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

Université de Liège - Belgique

Ph.D. dissertation abstract – Janvier 2002

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 questions à choix multiple (QCM) ce qui permet d'évaluer dans des délais raisonnables de grands groupes d'étudiants.

La technique des degrés de certitude associée aux QCM permet de dépasser le caractère « binaire » de l'évaluation des performances des étudiants (la proposition choisie est soit correcte, soit incorrecte) 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 correcte, nous permettons plus de nuances dans l'analyse de ses performances. A un extrême, 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* »).

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 d'une évaluation, lorsqu'il s'agit de mettre en évidence les QCM problématiques et, au sein de celles-ci, les propositions qui contiennent des anomalies.

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* ».

Pour mesurer la cohérence spectrale 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. Le *rpbis classique* permet d'évaluer dans quelle mesure la question discrimine les étudiants en fonction du critère du nombre de réponses correctes. Logiquement, on s'attend à ce que les sujets qui récoltent un nombre élevé de réponses correctes aient tendance à choisir la proposition correcte et les sujets qui récoltent un nombre moins élevé aient eux tendance à choisir une proposition incorrecte.

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, 1998). Le *rpbis S* a été développé en vue d'analyser la tendance à utiliser des certitudes plus élevées dans le cas d'une réponse correcte que 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 *C*ontrasté » qui 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 » il s'agit 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. Nous pouvons ainsi accroître la confiance dans les informations liées aux indices spectraux en ne prenant en compte que les données des étudiants qui commettent le moins d'erreurs dans leurs auto-estimations. Nous mesurons la quantité d'erreurs d'auto-estimations commises par les sujets à l'aide de l'indice de réalisme qui varie de 0 à 100 (Leclercq & al., 2000). L'appellation *rpbis SCT* désigne les *rpbis SC* calculés dans le cadre d'une *T*urbo analyse. Le mot « 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. Nous ajoutons à l'indice le seuil de réalisme utilisé pour sélectionner les données. Par exemple le *rpbis SCT80* a été calculé à partir des données des sujets dont le réalisme est supérieur ou égal à 80 (qui commettent entre 0% et 20% d'erreurs dans leurs auto-estimations).

En plus des *rpbis SC* et *rpbis SCT*, qui sont au cœur de cette recherche et qui permettent la détection d'anomalies à un niveau « *propositions* » au sein des QCM, 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 QCM, donc à un niveau « *questions* ». Il s'agit essentiellement d'une part de l'indice de Réalisation des prédictions par question (*Rq*) ou la quantité d'erreurs d'auto-estimations contenue dans les résultats d'une question et, d'autre part, de l'indice de Centration par question (*Cq*) ou le niveau global de sur ou sous-estimations pour une question. Enfin, nous avons aussi mesuré le Niveau de Cohérence Spectrale d'une question (*NCSq*) en nous basant sur les *rpbis SC* des différentes propositions d'une QCM. Le principe de la « *turbo analyse* » a aussi été appliqué au calcul de ces indices spectraux.

Nous avons mis ces instruments de détection des propositions posant des problèmes au sein de QCM suspectes à l'épreuve des données en calculant les indices spectraux au départ de plusieurs milliers de réponses et certitudes récoltées lors des dix tests standardisés du projet MOHICAN (Leclercq & al., 2001). Il s'agissait de dix épreuves de connaissance des principales matières de fin de l'enseignement secondaire qui ont été soumises à des groupes d'étudiants entrant en première année dans huit des neuf institutions universitaires de la Communauté Française de Belgique (le nombre d'étudiants interrogés variait entre 1.392 et 3.846 selon les tests). Ces épreuves standardisées étaient constituées de QCM pour lesquelles les étudiants furent invités à accompagner systématiquement le choix de chacune de leurs réponses d'un pourcentage de certitude. Il s'agissait donc pour ces 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 (chaque étudiant a reçu un feedback individualisé et les évaluateurs un feedback global), 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.

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 marquée, les étudiants ayant tendance à fournir des certitudes moins élevées pour la réponse considérée comme correcte et plus élevées pour les propositions incorrectes. 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 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 (les *rpbis classiques*) lorsqu'il s'agit de repérer les questions problématiques et en leur sein les propositions qui contiennent des anomalies ? La réponse doit être nuancée. Nous avons analysé les 173 QCM des 10 tests MOHICAN en utilisant les *rpbis spectraux* (*rpbis SC*, *rpbis SCT80* et *rpbis SCT90*) 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 signalées précédemment, 14 autres questions sont épinglées. Les *rpbis classiques* semblent indiquer des anomalies dans chacune de ces 14 QCM. Six QCM présentent des valeurs anormales aux *rpbis SC*. Une seule QCM obtient un *rpbis SCT80* anormal. Aucune obtient un *rpbis SCT90* anormal. Enfin, parmi ces 14 questions, seulement 3 sont pointées par les experts.

En ce qui concerne les trois questions signalées par les experts, ces derniers ont désigné un ensemble de propositions que seuls les *rpbis SC* signalent. 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.

Cette analyse montre que les différents types de rpbis déclenchent aussi ce que nous avons appelé des « 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 particulière. 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. Ces qualités de meilleure « détection » et de moins de « fausses alertes » sont cruciales lorsqu'il s'agit de mettre en évidence les QCM problématiques.

Lorsque nous corrigeons les anomalies contenues dans certaines propositions au sein des questions, nous pouvons désormais non seulement évaluer l'impact spectral de ces rectifications sur les 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 les plus problématiques du test de Connaissances en Histoire et Socio Economie et chiffré les gains en cohérence spectrale. L'amélioration de la cohérence spectrale de l'épreuve a aussi été mesurée en calculant la moyenne des valeurs obtenues aux indices spectraux des QCM. Ces indices moyennés ont ainsi permis d'évaluer l'impact spectral des rectifications effectuées sur les propositions des QCM à un troisième niveau, celui du test. En parallèle, nous avons aussi observé une amélioration de la fidélité du test à l'aide des indices classiques (alpha de Cronbach).

A l'aide des indices spectraux développés dans le cadre de notre thèse et utilisables à trois niveaux d'analyse spectrale : « PROPOSITIONS », « QCM » et « TEST », nous ouvrons une nouvelle voie pour l'analyse de la qualité des épreuves standardisées et leur régulation. Nous sommes en effet désormais 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 la contribution de notre thèse 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.

* * *