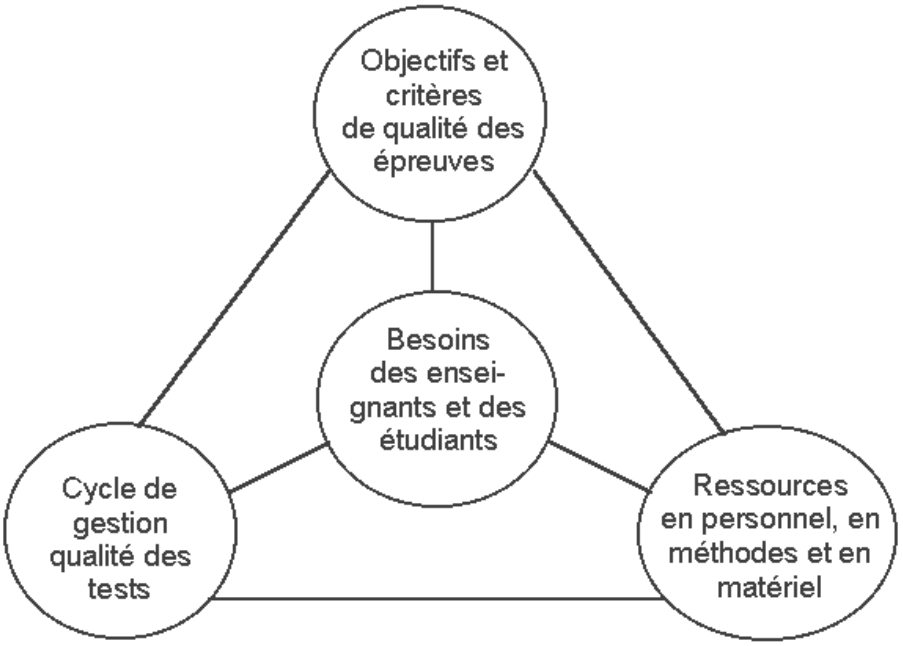
UNIVERSITE DE LIEGE
FACULTE DE PSYCHOLOGIE
ET DES SCIENCES DE L'EDUCATION

Construction et gestion qualité d'outils d'évaluation

Jean-Luc GILLES



Système Méthodologique d'Aide àla Réalisation de Tests

2001-2002

Les recommandations de la norme internationale ISO 9004-2[1] *(Gestion de la qualité et éléments de système qualité - Lignes directrices pour les services) mettent en évidence les facteurs clés d'un système qualité.*

Nous avons illustré ces facteurs clés transposés au contexte du Système Méthodologique d'Aide à la Réalisation de Tests de l'Université de Liège (SMART-ULg) à l'aide du schéma qui figure en couverture.

En haut du schéma nous avons placé les «Objectifs et critères de qualité des épreuves ».

Au centre : les « Besoins des enseignants et des étudiants », destinataires des services proposés par le SMART-ULg 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 « Cycle de gestion qualité des tests » et, d'autre part, les « Ressources en personnel, en méthodes et en matériel » nécessaires à la mise en œuvre du processus de réalisation des épreuves.

[1] *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.

Introduction

Le besoin :
des épreuves de qualité

Sommaire

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

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

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

D. Un exemple de régulation de la qualité des questions

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[2] :

	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.

[2] 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. 83) et les techniques de recueil (p. 88) des pourcentages de certitude qui sont à la base des indices de qualité spectrale des épreuves exposés dans les chapitres suivants (p. 149 et p. 175).

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[3]. 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.

[3] 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. 11) 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. 46), 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[4] du 29 mai 1999 : « Daniel, en deuxième licence philo et

[4] 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[5]. 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. 53), 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. 65).

[5] 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. 45) 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. 73). 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. 71).

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[6] « *…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. 86). 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[7] 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.

[6] Article signé par André Riche.

[7] 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.

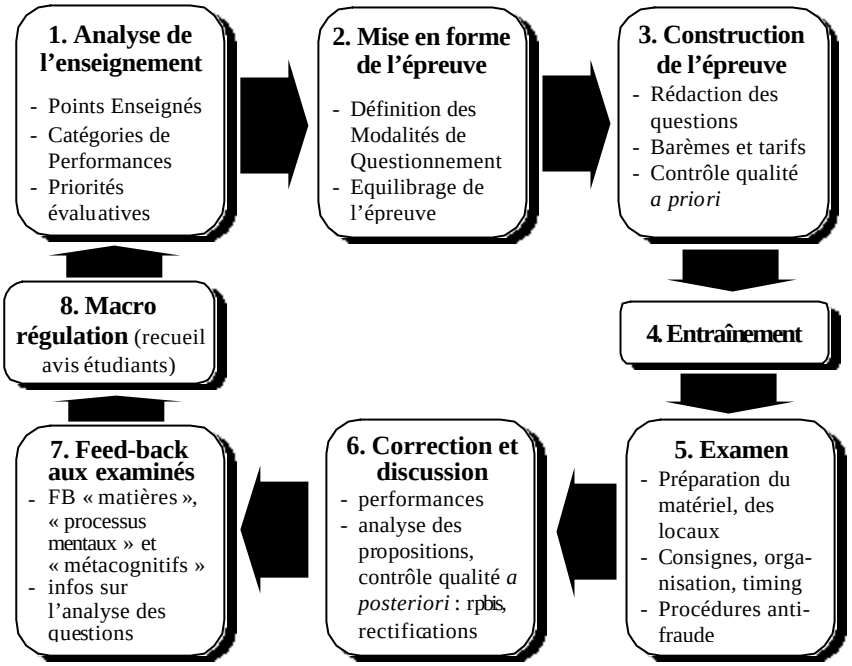
En fait, les questions à réponses ouvertes moyennes ou longues écrites améliorées à l'aide de techniques telles que les échelles descriptives, devraient être utilisées en complémentarité avec les épreuves standardisées ayant recours aux QCM. C'est ce que nous préconisons dans le cadre de l'aide méthodologique que nous proposons aux enseignants de l'Université de Liège via le Système Méthodologique d'Aide à la Réalisation de Tests (SMART) (voir p. 43). 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. 53) 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. 60).



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. 49) 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. 46)

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. 53) et les pourcentages de certitude (p. 54) 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. 60).

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. 49) 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.

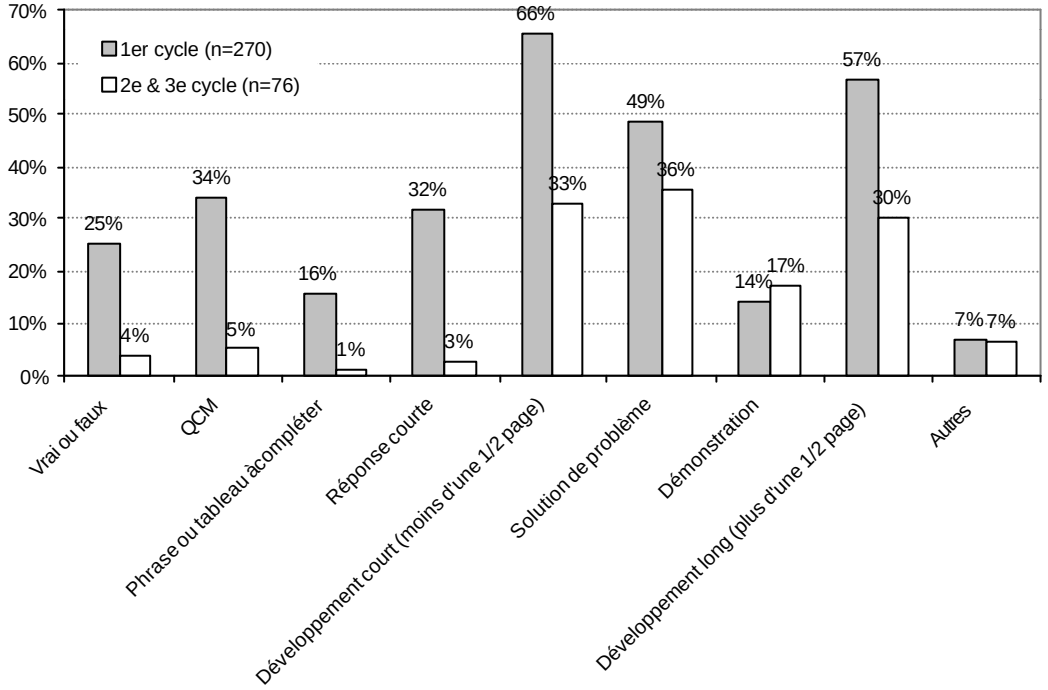
C. Fréquences des modalités de questionnement utilisées dans les examens universitaires

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[8] (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[9] (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 ½page)* », « *Solution de problème* », « *Démonstration* » et « *Développement long (plus d'une ½page)* » sont quant à elles du type Questions à Réponses Ouvertes Moyenne (QROM) ou Longues (QROL).

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

[9] 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%.

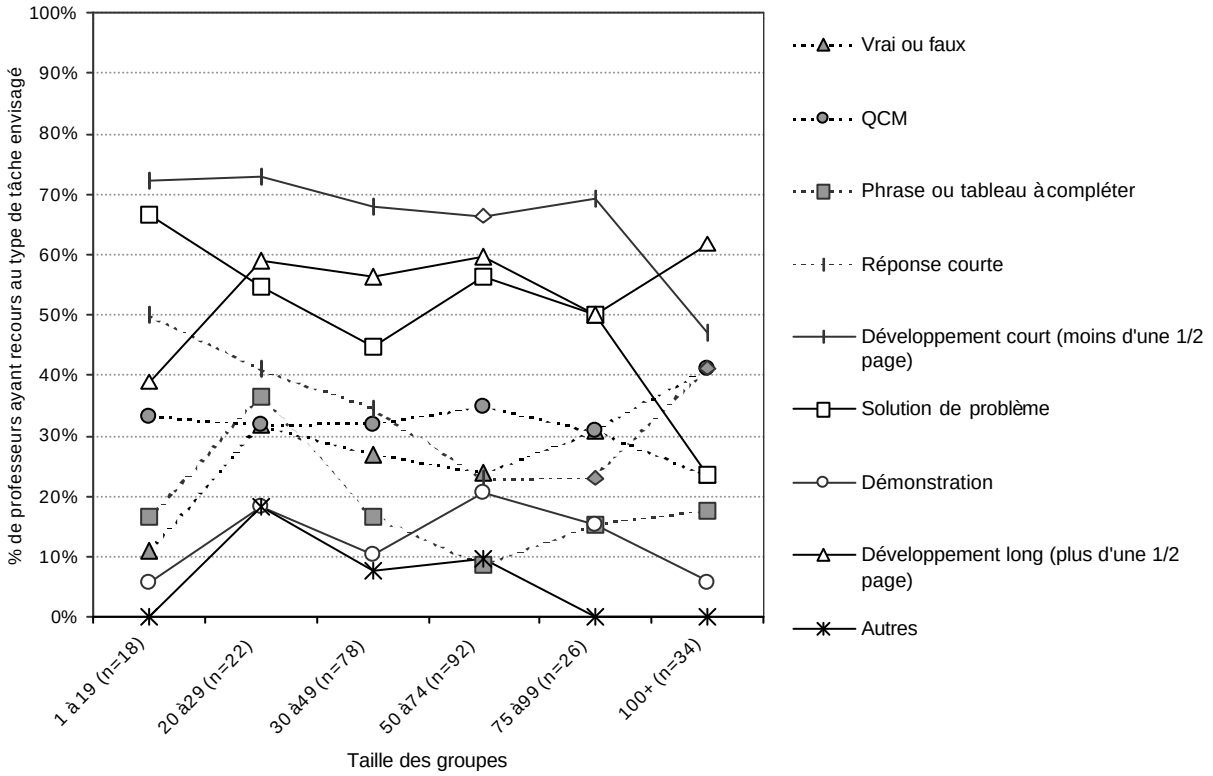
On voit que dans le 1er cycle toutes les modalités de questionnement sont utilisées. La plus fréquente est celle qui consiste à demander un développement court de moins d'une ½ page (66% des interrogés du 1er cycle l'utilisent). On remarque des pourcentages d'utilisation moins élevés pour les formes 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)



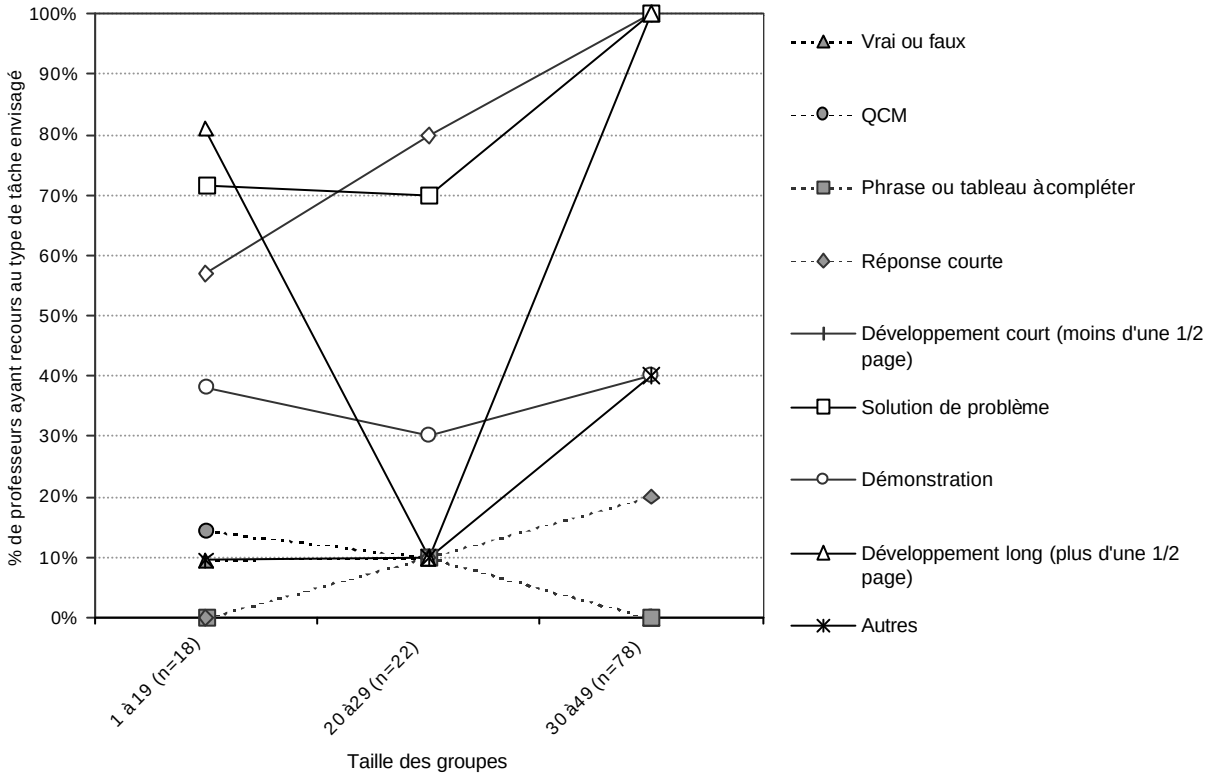
On remarque que même lorsque la taille des groupes est supérieure ou égale à 100 étudiants [100+ (*n*=34)], les questions à réponses ouvertes longues (« *Développement long (plus d'une ½ page)* ») et moyennes (« *Développement court (moins d'une ½ page)* ») sont les plus utilisées par les professeurs.

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[10] pour l'année académique 1997-1998.

[10] 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.

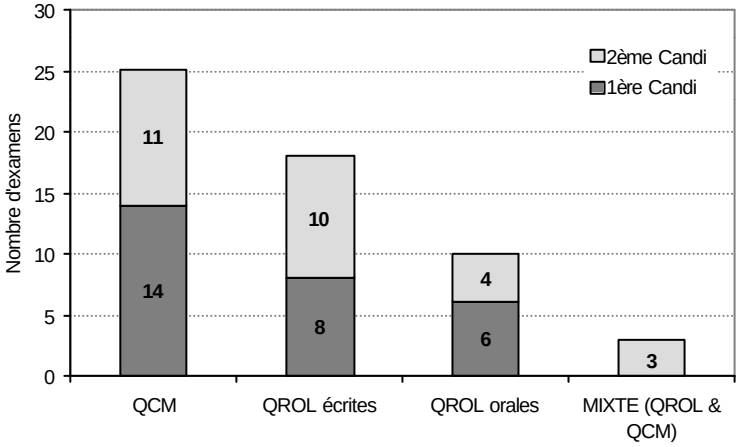
Ce travail a abouti à un rapport intitulé « *Observations sur les Etudes Universitaires à la Faculté (ŒUF)* ». Il montre notamment (*op. cit.*, pp. 14-16) que trois modalités d'examens sont essentiellement utilisées en candidatures :

- 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)

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.

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



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 [11]	% 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 <u>pourcentages de réussite</u> 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

[11] 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[12] 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[13] 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.

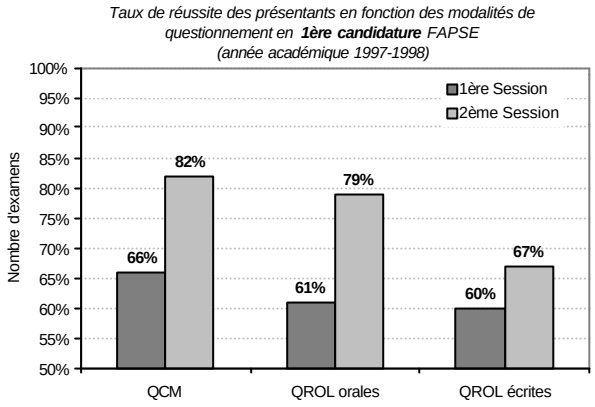
	QROM/QROL ORALES		QROM/QROL ECRITES		VF/QCM STANDARDISES	
	% réussite des présentants	Indice de réussite	% réussite des présentants	Indice de réussite	% réussite des présentants	Indice de réussite
1ère SESSION	-	-	57%	50,7	66%	64,6
2ème SESSION	78%	66,1	62%	52	85%	77,3

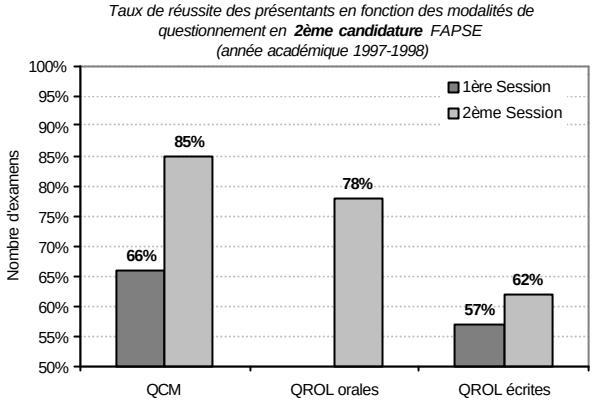
[12] Le cours 2C13 étant évalué à l'aide d'un examen dispensatoire écrit et d'un examen oral et le 2C14 par un QCM et une épreuve orale.

[13] 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[14] (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.

[14] 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[15]. « *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. 17).

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. 11) ;

[15] 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. 53, 54 et 83) ;
- 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[16] : « … *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.

***

[16] 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. 53).

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. 43) propose un processus en « spirale de qualité » (p. 60) 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[17] organise régulièrement en collaboration avec le Système Méthodologique d'Aide à la Réalisation de Tests (SMART, voir p. 43) 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.

[17] 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. 53). 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. 151) et est à la base des nouveaux indices de cohérence spectrale que nous proposons plus loin (pp. 158 et 164) 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. 156). 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[18] 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.

[18] 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.

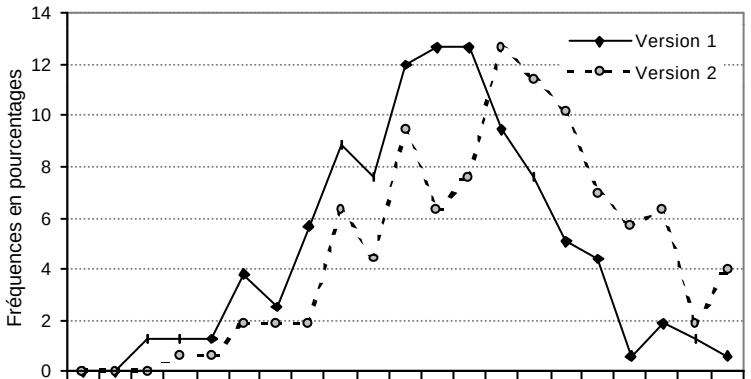
	0	1	2	3	4	5	6	7	8	9
Q1	0.0	2.5	5.7	15.1			**56.0**	15.1	4.4	0.6
rpbis	0.00	-0.02	-0.15	-0.18			**0.48**	-0.16	-0.33	-0.11
Cmoy	0.00	52.5	62.5	64.27			**59.13**	52.81	40.36	77.5
Q2	0.0	6.3	0.6	0.6	4.4		**87.4**	0.0	0.0	0.0
rpbis	0.00	-0.19	-0.2	-0.06	-0.03		**0.22**	0.00	0.00	0.00
Cmoy	0.00	60.0	12.5	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



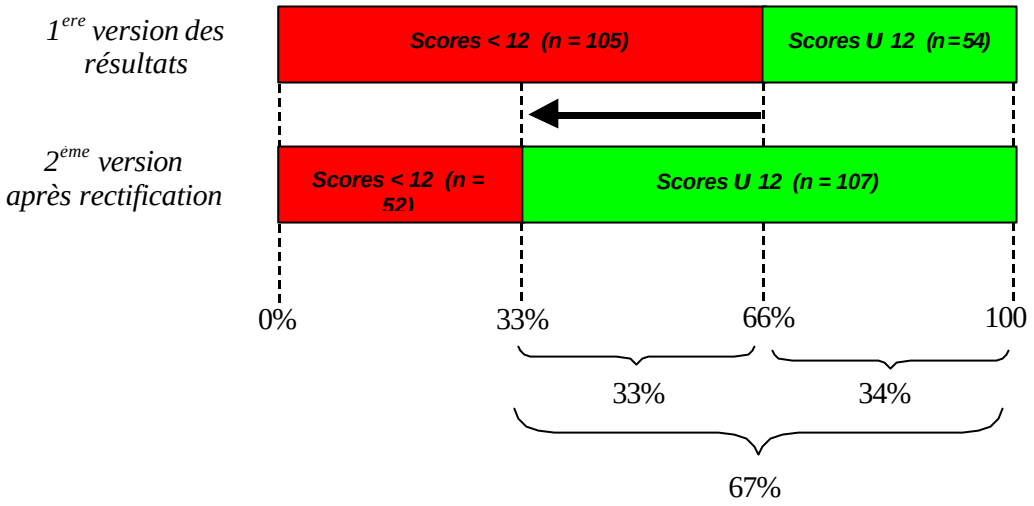
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Version 1	0	0	1	1	1	4	3	6	9	8	12	13	13	9	8	5	4	1	2	1	1
Version 2	0	0	0	1	1	2	2	2	6	4	9	6	8	13	11	10	7	6	6	2	4

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[19] ;
- 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.

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