-0,08	-0,08	-0,19	-0,09

4. Palier de Turbo analyse : T40

	OM	**P1**	P2	P3	P4	P5	P6	P7
4.1 N Rép. T40	280	**2541**	73	307	14	12	235	186
4.2 % Rép. T40	8%	**70%**	2%	8%	0%	0%	6%	5%
4.3 C. Moy. T40	25%	**65%**	31%	45%	20%	18%	38%	52%
4.4 rpbis SC T40	-0,33	**0,37**	-0,14	-0,17	-0,08	-0,08	-0,20	-0,09

5. Palier de Turbo analyse : T50

	OM	**P1**	P2	P3	P4	P5	P6	P7
5.1 N Rép. T50	275	**2488**	70	299	13	12	227	182
5.2 % Rép. T50	8%	**70%**	2%	8%	0%	0%	6%	5%
5.3 C. Moy. T50	25%	**65%**	29%	45%	14%	18%	38%	51%
5.4 rpbis SC T50	-0,33	**0,39**	-0,16	-0,18	-0,10	-0,08	-0,21	-0,10

6. Palier de Turbo analyse : T60

	OM	**P1**	P2	P3	P4	P5	P6	P7
6.1 N Rép. T60	257	**2318**	62	269	12	11	214	168
6.2 % Rép. T60	8%	**70%**	2%	8%	0%	0%	6%	5%
6.3 C. Moy. T60	25%	**67%**	27%	44%	15%	16%	37%	50%
6.4 rpbis SC T60	-0,34	**0,42**	-0,16	-0,19	-0,10	-0,09	-0,23	-0,11

7. Palier de Turbo analyse : T70

	OM	**P1**	P2	P3	P4	P5	P6	P7
7.1 N Rép. T70	195	**1840**	49	219	10	10	170	126
7.2 % Rép. T70	7%	**70%**	2%	8%	0%	0%	6%	5%
7.3 C. Moy. T70	26%	**69%**	27%	43%	12%	13%	34%	48%
7.4 rpbis SC T70	-0,35	**0,48**	-0,18	-0,23	-0,11	-0,11	-0,27	-0,14

8. Palier de Turbo analyse : T80

	OM	**P1**	P2	P3	P4	P5	P6	P7
8.1 N Rép. T80	122	**1073**	27	117	8	8	91	72
8.2 % Rép. T80	8%	**71%**	2%	8%	1%	1%	6%	5%
8.3 C. Moy. T80	27%	**73%**	23%	40%	12%	13%	35%	47%
8.4 rpbis SC T80	-0,39	**0,55**	-0,21	-0,28	-0,15	-0,14	-0,28	-0,17

9. Palier de Turbo analyse : T90

	OM	**P1**	P2	P3	P4	P5	P6	P7
9.1 N Rép. T90	44	**278**	7	32	1	2	23	19
9.2 % Rép. T90	11%	**68%**	2%	8%	0%	0%	6%	5%
9.3 C. Moy. T90	16%	**76%**	23%	37%	xxxx	0%	32%	46%
9.4 rpbis SC T90	-0,52	**0,64**	-0,19	-0,32	xxxx	xxxx	-0,30	-0,18

14. Question 16 de l'épreuve de Connaissances en Histoire et Socio Eco - [H]q16

a) Statistiques classiques

	OM	P1	P2	P3	**P4**	P5	P6
1) N Rép.	108	405	221	137	**328**	79	132
2) % Rép.	8%	29%	16%	10%	**23%**	6%	9%
3) rpbis	-0,12	0,06	-0,07	-0,07	**0,17**	0,00	-0,08

b) Statistiques spectrales

	OM	P1	P2	P3	**P4**	P5	P6
1) C Moy.	3%	48%	32%	29%	**35%**	29%	21%
2) rpbis SC	-0,27	0,19	-0,03	-0,05	**0,02**	-0,04	-0,13

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	P3	**P4**	P5	P6
1.1 N Rép. T10	108	404	221	137	**328**	79	132
1.2 % Rép. T10	8%	29%	16%	10%	**23%**	6%	9%
1.3 C. Moy. T10	3%	48%	32%	29%	**35%**	29%	21%
1.4 rpbis SC T10	-0,27	0,19	-0,03	-0,05	**0,02**	-0,04	-0,13

2. Palier de Turbo analyse : T20

	OM	P1	P2	P3	**P4**	P5	P6
2.1 N Rép. T20	108	404	221	137	**328**	79	132
2.2 % Rép. T20	8%	29%	16%	10%	**23%**	6%	9%
2.3 C. Moy. T20	3%	48%	32%	29%	**35%**	29%	21%
2.4 rpbis SC T20	-0,27	0,19	-0,03	-0,05	**0,02**	-0,04	-0,13

3. Palier de Turbo analyse : T30

	OM	P1	P2	P3	**P4**	P5	P6
3.1 N Rép. T30	108	403	221	137	**327**	79	131
3.2 % Rép. T30	8%	29%	16%	10%	**23%**	6%	9%
3.3 C. Moy. T30	3%	48%	32%	29%	**35%**	29%	20%
3.4 rpbis SC T30	-0,27	0,19	-0,03	-0,06	**0,03**	-0,04	-0,13

4. Palier de Turbo analyse : T40

	OM	P1	P2	P3	**P4**	P5	P6
4.1 N Rép. T40	107	402	221	137	**327**	78	131
4.2 % Rép. T40	8%	29%	16%	10%	**23%**	6%	9%
4.3 C. Moy. T40	3%	48%	32%	29%	**35%**	28%	21%
4.4 rpbis SC T40	-0,27	0,19	-0,03	-0,06	**0,03**	-0,05	-0,13

5. Palier de Turbo analyse : T50

	OM	P1	P2	P3	**P4**	P5	P6
5.1 N Rép. T50	106	400	218	137	**326**	78	129
5.2 % Rép. T50	8%	29%	16%	10%	**23%**	6%	9%
5.3 C. Moy. T50	2%	48%	32%	29%	**35%**	28%	21%
5.4 rpbis SC T50	-0,28	0,18	-0,04	-0,06	**0,03**	-0,05	-0,13

6. Palier de Turbo analyse : T60

	OM	P1	P2	P3	**P4**	P5	P6
6.1 N Rép. T60	104	391	213	131	**312**	77	124
6.2 % Rép. T60	8%	29%	16%	10%	**23%**	6%	9%
6.3 C. Moy. T60	2%	47%	32%	28%	**36%**	28%	21%
6.4 rpbis SC T60	-0,29	0,17	-0,05	-0,07	**0,04**	-0,05	-0,13

7. Palier de Turbo analyse : T70

	OM	P1	P2	P3	**P4**	P5	P6
7.1 N Rép. T70	101	346	184	114	**259**	65	106
7.2 % Rép. T70	9%	29%	16%	10%	**22%**	6%	9%
7.3 C. Moy. T70	1%	46%	31%	28%	**39%**	29%	20%
7.4 rpbis SC T70	-0,35	0,11	-0,08	-0,10	**0,09**	-0,07	-0,17

8. Palier de Turbo analyse : T80

	OM	P1	P2	P3	**P4**	P5	P6
8.1 N Rép. T80	71	193	92	69	**133**	39	59
8.2 % Rép. T80	11%	29%	14%	11%	**20%**	6%	9%
8.3 C. Moy. T80	1%	44%	35%	29%	**42%**	28%	22%
8.4 rpbis SC T80	-0,44	0,03	-0,09	-0,14	**0,16**	-0,12	-0,20

9. Palier de Turbo analyse : T90

	OM	P1	P2	P3	**P4**	P5	P6
9.1 N Rép. T90	24	39	12	12	**18**	7	11
9.2 % Rép. T90	20%	32%	10%	10%	**15%**	6%	9%
9.3 C. Moy. T90	0%	39%	33%	24%	**53%**	27%	23%
9.4 rpbis SC T90	xxxx	-0,11	-0,16	-0,27	**0,29**	-0,14	-0,32

15. Question 24 de l'épreuve de Connaissances artistiques - [A]q24

a) Statistiques classiques

	OM	**P1**	P2	P3	P4	P5	P6	P7
1) N Rép.	248	**233**	88	362	173	135	141	12
2) % Rép.	18%	**17%**	6%	26%	12%	10%	10%	1%
3) rpbis	-0,16	**0,16**	-0,03	0,05	0,00	-0,03	0,00	-0,04

b) Statistiques spectrales

	OM	**P1**	P2	P3	P4	P5	P6	P7
1) C Moy.	0%	**21%**	11%	18%	14%	24%	10%	25%
2) rpbis SC	-0,33	**0,13**	-0,10	-0,06	-0,10	0,03	-0,14	0,01

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	**P1**	P2	P3	P4	P5	P6	P7
1.1 N Rép. T10	248	**233**	88	362	173	135	141	12
1.2 % Rép. T10	18%	**17%**	6%	26%	12%	10%	10%	1%
1.3 C. Moy. T10	0%	**21%**	11%	18%	14%	24%	10%	25%
1.4 rpbis SC T10	-0,33	**0,13**	-0,10	-0,06	-0,10	0,03	-0,14	0,01

2. Palier de Turbo analyse : T20

	OM	**P1**	P2	P3	P4	P5	P6	P7
2.1 N Rép. T20	248	**233**	88	362	173	135	141	12
2.2 % Rép. T20	18%	**17%**	6%	26%	12%	10%	10%	1%
2.3 C. Moy. T20	0%	**21%**	11%	18%	14%	24%	10%	25%
2.4 rpbis SC T20	-0,33	**0,13**	-0,10	-0,06	-0,10	0,03	-0,14	0,01

3. Palier de Turbo analyse : T30

	OM	**P1**	P2	P3	P4	P5	P6	P7
3.1 N Rép. T30	248	**232**	88	362	172	134	141	12
3.2 % Rép. T30	18%	**17%**	6%	26%	12%	10%	10%	1%
3.3 C. Moy. T30	0%	**22%**	11%	18%	14%	24%	10%	25%
3.4 rpbis SC T30	-0,33	**0,13**	-0,10	-0,06	-0,10	0,02	-0,14	0,01

4. Palier de Turbo analyse : T40

	OM	**P1**	P2	P3	P4	P5	P6	P7
4.1 N Rép. T40	248	**232**	88	361	172	134	141	12
4.2 % Rép. T40	18%	**17%**	6%	26%	12%	10%	10%	1%
4.3 C. Moy. T40	0%	**22%**	11%	18%	14%	24%	10%	25%
4.4 rpbis SC T40	-0,33	**0,13**	-0,10	-0,07	-0,10	0,02	-0,14	0,01

5. Palier de Turbo analyse : T50

	OM	**P1**	P2	P3	P4	P5	P6	P7
5.1 N Rép. T50	246	**228**	88	359	169	131	140	12
5.2 % Rép. T50	18%	**17%**	6%	26%	12%	10%	10%	1%
5.3 C. Moy. T50	0%	**21%**	11%	17%	13%	22%	10%	25%
5.4 rpbis SC T50	-0,34	**0,14**	-0,11	-0,07	-0,12	0,01	-0,15	0,01

6. Palier de Turbo analyse : T60

	OM	**P1**	P2	P3	P4	P5	P6	P7
6.1 N Rép. T60	242	**221**	88	348	168	129	137	10
6.2 % Rép. T60	18%	**16%**	7%	26%	13%	10%	10%	1%
6.3 C. Moy. T60	0%	**21%**	11%	16%	12%	21%	10%	24%
6.4 rpbis SC T60	-0,34	**0,14**	-0,10	-0,08	-0,12	0,00	-0,14	0,01

7. Palier de Turbo analyse : T70

	OM	**P1**	P2	P3	P4	P5	P6	P7
7.1 N Rép. T70	232	**193**	76	308	155	118	124	9
7.2 % Rép. T70	19%	**16%**	6%	25%	13%	10%	10%	1%
7.3 C. Moy. T70	0%	**23%**	9%	15%	11%	19%	9%	16%
7.4 rpbis SC T70	-0,40	**0,20**	-0,14	-0,14	-0,16	-0,04	-0,19	-0,03

8. Palier de Turbo analyse : T80

	OM	**P1**	P2	P3	P4	P5	P6	P7
8.1 N Rép. T80	183	**95**	48	189	102	78	76	7
8.2 % Rép. T80	24%	**12%**	6%	24%	13%	10%	10%	1%
8.3 C. Moy. T80	0%	**28%**	11%	15%	12%	19%	8%	20%
8.4 rpbis SC T80	-0,56	**0,28**	-0,20	-0,26	-0,26	-0,13	-0,27	-0,03

9. Palier de Turbo analyse : T90

	OM	**P1**	P2	P3	P4	P5	P6	P7
9.1 N Rép. T90	69	**18**	7	45	26	18	15	3
9.2 % Rép. T90	34%	**9%**	3%	22%	13%	9%	7%	1%
9.3 C. Moy. T90	0%	**37%**	4%	16%	3%	25%	18%	20%
9.4 rpbis SC T90	xxxx	**0,40**	-0,32	-0,40	-0,52	-0,17	-0,30	-0,18

16. Question 10 de l'épreuve de Biologie - [B]q10

a) Statistiques classiques

	OM	P1	P2	**P3**	P4	P5	P6	P7
1) N Rép.	23	8	11	**2433**	8	11	5	8
2) % Rép.	1%	0%	0%	**97%**	0%	0%	0%	0%
3) rpbis	-0,13	-0,06	-0,05	**0,18**	-0,06	-0,05	-0,04	-0,05

b) Statistiques spectrales

	OM	P1	P2	**P3**	P4	P5	P6	P7
1) C Moy.	26%	38%	73%	**87%**	53%	65%	32%	55%
2) rpbis SC	-0,20	-0,09	-0,03	**0,23**	-0,07	-0,05	-0,08	-0,06

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	**P3**	P4	P5	P6	P7
1.1 N Rép. T10	23	8	11	**2433**	8	11	5	8
1.2 % Rép. T10	1%	0%	0%	**97%**	0%	0%	0%	0%
1.3 C. Moy. T10	26%	38%	73%	**87%**	53%	65%	32%	55%
1.4 rpbis SC T10	-0,20	-0,09	-0,03	**0,23**	-0,07	-0,05	-0,08	-0,06

2. Palier de Turbo analyse : T20

	OM	P1	P2	**P3**	P4	P5	P6	P7
2.1 N Rép. T20	23	8	11	**2429**	8	11	5	8
2.2 % Rép. T20	1%	0%	0%	**97%**	0%	0%	0%	0%
2.3 C. Moy. T20	26%	38%	73%	**87%**	53%	65%	32%	55%
2.4 rpbis SC T20	-0,20	-0,09	-0,03	**0,23**	-0,07	-0,05	-0,08	-0,06

3. Palier de Turbo analyse : T30

	OM	P1	P2	**P3**	P4	P5	P6	P7
3.1 N Rép. T30	23	7	11	**2425**	8	10	5	8
3.2 % Rép. T30	1%	0%	0%	**97%**	0%	0%	0%	0%
3.3 C. Moy. T30	26%	34%	73%	**87%**	53%	62%	32%	55%
3.4 rpbis SC T30	-0,20	-0,10	-0,03	**0,24**	-0,07	-0,05	-0,08	-0,06

4. Palier de Turbo analyse : T40

	OM	P1	P2	**P3**	P4	P5	P6	P7
4.1 N Rép. T40	23	7	10	**2406**	8	9	5	7
4.2 % Rép. T40	1%	0%	0%	**97%**	0%	0%	0%	0%
4.3 C. Moy. T40	26%	34%	74%	**88%**	53%	58%	32%	49%
4.4 rpbis SC T40	-0,20	-0,10	-0,03	**0,25**	-0,07	-0,06	-0,09	-0,07

5. Palier de Turbo analyse : T50

	OM	P1	P2	**P3**	P4	P5	P6	P7
5.1 N Rép. T50	20	7	8	**2315**	7	9	4	5
5.2 % Rép. T50	1%	0%	0%	**97%**	0%	0%	0%	0%
5.3 C. Moy. T50	20%	34%	68%	**88%**	46%	58%	15%	28%
5.4 rpbis SC T50	-0,22	-0,10	-0,04	**0,28**	-0,08	-0,06	-0,10	-0,10

6. Palier de Turbo analyse : T60

	OM	P1	P2	**P3**	P4	P5	P6	P7
6.1 N Rép. T60	17	5	8	**2055**	4	8	4	4
6.2 % Rép. T60	1%	0%	0%	**98%**	0%	0%	0%	0%
6.3 C. Moy. T60	6%	40%	69%	**90%**	15%	53%	15%	35%
6.4 rpbis SC T60	-0,28	-0,09	-0,05	**0,33**	-0,12	-0,08	-0,12	-0,09

7. Palier de Turbo analyse : T70

	OM	P1	P2	**P3**	P4	P5	P6	P7
7.1 N Rép. T70	12	4	5	**1569**	4	5	4	3
7.2 % Rép. T70	1%	0%	0%	**98%**	0%	0%	0%	0%
7.3 C. Moy. T70	0%	20%	60%	**91%**	20%	28%	20%	30%
7.4 rpbis SC T70	xxxx	-0,10	-0,07	**0,40**	-0,15	-0,13	-0,15	-0,12

8. Palier de Turbo analyse : T80

	OM	P1	P2	**P3**	P4	P5	P6	P7
8.1 N Rép. T80	10	1	1	**834**	3	3	1	1
8.2 % Rép. T80	1%	0%	0%	**98%**	0%	0%	0%	0%
8.3 C. Moy. T80	0%	xxxx	xxxx	**92%**	20%	33%	0%	20%
8.4 rpbis SC T80	xxxx	xxxx	xxxx	**0,50**	-0,17	-0,14	xxxx	-0,10

9. Palier de Turbo analyse : T90

	OM	P1	P2	**P3**	P4	P5	P6	P7
9.1 N Rép. T90	4	0	0	**197**	2	2	1	0
9.2 % Rép. T90	2%	0%	0%	**96%**	1%	1%	0%	0%
9.3 C. Moy. T90	0%	xxxx	xxxx	**92%**	60%	0%	0%	xxxx
9.4 rpbis SC T90	xxxx	xxxx	xxxx	**0,62**	-0,21	xxxx	xxxx	xxxx

17. Question 1 de l'épreuve de Chimie - [Ch]q1

a) Statistiques classiques

	OM	P1	P2	P3	**P4**	P5	P6	P7
1) N Rép.	16	32	18	81	**2244**	26	82	2
2) % Rép.	1%	1%	1%	3%	**90%**	1%	3%	0%
3) rpbis	-0,13	-0,06	-0,10	-0,16	**0,32**	-0,08	-0,19	-0,02

b) Statistiques spectrales

	OM	P1	P2	P3	**P4**	P5	P6	P7
1) C Moy.	50%	63%	47%	70%	**82%**	65%	46%	100%
2) rpbis SC	-0,08	-0,07	-0,10	-0,07	**0,23**	-0,06	-0,20	0,02

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	P3	**P4**	P5	P6	P7
1.1 N Rép. T10	16	32	18	81	**2243**	26	82	2
1.2 % Rép. T10	1%	1%	1%	3%	**90%**	1%	3%	0%
1.3 C. Moy. T10	50%	63%	47%	70%	**82%**	65%	46%	100%
1.4 rpbis SC T10	-0,08	-0,07	-0,10	-0,07	**0,23**	-0,06	-0,20	0,02

2. Palier de Turbo analyse : T20

	OM	P1	P2	P3	**P4**	P5	P6	P7
2.1 N Rép. T20	16	31	18	81	**2235**	26	82	2
2.2 % Rép. T20	1%	1%	1%	3%	**90%**	1%	3%	0%
2.3 C. Moy. T20	50%	61%	47%	70%	**82%**	65%	46%	100%
2.4 rpbis SC T20	-0,08	-0,07	-0,10	-0,07	**0,24**	-0,06	-0,21	0,02

3. Palier de Turbo analyse : T30

	OM	P1	P2	P3	**P4**	P5	P6	P7
3.1 N Rép. T30	15	31	18	80	**2222**	26	82	2
3.2 % Rép. T30	1%	1%	1%	3%	**90%**	1%	3%	0%
3.3 C. Moy. T30	47%	61%	47%	70%	**83%**	65%	46%	100%
3.4 rpbis SC T30	-0,09	-0,08	-0,10	-0,08	**0,25**	-0,06	-0,21	0,02

4. Palier de Turbo analyse : T40

	OM	P1	P2	P3	**P4**	P5	P6	P7
4.1 N Rép. T40	14	29	18	78	**2194**	25	80	2
4.2 % Rép. T40	1%	1%	1%	3%	**90%**	1%	3%	0%
4.3 C. Moy. T40	43%	60%	47%	69%	**84%**	63%	46%	100%
4.4 rpbis SC T40	-0,10	-0,09	-0,11	-0,09	**0,27**	-0,07	-0,23	0,02

5. Palier de Turbo analyse : T50

	OM	P1	P2	P3	**P4**	P5	P6	P7
5.1 N Rép. T50	14	26	18	71	**2154**	24	74	2
5.2 % Rép. T50	1%	1%	1%	3%	**90%**	1%	3%	0%
5.3 C. Moy. T50	43%	62%	47%	68%	**84%**	63%	42%	100%
5.4 rpbis SC T50	-0,11	-0,08	-0,11	-0,09	**0,28**	-0,07	-0,24	0,02

6. Palier de Turbo analyse : T60

	OM	P1	P2	P3	**P4**	P5	P6	P7
6.1 N Rép. T60	13	23	14	59	**1960**	20	70	2
6.2 % Rép. T60	1%	1%	1%	3%	**91%**	1%	3%	0%
6.3 C. Moy. T60	38%	63%	41%	65%	**86%**	55%	41%	100%
6.4 rpbis SC T60	-0,13	-0,08	-0,13	-0,12	**0,34**	-0,10	-0,28	0,02

7. Palier de Turbo analyse : T70

	OM	P1	P2	P3	**P4**	P5	P6	P7
7.1 N Rép. T70	12	17	11	38	**1622**	14	50	1
7.2 % Rép. T70	1%	1%	1%	2%	**92%**	1%	3%	0%
7.3 C. Moy. T70	33%	59%	32%	51%	**87%**	70%	38%	100%
7.4 rpbis SC T70	-0,17	-0,11	-0,17	-0,20	**0,42**	-0,06	-0,31	0,01

8. Palier de Turbo analyse : T80

	OM	P1	P2	P3	**P4**	P5	P6	P7
8.1 N Rép. T80	9	7	5	20	**942**	6	26	0
8.2 % Rép. T80	1%	1%	0%	2%	**93%**	1%	3%	0%
8.3 C. Moy. T80	13%	46%	20%	52%	**89%**	70%	32%	xxxx
8.4 rpbis SC T80	-0,25	-0,14	-0,19	-0,21	**0,51**	-0,06	-0,37	xxxx

9. Palier de Turbo analyse : T90

	OM	P1	P2	P3	**P4**	P5	P6	P7
9.1 N Rép. T90	4	2	2	5	**260**	2	9	0
9.2 % Rép. T90	1%	1%	1%	2%	**92%**	1%	3%	0%
9.3 C. Moy. T90	0%	30%	0%	40%	**92%**	30%	22%	xxxx
9.4 rpbis SC T90	xxxx	-0,19	xxxx	-0,24	**0,70**	-0,19	-0,44	xxxx

18. Question 17 de l'épreuve de Mathématique - [M]q17

a) Statistiques classiques

	OM	P1	P2	P3	**P4**	P5	P6	P7
1) N Rép.	81	98	113	199	**1776**	149	99	1
2) % Rép.	3%	4%	4%	8%	**71%**	6%	4%	0%
3) rpbis	-0,07	-0,19	-0,22	-0,15	**0,43**	-0,15	-0,12	-0,03

b) Statistiques spectrales

	OM	P1	P2	P3	**P4**	P5	P6	P7
1) C Moy.	40%	57%	60%	68%	**78%**	64%	49%	0%
2) rpbis SC	-0,20	-0,13	-0,12	-0,08	**0,27**	-0,10	-0,17	xxxx

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	P3	**P4**	P5	P6	P7
1.1 N Rép. T10	81	98	113	199	**1774**	149	99	1
1.2 % Rép. T10	3%	4%	4%	8%	**71%**	6%	4%	0%
1.3 C. Moy. T10	40%	57%	60%	68%	**78%**	64%	49%	0%
1.4 rpbis SC T10	-0,20	-0,13	-0,12	-0,09	**0,27**	-0,10	-0,17	xxxx

2. Palier de Turbo analyse : T20

	OM	P1	P2	P3	**P4**	P5	P6	P7
2.1 N Rép. T20	81	97	113	199	**1767**	149	98	1
2.2 % Rép. T20	3%	4%	5%	8%	**71%**	6%	4%	0%
2.3 C. Moy. T20	40%	57%	60%	68%	**79%**	64%	50%	0%
2.4 rpbis SC T20	-0,21	-0,13	-0,12	-0,09	**0,28**	-0,10	-0,17	xxxx

3. Palier de Turbo analyse : T30

	OM	P1	P2	P3	**P4**	P5	P6	P7
3.1 N Rép. T30	81	97	113	199	**1746**	145	97	1
3.2 % Rép. T30	3%	4%	5%	8%	**70%**	6%	4%	0%
3.3 C. Moy. T30	40%	57%	60%	68%	**80%**	66%	50%	0%
3.4 rpbis SC T30	-0,22	-0,14	-0,13	-0,10	**0,29**	-0,10	-0,18	xxxx

4. Palier de Turbo analyse : T40

	OM	P1	P2	P3	**P4**	P5	P6	P7
4.1 N Rép. T40	81	95	109	199	**1724**	143	96	1
4.2 % Rép. T40	3%	4%	4%	8%	**70%**	6%	4%	0%
4.3 C. Moy. T40	40%	58%	59%	68%	**81%**	66%	51%	0%
4.4 rpbis SC T40	-0,23	-0,14	-0,14	-0,11	**0,31**	-0,11	-0,18	xxxx

5. Palier de Turbo analyse : T50

	OM	P1	P2	P3	**P4**	P5	P6	P7
5.1 N Rép. T50	80	93	107	195	**1706**	140	92	1
5.2 % Rép. T50	3%	4%	4%	8%	**71%**	6%	4%	0%
5.3 C. Moy. T50	40%	58%	58%	68%	**81%**	68%	52%	0%
5.4 rpbis SC T50	-0,24	-0,15	-0,15	-0,11	**0,32**	-0,11	-0,18	xxxx

6. Palier de Turbo analyse : T60

	OM	P1	P2	P3	**P4**	P5	P6	P7
6.1 N Rép. T60	76	89	95	180	**1661**	131	83	1
6.2 % Rép. T60	3%	4%	4%	8%	**72%**	6%	4%	0%
6.3 C. Moy. T60	39%	59%	59%	69%	**83%**	67%	51%	0%
6.4 rpbis SC T60	-0,26	-0,15	-0,15	-0,12	**0,34**	-0,12	-0,20	xxxx

7. Palier de Turbo analyse : T70

	OM	P1	P2	P3	**P4**	P5	P6	P7
7.1 N Rép. T70	64	59	76	144	**1489**	95	70	1
7.2 % Rép. T70	3%	3%	4%	7%	**75%**	5%	4%	0%
7.3 C. Moy. T70	41%	53%	60%	67%	**84%**	66%	48%	0%
7.4 rpbis SC T70	-0,26	-0,18	-0,15	-0,14	**0,38**	-0,13	-0,23	xxxx

8. Palier de Turbo analyse : T80

	OM	P1	P2	P3	**P4**	P5	P6	P7
8.1 N Rép. T80	38	30	36	89	**1051**	42	38	1
8.2 % Rép. T80	3%	2%	3%	7%	**79%**	3%	3%	0%
8.3 C. Moy. T80	40%	49%	61%	65%	**86%**	63%	45%	0%
8.4 rpbis SC T80	-0,29	-0,20	-0,15	-0,20	**0,45**	-0,15	-0,25	xxxx

9. Palier de Turbo analyse : T90

	OM	P1	P2	P3	**P4**	P5	P6	P7
9.1 N Rép. T90	6	2	11	9	**326**	9	8	0
9.2 % Rép. T90	2%	1%	3%	2%	**88%**	2%	2%	0%
9.3 C. Moy. T90	20%	60%	64%	69%	**90%**	60%	50%	xxxx
9.4 rpbis SC T90	-0,34	-0,10	-0,20	-0,14	**0,48**	-0,22	-0,26	xxxx

19. Question [H]q3 Après rectification

a) Statistiques classiques

	OM	P1	P2	P3	P4	P5	**P6**
1) N Rép.	86	0	132	197	104	258	**629**
2) % Rép.	6%	0%	9%	14%	7%	18%	**45%**
3) rpbis	-0,22	xxxx	-0,06	-0,14	-0,09	-0,20	**0,45**

b) Statistiques spectrales

	OM	P1	P2	P3	P4	P5	**P6**
1) C Moy.	4%	xxxx	30%	31%	30%	42%	**53%**
2) rpbis SC	-0,33	xxxx	-0,19	-0,21	-0,17	-0,12	**0,30**

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	P3	P4	P5	**P6**
1.1 N Rép. T10	86	0	132	197	104	258	**628**
1.2 % Rép. T10	6%	0%	9%	14%	7%	18%	**45%**
1.3 C. Moy. T10	4%	xxxx	30%	31%	30%	42%	**53%**
1.4 rpbis SC T10	-0,33	xxxx	-0,19	-0,21	-0,17	-0,12	**0,30**

2. Palier de Turbo analyse : T20

	OM	P1	P2	P3	P4	P5	**P6**
2.1 N Rép. T20	86	0	132	196	104	258	**628**
2.2 % Rép. T20	6%	0%	9%	14%	7%	18%	**45%**
2.3 C. Moy. T20	4%	xxxx	30%	31%	30%	42%	**53%**
2.4 rpbis SC T20	-0,33	xxxx	-0,19	-0,21	-0,17	-0,12	**0,31**

3. Palier de Turbo analyse : T30

	OM	P1	P2	P3	P4	P5	**P6**
3.1 N Rép. T30	86	0	132	196	104	257	**626**
3.2 % Rép. T30	6%	0%	9%	14%	7%	18%	**45%**
3.3 C. Moy. T30	4%	xxxx	30%	31%	30%	42%	**53%**
3.4 rpbis SC T30	-0,33	xxxx	-0,19	-0,22	-0,17	-0,12	**0,31**

4. Palier de Turbo analyse : T40

	OM	P1	P2	P3	P4	P5	**P6**
4.1 N Rép. T40	85	0	131	196	102	257	**625**
4.2 % Rép. T40	6%	0%	9%	14%	7%	18%	**45%**
4.3 C. Moy. T40	4%	xxxx	30%	31%	29%	42%	**53%**
4.4 rpbis SC T40	-0,33	xxxx	-0,19	-0,22	-0,18	-0,13	**0,31**

5. Palier de Turbo analyse : T50

	OM	P1	P2	P3	P4	P5	**P6**
5.1 N Rép. T50	82	0	130	196	102	256	**621**
5.2 % Rép. T50	6%	0%	9%	14%	7%	18%	**45%**
5.3 C. Moy. T50	3%	xxxx	30%	31%	29%	42%	**53%**
5.4 rpbis SC T50	-0,34	xxxx	-0,19	-0,22	-0,18	-0,13	**0,31**

6. Palier de Turbo analyse : T60

	OM	P1	P2	P3	P4	P5	**P6**
6.1 N Rép. T60	81	0	124	188	98	248	**602**
6.2 % Rép. T60	6%	0%	9%	14%	7%	18%	**45%**
6.3 C. Moy. T60	3%	xxxx	31%	31%	28%	42%	**54%**
6.4 rpbis SC T60	-0,35	xxxx	-0,19	-0,23	-0,19	-0,13	**0,33**

7. Palier de Turbo analyse : T70

	OM	P1	P2	P3	P4	P5	**P6**
7.1 N Rép. T70	68	0	113	166	87	215	**542**
7.2 % Rép. T70	6%	0%	9%	14%	7%	18%	**45%**
7.3 C. Moy. T70	2%	xxxx	30%	31%	28%	40%	**58%**
7.4 rpbis SC T70	-0,36	xxxx	-0,22	-0,26	-0,21	-0,19	**0,38**

8. Palier de Turbo analyse : T80

	OM	P1	P2	P3	P4	P5	**P6**
8.1 N Rép. T80	41	0	69	99	54	119	**337**
8.2 % Rép. T80	6%	0%	10%	14%	7%	17%	**47%**
8.3 C. Moy. T80	1%	xxxx	29%	33%	31%	36%	**62%**
8.4 rpbis SC T80	-0,41	xxxx	-0,28	-0,29	-0,25	-0,27	**0,47**

9. Palier de Turbo analyse : T90

	OM	P1	P2	P3	P4	P5	**P6**
9.1 N Rép. T90	9	0	13	18	14	28	**62**
9.2 % Rép. T90	6%	0%	9%	12%	10%	19%	**43%**
9.3 C. Moy. T90	0%	xxxx	32%	45%	42%	33%	**71%**
9.4 rpbis SC T90	xxxx	xxxx	-0,30	-0,26	-0,27	-0,44	**0,55**

20. Question [H]q20 – Après rectification

a) Statistiques classiques

	OM	P1	P2	P3	P4	P5	**P6**
1) N Rép.	58	263	39	256	20	7	**744**
2) % Rép.	4%	19%	3%	18%	1%	0%	**53%**
3) rpbis	-0,09	-0,22	-0,08	-0,06	-0,01	-0,09	**0,31**

b) Statistiques spectrales

	OM	P1	P2	P3	P4	P5	**P6**
1) C Moy.	7%	30%	24%	46%	38%	37%	**82%**
2) rpbis SC	-0,38	-0,52	-0,25	-0,35	-0,13	-0,08	**0,59**

c) Statistiques spectrales après turbo analyse

1. Palier de Turbo analyse : T10

	OM	P1	P2	P3	P4	P5	**P6**
1.1 N Rép. T10	58	263	39	256	20	7	**743**
1.2 % Rép. T10	4%	19%	3%	18%	1%	0%	**53%**
1.3 C. Moy. T10	7%	30%	24%	46%	38%	37%	**83%**
1.4 rpbis SC T10	-0,38	-0,52	-0,25	-0,36	-0,13	-0,08	**0,59**

2. Palier de Turbo analyse : T20

	OM	P1	P2	P3	P4	P5	**P6**
2.1 N Rép. T20	58	263	39	256	20	7	**742**
2.2 % Rép. T20	4%	19%	3%	18%	1%	0%	**53%**
2.3 C. Moy. T20	7%	30%	24%	46%	38%	37%	**83%**
2.4 rpbis SC T20	-0,38	-0,52	-0,25	-0,36	-0,13	-0,08	**0,59**

3. Palier de Turbo analyse : T30

	OM	P1	P2	P3	P4	P5	**P6**
3.1 N Rép. T30	58	262	39	256	20	6	**741**
3.2 % Rép. T30	4%	19%	3%	18%	1%	0%	**53%**
3.3 C. Moy. T30	7%	30%	24%	46%	38%	27%	**83%**
3.4 rpbis SC T30	-0,38	-0,52	-0,25	-0,36	-0,14	-0,09	**0,60**

4. Palier de Turbo analyse : T40

	OM	P1	P2	P3	P4	P5	**P6**
4.1 N Rép. T40	58	260	39	256	20	6	**738**
4.2 % Rép. T40	4%	19%	3%	18%	1%	0%	**53%**
4.3 C. Moy. T40	7%	30%	24%	46%	38%	27%	**83%**
4.4 rpbis SC T40	-0,39	-0,53	-0,25	-0,36	-0,14	-0,09	**0,61**

5. Palier de Turbo analyse : T50

	OM	P1	P2	P3	P4	P5	**P6**
5.1 N Rép. T50	57	260	39	253	20	6	**733**
5.2 % Rép. T50	4%	19%	3%	18%	1%	0%	**53%**
5.3 C. Moy. T50	7%	30%	24%	46%	38%	27%	**83%**
5.4 rpbis SC T50	-0,39	-0,54	-0,25	-0,36	-0,14	-0,10	**0,61**

6. Palier de Turbo analyse : T60

	OM	P1	P2	P3	P4	P5	**P6**
6.1 N Rép. T60	55	248	37	245	19	5	**715**
6.2 % Rép. T60	4%	18%	3%	18%	1%	0%	**53%**
6.3 C. Moy. T60	7%	30%	22%	46%	39%	24%	**84%**
6.4 rpbis SC T60	-0,39	-0,55	-0,26	-0,38	-0,14	-0,09	**0,63**

7. Palier de Turbo analyse : T70

	OM	P1	P2	P3	P4	P5	**P6**
7.1 N Rép. T70	47	225	32	207	17	3	**646**
7.2 % Rép. T70	4%	19%	3%	17%	1%	0%	**54%**
7.3 C. Moy. T70	6%	30%	23%	45%	44%	20%	**86%**
7.4 rpbis SC T70	-0,40	-0,56	-0,27	-0,39	-0,13	-0,09	**0,65**

8. Palier de Turbo analyse : T80

	OM	P1	P2	P3	P4	P5	**P6**
8.1 N Rép. T80	27	131	19	114	14	3	**404**
8.2 % Rép. T80	4%	18%	3%	16%	2%	0%	**56%**
8.3 C. Moy. T80	4%	30%	19%	42%	40%	20%	**88%**
8.4 rpbis SC T80	-0,43	-0,61	-0,30	-0,47	-0,18	-0,12	**0,75**

9. Palier de Turbo analyse : T90

	OM	P1	P2	P3	P4	P5	**P6**
9.1 N Rép. T90	4	19	2	18	3	1	**97**
9.2 % Rép. T90	3%	13%	1%	12%	2%	1%	**67%**
9.3 C. Moy. T90	0%	23%	20%	31%	20%	60%	**85%**
9.4 rpbis SC T90	xxxx	-0,60	-0,25	-0,47	-0,24	-0,06	**0,79**

E. Table du t de Student

Ddl	0,10	0,05	0,02	0,01	0,001
1	6,314	12,706	31,821	63,657	636,619
2	2,920	4,303	6,965	9,925	31,598
3	2,353	3,182	4,541	5,841	12,924
4	2,132	2,776	3,747	4,604	8,610
5	2,015	2,571	3,365	4,032	6,869
6	1,943	2,447	3,143	3,707	5,959
7	1,895	2,365	2,998	3,499	5,408
8	1,860	2,306	2,896	3,355	5,041
9	1,833	2,262	2,821	3,250	4,781
10	1,812	2,228	2,764	3,169	4,587
11	1,796	2,201	2,718	3,106	4,437
12	1,782	2,179	2,681	3,055	4,318
13	1,771	2,160	2,650	3,012	4,221
14	1,761	2,145	2,624	2,977	4,140
15	1,753	2,131	2,602	2,947	4,073
16	1,746	2,120	2,583	2,921	4,015
17	1,740	2,110	2,567	2,898	3,965
18	1,734	2,101	2,552	2,878	3,922
19	1,729	2,093	2,539	2,861	3,883
20	1,725	2,086	2,528	2,845	3,850
21	1,721	2,080	2,518	2,831	3,819
22	1,717	2,074	2,508	2,819	3,792
23	1,714	2,069	2,500	2,807	3,767
24	1,711	2,064	2,492	2,797	3,745
25	1,708	2,060	2,485	2,787	3,725
26	1,706	2,056	2,479	2,779	3,707
27	1,703	2,052	2,473	2,771	3,690
28	1,701	2,048	2,467	2,763	3,674
29	1,699	2,045	2,462	2,756	3,659
30	1,697	2,042	2,457	2,750	3,646
∞	1,645	1,960	2,326	2,576	3,291

F. Tableaux des valeurs obtenues par les 173 QCM aux rpbis classiques, rpbis SC, rpbis SCT80 et rpbis SCT90

QCM	NC/q	rpbis classiques							NC/q	rpbis Spectraux Contrastés (*rpbis SC*)							NC/q	rpbis SC calculés au palier turbo T80 (*rpbis SCT80*)							NC/q	rpbis SC calculés au palier turbo T90 (*rpbis SCT90*)						
	70	OM	P1	P2	P3	P4	P5	P6	P7	70	OM	P1	P2	P3	P4	P5	P6	P7	80	OM	P1	P2	P3	P4	P5	P6	P7	90	OM	P1	P2	P3	P4	P5	P6	P7
Mq1	0,465	-0,20	-0,12	**0,33**	-0,08	-0,19	-0,05	-0,01	-0,04	0,568	-0,26	-0,21	**0,35**	-0,13	-0,23	-0,10	-0,06	-0,03	0,737	-0,35	-0,28	**0,47**	-0,14	-0,27	-0,15	-0,10	-0,04	0,912	-0,51	-0,33	**0,58**	-0,18	-0,27	-0,13	xxxx	-0,04
Mq2	0,173	-0,15	**0,16**	-0,06	-0,08	0,02	-0,13	-0,03	-0,01	0,248	-0,23	**0,15**	-0,06	-0,12	-0,09	-0,12	-0,14	-0,05	0,421	-0,43	-0,34	**0,24**	-0,11	-0,22	-0,17	-0,16	-0,02	0,670	-0,58	**0,38**	-0,11	-0,21	-0,28	-0,25	-0,17	-0,09
Mq3	0,200	-0,15	-0,10	0,04	**0,17**	-0,09	-0,04	-0,02	-0,03	0,291	-0,33	-0,12	-0,02	**0,18**	-0,09	-0,10	-0,12	-0,05	0,450	-0,43	-0,19	-0,06	**0,27**	-0,17	-0,16	-0,12	-0,07	0,604	-0,58	-0,27	**0,35**	-0,19	-0,20	-0,19	xxxx	-0,04
Mq4	0,334	-0,14	-0,16	-0,05	-0,07	**0,26**	-0,13	-0,05	-0,05	0,378	-0,25	-0,12	-0,13	-0,12	**0,25**	-0,14	-0,09	0,02	0,570	-0,36	-0,16	-0,18	-0,20	**0,37**	-0,19	-0,18	-0,12	0,646	-0,47	-0,11	-0,18	-0,20	**0,42**	-0,22	-0,19	-0,04
Mq5	0,165	-0,15	-0,07	0,01	-0,07	0,10	**0,15**	0,02	-0,06	0,217	-0,23	-0,03	0,00	-0,05	-0,04	**0,01**	-0,07	-0,05	0,337	-0,29	-0,15	-0,04	-0,11	-0,12	**0,21**	0,01	-0,08	0,469	-0,56	-0,18	-0,06	-0,09	-0,15	**0,13**	-0,01	xxxx
Mq6	0,428	-0,18	-0,06	-0,14	-0,18	**0,26**	-0,01	-0,04	-0,01	0,204	-0,17	-0,06	-0,10	-0,07	-0,05	0,02	-0,07	-0,05	0,337	-0,29	-0,07	-0,10	-0,12	**0,21**	0,02	-0,08	-0,04	0,469	-0,56	xxxx	-0,09	-0,15	-0,17	**0,29**	xxxx	xxxx
Mq7	0,351	-0,16	-0,05	-0,12	**0,28**	-0,08	-0,06	-0,02	-0,06	0,455	-0,32	-0,14	-0,13	**0,28**	-0,19	-0,16	-0,17	-0,07	0,659	-0,43	-0,22	-0,18	**0,41**	-0,27	-0,21	-0,19	-0,08	0,763	-0,57	-0,28	-0,17	**0,47**	-0,25	-0,20	-0,20	-0,06
Mq8	0,330	-0,22	-0,16	-0,02	**0,26**	-0,16	-0,02	-0,05	-0,04	0,308	-0,24	-0,13	-0,14	-0,13	**0,21**	-0,16	-0,04	-0,05	0,495	-0,36	-0,22	-0,11	-0,18	-0,16	**0,34**	-0,22	-0,11	0,680	-0,51	-0,45	-0,16	-0,17	-0,15	-0,16	-0,17	-0,09
Mq9	0,218	-0,13	-0,05	-0,04	-0,06	**0,18**	-0,02	0,03	-0,03	0,207	-0,28	-0,09	-0,07	-0,07	**0,13**	-0,07	-0,10	0,01	0,385	-0,40	-0,13	-0,14	-0,15	**0,24**	-0,12	-0,14	-0,04	0,646	-0,58	-0,21	-0,25	-0,18	**0,40**	-0,18	-0,14	-0,14
Mq10	0,821	-0,19	-0,10	-0,06	-0,38	-0,02	**0,54**	-0,06	-0,04	0,505	-0,29	-0,16	-0,15	-0,22	-0,12	**0,30**	-0,16	-0,07	0,767	-0,41	-0,20	-0,19	-0,14	-0,15	**0,24**	-0,12	-0,14	0,970	-0,59	-0,25	-0,14	-0,42	-0,17	**0,59**	-0,24	0,01
Mq11	0,488	-0,22	-0,15	-0,11	-0,09	**0,37**	0,01	-0,05	-0,04	0,498	-0,31	-0,16	-0,20	-0,16	-0,05	**0,31**	-0,19	-0,06	0,670	-0,43	-0,22	-0,26	-0,21	-0,18	**0,43**	-0,23	-0,10	0,805	-0,57	-0,18	-0,24	-0,25	-0,21	**0,52**	-0,28	-0,10
Mq12	0,335	-0,19	-0,13	-0,08	-0,01	0,03	**0,29**	-0,05	-0,04	0,065	-0,25	-0,11	-0,07	-0,10	0,09	-0,10	**0,05**	-0,03	0,250	-0,38	-0,18	-0,17	-0,18	0,01	-0,16	**0,15**	-0,06	0,488	xxxx	-0,15	-0,27	-0,29	-0,12	-0,19	**0,29**	-0,06
Mq13	0,441	-0,19	-0,11	-0,10	-0,06	**0,34**	-0,08	-0,04	-0,04	0,411	-0,35	-0,17	-0,11	-0,09	-0,11	**0,24**	-0,17	-0,04	0,612	-0,45	-0,23	-0,14	-0,13	-0,13	**0,35**	-0,26	-0,02	0,839	-0,67	-0,29	-0,18	-0,12	-0,16	**0,48**	-0,32	0,01
Mq14	0,452	-0,19	-0,15	**0,26**	-0,14	-0,03	-0,11	0,02	-0,06	0,376	-0,28	-0,12	**0,26**	-0,09	-0,10	-0,08	-0,11	-0,06	0,533	-0,35	-0,17	**0,36**	-0,12	-0,15	-0,12	-0,13	-0,09	0,780	-0,56	-0,23	**0,52**	-0,11	-0,13	-0,21	-0,16	-0,09
Mq15	0,463	-0,21	-0,16	-0,14	-0,09	**0,33**	-0,10	-0,10	-0,01	0,573	-0,32	-0,16	-0,13	-0,11	**0,39**	-0,08	-0,19	-0,02	0,806	-0,42	-0,20	-0,14	-0,10	**0,54**	-0,13	-0,29	-0,07	0,963	-0,55	-0,24	-0,21	-0,06	**0,63**	-0,09	-0,26	-0,09
Mq16	0,493	-0,22	-0,14	-0,09	**0,39**	-0,08	-0,13	-0,03	-0,04	0,526	-0,33	-0,13	-0,17	-0,06	**0,34**	-0,08	-0,18	-0,04	0,733	-0,43	-0,23	-0,21	-0,21	**0,47**	-0,23	-0,21	-0,08	0,965	-0,66	-0,30	-0,24	-0,21	**0,61**	-0,24	-0,26	-0,10
Mq17	0,397	-0,18	0,05	-0,14	-0,10	-0,14	-0,05	**0,31**	-0,01	0,223	-0,31	0,00	-0,10	-0,17	**0,34**	-0,18	-0,13	-0,08	0,427	-0,43	-0,04	-0,20	-0,21	**0,47**	-0,13	-0,04	-0,07	0,731	-0,59	-0,11	-0,30	-0,18	-0,24	**0,45**	-0,35	-0,05
Mq18	0,540	-0,20	**0,41**	-0,20	-0,14	-0,10	-0,07	-0,06	-0,07	0,557	-0,35	**0,34**	-0,21	-0,18	-0,13	-0,08	-0,23	-0,07	0,785	-0,45	**0,48**	-0,31	-0,27	-0,18	-0,11	-0,31	-0,09	0,965	xxxx	**0,62**	-0,37	-0,30	-0,23	-0,13	-0,38	-0,14
Mq19	0,616	-0,19	**0,42**	-0,13	-0,25	-0,18	-0,10	-0,04	-0,04	0,547	-0,33	-0,11	-0,20	-0,14	-0,09	-0,17	-0,05	-0,05	0,765	-0,50	-0,16	-0,31	-0,16	-0,19	-0,15	-0,20	-0,05	0,896	-0,54	**0,58**	-0,32	-0,20	-0,09	-0,15	-0,06	-0,03
Mq20	0,439	-0,18	-0,05	-0,10	0,01	-0,15	-0,12	**0,35**	-0,03	0,458	-0,38	-0,18	-0,17	-0,12	-0,17	-0,15	**0,28**	-0,05	0,630	-0,48	-0,24	-0,21	-0,21	-0,17	-0,22	-0,19	**0,39**	0,887	-0,67	-0,32	-0,24	-0,27	-0,23	-0,25	**0,55**	-0,03
Mq21	0,507	-0,23	-0,21	-0,09	**0,38**	-0,06	-0,03	-0,04	-0,05	0,504	-0,35	-0,20	-0,14	**0,32**	-0,11	-0,15	-0,17	-0,09	0,727	-0,49	-0,28	-0,21	**0,47**	-0,14	-0,21	-0,21	-0,12	0,943	-0,67	-0,35	-0,23	**0,60**	-0,16	-0,17	-0,23	-0,17
Mq22	0,594	-0,27	-0,20	**0,41**	-0,17	-0,11	-0,14	-0,06	-0,02	0,522	-0,31	-0,23	**0,29**	-0,23	-0,14	-0,09	-0,14	-0,06	0,776	-0,44	-0,31	**0,43**	-0,34	-0,24	-0,19	-0,15	-0,09	0,973	xxxx	-0,29	**0,57**	-0,23	-0,16	-0,23	-0,17	-0,23
Mq23	0,559	-0,29	**0,40**	-0,19	-0,13	-0,17	-0,11	-0,04	-0,03	0,568	-0,31	**0,32**	-0,21	-0,25	-0,23	-0,10	-0,06	-0,03	0,832	-0,43	**0,48**	-0,27	-0,37	-0,24	-0,22	-0,17	-0,12	1,059	-0,66	**0,62**	-0,29	-0,43	-0,31	-0,32	-0,23	-0,17
Mq24	0,629	-0,28	**0,42**	-0,19	-0,19	-0,13	-0,20	-0,09	-0,04	0,743	-0,32	**0,45**	-0,25	-0,34	-0,17	-0,23	-0,14	-0,07	0,962	-0,42	**0,58**	-0,31	-0,41	-0,27	-0,23	-0,21	-0,14	1,007	xxxx	**0,66**	-0,31	-0,39	-0,27	-0,29	-0,23	-0,17
Mq25	0,682	-0,30	-0,24	**0,48**	-0,16	-0,14	-0,11	-0,06	-0,06	0,661	-0,34	-0,26	**0,37**	-0,30	-0,23	-0,16	-0,20	-0,05	0,949	-0,45	-0,36	**0,54**	-0,43	-0,32	-0,27	-0,24	-0,17	1,031	xxxx	-0,29	**0,63**	-0,37	-0,29	-0,32	-0,24	-0,20
Mq26	0,714	-0,28	**0,48**	-0,23	-0,22	-0,18	-0,14	-0,08	-0,07	0,720	-0,37	**0,43**	-0,24	-0,30	-0,23	-0,18	-0,20	-0,07	0,961	-0,48	**0,58**	-0,31	-0,39	-0,27	-0,23	-0,21	-0,14	1,060	xxxx	**0,69**	-0,30	-0,40	-0,27	-0,29	-0,23	-0,17
Mq27	0,162	-0,24	0,13	-0,09	**0,13**	-0,07	-0,07	-0,03	-0,03	0,064	-0,27	-0,14	-0,06	**0,04**	-0,07	-0,05	-0,06	-0,03	0,218	-0,37	-0,12	-0,13	**0,13**	-0,09	-0,08	-0,07	-0,04	0,599	-0,61	-0,03	-0,21	**0,33**	-0,31	-0,31	-0,23	-0,16
Mq28	0,693	-0,26	**0,46**	-0,16	-0,27	**0,45**	-0,22	-0,12	-0,06	0,798	-0,36	**0,49**	-0,24	-0,34	-0,24	-0,13	-0,16	-0,07	1,015	-0,47	**0,62**	-0,30	-0,40	-0,27	-0,23	-0,21	-0,14	1,185	-0,65	**0,72**	-0,33	-0,31	-0,31	-0,31	-0,23	-0,17
Mq29	0,686	-0,26	-0,26	-0,17	**0,45**	-0,22	-0,12	-0,06	-0,04	0,546	-0,34	-0,17	-0,18	**0,34**	-0,20	-0,16	-0,17	-0,07	0,780	-0,45	-0,26	-0,22	**0,48**	-0,23	-0,22	-0,19	-0,14	0,860	xxxx	-0,24	-0,25	**0,62**	-0,22	-0,22	-0,19	-0,14
Mq30	0,782	-0,26	-0,32	**0,51**	-0,22	-0,16	-0,13	-0,06	-0,05	0,741	-0,35	-0,36	**0,45**	-0,20	-0,16	-0,20	-0,21	-0,08	0,964	-0,47	-0,43	-0,32	-0,41	-0,29	-0,25	-0,23	-0,16	1,236	-0,71	-0,40	**0,76**	-0,33	**0,53**	-0,39	-0,27	-0,21
Mq31	0,621	-0,24	-0,04	-0,30	**0,40**	-0,18	-0,14	-0,05	-0,04	0,343	-0,29	-0,07	-0,14	**0,27**	-0,15	-0,12	-0,14	-0,05	0,579	-0,40	-0,11	-0,25	**0,35**	-0,23	-0,19	-0,17	-0,12	0,892	-0,59	-0,22	-0,33	**0,53**	-0,39	-0,39	-0,27	-0,21
Mq32	0,732	-0,26	**0,48**	-0,15	-0,32	-0,18	-0,13	-0,06	-0,05	0,779	-0,31	**0,46**	-0,22	-0,38	-0,22	-0,16	-0,20	-0,07	1,029	-0,44	**0,62**	-0,27	-0,48	-0,29	-0,25	-0,23	-0,16	1,001	xxxx	**0,66**	-0,22	-0,39	-0,27	-0,29	-0,23	-0,17
Mq33	0,687	-0,28	-0,27	**0,46**	-0,18	-0,15	-0,10	-0,06	-0,04	0,697	-0,34	-0,29	**0,42**	-0,25	-0,18	-0,16	-0,18	-0,06	0,918	-0,46	-0,39	**0,57**	-0,48	-0,29	-0,25	-0,23	-0,16	1,126	-0,70	-0,36	**0,70**	-0,28	-0,27	-0,29	-0,23	-0,17
Mq34	0,638	-0,26	**0,52**	-0,15	-0,36	-0,16	-0,13	-0,06	-0,05	0,664	-0,33	**0,39**	-0,14	-0,30	-0,18	-0,17	-0,19	-0,06	0,993	-0,45	**0,58**	-0,17	-0,45	-0,29	-0,25	-0,23	-0,16	1,129	-0,66	**0,68**	-0,21	-0,41	-0,31	-0,31	-0,23	-0,17
Mq35	0,732	-0,24	-0,30	-0,15	**0,46**	-0,16	-0,13	-0,06	-0,05	0,407	-0,32	-0,16	-0,12	**0,24**	-0,15	-0,12	-0,13	-0,05	0,717	-0,43	-0,30	-0,18	**0,42**	-0,27	-0,23	-0,21	-0,14	0,946	-0,62	-0,36	-0,13	**0,56**	-0,41	-0,40	-0,29	-0,22
Mq36	0,780	-0,27	**0,51**	-0,16	-0,30	-0,18	-0,14	-0,06	-0,05	0,731	-0,35	**0,41**	-0,22	-0,35	-0,18	-0,15	-0,18	-0,06	1,015	-0,48	**0,58**	-0,28	-0,47	-0,32	-0,27	-0,24	-0,17	1,191	-0,67	**0,70**	-0,28	-0,49	-0,36	-0,37	-0,27	-0,22
Mq37	0,604	-0,25	-0,09	**0,39**	-0,22	-0,17	-0,13	-0,06	-0,05	0,504	-0,38	-0,15	**0,29**	-0,21	-0,16	-0,13	-0,16	-0,07	0,785	-0,50	-0,20	**0,45**	-0,33	-0,25	-0,22	-0,19	-0,14	0,984	xxxx	-0,18	**0,61**	-0,40	-0,29	-0,33	-0,25	-0,19
Mq38	0,221	-0,18	**0,22**	0,05	-0,11	-0,14	-0,11	-0,08	-0,08	0,353	-0,41	-0,19	**0,28**	-0,18	-0,16	-0,07	-0,21	-0,08	0,514	-0,46	**0,28**	-0,12	-0,27	-0,24	-0,14	-0,30	-0,11	0,866	-0,65	**0,45**	-0,43	-0,49	-0,25	-0,22	-0,35	xxxx
Mq39	0,260	-0,17	-0,08	**0,22**	-0,03	-0,08	0,04	-0,01	-0,02	0,461	-0,41	-0,19	**0,28**	-0,18	-0,16	-0,07	-0,21	-0,08	0,699	-0,52	-0,26	**0,40**	-0,27	-0,24	-0,14	-0,30	-0,11	0,972	-0,72	-0,28	**0,53**	-0,43	-0,49	-0,25	-0,22	-0,35
Mq40	0,468	-0,21	-0,10	-0,14	**0,37**	-0,05	-0,08	-0,03	-0,01	0,610	-0,42	-0,23	-0,22	**0,37**	-0,19	-0,13	-0,28	-0,07	0,820	-0,53	-0,29	-0,30	**0,49**	-0,29	-0,25	-0,23	-0,16	1,061	-0,71	-0,38	-0,34	**0,64**	-0,37	-0,12	-0,37	-0,07
Mq41	0,655	-0,15	-0,13	**0,41**	-0,20	-0,10	-0,13	-0,04	-0,03	0,663	-0,31	-0,14	**0,42**	-0,23	-0,22	-0,14	-0,08	-0,04	0,847	-0,39	-0,19	**0,54**	-0,29	-0,31	-0,26	-0,30	-0,14	0,967	-0,54	-0,21	**0,62**	-0,31	-0,12	-0,29	-0,36	-0,15
Mq42	0,467	-0,16	-0,14	-0,11	-0,16	**0,34**	-0,01	-0,03	-0,03	0,724	-0,33	-0,23	-0,23	-0,31	**0,46**	-0,20	-0,24	-0,06	0,889	-0,43	-0,28	-0,28	-0,37	**0,58**	-0,31	-0,28	-0,18	0,960	-0,60	-0,30	-0,29	-0,38	**0,67**	-0,27	-0,27	-0,21
Mq43	0,095	-0,21	0,06	0,03	**0,09**	-0,03	-0,05	-0,05	0,02	0,172	-0,31	0,04	0,00	**0,11**	-0,12	-0,08	-0,19	0,00	0,331	-0,42	-0,04	-0,02	**0,21**	-0,19	-0,11	-0,24	-0,04	0,539	-0,61	-0,07	-0,10	**0,32**	-0,27	-0,13	-0,29	-0,09
Mq44	0,342	-0,20	**0,28**	0,04	-0,15	-0,08	-0,02	-0,04	-0,03	0,372	-0,34	**0,23**	-0,04	-0,11	-0,15	-0,14	-0,17	-0,04	0,575	-0,46	**0,35**	-0,10	-0,18	-0,23	-0,20	-0,17	-0,05	0,758	-0,61	-0,26	-0,15	**0,46**	-0,31	-0,27	-0,31	-0,05
Mq45	0,160	-0,19	-0,16	0,06	-0,03	-0,11	0,00	-0,06	-0,06	0,163	-0,30	-0,14	0,03	0,03	-0,07	**0,11**	-0,07	-0,05	0,361	-0,42	-0,10	-0,06	-0,12	**0,21**	-0,10	-0,19	0,00	0,649	-0,64	-0,11	-0,21	-0,11	**0,37**	-0,09	-0,14	-0,05
Mq1	0,637	-0,07	-0,15	-0,11	-0,15	**0,18**	-0,11	0,00	-0,06	0,178	-0,17	-0,03	-0,06	0,00	**0,11**	-0,07	-0,15	-0,02	0,229	-0,28	-0,13	-0,11	-0,09	-0,06	**0,14**	-0,09	-0,06	0,570	xxxx	-0,19	-0,12	-0,14	-0,12	-0,27	**0,33**	-0,12
Mq2	0,720	-0,12	**0,44**	-0,31	-0,27	-0,11	-0,14	-0,08	-0,06	0,316	-0,10	**0,19**	-0,14	-0,08	-0,09	-0,04	-0,13	-0,02	0,769	xxxx	**0,43**	-0,37	-0,19	-0,11	-0,09	-0,09	-0,04	0,909	xxxx	**0,50**	-0,44	-0,25	-0,06	-0,12	-0,11	xxxx
Mq3	0,697	-0,08	**0,39**	-0,21	-0,26	-0,13	-0,05	-0,12	-0,05	0,342	-0,07	**0,23**	-0,09	-0,09	-0,11	-0,04	-0,08	-0,04	0,611	-0,08	**0,40**	-0,17	-0,24	-0,08	-0,12	-0,25	-0,13	0,786	-0,12	**0,50**	-0,23	-0,29	xxxx	-0,30	-0,18	xxxx
Mq4	0,588	-0,05	-0,19	-0,11	-0,14	-0,17	-0,11	-0,14	-0,11	0,297	-0,07	-0,15	-0,09	-0,10	-0,11	-0,10	-0,18	**0,18**	0,604	-0,13	-0,29	-0,10	-0,17	-0,24	-0,08	-0,12	-0,25	0,762	-0,11	-0,26	-0,06	-0,23	-0,31	-0,33	-0,01	**0,47**
ISq5	0,571	-0,13	-0,11	-0,14	**0,41**	-0,18	-0,11	-0,06	-0,06	0,338	-0,25	-0,08	-0,16	-0,23	-0,11	-0,09	-0,07	xxxx	0,670	-0,43	-0,22	-0,26	-0,22	-0,26	-0,24	-0,08	xxxx	0,815	-0,36	-0,12	-0,28	-0,36	-0,12	-0,26	-0,17	xxxx
ISq6	0,705	-0,13	-0,08	-0,28	**0,48**	-0,27	-0,11	-0,10	-0,02	0,417	-0,29	-0,11	-0,12	**0,27**	-0,15	-0,05	-0,11	-0,01	0,701	-0,31	-0,13	-0,26	**0,45**	-0,28	-0,12	-0,17	-0,05	0,705	-0,28	-0,13	-0,19	**0,46**	-0,31	-0,16	-0,19	xxxx
ISq7	0,576	-0,13	-0,13	-0,14	**0,40**	-0,18	-0,13	-0,02	-0,02	0,346	-0,17	-0,10	-0,13	**0,24**	-0,09	-0,12	-0,10	-0,07	0,601	-0,31	-0,18	-0,21	**0,39**	-0,23	-0,18	-0,15	xxxx	0,484	-0,01	-0,18	xxxx	**0,34**	-0,16	-0,21	-0,07	xxxx
ISq8	0,643	-0,13	-0,27	-0,22	**0,46**	-0,15	-0,10	-0,12	-0,05	0,444	-0,16	-0,13	-0,15	**0,31**	-0,11	-0,10	-0,16	-0,07	0,661	-0,16	-0,26	-0,22	**0,47**	-0,18	-0,16	-0,23	-0,11	0,705	-0,13	-0,25	-0,25	**0,50**	-0,27	-0,20	-0,23	-0,12
ISq9	0,523	-0,09	-0,26	**0,38**	-0,13	-0,04	-0,11	-0,16	-0,01	0,233	-0,17	-0,02	-0,15	-0,08	-0,07	-0,03	-0,14	-0,02	0,521	-0,19	-0,15	-0,38	**0,58**	-0,18	-0,17	-0,05	-0,04	0,693	-0,17	-0,22	**0,44**	-0,20	-0,24	-0,01	-0,31	-0,10
ISq10	0,671	-0,10	-0,15	-0,25	**0,46**	-0,19	-0,12	-0,21	-0,16	0,602	-0,20	-0,07	-0,28	**0,36**	-0,07	-0,08	-0,26	-0,05	0,971	-0,23	-0,19	-0,15	-0,38	**0,58**	-0,22	-0,15	-0,10	0,971	xxxx	-0,25	**0,62**	-0,34	-0,15	**0,39**	-0,13	xxxx
ISq11	0,577	-0,15	-0,18	-0,20	-0,19	**0,41**	-0,15	-0,13	-0,10	0,409	-0,21	-0,12	-0,10	**0,29**	-0,07	-0,08	-0,16	-0,05	0,547	-0,17	-0,15	-0,22	-0,15	**0,38**	-0,08	-0,21	xxxx	0,569	-0,11	-0,18	-0,34	-0,15	**0,39**	-0,14	xxxx	xxxx
ISq12	0,594	-0,14	-0,12	**0,42**	-0,21	-0,18	-0,22	-0,10	-0,04	0,379	-0,22	-0,10	**0,26**	-0,05	-0,11	-0,08	-0,09	-0,05	0,564	-0,22	-0,17	**0,39**	-0,13	-0,18	-0,12	-0,21	xxxx	0,675	xxxx	-0,05	**0,46**	-0,23	-0,21	-0,22	-0,21	xxxx
ICoq1	0,606	-0,21	-0,19	-0,16	**0,47**	-0,11	-0,05	-0,06	-0,05	0,272	-0,21	-0,08	-0,06	-0,08	-0,04	-0,10	-0,05	-0,05	0,839	-0,51	-0,25	-0,28	-0,23	-0,24	-0,14	-0,27	-0,14	0,995	-0,63	-0,34	-0,25	-0,06	-0,12	-0,11	-0,12	xxxx
ICoq2	0,757	-0,29	-0,25	-0,17	-0,18	**0,55**	-0,16	-0,14	-0,07	0,609	-0,40	-0,14	-0,10	-0,08	-0,06	-0,05	-0,16	-0,05	1,130	-0,61	-0,31	-0,28	-0,33	**0,74**	-0,23	-0,20	-0,13	1,143	xxxx	-0,30	-0,30	-0,30	-0,30	-0,30	-0,32	-0,19
ICoq3	0,546	-0,25	-0,09	-0,09	-0,12	-0,13	0,02	**0,44**	-0,02	0,060	-0,31	0,18	-0,05	-0,07	-0,07	**0,01**	0,00	-0,06	0,752	-0,88	-0,22	-0,26	-0,30	-0,30	-0,14	**0,42**	-0,09	0,925	xxxx	-0,30	-0,30	-0,30	-0,34	-0,29	-0,16	**0,49**
ICoq4	0,564	-0,10	-0,10	-0,08	-0,13	-0,07	-0,07	-0,10	**0,48**	0,074	-0,20	-0,15	-0,03	-0,09	0,05	-0,06	-0,12	**0,05**	0,845	-0,60	-0,33	-0,31	-0,36	-0,34	-0,27	-0,27	**0,48**	1,042	-0,76	-0,42	-0,34	-0,46	-0,42	-0,36	-0,35	**0,58**
ICoq5	0,661	-0,17	-0,08	-0,14	-0,08	-0,12	-0,04	-0,05	**0,54**	0,316	-0,39	-0,14	-0,11	-0,13	-0,16	-0,11	-0,16	**0,17**	1,071	-0,76	-0,29	-0,46	-0,40	-0,41	-0,34	-0,32	**0,60**	1,048	xxxx	-0,37	-0,46	-0,42	-0,40	-0,27	-0,33	**0,64**
ICoq6	0,733	-0,26	**0,53**	-0,16	-0,22	-0,14	-0,24	-0,12	-0,04	0,704	-0,36	**0,48**	-0,16	-0,20	-0,15	-0,13	-0,20	-0,06	1,101	-0,96	**0,73**	-0,27	-0,31	-0,22	-0,32	-0,24	-0,13	1,136	xxxx	**0,81**	-0,31	-0,37	-0,33	-0,19	xxxx	xxxx

QCM	NCIq	rpbis classiques							NCSq	rpbis Spectraux Contrastés (rpbis SC)							NCSq	rpbis SC calculés au palier turbo T80 (rpbis SC780)							NCSq	rpbis SC calculés au palier turbo T90 (rpbis SC790)						
	70	OM	P1	P2	P3	P4	P5	P6	P7	70	OM	P1	P2	P3	P4	P5	P6	P7	780	OM	P1	P2	P3	P4	P5	P6	P7	790	OM	P1	P2	P3	P4	P5	P6	P7
[G]q1	0,200	-0,19	0,03	-0,07	-0,04	**0,20**	-0,02	0,05	-0,07	0,342	-0,46	-0,09	-0,15	-0,13	**0,17**	-0,10	-0,15	-0,13	0,726	-0,60	-0,27	-0,28	-0,26	**0,33**	-0,26	-0,45	-0,23	0,892	-0,68	-0,30	-0,29	-0,34	**0,39**	-0,30	-0,59	-0,18
[G]q2	0,614	-0,15	-0,09	-0,12	-0,11	**0,42**	-0,25	-0,19	-0,04	0,477	-0,25	-0,09	-0,06	-0,09	**0,31**	-0,20	-0,19	-0,08	0,604	-0,30	-0,14	-0,14	-0,15	**0,53**	-0,34	-0,30	-0,10	0,972	-0,43	-0,17	-0,13	-0,20	**0,64**	-0,37	-0,56	xxxx
[G]q3	0,492	-0,24	**0,39**	-0,05	-0,09	-0,12	-0,14	-0,07	-0,04	0,215	-0,31	**0,13**	-0,02	-0,08	0,00	-0,07	-0,12	0,01	0,637	-0,47	**0,38**	-0,18	-0,19	-0,15	-0,24	-0,32	-0,06	0,878	-0,63	**0,52**	-0,24	-0,17	-0,29	-0,24	-0,41	-0,03
[G]q4	0,675	-0,25	-0,12	-0,11	**0,53**	-0,15	-0,14	-0,13	-0,08	0,595	-0,45	-0,17	-0,17	**0,38**	-0,16	-0,18	-0,25	-0,10	0,969	-0,56	-0,31	-0,27	**0,62**	-0,34	-0,25	-0,36	-0,16	1,090	-0,67	-0,31	-0,29	**0,68**	-0,41	-0,17	-0,38	-0,07
[G]q5	0,792	-0,23	-0,08	-0,10	**0,56**	-0,13	-0,30	-0,14	-0,07	0,366	-0,29	-0,10	-0,09	**0,24**	-0,11	-0,10	-0,15	-0,05	0,796	-0,35	-0,20	-0,18	**0,51**	-0,23	-0,32	-0,25	-0,11	0,992	-0,53	-0,15	-0,18	**0,61**	-0,21	-0,40	-0,24	-0,10
[G]q6	0,703	-0,34	-0,10	-0,09	-0,12	-0,20	**0,52**	-0,13	-0,07	0,765	-0,48	-0,15	-0,17	-0,21	-0,20	**0,50**	-0,24	-0,08	1,045	-0,52	-0,19	-0,25	-0,31	-0,30	**0,69**	-0,36	-0,13	1,169	-0,64	-0,21	-0,21	-0,34	-0,30	**0,75**	-0,58	-0,14
[G]q7	0,519	-0,24	-0,04	**0,41**	-0,04	-0,01	-0,05	-0,10	-0,04	0,580	-0,43	-0,16	**0,32**	-0,17	-0,15	-0,10	-0,25	-0,16	0,970	-0,64	-0,32	**0,51**	-0,38	-0,36	-0,18	-0,47	-0,31	1,253	-0,82	-0,44	**0,66**	-0,45	-0,47	-0,17	-0,58	-0,36
[G]q8	0,705	-0,33	-0,12	**0,53**	-0,15	-0,13	-0,09	-0,13	-0,03	0,823	-0,50	-0,21	**0,53**	-0,26	-0,27	-0,18	-0,24	-0,04	1,057	-0,63	-0,29	**0,71**	-0,39	-0,38	-0,26	-0,24	-0,09	1,163	-0,64	-0,29	**0,80**	-0,45	-0,47	-0,24	-0,45	-0,05
[G]q9	0,464	-0,21	-0,24	**0,35**	-0,13	-0,09	-0,13	0,04	-0,01	0,324	-0,26	-0,10	**0,20**	-0,15	-0,05	-0,07	-0,03	-0,03	0,725	-0,38	-0,30	**0,44**	-0,32	-0,26	-0,24	-0,10	-0,03	0,883	-0,57	-0,29	**0,54**	-0,34	-0,36	-0,24	-0,23	-0,10
[G]q10	0,545	-0,27	**0,39**	-0,08	-0,16	-0,13	-0,05	-0,11	-0,07	0,568	-0,33	**0,37**	-0,14	-0,16	-0,07	-0,08	-0,08	-0,08	0,836	-0,39	**0,55**	-0,21	-0,28	-0,17	-0,23	-0,32	-0,11	0,996	-0,52	**0,64**	-0,19	-0,32	-0,23	-0,40	-0,31	-0,18
[H]q1	0,552	-0,12	-0,11	**0,43**	-0,05	-0,19	-0,11	-0,04	-0,07	0,765	-0,48	-0,15	-0,17	-0,21	-0,20	**0,50**	-0,24	-0,08	1,045	-0,52	-0,19	-0,25	-0,31	-0,30	**0,69**	-0,36	-0,13	1,169	-0,64	-0,21	-0,21	-0,34	-0,30	**0,75**	-0,58	-0,14
[H]q2	0,565	-0,25	-0,06	-0,11	-0,21	-0,08	**0,43**	-0,16		0,773	-0,32	-0,22	-0,20	-0,33	-0,17	**0,51**	-0,23		0,963	-0,46	-0,27	-0,25	-0,40	-0,19	**0,63**	-0,35		1,002	-0,64	-0,26	-0,34	-0,38	-0,06	**0,69**	-0,36	-0,95	
[H]q3	-0,069	-0,21	0,36	-0,02	-0,10	-0,06	-0,17	**0,01**		-0,389	-0,14	0,49	0,05	0,07	0,04	0,19	**0,15**		-0,187	-0,27	0,38	0,00	0,02	-0,03	0,06	**-0,06**		0,399	-0,70	-0,03	-0,18	-0,17	-0,14	-0,27	**0,15**		
[H]q4	0,495	-0,18	-0,13	-0,18	-0,12	-0,07	**0,38**	-0,10		0,475	-0,30	-0,10	-0,17	-0,22	-0,08	**0,31**	-0,13		0,765	-0,43	-0,19	-0,22	-0,29	-0,19	**0,50**	-0,23		0,997	-0,64	-0,19	-0,30	-0,32	-0,27	**0,71**	-0,33		
[H]q5	0,558	-0,19	-0,12	**0,40**	-0,13	-0,17	-0,18	-0,08		0,752	xxxx	-0,20	**0,51**	-0,13	-0,29	-0,27	-0,18		0,878	xxxx	-0,24	**0,60**	-0,14	-0,31	-0,22	-0,17		1,057	xxxx	-0,22	**0,75**	-0,14	-0,35	-0,39	-0,21		
[H]q6	0,400	-0,14	-0,11	-0,12	**0,31**	-0,01	-0,17	-0,01		0,247	-0,31	-0,02	-0,02	**0,16**	-0,07	-0,14	-0,10	-0,04	0,462	-0,48	-0,12	-0,04	**0,28**	-0,11	-0,24	-0,14		0,671	-0,78	-0,24	xxxx	**0,48**	-0,20	-0,32	-0,20		
[H]q7	0,304	-0,16	**0,24**	-0,09	-0,09	-0,03	-0,01	-0,11		0,534	xxxx	**0,35**	-0,10	-0,14	-0,10	-0,19	-0,24		0,795	xxxx	**0,52**	-0,17	-0,18	-0,30		0,961	xxxx	**0,63**	-0,19	-0,14	-0,14	-0,37	-0,39		
[H]q8	0,506	-0,14	**0,34**	-0,22	-0,06	-0,08	-0,10	-0,11		0,556	-0,28	**0,35**	-0,25	-0,08	-0,16	-0,03	-0,03		0,774	xxxx	**0,50**	-0,32	-0,10	-0,20	-0,10	-0,06		0,966	xxxx	**0,64**	-0,39	-0,11	-0,19	-0,11	xxxx		
[H]q9	0,552	-0,09	-0,14	-0,10	-0,06	-0,11	**0,33**	-0,24		0,297	-0,20	-0,06	-0,12	-0,15	**0,19**	-0,10			0,502	-0,27	-0,07	-0,12	-0,05	-0,18	**0,30**	-0,21		0,730	xxxx	xxxx	-0,49	**0,72**	-0,26	-0,33	-0,30	-0,34		
[H]q10	0,667	-0,21	-0,19	**0,51**	-0,11	-0,17		**0,33**		0,917	-0,40	-0,37	**0,56**	-0,33	-0,36	-0,15	-0,22	-0,23	0,866	-0,44	-0,25	-0,42	-0,33	-0,29	-0,32	**0,41**		1,081	xxxx	-0,38	-0,54	-0,51	-0,36	-0,33	**0,59**		
[H]q11	0,378	-0,15	-0,11	-0,14	**0,05**	-0,11	-0,14			0,473	-0,27	-0,20	-0,31	**0,42**	-0,17	-0,22	-0,24	-0,22	0,767	-0,53	-0,47	-0,25	-0,42	-0,33	-0,29	-0,32		1,125	xxxx	-0,20	-0,49	**0,72**	-0,34	-0,44	-0,34		
[H]q12	0,521	-0,20	-0,08	-0,13	**0,40**	-0,08	-0,17	-0,11		0,654	-0,30	-0,18	-0,27	**0,42**	-0,26	-0,27	-0,22	-0,20	0,866	-0,44	-0,22	-0,34	**0,56**	-0,30	-0,31	-0,21		1,125	xxxx	-0,20	-0,49	**0,72**	-0,34	-0,44	-0,34		
[H]q13	0,511	-0,12	-0,07	-0,09	-0,11	**0,35**	-0,21	-0,16		0,796	-0,18	-0,14	-0,14	-0,22	**0,53**	-0,35	-0,25		0,959	-0,32	-0,18	-0,10	-0,27	**0,64**	-0,41	-0,25		1,039	xxxx	-0,21	xxxx	-0,35	**0,71**	-0,36	-0,29		
[H]q14	0,626	-0,16	-0,09	-0,25	-0,12	**0,42**	-0,13			0,610	-0,19	-0,10	-0,25	-0,08	-0,19	**0,39**	-0,22		0,803	-0,24	-0,06	-0,35	-0,17	-0,27	**0,50**	-0,21		1,040	xxxx	-0,41	-0,15	-0,44	**0,65**	-0,28			
[H]q15	0,470	-0,12	-0,12	0,00	-0,19	**0,34**	-0,12	-0,02		0,446	-0,25	-0,15	-0,06	-0,27	**0,27**	-0,15	-0,14		0,660	-0,35	-0,25	-0,09	-0,34	**0,40**	-0,20	-0,16		0,901	xxxx	-0,38	xxxx	-0,34	**0,55**	-0,18	-0,09		
[H]q16	0,198	-0,12	0,06	-0,07	-0,07	**0,17**	0,00	-0,08		0,008	-0,27	0,19	-0,03	-0,05	**0,02**	-0,04	-0,13		0,272	-0,44	0,03	-0,09	-0,14	**0,16**	-0,12	-0,20		0,464	xxxx	-0,11	-0,16	-0,27	**0,29**	-0,14	-0,32		
[H]q17	0,623	-0,22	-0,14	**0,47**	-0,19	-0,10	-0,10	-0,16		0,842	-0,39	-0,33	**0,55**	-0,33	-0,24	-0,15	-0,28		1,016	-0,49	-0,38	**0,65**	-0,42	-0,25	-0,12	-0,33		0,973	xxxx	-0,32	**0,68**	-0,35	-0,27	-0,09	-0,33		
[H]q18	0,500	-0,14	-0,09	**0,38**	-0,13	-0,11	-0,13	-0,07		0,865	-0,44	-0,16	**0,53**	-0,33	-0,31	-0,33	-0,26		1,051	-0,56	-0,16	**0,64**	-0,40	-0,37	-0,38	-0,29		1,098	-0,75	xxxx	**0,62**	-0,28	-0,31	-0,22	-0,20	0,02	
[H]q19	0,483	-0,18	-0,12	**0,35**	-0,15	-0,11	-0,14	-0,07		0,583	-0,31	-0,15	**0,39**	-0,14	-0,12	-0,24	-0,12		0,831	-0,44	-0,19	**0,55**	-0,20	-0,20	-0,30	-0,17		0,693	xxxx	-0,22	**0,56**	-0,12	-0,22	-0,20	0,02		
[H]q20	0,066	-0,07	-0,18	-0,06	**0,10**	0,00	-0,09	0,14		0,418	-0,20	-0,16	-0,10	**0,17**	-0,03	-0,02	-0,24		-0,254	xxxx	-0,24	-0,17	**0,07**	-0,05	-0,06	-0,06		0,106	xxxx	-0,22	-0,18	**0,01**	-0,13	0,03	0,25		
[H]q21	0,362	-0,17	0,00	-0,07	-0,02	**0,29**	-0,06	-0,04		0,701	-0,48	-0,35	-0,28	-0,19	**0,38**	-0,24	-0,27		1,034	-0,54	-0,46	-0,34	**0,59**	-0,35	-0,39			1,249	-0,66	-0,56	-0,52	-0,35	**0,76**	-0,39	-0,36		
[H]q22	0,527	-0,19	**0,44**	-0,05	-0,10	-0,06	-0,14	-0,11		0,622	xxxx	**0,35**	-0,31	-0,33	-0,16	-0,09	-0,17		0,925	xxxx	**0,52**	-0,48	-0,44	-0,29	-0,18	-0,21		1,021	xxxx	**0,62**	-0,47	-0,30	-0,35	-0,29	-0,22		
[H]q23	0,324	-0,10	-0,09	-0,09	-0,07	**0,23**	-0,10	-0,14		0,633	-0,23	-0,19	**0,49**	-0,15	-0,19	-0,16	-0,15		0,820	xxxx	-0,27	-0,17	**0,61**	-0,25	-0,23	-0,22		0,991	xxxx	xxxx	-0,36	-0,17	**0,71**	-0,19	-0,13	xxxx	
[H]q24	0,378	-0,16	-0,06	-0,07	**0,28**	-0,13	-0,11	-0,10		0,661	-0,34	-0,22	-0,21	**0,44**	-0,23	-0,16	-0,15		0,877	-0,46	-0,27	-0,30	**0,50**	-0,25	-0,23	-0,22		1,033	xxxx	-0,36	-0,17	**0,71**	-0,19	-0,13	xxxx		
[H]q25	0,345	-0,12	-0,06	0,07	**0,28**	-0,08	-0,18	-0,14		0,103	-0,29	-0,29	0,04	0,01	**0,06**	-0,07	0,01		0,453	-0,47	0,04	-0,17	**0,26**	-0,19	-0,09	-0,19		0,522	xxxx	0,02	-0,15	**0,35**	-0,17	-0,21	-0,25		
[A]q1	0,412	-0,25	-0,10	-0,08	**0,31**	-0,11	-0,16	-0,05	0,03	0,469	-0,29	-0,20	-0,16	-0,12	**0,29**	-0,07	0,01	xxxx	0,728	-0,40	-0,29	-0,25	-0,15	-0,12	-0,10	xxxx		0,869	xxxx	xxxx	-0,41	-0,32	**0,57**	-0,20	xxxx	xxxx	
[A]q2	0,353	-0,27	-0,04	-0,10	**0,27**	-0,08	-0,10	-0,02	-0,11	0,407	-0,28	-0,15	-0,16	**0,25**	-0,08	-0,03	-0,12	-0,04	0,723	-0,42	-0,30	-0,26	**0,43**	-0,10	-0,11	-0,02	xxxx	0,924	-0,62	-0,34	-0,21	**0,55**	-0,17	-0,12	-0,05	xxxx	
[A]q3	0,682	-0,26	-0,11	-0,17	-0,20	-0,14	-0,06	**0,51**	-0,03	0,730	-0,37	-0,24	-0,30	-0,27	-0,19	-0,16	**0,45**	-0,02	0,990	-0,49	-0,31	-0,40	-0,35	-0,23	-0,20	**0,62**	-0,05	1,139	xxxx	-0,28	-0,44	-0,44	-0,23	-0,17	**0,76**	xxxx	
[A]q4	0,516	-0,23	-0,11	-0,11	-0,12	-0,09	-0,11	**0,40**	-0,03	0,706	-0,39	-0,31	-0,32	-0,21	-0,20	-0,16	**0,40**	-0,02	0,889	-0,50	-0,38	-0,39	-0,25	-0,18	-0,18	**0,51**	0,00	1,015	xxxx	-0,37	-0,41	-0,32	xxxx	-0,21	**0,63**	xxxx	
[A]q5	0,617	-0,28	-0,12	**0,46**	-0,18	-0,09	-0,15	-0,03	-0,06	0,702	-0,37	-0,23	**0,46**	-0,26	-0,18	-0,20	-0,21	-0,19	0,919	-0,45	-0,29	**0,60**	-0,35	-0,22	-0,30	-0,20	-0,03	0,930	xxxx	-0,28	**0,66**	-0,29	-0,25	-0,29	-0,14	xxxx	
[A]q6	0,410	-0,26	-0,15	-0,17	**0,33**	-0,12	-0,07	0,09	-0,04	0,520	-0,37	-0,19	-0,17	**0,32**	-0,20	-0,05	-0,19	-0,08	0,741	-0,48	-0,25	-0,23	**0,45**	-0,31	-0,21	-0,08	**0,26**	1,019	-0,72	-0,29	-0,34	**0,60**	-0,32	-0,02	-0,36	xxxx	
[A]q7	0,349	-0,26	-0,10	-0,04	-0,03	0,04	-0,12	**0,30**	0,01	0,297	-0,33	-0,08	-0,12	-0,11	-0,13	-0,05	**0,17**	0,00	0,472	-0,46	-0,14	-0,19	-0,17	-0,12	-0,18	-0,15	xxxx	0,917	-0,86	-0,21	-0,41	-0,27	-0,40	-0,21	**0,45**	xxxx	
[A]q8	0,434	-0,27	**0,32**	-0,10	-0,14	-0,09	-0,10	0,00	0,03	0,329	-0,37	**0,22**	-0,16	-0,12	-0,13	-0,06	**0,17**	0,04	0,587	-0,47	**0,37**	-0,22	-0,17	-0,13	-0,21	-0,08	**0,26**	0,695	xxxx	**0,49**	-0,25	-0,16	xxxx	-0,20	-0,17	xxxx	
[A]q9	0,510	-0,22	**0,36**	-0,14	-0,17	-0,15	-0,05	-0,13	-0,02	0,844	-0,29	**0,57**	-0,35	-0,32	-0,18	-0,02	-0,15	-0,06	1,086	-0,45	**0,70**	-0,42	-0,33	-0,18	-0,19	**0,80**	-0,09	1,338	xxxx	**0,89**	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	
[A]q10	0,562	-0,29	-0,10	-0,19	-0,10	**0,41**	-0,13	-0,07	-0,02	0,990	-0,46	-0,19	-0,31	-0,20	-0,19	**0,67**	-0,15	-0,06	1,203	-0,59	-0,22	-0,33	-0,18	-0,19	**0,80**	-0,09	-0,01	1,469	-0,77	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	
[A]q11	0,717	-0,18	-0,19	-0,28	-0,11	-0,07	-0,16	**0,39**	-0,08	0,332	-0,28	-0,07	-0,09	-0,10	-0,07	-0,08	-0,12	-0,04	0,628	-0,43	-0,09	-0,18	-0,06	-0,14	-0,21	-0,16	**0,42**	0,941	-0,63	-0,10	-0,26	-0,07	-0,24	-0,24	xxxx	**0,62**	
[A]q12	0,572	-0,30	-0,12	-0,11	-0,16	-0,09	-0,08	-0,02	-0,08	0,851	-0,44	-0,16	-0,10	-0,24	-0,17	**0,57**	-0,19	-0,08	1,069	-0,60	-0,21	xxxx	-0,26	-0,19	**0,69**	-0,21	-0,10	1,373	-0,71	-0,33	xxxx	-0,21	-0,16	xxxx	xxxx	xxxx	
[A]q13	0,619	-0,23	-0,14	-0,15	-0,14	-0,11	-0,06	**0,48**	0,02	0,612	-0,38	-0,18	-0,29	-0,27	-0,27	-0,27	-0,16	-0,25	0,815	-0,51	-0,26	-0,29	-0,30	-0,31	xxxx	-0,36	**0,51**	0,890	xxxx	-0,27	-0,23	-0,37	-0,12	xxxx	xxxx	**0,62**	
[A]q14	0,108	-0,18	0,03	-0,01	**0,09**	0,08	0,01	-0,06	0,02	-0,003	-0,30	0,04	-0,16	0,03	0,20	0,01	-0,04	0,04	0,107	xxxx	-0,04	-0,01	**0,11**	0,10	-0,06	-0,01	-0,01	0,499	xxxx	-0,20	-0,06	**0,32**	-0,13	-0,26	-0,23	-0,06	
[A]q15	0,209	-0,17	0,07	-0,05	-0,06	0,01	-0,06	**0,29**	-0,01	0,383	-0,45	-0,16	-0,06	-0,19	-0,33	-0,13	-0,30	**0,16**	0,710	-0,71	-0,38	-0,30	-0,32	-0,51	-0,44	**0,27**		1,069	-1,10	-0,62	-0,41	-0,33	-0,59	-0,26	-0,55	**0,37**	
[A]q16	0,338	-0,18	-0,08	-0,10	-0,11	0,02	-0,03	0,01	**0,29**	0,452	-0,42	-0,12	-0,23	-0,30	-0,29	-0,20	**0,24**		0,799	-0,62	-0,32	-0,39	-0,46	-0,40	-0,31	**0,41**		1,084	-0,92	-0,37	-0,46	-0,57	-0,42	-0,37	**0,55**		
[A]q17	0,430	-0,18	-0,09	**0,35**	-0,12	-0,03	-0,03	0,01	-0,01	0,541	-0,42	-0,20	**0,31**	-0,23	-0,30	-0,19	-0,17	-0,16	0,807	-0,60	-0,29	**0,45**	-0,35	-0,27	-0,24	xxxx		1,097	-0,85	-0,33	**0,57**	-0,44	-0,36	-0,33	-0,09	xxxx	
[A]q18	0,324	-0,24	0,03	-0,11	**0,26**	0,00	-0,09	-0,01	-0,06	0,537	-0,43	-0,17	-0,21	**0,32**	-0,13	-0,12	-0,17	xxxx	0,827	-0,59	-0,28	**0,48**	-0,27	-0,23	-0,18	xxxx		1,187	-0,80	-0,44	-0,36	**0,65**	-0,31	-0,29	-0,23	xxxx	
[A]q19	0,436	-0,17	-0,10	-0,16	**0,33**	-0,02	-0,17	**0,39**	-0,04	0,468	-0,29	-0,13	-0,04	**0,31**	-0,18	-0,20	-0,22	-0,07	0,682	-0,40	-0,21	-0,11	**0,44**	-0,25	-0,25	-0,04		0,928	-0,55	-0,33	-0,08	**0,57**	-0,32	-0,36	-0,22	-0,09	
[A]q20	0,517	-0,24	-0,07	-0,12	-0,16	-0,07	**0,39**	-0,04	-0,02	0,986	-0,51	-0,31	-0,21	-0,32	-0,36	**0,62**	-0,31	0,02	1,154	-0,62	-0,32	-0,23	-0,39	-0,39	**0,71**	-0,32	xxxx	1,431	-0,85	-0,39	-0,29	**0,83**	-0,31	xxxx		
[A]q21	0,644	-0,30	-0,14	**0,46**	-0,13	-0,15	-0,16	-0,10	0,02	1,043	-0,46	-0,22	**0,71**	-0,21	-0,32	-0,29	-0,34	-0,28	1,183	-0,58	-0,24	**0,79**	-0,31	-0,30	-0,32	**0,73**	xxxx	1,386	-0,72	-0,15	**0,95**	-0,33	-0,35	**0,80**	xxxx	xxxx	
[A]q22	0,546	-0,23	-0,14	-0,08	-0,18	-0,12	**0,38**	-0,08	-0,12	0,866	-0,41	-0,14	-0,13	-0,24	-0,29	**0,59**	-0,16	-0,13	1,133	-0,58	-0,15	-0,19	-0,25	-0,32	**0,73**	xxxx	xxxx	1,300	-0,65	xxxx	-0,20	-0,33	-0,35	**0,80**	xxxx	xxxx	
[A]q23	0,553	-0,24	-0,10	-0,05	-0,05	-0,14	-0,10	**0,44**	0,03	0,639	-0,40	-0,19	-0,28	-0,10	-0,06	-0,10	-0,23	-0,14	0,988	-0,56	-0,34	-0,36	-0,20	-0,38	-0,44	-0,17	**0,58**	1,103	xxxx	-0,40	-0,32	-0,40	-0,52	-0,17	-0,30	-0,18
[A]q24	0,187	-0,16	**0,16**	-0,03	0,00	-0,03	0,00	-0,04	-0,04	0,253	-0,33	**0,13**	-0,10	-0,06	-0,10	-0,03	-0,14	0,01	0,600	-0,56	**0,28**	-0,20	-0,26	-0,13	-0,27	-0,03		0,767	xxxx	**0,40**	-0,32	-0,40	-0,52	-0,17	-0,30	-0,18
[A]q25	0,620	-0,25	-0,12	-0,12	-0,11	-0,09	-0,15	**0,48**	-0,04	1,028	-0,47	-0,41	-0,24	-0,34	-0,36	-0,38	**0,65**		1,257	-0,61	-0,47	-0,27	-0,37	-0,41	-0,48	**0,79**		1,373	-0,76	-0,37	-0,19	-0,33	-0,40	-0,41	**0,84**		

	NCIq	rpbis classiques						NCSq	rpbis Spectraux Contrastés (rpbis SC)							NCSq	rpbis SC calculés au palier turbo T80 (rpbis SCT80)							NCSq	rpbis SC calculés au palier turbo T90 (rpbis SCT90)						
QCM	70	OM	P1	P2	P3	P4	P5	P6	P7	70	OM	P1	P2	P3	P4	P5	P6	P7	80	OM	P1	P2	P3	P4	P5	P6	P7	90	OM	P1	P2	P3	P4	P5	P6	P7
[Mq1	0,582	-0,07	-0,19	-0,22	-0,15	0,43	-0,15	-0,12	-0,03	0,391	-0,20	-0,13	-0,12	-0,08	0,27	-0,17	xxxx	0,665	-0,29	-0,20	0,45	-0,15	-0,20	-0,25	xxxx	0,696	-0,34	-0,10	-0,20	-0,14	0,48	-0,22	-0,26	xxxx
[Mq2	0,553	-0,14	-0,18	0,40	-0,14	-0,17	-0,14	-0,13	-0,06	0,479	-0,31	-0,12	0,33	-0,08	-0,15	-0,09	-0,21	-0,04	0,732	-0,44	-0,19	0,51	-0,14	-0,20	-0,11	-0,26	-0,06	0,795	xxxx	-0,19	0,62	-0,01	-0,20	-0,21	-0,22	xxxx
[Mq3	0,689	-0,11	-0,30	0,49	-0,15	-0,11	-0,13	-0,22	0,01	0,537	-0,22	-0,21	0,36	-0,12	-0,09	-0,09	-0,26	xxxx	0,785	xxxx	-0,41	-0,13	-0,12	-0,17	0,50	-0,32	xxxx	0,947	xxxx	-0,27	0,65	-0,23	-0,32	-0,36	-0,34	xxxx
[Mq4	0,625	-0,11	-0,26	-0,15	-0,19	0,44	-0,17	0,05	-0,05	0,484	-0,18	-0,23	-0,12	-0,07	-0,09	0,32	-0,20	xxxx	0,773	xxxx	-0,41	-0,13	-0,08	-0,12	0,50	-0,27	xxxx	0,947	xxxx	-0,44	-0,13	-0,11	0,02	0,59	-0,40	xxxx
[Mq5	0,405	-0,15	0,34	-0,09	-0,14	-0,22	-0,17	0,05	-0,05	0,400	-0,41	0,25	-0,06	-0,17	-0,10	-0,14	-0,04	0,635	-0,48	0,39	-0,14	-0,17	-0,17	-0,22	-0,27	0,812	-0,53	0,50	-0,16	-0,20	-0,20	-0,25	-0,35	xxxx	xxxx
[Mq6	0,451	-0,12	0,32	-0,03	-0,07	-0,07	-0,13	-0,01	-0,03	0,536	-0,37	-0,13	0,36	-0,07	-0,10	-0,12	-0,19	-0,06	0,782	-0,43	-0,16	0,52	-0,14	-0,17	-0,18	-0,30	-0,06	1,028	-0,41	-0,13	0,67	xxxx	-0,27	-0,25	-0,45	-0,06
[Mq7	0,638	-0,15	-0,16	0,46	-0,13	-0,22	-0,17	-0,14	-0,02	0,749	-0,41	-0,24	0,51	-0,20	-0,17	-0,25	-0,25	-0,04	1,003	-0,52	-0,28	0,68	-0,17	-0,24	-0,33	-0,34	-0,06	1,247	-0,61	-0,26	0,83	xxxx	-0,29	-0,39	-0,34	xxxx
[Mq8	0,665	-0,15	-0,23	-0,05	-0,18	0,46	-0,13	-0,23	0,00	0,687	-0,29	-0,24	-0,11	-0,18	0,46	-0,16	-0,28	-0,03	0,886	xxxx	-0,28	-0,16	-0,23	0,61	-0,22	-0,39	-0,06	1,038	xxxx	-0,30	xxxx	-0,34	0,71	-0,29	-0,36	xxxx
[Mq9	0,566	-0,12	-0,06	-0,04	-0,22	0,43	-0,10	-0,07	-0,05	0,603	-0,31	-0,19	-0,16	0,33	-0,37	-0,09	-0,24	-0,04	0,814	-0,35	-0,17	-0,22	0,46	-0,48	-0,14	-0,32	xxxx	0,981	xxxx	-0,21	-0,26	0,59	-0,52	-0,21	-0,40	xxxx
[Mq10	0,513	-0,07	-0,11	-0,09	-0,16	-0,14	0,35	-0,21	-0,02	0,624	-0,26	-0,04	-0,07	-0,22	-0,11	-0,16	-0,09	-0,02	1,064	-0,39	xxxx	-0,16	-0,32	-0,11	0,68	-0,49	xxxx	1,117	-0,42	xxxx	-0,25	-0,15	-0,28	0,73	-0,47	xxxx
[Mq11	0,441	-0,07	-0,20	0,26	-0,04	-0,07	-0,12	0,09	xxxx	0,226	-0,10	-0,09	0,14	-0,06	-0,04	-0,03	-0,04	-0,02	0,515	-0,20	-0,22	0,31	0,02	-0,06	xxxx	-0,19	xxxx	0,663	xxxx	-0,21	0,44	xxxx	-0,13	-0,21	-0,38	xxxx
[Mq12	0,513	-0,11	0,36	-0,12	-0,10	-0,15	-0,10	-0,24	-0,04	0,572	-0,28	0,38	-0,14	-0,08	-0,15	-0,13	-0,30	-0,06	0,749	xxxx	0,50	-0,19	-0,11	-0,10	-0,10	-0,44	xxxx	0,844	xxxx	0,60	-0,18	xxxx	-0,13	-0,21	-0,38	xxxx
[Mq13	0,577	-0,12	-0,18	-0,17	-0,09	-0,20	0,43	-0,07	-0,02	0,530	-0,31	-0,18	-0,16	-0,20	-0,13	0,35	-0,21	-0,08	0,783	-0,38	-0,27	-0,24	-0,27	-0,17	0,52	-0,30	-0,10	0,870	-0,51	-0,18	-0,26	-0,28	-0,15	0,59	-0,33	-0,02
[Mq14	0,559	-0,12	-0,14	-0,13	0,39	-0,26	-0,01	-0,06	-0,05	0,368	-0,25	-0,14	-0,12	0,24	-0,11	-0,11	-0,12	-0,11	0,593	-0,35	-0,16	-0,17	0,39	-0,24	-0,15	-0,15	-0,14	0,777	-0,46	-0,21	-0,24	0,53	-0,28	-0,12	-0,19	-0,22
[Mq15	0,568	-0,07	0,38	-0,21	-0,11	-0,17		-0,22	-0,02	0,379	-0,12	0,25	-0,08	-0,09	-0,12		-0,20	xxxx	0,664	-0,18	0,45	-0,19	-0,21	-0,18		-0,30	xxxx	0,732	xxxx	0,50	-0,17	-0,27	-0,14		-0,35	xxxx
[Mq16	0,659	-0,21	0,48	-0,10	-0,19	-0,17	-0,08	-0,25	-0,04	0,657	-0,38	0,43	-0,13	-0,23	-0,22	-0,04	-0,29	-0,06	0,865	-0,41	0,58	-0,22	-0,28	-0,28	-0,07	-0,36	-0,11	0,954	xxxx	0,64	-0,21	-0,30	-0,28	-0,11	-0,45	xxxx
[Mq17	0,563	-0,05	0,37	-0,36	-0,09	-0,01	0,18	-0,08	-0,02	0,154	-0,23	0,09	-0,08	-0,10	-0,27	-0,15	-0,27	-0,10	0,417	-0,29	0,23	-0,25	-0,16	-0,17	-0,06	-0,27	-0,04	0,692	-0,33	0,40	-0,37	-0,16	-0,10	-0,23	-0,28	-0,09
[Mq18	0,667	-0,21	-0,17	-0,17	-0,15	-0,12	0,48	-0,19	-0,06	0,707	-0,46	-0,15	-0,15	-0,27	-0,15	0,47	-0,27	-0,10	0,971	-0,51	-0,25	-0,23	-0,32	-0,19	0,65	-0,34	-0,08	0,991	xxxx	-0,26	xxxx	-0,26	-0,10	-0,26	0,66	-0,49
[Mq19	0,551	-0,14	-0,14	-0,12	-0,24	-0,11	0,41	-0,11	-0,04	0,697	-0,37	-0,15	-0,19	-0,22	-0,15	0,43	-0,30	-0,04	0,926	-0,42	-0,21	-0,21	-0,28	-0,18	0,58	-0,40	-0,05	1,084	xxxx	-0,26	-0,24	-0,22	xxxx	0,66	-0,49	-0,03
[Mq20	0,581	-0,13	0,40	-0,13	-0,17	-0,19	-0,20	0,01	-0,04	0,495	-0,25	0,33	-0,12	-0,13	-0,12	-0,01	-0,01	-0,04	0,730	-0,34	0,50	-0,16	-0,16	-0,20	-0,03	-0,30	-0,03	0,811	xxxx	0,59	-0,19	-0,25	-0,14	-0,28	-0,08	
[Mq21	0,839	-0,07	-0,08	-0,40	-0,12	-0,06	0,49	-0,08	0,00	0,420	-0,22	-0,13	-0,17	-0,12	-0,09	0,25	-0,17	-0,02	0,693	-0,27	-0,18	-0,29	-0,16	-0,12	0,42	-0,27	-0,07	0,861	xxxx	-0,26	-0,34	-0,10	-0,18	0,54	-0,32	-0,15
[Mq22	0,749	-0,07	0,49	-0,22	-0,35	-0,09	-0,10	-0,03	0,00	0,449	-0,21	0,29	-0,13	-0,17	-0,11	-0,14	-0,19	-0,04	0,734	-0,30	0,48	-0,17	-0,28	-0,08	-0,23	-0,29	-0,08	0,771	-0,29	0,52	-0,22	-0,28	xxxx	-0,18	-0,34	-0,09
[Bq1	0,581	-0,14	-0,18	-0,12	-0,13	-0,16	0,45	-0,04	-0,02	0,557	-0,27	-0,23	-0,19	-0,21	-0,23	0,35	-0,11	-0,03	0,916	-0,42	-0,35	-0,30	-0,34	-0,37	0,58	-0,18	xxxx	1,135	-0,59	-0,31	-0,35	-0,46	-0,43	0,70	-0,29	xxxx
[Bq2	0,633	-0,16	0,46	-0,19	-0,13	-0,20	-0,12	-0,03	0,07	0,395	-0,20	0,26	-0,13	-0,26	-0,12	-0,18	-0,16	-0,04	0,829	-0,37	0,55	-0,26	-0,23	-0,31	-0,27	-0,28	-0,08	1,044	xxxx	0,69	-0,36	xxxx	-0,37	-0,27	-0,35	-0,10
[Bq3	0,615	-0,17	0,50	-0,14	-0,14	-0,12	-0,11	-0,03	-0,07	0,656	-0,39	0,38	-0,26	-0,29	-0,32	-0,20	-0,23	-0,09	1,080	xxxx	0,64	-0,44	-0,48	-0,46	-0,32	-0,43	-0,22	1,163	xxxx	0,68	-0,46	-0,58	-0,45	-0,37	-0,46	-0,30
[Bq4	0,721	-0,12	-0,05	-0,21	-0,29	0,48	-0,02	-0,08	-0,03	0,242	-0,18	-0,05	-0,11	-0,09	0,15	-0,06	-0,17	-0,04	0,839	-0,36	-0,14	-0,32	-0,37	0,51	-0,12	-0,17	-0,05	1,083	xxxx	-0,18	-0,44	-0,49	0,65	-0,05	xxxx	
[Bq5	0,768	-0,13	0,02	-0,10	0,24	-0,12		-0,05	0,00	0,068	-0,26	0,05	-0,09	0,05	-0,06		-0,09	-0,02	0,560	-0,47	-0,24	-0,32	-0,26	0,29		-0,21	-0,10	0,807	xxxx	-0,42	-0,47	0,40	-0,33		-0,35	-0,07
[Bq6	0,378	-0,18	-0,16	0,29	-0,05	-0,11		-0,07	-0,07	0,059	-0,27	-0,07	0,06	0,09	-0,12		-0,06	-0,07	0,541	xxxx	-0,22	0,38	-0,13	-0,24		-0,17	-0,11	0,777	xxxx	-0,28	0,53	-0,21	-0,31		-0,30	-0,20
[Bq7	0,702	-0,17	-0,13	-0,11	0,46	-0,31	-0,04	-0,11	-0,05	0,410	-0,27	-0,16	-0,10	-0,13	-0,03	-0,14	-0,16	-0,03	0,917	-0,38	-0,25	-0,17	0,58	-0,36	-0,10	-0,14	-0,10	1,032	xxxx	0,64	-0,42	-0,31		-0,32	-0,50	-0,18
[Bq8	0,535	-0,18	-0,13	0,43	-0,12	-0,12	-0,01	-0,08	-0,02	0,401	-0,29	-0,18	0,24	-0,08	-0,14	-0,16	-0,28	-0,04	0,844	-0,46	-0,35	0,51	-0,28	-0,28	-0,27	-0,45	-0,14	0,975	xxxx	-0,34	0,60	-0,34	-0,36	-0,32	-0,50	-0,18
[Bq9	0,593	-0,14	-0,02	0,42	-0,27	-0,08	-0,08	-0,08	-0,02	0,351	-0,26	-0,13	0,22	-0,14	-0,03	-0,16	-0,03	-0,06	0,850	xxxx	-0,23	0,50	-0,12	0,50	-0,27	-0,29	-0,12	0,955	-0,53	-0,19	0,57	-0,44	-0,31	-0,42	-0,22	
[Bq10	0,254	-0,13	-0,06	0,05	0,18	-0,06	-0,05	-0,04	-0,05	0,335	-0,20	-0,09	-0,03	0,23	-0,07	-0,06	-0,08	-0,06	0,650	xxxx	-0,23	0,50	-0,37	0,51	-0,14	-0,27	-0,10	0,833	xxxx	-0,19	0,57	-0,21	-0,31	-0,42	-0,22	
[Chq1	0,461	-0,13	-0,06	-0,10	0,32	-0,16	-0,08	-0,19	-0,06	0,342	-0,08	-0,07	-0,10	0,07	0,23	-0,06	-0,20	0,02	0,759	-0,25	-0,14	-0,19	-0,21	0,51	-0,06	-0,37	-0,11	1,030	xxxx	-0,19	xxxx	-0,24	0,70	-0,19	-0,44	xxxx
[Chq2	0,571	-0,17	-0,08	0,39	-0,20	-0,06	-0,14	-0,06	-0,02	0,618	-0,22	-0,11	0,42	-0,23	-0,07	-0,18	-0,21	xxxx	1,077	-0,40	-0,16	0,72	-0,36	-0,15	-0,29	-0,39	xxxx	1,265	xxxx	0,86	-0,44	-0,20	-0,20	-0,43	xxxx	
[Chq3	0,623	-0,22	0,48	-0,18	-0,12	-0,08	-0,14	-0,16	0,02	0,462	-0,34	0,29	-0,16	-0,26	-0,13	-0,13	-0,22	0,02	0,956	-0,56	0,60	-0,36	-0,30	-0,18	-0,26	-0,41	-0,09	1,203	-0,72	0,75	-0,42	-0,31	-0,20	-0,45	-0,17	
[Chq4	0,567	-0,20	-0,12	-0,10	-0,12	0,46	-0,08	-0,05	-0,05	0,576	-0,45	-0,22	-0,16	-0,26	0,35	-0,17	-0,24	-0,14	1,098	-0,69	-0,39	-0,35	-0,41	0,67	-0,29	-0,45	-0,31	1,392	-0,87	-0,42	-0,50	-0,43	0,82	-0,32	-0,55	-0,31
[Chq5	0,716	-0,20	-0,11	-0,13	-0,19	0,54	-0,21	-0,17	-0,02	0,730	-0,38	-0,17	-0,21	-0,31	0,48	-0,21	-0,24	-0,07	1,099	-0,52	-0,27	-0,27	-0,41	0,72	-0,34	-0,39	-0,10	1,301	-0,67	-0,23	-0,28	-0,36	0,83	-0,38	-0,52	
[Chq6	0,528	-0,19	-0,06	-0,10	-0,19	-0,08	0,37	-0,10	0,01	0,177	-0,26	-0,06	-0,04	-0,04	-0,11	0,12	-0,10	-0,03	0,627	-0,42	-0,10	-0,16	-0,27	-0,20	0,38	-0,12	-0,07	0,935	xxxx	-0,10	-0,19	-0,42	-0,21	0,59	-0,19	xxxx
[Chq7	0,568	-0,17	-0,10	-0,10	-0,13	0,45	-0,12	-0,09	-0,05	0,575	-0,42	-0,14	-0,21	-0,20	0,35	-0,17	-0,27	-0,10	1,045	-0,61	-0,21	-0,32	-0,40	0,64	-0,33	-0,45	-0,17	1,196	xxxx	-0,27	-0,37	-0,38	0,80	-0,34	-0,54	-0,20
[Chq8	0,683	-0,21	-0,20	0,51	-0,22	-0,10	-0,07	-0,09	-0,03	0,644	-0,36	-0,25	0,40	-0,26	-0,15	-0,18	-0,19	-0,04	1,013	-0,52	-0,37	0,64	-0,42	-0,21	-0,26	-0,33	xxxx	1,133	xxxx	-0,42	0,76	-0,37	-0,25	-0,37	-0,31	xxxx
[Pq1	0,701	-0,11	0,51	-0,03	-0,27	-0,10	-0,19	-0,10	-0,04	0,522	-0,25	0,31	-0,09	-0,25	-0,10	-0,20	-0,19	-0,04	0,850	-0,31	0,54	-0,13	-0,40	-0,17	-0,32	-0,35	-0,10	0,930	-0,33	0,60	-0,09	-0,42	-0,11	-0,28	-0,41	xxxx
[Pq2	0,570	-0,21	-0,09	-0,09	0,39	-0,22	-0,04	-0,18	-0,05	0,468	-0,32	0,26	-0,10	-0,04	0,34	-0,16	-0,01	-0,08	0,764	-0,44	-0,07	-0,16	0,55	-0,26	-0,21	-0,27	-0,12	0,887	-0,50	xxxx	-0,33	-0,40	-0,30	-0,25	xxxx	xxxx
[Pq3	0,541	-0,16	0,39	-0,12	-0,20	-0,11	-0,09	-0,01	0,01	0,354	-0,25	0,26	-0,10	-0,16	-0,09	-0,13	-0,26	-0,01	0,736	-0,17	0,51	-0,14	0,58	-0,22	-0,37	-0,29	-0,25	0,904	-0,35	-0,17	0,61	-0,23	-0,30	-0,23	xxxx	xxxx
[Pq4	0,548	-0,16	-0,12	0,37	-0,07	-0,25	-0,13	-0,09	-0,02	0,512	-0,15	0,35	-0,06	-0,22	-0,16	-0,20	-0,26	-0,02	0,900	-0,17	-0,14	0,58	-0,22	-0,37	-0,29	-0,25	xxxx	0,904	-0,35	-0,17	0,61	-0,23	-0,30	-0,23	xxxx	xxxx
[Pq5	0,601	-0,17	-0,06	0,45	-0,08	-0,08	-0,23	-0,09	-0,01	0,466	-0,37	-0,12	0,31	-0,17	-0,13	-0,17	-0,20	-0,05	0,870	-0,48	-0,22	0,58	-0,33	-0,33	-0,40	-0,36	-0,07	1,106	xxxx	-0,23	0,76	-0,35	-0,40	-0,52	-0,42	xxxx
[Pq6	0,629	-0,14	-0,15	-0,11	-0,21	0,46	-0,17	-0,13	-0,02	0,472	-0,23	-0,12	-0,17	0,30	-0,20	-0,12	-0,02	0,796	xxxx	-0,21	-0,27	-0,27	-0,33	0,52	-0,30	-0,18	xxxx	0,932	xxxx	-0,26	-0,26	-0,40	0,60	-0,25	-0,16	xxxx
[Pq7	0,576	-0,14	-0,08	0,43	-0,20	-0,09	-0,12	0,01	-0,06	0,462	-0,31	-0,12	0,27	-0,23	-0,13	-0,27	-0,23	-0,06	0,792	-0,31	-0,20	0,45	-0,40	-0,17	-0,26	-0,38	xxxx	1,004	-0,51	-0,25	0,60	-0,42	-0,12	-0,30	-0,41	xxxx
[Pq8	0,647	-0,24	0,47	-0,13	-0,15	-0,22	-0,09	-0,14	-0,06	0,635	-0,37	0,47	-0,16	-0,19	-0,24	-0,12	-0,20	0,02	0,917	-0,39	0,66	-0,28	-0,25	-0,37	-0,17	-0,31	xxxx	1,032	xxxx	0,73	-0,25	-0,28	-0,38	-0,16	-0,31	xxxx
[Pq9	0,490	-0,15	-0,09	0,34	-0,19	-0,14	-0,11	-0,12	-0,04	0,405	-0,14	-0,10	0,29	-0,11	-0,14	-0,18	-0,14	-0,02	0,690	-0,19	-0,03	0,47	-0,16	-0,22	-0,18	-0,31	-0,08	0,773	-0,23	xxxx	0,57	-0,14	-0,22	-0,18	-0,46	xxxx
[Pq10	0,573	-0,16	0,34	-0,10	-0,10	-0,06	-0,25	-0,07	-0,02	0,437	-0,18	0,29	-0,05	-0,07	-0,06	-0,21	-0,10	xxxx	0,687	-0,22	0,45	-0,06	-0,14	-0,12	-0,34	-0,14	xxxx	0,891	xxxx	0,53	xxxx	xxxx	-0,29	xxxx	xxxx	xxxx

TABLE DES MATIERES