CHAPITRE 1. QUAND LA COUTUME TIENT LIEU DE COMPÉTENCE : LES PRATIQUES D'ÉVALUATION DES ACQUIS À L'UNIVERSITÉ

Marc Romainville

in Nicole Rege Colet et Marc Romainville , *La pratique enseignante en mutation à l'université*

De Boeck Université | *Perspectives en éducation et formation*

2006
pages 19 à 40

CHAPITRE 1

Quand la coutume tient lieu de compétence : les pratiques d'évaluation des acquis à l'université

Marc Romainville [1]

Bien qu'elles occupent une place importante dans l'exercice du métier d'enseignant-chercheur et qu'elles conditionnent, dans une large mesure, l'apprentissage des étudiants, les pratiques d'évaluation dans le supérieur restent assez mal connues. Le présent chapitre vise à lever un coin du voile qui recouvre pudiquement ces pratiques. Pour l'essentiel, il se base sur un rapport que l'auteur a réalisé à la demande du Haut Conseil de l'Évaluation de l'École et dont la version longue et complète est disponible à l'adresse suivante : http://cisad.adc.education.fr/hcee.

Plus précisément, ce chapitre est une version courte et remaniée de la deuxième partie de ce rapport, qui en comptait quatre. De manière à situer ce chapitre dans le contexte général de la mission commanditée par le Haut Conseil, un bref aperçu des autres questions abordées par le rapport est présenté ci-dessous.

Le rapport s'ouvre sur une constatation : les connaissances dont on dispose sur les acquis des étudiants sont très lacunaires et fragmentaires. On

1 Facultés universitaires de Namur (Belgique)

ne sait, en général, pas trop bien ce que connaissent et ce que savent faire les étudiants à la sortie de leurs études supérieures. Cette ignorance locale, nationale et internationale contrecarre toute tentative d'évaluation des hautes ambitions des formations du supérieur. De manière à comprendre cet état de fait, la deuxième partie, synthétisée dans le présent chapitre, s'interroge sur les pratiques d'évaluation en cours à l'université et cherche à dégager les traits dominants de ces pratiques, dont la grande hétérogénéité et l'absence d'explicitation. La troisième partie dresse un inventaire des effets dommageables des pratiques actuelles d'évaluation, comme la transformation, aux yeux de l'étudiant, de l'examen universitaire en « jeu du chat et de la souris ». La quatrième partie décrit les pistes actuellement explorées pour améliorer ces pratiques, comme les épreuves intégrées et le portfolio.

Les **pratiques réelles d'évaluation** des acquis des étudiants sont **encore mal connues**. Par exemple, l'inventaire pourtant minutieux des tâches du professeur d'université réalisé par Bertrand et ses collègues (1994) ne fournit pas d'indications précises sur la part du travail professoral consacrée aux tâches d'évaluation et sur les comportements qu'elles impliquent. De même, on connaît mal la perception qu'ont les étudiants des pratiques d'évaluation. On sait quand même qu'une plus grande cohérence des pratiques d'évaluation et de notation constitue une de leurs principales revendications quand ils sont interrogés à propos des lacunes de l'enseignement universitaire (Blais, Laurier, Van der Maren, Gervais, Lévesque & Pelletier, 1997 ; Gibbs, 1995). De même, quel que soit leur profil d'apprentissage, ils se montrent fort critiques sur les examens traditionnels. Ces derniers font, à leurs yeux, trop souvent appel à la seule mémorisation, pas assez à l'intégration des connaissances et les incitent à développer des apprentissages superficiels, qui ne leur seront d'aucune utilité dans leur vie professionnelle (Blais *et al.*, 1997 ; Lindblom-Ylänne & Lonka, 2001).

Il existe cependant des études, mais en nombre limité, qui **analysent minutieusement les pratiques d'évaluation des acquis des étudiants**. Citons-en quatre, parmi celles qui seront le plus abondamment utilisées dans ce chapitre. La première est une recherche de Warren Piper (1994), qui porte principalement sur les évaluateurs externes des universités anglaises (*external examiners*). La deuxième étude a été produite par l'Agence de modernisation des universités qui a consacré une de ses Rencontres à « L'organisation et la préparation des examens ». La troisième étude est une enquête fouillée sur les pratiques d'évaluation des apprentissages menée à l'Université de Montréal et dans ses écoles affiliées, par questionnaire auprès de 643 professeurs et par entretiens auprès de plusieurs échantillons d'enseignants et d'étudiants (Blais *et al.*, 1997). Enfin, un petit nombre d'études récentes commencent à investiguer, dans le détail et de manière presque clinique, les comportements évaluatifs des enseignants du supérieur (York, Bridges & Woolf, 2000). Par la technique de la réflexion parlée ou par celle des entretiens de groupe, on cherche à savoir quels sont les comportements courants des enseignants du supérieur quand ils sont confrontés à des tâches authentiques d'évaluation.

Par exemple, comment élaborent-ils leur jugement face à une copie ? Selon une démarche « platonicienne » de comparaison entre une copie modèle qu'ils ont en tête et les copies réelles ? Selon une approche intuitive, en regard de leur expérience antérieure de correcteur d'épreuves équivalentes ? En référence à une grille explicite de critères dont la pondération a été élaborée à l'avance ?

Quelles conclusions d'ensemble peut-on tirer de ces diverses études et des entretiens menés dans le cadre de la mission du Haut Conseil ?

1. UN ALOURDISSEMENT DES TÂCHES D'ÉVALUATION, AYANT DES EFFETS NÉGATIFS SUR SA QUALITÉ

La première conclusion qui se dégage est que l'évaluation des acquis des étudiants est devenue une **composante de plus en plus lourde du métier d'enseignant-chercheur**. Les tâches d'évaluation occupent une place grandissante au sein de leurs activités. Les enseignants déplorent que le temps consacré à l'évaluation grignote progressivement celui qui est dévolu à l'enseignement. La répartition entre 24 à 25 semaines d'enseignement et 12 à 13 semaines d'évaluation a souvent été évoquée. On la juge inadéquate, sans compter que des épreuves de contrôle continu sont aussi organisées au cours de la période dite d'enseignement. La massification, la modularisation et la semestrialisation sont souvent montrées du doigt comme étant les principales responsables de cette augmentation considérable du poids des examens.

Outre l'alourdissement des tâches d'évaluation qu'elle entraîne pour les enseignants-chercheurs, cette évolution semble aussi avoir produit des **effets négatifs sur la qualité de l'évaluation** : diminution des possibilités laissées aux étudiants de choisir leur formule d'examens, non-tenue de certaines réunions de jury, modification de la nature même des épreuves : régression, voire disparition des oraux et diminution du contrôle continu (Poirier, 2001 ; Girod de l'Ain, 1997). La multiplication d'examens déconnectés les uns des autres aboutirait au développement d'un *« travail intellectuel morcelé et concentré sur les prochains "obstacles" au détriment de la réflexion et de l'approfondissement, attributs fondamentaux d'études universitaires »* (Girod de l'Ain, 1997, p. 108).

Une même tendance est observée au Royaume-Uni (Montgomery, 1995), en Allemagne (Kehm, 2001) et au Québec (Blais *et al.*, 1997). D'après cette dernière étude, plus les effectifs du premier cycle sont nombreux, plus l'examen final écrit constitue le mode d'évaluation privilégié. Les enseignants veillent cependant à ce que l'examen écrit final comporte des tâches complexes, comme la résolution de problèmes ou le développement long ou court. Ils n'attribuent généralement pas un poids démesuré aux tâches les plus simples (QCM, Vrai-faux). On note quand même que 9 % de ceux qui ont recours aux QCM leur confèrent un poids supérieur à 80 %. Les modalités d'évaluation sont aussi davantage diversifiées pour un petit groupe (exposé, examen inter-

médiaire, travail de recherche…) que pour un grand groupe. Un lien entre la taille des groupes et le type d'habilités mesurées semble, de plus, établi par cette enquête : les enseignants évaluent davantage la mémorisation au sein des grands groupes de premier cycle et cette habilité occupe ensuite une part de plus en plus ténue dans les examens des cycles supérieurs. Les enseignants soulignent aussi que l'accroissement du nombre d'étudiants rend plus problématique la réalisation d'une évaluation personnalisée.

Bref, si les pratiques effectives d'évaluation sont encore assez mal connues, on sait au moins qu'elles représentent une part importante du travail professoral. On sait aussi que cette part aurait tendance à s'accroître, au grand dam des enseignants-chercheurs, suite à l'explosion des effectifs et aux réformes des structures de l'enseignement universitaire (semestrialisation et modularisation). On redoute enfin que cette évolution ait eu des conséquences importantes sur la qualité des pratiques d'évaluation, sur la nature des tâches proposées aux étudiants et, en définitive, sur le type de connaissances et de compétences qu'ils développent.

2. DES PRATIQUES D'ÉVALUATION PEU STANDARDISÉES

La deuxième conclusion majeure qui se dégage des études portant sur les pratiques d'évaluation dans le supérieur est la **grande hétérogénéité** de ces pratiques. À cause de la tradition d'autonomie des établissements d'enseignement universitaire et de la « liberté académique » dont jouissent leurs membres, l'évaluation des acquis, en ce compris la délivrance des diplômes, est de la compétence de chaque établissement et de chacun des enseignants, ces derniers évaluant leurs propres étudiants.

Bien sûr, des textes légaux encadrent l'évaluation des acquis dans l'enseignement universitaire et assurent un minimum de standardisation des procédures générales. Ainsi, une circulaire du 1er mars 2000 relative à l'organisation des examens dans les établissements publics de l'enseignement supérieur français rappelle les réglementations en vigueur en matière d'examens universitaires. Concernant l'organisation des examens, la circulaire précise les règles qui régissent les aspects suivants : la convocation des étudiants, la constitution du jury, l'élaboration d'un procès-verbal pour chaque épreuve, les comportements à adopter en cas de fraude, la communication des résultats aux étudiants, les conditions de réussite de l'année et de l'obtention du diplôme et l'accès aux copies. Mais le texte de la circulaire insiste surtout sur la nécessité que chaque établissement définisse ses règles propres, les explicite et les fasse connaître dans la plus grande transparence et tôt dans l'année aux étudiants, étant donné qu'un grand nombre de décisions cruciales sont en fait prises au niveau de chaque établissement, voire de chaque unité d'enseignement. Ainsi, la loi du 26 janvier 1984 ne trace qu'un cadre très général et pour le moins ouvert sur le rapport entre évaluation continue et terminale :

« les aptitudes et l'acquisition des connaissances sont appréciées, soit par un contrôle continu et régulier, soit par un examen terminal, soit par ces deux modes de contrôle combinés ». Reste que la formule permet localement de nombreux cas de figure… De plus, ces textes définissent les règles générales à suivre quant à l'organisation des épreuves, mais ils ne disent rien de leur contenu.

Les pratiques d'évaluation se différencient d'abord en regard de **l'importance relative** qu'elles accordent au **contrôle continu** par rapport au contrôle terminal et en regard de la **nature** des épreuves de contrôle continu. Comme nous l'avons noté ci-dessus, la part relative du contrôle continu dans l'évaluation semble tributaire de la taille du groupe d'étudiants : plus les étudiants sont nombreux, plus les enseignants réduisent la part du contrôle continu, face à la lourdeur des corrections (Gibbs & Lucas, 1997). Or les travaux exigés des étudiants tout au long du cours semblent garantir un certain niveau dans la qualité des apprentissages : ils incitent les étudiants à privilégier le travail régulier plutôt que le bourrage de crâne de dernière minute, ils leur fournissent des feedbacks formateurs… Les résultats des étudiants au contrôle continu semblent d'ailleurs, davantage que leurs résultats aux examens, corrélés aux mesures d'acquis effectuées à long terme, sans doute parce qu'ils se rapprochent des tâches qu'ils seront amenés à exécuter dans leur vie professionnelle.

Une autre dimension de l'hétérogénéité des pratiques d'évaluation a trait aux **spécificités disciplinaires** : l'évaluation des acquis se réalise différemment selon la filière d'études (Brown & Glasner, 1999 ; Heywood, 2000 ; Warren Piper, 1994). Ainsi, les épreuves de calcul numérique semblent plus fréquentes en sciences, les travaux pratiques aussi. Le recours à des performances non langagières s'observe surtout en faculté d'ingénieur et de technologie. Plus curieusement, l'examen oral semble privilégié en sciences. Les enseignants de ces filières estiment qu'il s'agit d'une méthode valide qui permet de discriminer fortement les étudiants, sans doute parce que l'oral permet de reposer des questions précises et de « tester » ainsi la compréhension fine de concepts spécifiques. À l'inverse, les enseignants des facultés des sciences humaines et sociales et de lettres émettent des doutes sur la validité de cette forme d'examen et privilégient des épreuves écrites à réponses ouvertes (l'essai, la dissertation et les questions à développement long ou court).

Les spécificités disciplinaires s'observent jusque dans les distributions des notes des étudiants : ces distributions sont en effet différentes d'une discipline à l'autre, en termes de moyenne, d'écart-type et d'allure générale de la courbe (Bridges, Bourdillon, Collymore, Cooper, Fox, Haines, Turner, Woolf & Yorke, 1999 ; York *et al.*, 2000). Ainsi, les disciplines scientifiques ont tendance à recourir à une large palette de notes alors que l'empan des notes est considérablement plus restreint dans les autres filières. En conséquence, la proportion des mentions y est dès lors aussi fort différente : ainsi, dans les universités anglaises, alors que 21 % des étudiants en mathématiques obtiennent la meilleure mention, ils ne sont que 3,7 % a y avoir accès dans les études juri-

diques. Cette différence soulève un problème d'équité entre étudiants de différentes filières : ils n'ont pas, au départ, la même probabilité d'obtenir des mentions. En cherchant à comprendre ce qui pourrait expliquer ces conventions implicites de notation spécifiques à chaque filière, Bridges *et al.* (1999) montrent que la nature des connaissances et des compétences qui font l'objet de l'évaluation peut être mise en parallèle avec l'usage de l'échelle de notation. En histoire par exemple, les enseignants déclarent chercher à mesurer la maîtrise de démarches assez générales, telles que la critique des sources. Il leur semble que les extrémités de l'échelle constituent des « territoires peu sûrs », qu'ils préfèrent, dès lors, ne pas explorer l'excellence et, à l'autre bout de l'échelle, la médiocrité dans la maîtrise de ces démarches étant rarement atteintes. Par contre, des enseignants de mathématiques et d'informatique déclarent procéder à une addition de réponses correctes ou incorrectes et s'autorisent alors à recourir à l'ensemble des échelons de l'échelle de notation, y compris dans ses deux extrémités.

L'absence de standardisation des pratiques d'évaluation s'observe aussi à **l'intérieur d'une même filière**, même dans le cas de diplômes nationaux. Les rapports transversaux du Conseil national d'évaluation fournissent de belles illustrations de l'hétérogénéité des pratiques. À titre d'exemple, le rapport portant sur les formations offertes par les 24 UFR de pharmacie aboutit à la conclusion suivante, en ce qui concerne le contrôle des connaissances : *« chaque UFR organise librement, dans le cadre de son autonomie, la préparation au concours et décide des épreuves, ce qui pose la question de l'homogénéité des résultats »* (CNE, 1998a, p. 34).

L'étude de Jarousse et Michaux (2001) montre également que les modalités d'évaluation varient considérablement d'une filière à l'autre, mais aussi d'une université à l'autre pour une même filière. D'après leur enquête auprès de 155 responsables de DEUG, les pratiques d'évaluation des acquis se différencient sur les points suivants :

- Les conditions générales de certification : certification annuelle avec capitalisation modulaire (avec ou sans condition de moyenne et compensation entre les modules), capitalisation modulaire avec compensation à l'issue des deux années, validation semestrielle des acquis…
- Les conditions de rattrapage : place de la session de rattrapage (en septembre ou juste après la session terminale), modalités de conservation des notes (condition de moyenne, choix de la note conservée : la meilleure des deux notes, celle de la session de rattrapage)…
- La pondération des modules et du contrôle continu. Ici aussi, c'est la diversité qui domine. Le contrôle continu se voit accorder une importance tantôt proportionnelle à sa part dans le volume d'enseignement, tantôt inférieure, tantôt supérieure. Son poids dans l'évaluation peut ainsi passer de 35 % dans une filière à 59 % dans une autre. Dans le même sens, le poids des modules obligatoires et optionnels est aussi très variable : pour une même filière, ils ont un

poids équivalent dans tel établissement, alors que les modules obligatoires ont un coefficient de 3 dans tel autre.

En bout de course, l'évaluation des acquis des étudiants apparaît donc comme entièrement décentralisée : « *aucune épreuve commune ne permet de comparer sur une base externe ce que les étudiants ont effectivement appris durant leur cursus* » (Duru-Bellat, Jarousse, Leroy-Audouin & Michaut, 2000, p. 138). Il en découle que la définition même de la réussite est contextualisée : « *Tant les notes que les décisions de passage (…) sont entièrement fabriquées au sein de chaque filière* » (*ibid.*, p. 138). Chaque équipe d'enseignants (voire chaque enseignant tant la notion d'équipe reste, à l'université, une chimère) développe des pratiques d'évaluation spécifiques, qui tiennent probablement compte des caractéristiques des étudiants accueillis et qui visent implicitement à respecter certaines normes docimologiques tacites, par exemple une moyenne du groupe d'étudiants qui ne soit ni trop basse ni trop élevée.

Bien sûr, il n'y a pas de règle sans exception. L'Association nationale des directeurs de maîtrise en sciences de gestion a institué, depuis plusieurs années, des épreuves communes d'admission à la Maîtrise en Sciences de gestion. Ces épreuves, dénommées Message, constituent le socle commun des examens d'admission pratiqués au sein des différents établissements, chacun de ces établissements disposant de la liberté d'ajouter localement des épreuves spécifiques. Un des objectifs affichés de cette collaboration est d'assurer une plus grande homogénéité des procédures de recrutement. Ces épreuves sont décrites sur le site http://www.msgfrance.org/message.htm.

Plusieurs documents consultés s'interrogent sur les multiples causes de cette absence de standardisation des pratiques d'évaluation des acquis. Citons les deux principales.

2.1 Absence de standardisation des objectifs, des contenus et des méthodes

Le premier facteur explicatif du manque de standardisation des pratiques d'évaluation réside bien sûr dans l'absence de standardisation des objectifs, des contenus et des méthodes des formations elles-mêmes. Autrement dit, si les acquis des étudiants sont évalués de manière très diversifiée, c'est parce que les compétences attendues aux différents paliers de la formation n'ont pas fait l'objet d'une définition explicite, précise et standardisée. Dit plus crûment, si nous ne savons pas ce que les étudiants ont appris et si l'impression qui se dégage est qu'ils sont évalués différemment, c'est notamment parce qu'ils ont appris, pour un même intitulé de diplôme, des choses extrêmement différentes selon l'établissement dans lequel ils ont poursuivi leurs études.

Ce phénomène s'explique par la très large autonomie dont jouissent les établissements et les enseignants du supérieur quant à la définition du contenu précis des cursus. En France, l'autonomie croissante dont ont bénéficié

les universités en matière d'organisation pédagogique et les effets de cette autonomie ont été particulièrement bien décrits récemment par Musselin (2001), Jarousse et Michaut (2001) et Duru-Bellat *et al.* (2000).

Dans sa description de la lente émergence des universités comme entités intermédiaires entre l'État et les facultés, Musselin (2001) analyse comment, malgré les deux principes d'uniformisation et d'égalitarisme reconnus par la loi et auxquels adhèrent les acteurs de l'université, les formations sont devenues, dans les faits, de plus en plus hétérogènes. La palette des formations s'est considérablement étendue, le public lui-même est devenu plus hétérogène et surtout, pour s'y adapter, les pratiques locales se sont diversifiées. Musselin (2001) décrit aussi comment les autorités de tutelle ont cherché en vain à concilier l'augmentation sans précédent du nombre de cursus et la garantie d'équivalence entre ceux-ci. Ni la multiplication des maquettes en proportion des nouveaux cursus, ni la définition très ouverte de ces maquettes ne constituent des solutions satisfaisantes. Elles se contentent en effet, pour les diplômes nationaux, de préciser le nombre minimal d'heures de cours, les règles générales de l'évaluation des connaissances et une liste de sujets obligatoires à traiter (Duru-Bellat *et al.*, 2000). Définie de manière tellement large, elle n'offre plus la garantie d'équivalence pour laquelle elle avait été créée : *« dans un cas comme dans l'autre, le maintien des maquettes nationales est peu compatible avec la diversité croissante des filières et rend de plus en plus illusoire la garantie de l'homogénéité des cursus d'un bout du territoire à l'autre »* (Musselin, 2001, p. 100).

Jarousse et Michaut (2001) ont aussi mis en évidence l'hétérogénéité considérable des modes d'organisation pédagogique des premiers cycles, qui se différencient quant à l'organisation générale des études (ex. la proportion des divers types d'enseignements (magistral, TP, TD…), quant au contenu lui-même du curriculum (ex. importance accordée aux différentes disciplines, notamment en termes de volume horaire) et quant aux dispositifs d'aide à la réussite mis à la disposition des étudiants (ex. forme des tutorats et volume horaire associé).

Une conséquence importante de cette grande diversité réside dans le fait que les étudiants inscrits dans une même filière sont confrontés à des conditions d'enseignement très différentes selon l'université fréquentée. Il est dès lors probable qu'ils y développent des compétences et des connaissances différentes, qui sont évaluées de manière spécifique et non standardisée par rapport à d'autres sites.

Dans ce domaine aussi, les rapports transversaux du Conseil national d'évaluation sont très instructifs. Par exemple, celui qui porte sur les formations de pharmaciens relève que, même pour ce diplôme national avec concours, l'arrêté qui fixe le contenu des cursus, s'il évoque bien *« une mosaïque de plus d'un millier de notions qui doivent être obligatoirement étudiées »,* ne précise *« ni le poids relatif, ni le poids absolu des différentes matières ou disciplines, ni la chronologie des enseignements »* (CNE,

1998a, p. 36). Le rapport regrette aussi que les objectifs soient assez mal définis et le soient en termes trop généraux. Une conséquence importante est que les solutions retenues par les différentes UFR pour poursuivre une même finalité peuvent varier considérablement, au point sans doute de ne pas assurer les mêmes acquis à tous les étudiants. Ainsi, pour développer les capacités de synthèse de l'étudiant, telle UFR multiplie les renvois bibliographiques en maintenant le volume horaire constant, telle autre procède à une diminution de la charge horaire de 20 % et introduit des enseignements thématiques (CNE, 1998a, p. 37). Le rapport Fitoussi (2001) sur l'enseignement supérieur de l'économie évoque, lui aussi, cette question de l'absence de standardisation des niveaux d'études.

Cette tendance à douter de l'équivalence des titres nationaux s'observe aussi au Royaume-Uni. Selon le rapport de Warren Piper (1994), la large autonomie accordée aux universités dans l'évaluation des acquis des étudiants remet en cause l'équivalence de niveau des différents diplômes. Celle-ci est décrite comme une « fiction polie », même si des instituts de contrôle de la qualité sont censés la garantir. Une tension est vécue entre des exigences de diversité des programmes (notamment pour répondre à des publics variés) et des exigences d'équité.

2.2 Une longue tradition de « liberté académique »

Un deuxième facteur explicatif de l'absence de standardisation des pratiques d'évaluation des acquis réside dans la grande autonomie dont jouit, en sus de l'autonomie institutionnelle, chaque enseignant-chercheur dans l'organisation de son enseignement et dans la mise au point du contenu de ses examens. Quand bien même la description officielle d'un même curriculum serait semblable dans deux établissements, les enseignants à qui sont attribués les cours se retrouvent pratiquement libres de déterminer leur contenu et leur forme. Le curriculum réel peut dès lors être très différent, pour un même descriptif officiel de cours. La « liberté académique » n'est pas un vain mot. Dans la conception humboldtienne de l'université moderne, la notion de liberté académique représente une exigence d'autonomie institutionnelle de l'Université, d'indépendance de cette dernière par rapport à l'État et aux autorités publiques. Cette autonomie est considérée comme une condition indispensable au développement d'une recherche de qualité et créative, parce que cette dernière peut alors se déployer hors des contraintes limitantes et à court terme du pouvoir (Renaut, 1995). Cette exigence d'autonomie s'étend aussi à l'individu, qui doit pouvoir, au sein de l'Université, poursuivre sa recherche *« dans la solitude et la liberté », « sans qu'une contrainte ou un but déterminé lui soit imposé »* (Renaut, 1995, p. 130). On comprend bien comment cette tradition de liberté académique a laissé de profondes traces, y compris en qui concerne la liberté individuelle revendiquée par chaque enseignant d'organiser son enseignement et ses examens comme il l'entend (Dejean, 2002 ; Lahire, 1997).

3. UNE ÉVALUATION NORMATIVE QUI NE SE FONDE PAS SUR DES OBJECTIFS EXPLICITES DE FORMATION

L'hétérogénéité des pratiques d'évaluation n'est peut-être pas condamnable en soi, si l'on met de côté les problèmes d'équité et d'équivalence des titres. Elle pourrait même se justifier en termes d'adaptation pédagogique aux caractéristiques des « nouveaux » étudiants que l'université de masse a en charge d'accueillir ou en termes d'adéquation des compétences de sortie au contexte socioprofessionnel régional. Si le fait d'évaluer, pour une même filière, des acquis différents au sein d'établissements autonomes pourrait donc être encore acceptable, on voit mal ce qui pourrait justifier le fait de ne pas savoir précisément quels acquis font l'objet de l'évaluation. L'absence d'explicitation de ce que l'on mesure est en effet gênante à plus d'un titre : les étudiants manquent de repères sur les manières d'étudier les plus propices à les préparer aux examens et les futurs employeurs ne disposent pas d'inventaires précis des compétences de ceux qui se présentent sur le marché de l'emploi.

Par ailleurs, l'explicitation des objectifs de formation devrait, en principe, être au cœur du processus évaluatif : évaluer revient à recueillir de l'information de manière systématique, fidèle et valide pour prendre des décisions, en particulier pour juger de l'atteinte des objectifs de formation par les étudiants. L'évaluation suppose donc que les objectifs des cursus aient été préalablement explicités en termes de connaissances et de compétences, pour qu'elle puisse se réaliser en regard de ces objectifs. Or la troisième conclusion majeure qui s'impose à la lecture des études sur les pratiques d'évaluation est que le système des examens universitaires fonctionne la plupart du temps en l'absence de définition, même locale, des connaissances et des compétences qui sont attendues en fin de formation.

Le type d'évaluation qui se développe dans l'enseignement universitaire est essentiellement normatif : l'évaluation s'attache principalement à classer les étudiants les uns par rapport aux autres. L'attribution contextualisée des notes et le caractère local des décisions de réussite participent au développement d'une évaluation qui ne cherche pas prioritairement à mesurer les compétences acquises par les étudiants en fonction d'objectifs explicites de formation, mais bien à les distinguer les uns des autres L'ajustement local des épreuves permet ainsi aux enseignants de conserver, d'année en année, une distribution de leurs notes jugée acceptable. Ce phénomène a bien été décrit aux premier et second degrés sous le nom d'effet « Posthumus » : un enseignant a tendance à ajuster le niveau de son enseignement et de son évaluation de façon à conserver d'année en année approximativement la même distribution gaussienne de ses notes. Sans doute cet effet est-il également à l'œuvre dans le supérieur, avec des normes implicites différentes, cela va sans dire. Ainsi, une distribution jugée acceptable en première année du premier cycle se rapproche davantage de la courbe en i que de la distribution normale…

Lors des entretiens et de l'étude de cas mené dans le cadre de la mission du Haut Conseil, des pratiques d'évaluation ouvertement normatives ont été rapportées. Ainsi, il est fréquent que des notes provisoires soient d'abord attribuées à une série de copies, notes qui sont ensuite adaptées selon les résultats de l'ensemble du groupe, de manière à aboutir à cette fameuse norme implicite. Autre exemple, la correction s'effectue parfois en recherchant d'abord une très bonne et une très mauvaise copie puis en ajustant ensuite les autres sur ces deux extrêmes. Plus clairement encore, des programmes informatiques de corrections automatiques de questionnaires à choix multiples permettent d'ajuster le tarif de manière à obtenir la moyenne souhaitée. L'enquête de Blais *et al.* (1997) rapporte aussi le souci de certains enseignants de rechercher une « moyenne historique, raisonnable ».

À côté des problèmes d'équité que pose l'évaluation normative (les standards de qualité selon lesquels sont jugés les étudiants dépendent du groupe dans lequel ils se trouvent), cette évaluation pêche aussi par un manque de transparence sur les acquis. En effet, elle permet juste de savoir quels étudiants ont appris le plus ou le mieux, mais l'évaluation normative ne nous renseigne pas sur ce qu'ils ont acquis, ni même sur le fait qu'ils aient appris beaucoup ou peu. D'ailleurs, les gains en termes de connaissances (par rapport à leur niveau de départ) des étudiants en échec se révèlent parfois identiques à ceux des étudiants qui obtiennent des mentions (Astin, 1991). L'évaluation normative fournit plutôt une indication sur la manière dont les étudiants se classent les uns par rapport aux autres, à un moment donné.

Un autre reproche que l'on peut adresser à l'évaluation normative est qu'elle crée ce que les économistes appellent un « bien rare », puisque seul un pourcentage minime et prédéfini d'étudiants se trouveront en tête du groupe, indépendamment des efforts produits par les étudiants et de l'importance de leurs acquis. La rareté exacerbant l'importance d'un bien, les performances moyennes sont alors considérées comme médiocres : l'évaluation normative garantit en quelque sorte qu'une majorité d'étudiants ne soient pas satisfaits de leurs résultats…

Comment expliquer que l'évaluation des acquis soit essentiellement normative et se réalise peu en regard d'objectifs de formation ?

3.1 Des formations qui ne sont pas définies en termes de compétences à acquérir

Le premier facteur explicatif et le plus évident est que les formations elles-mêmes ne sont pas prioritairement conçues autour d'objectifs explicites d'acquisition de connaissances et de compétences. Pour comprendre cet état de fait, il faudrait analyser plus longuement les modes de constitution des programmes de l'enseignement supérieur. De nombreux indices incitent à penser que la confection des programmes (le choix des matières, leur contenu, leur articulation et leur répartition horaire) est souvent plus *« tributaire de la*

configuration des opportunités et des intérêts internes à l'institution » (Hutmacher, 2001, p. 43) que d'un inventaire méthodique des compétences à développer, par exemple sur la base d'une analyse des conditions d'exercice des métiers auxquels donne accès le diplôme. Comme le signale cet auteur, il s'agit sans doute d'un des obstacles majeurs à une véritable professionnalisation des formations supérieures.

Cette tendance à construire des plans de formation selon les opportunités, voire selon les personnes (leur profil, leurs intérêts de recherche, leur champ de compétences, etc.) qui en auront la charge a été particulièrement bien décrite par Kletz et Pallez (2001) dans l'étude qu'ils ont consacrée aux mécanismes de constitution de l'offre de formation des établissements universitaires. Pour notre propos, il faut retenir de cette étude que la création d'offres de formation trouve plus fréquemment son origine dans une initiative personnelle, liée parfois à des enjeux de carrière strictement individuels, que dans une analyse détaillée des besoins des formés ou encore un inventaire des compétences à développer chez ces formés en regard d'inadéquations qui auraient été observées entre leurs acquis et les besoins du marché du travail.

D'ailleurs, la logique qui préside encore de manière dominante à la confection des programmes de formation universitaire est une logique de contenus, interne au monde académique. Un programme reste un assemblage de matières et de cours, une mosaïque dont le découpage et l'agencement sont liés aux disciplines scientifiques de référence et à leur mode de structuration. Ainsi, le programme d'histoire est articulé autour des grandes spécialités reconnues qui ordonnent ce champ scientifique et la communauté de ses chercheurs. Une étude portant sur les processus de développement de nouveaux cursus confirme la prédominance de cette logique de discipline : *« les nouveaux projets étaient toujours lancés à l'initiative d'un universitaire ou d'un petit noyau de collègues, en fonction de leurs propres préoccupations et intérêts »* (Mignot-Gérard et Musselin, 2002, p. 12). Concevoir des formations en termes d'acquis représenterait une tout autre entrée : qu'est-ce que les étudiants doivent savoir et savoir faire à la sortie du programme ? Quel type de connaissances et de compétences cherche-t-on à leur faire mobiliser et face à quel type de situations ?

Le fait que les formations ne soient pas construites autour d'une explicitation de compétences à acquérir mais qu'elles s'élaborent selon des logiques internes à l'institution, selon son histoire, voire selon ses stratégies de positionnement concurrentiel, apparaît aussi dans de nombreux passages des rapports du CNE. Ainsi, l'évaluation des différentes UFR de pharmacie (CNE, 1998b) signale l'existence d'un déséquilibre dans la formation dispensée par telle unité, déséquilibre attribuable au fait que les enseignements des disciplines dont cette unité est spécialiste sont privilégiés au détriment d'autres disciplines. Un rapport d'évaluation institutionnelle d'une université déplore l'absence de réflexion sur l'offre de formation et conclut que *« l'offre de formation apparaît ainsi davantage comme la résultante d'une course pour*

augmenter au maximum les effectifs étudiants, que comme l'aboutissement d'une politique mûrie et réfléchie » (CNE, 2001, p. 74).

Dans son analyse des procédures d'habilitation des diplômes, Dejean (2002) montre bien comment même la rubrique « objectifs de formation » contient parfois un simple énoncé des contenus qui seront abordés dans la formation. Et, toujours d'après Dejean, un des mérites du développement des pratiques d'évaluation des enseignements résiderait précisément dans le fait qu'elles entraîneraient inévitablement un effort de formulation des objectifs de formation, puisque l'évaluation des enseignements doit se réaliser, selon l'arrêté qui la met en place, en relation avec les objectifs de ces enseignements.

Sans doute devrait-on aussi se poser la question plus générale (mais qui déborde alors le cadre de ce chapitre) de savoir si l'absence d'explicitation des objectifs à atteindre au terme des formations universitaires n'est pas à mettre en relation avec la difficulté que rencontre actuellement l'université de masse à définir ses missions, surtout pour le premier cycle. En effet, identifier les objectifs explicites des différentes formations exigerait de se positionner clairement au sein des débats qui animent actuellement le monde universitaire quant à ses missions et quant à la conception même de l'université. Le titre quelque peu embarrassé d'un ouvrage de Renaut (2002) est représentatif de l'importance et de l'urgence de ces débats : *Que faire des universités ?* Le relatif mutisme sur les connaissances et compétences qu'il s'agirait de faire acquérir aux étudiants est peut-être le reflet des tensions et des hésitations actuelles entre plusieurs modèles de l'Université.

Pour en revenir aux pratiques d'évaluation, si nous ne savons pas ce que les étudiants ont appris dans leur cursus et si l'évaluation se réalise sans trop expliciter les acquis attendus, ce serait parce que nous ne savons pas ce qu'ils auraient dû y apprendre et parce que l'université n'est pas au clair par rapport à ce qu'ils devraient y apprendre. L'explicitation des acquis attendus renvoie donc à la question fondamentale des finalités de la formation universitaire et notamment de son premier cycle : s'agit-il d'une formation propédeutique, par exemple en termes de maîtrise d'une méthodologie et d'élaboration d'un projet personnel ? D'une pré-professionnalisation ? Le premier cycle vise-t-il à développer une culture générale ? Une culture générale de sa discipline et de sa spécialité, comme le suggère Renaut (2002) ? Il nous a d'ailleurs été dit à plusieurs reprises lors des entretiens que *« l'on ne sait pas bien ce que doit acquérir tel ou tel type d'étudiants de premier cycle »* et que cette ignorance est explicable par le caractère très général des objectifs qui y sont poursuivis et par le fait qu'un consensus ne se dégage pas sur cette question.

À l'appui de cette hypothèse, on notera que le déficit d'explicitation des objectifs en termes de connaissances et de compétences à acquérir est davantage le fait du premier cycle que des cycles ultérieurs. On peut en effet comprendre que, si les maquettes des maîtrises et des nouvelles licences pro-

fessionnelles contiennent des informations de ce type, ces dernières font davantage défaut pour les diplômes de premier cycle, dont les objectifs peuvent apparaître comme tellement généraux et génériques, comme la culture générale et la méthodologie de travail, qu'une explicitation détaillée des acquis attendus semble superflue et très délicate à réaliser.

3.2 Une absence de reconnaissance du caractère « professionnel » de l'acte d'évaluation

Malgré les réserves exprimées par les examinateurs externes interrogés par Warren Piper (1994), les enseignants semblent faire preuve, dans la même enquête, d'une grande confiance dans leur jugement global intuitif et dans leur compréhension partagée et tacite des critères d'évaluation. Ils se disent convaincus que leurs jugements se rejoignent et marquent leur étonnement face aux études docimologiques qui identifient, au contraire, des problèmes de fidélité et de validité dans la passation des examens. Deux tiers estiment, dans le même sens, qu'il y a peu de variation dans les niveaux des diplômes décernés par les différentes universités. Poirier (2001) note aussi que, si l'anonymat des copies n'est pas souvent garanti dans les faits, c'est notamment parce que les enseignants refusent de mettre en cause « l'honnêteté de leur évaluation ». Blais *et al.* (1997) rapportent, dans le même sens, que 89 % des enseignants interrogés estiment qu'ils disposent des connaissances et des compétences docimologiques nécessaires à la bonne évaluation de leurs étudiants. Selon Warren Piper (1994), cette confiance aveugle provient du fait que les professeurs d'université s'identifient davantage à leur discipline scientifique qu'à leur rôle d'enseignant ou d'évaluateur. En tant que membres d'une communauté scientifique, ils estiment partager une même idée de ce que représentent les critères de qualité et d'excellence au sein de leur discipline. Puisque les étudiants sont jugés sur ces critères qui constituent le ciment même de leur appartenance à une communauté de chercheurs, l'évaluation ne devrait dès lors pas soulever de problèmes docimologiques importants.

Une citation, extraite d'un entretien rapporté par Dejean (2002, p. 29), témoigne de cette confiance en un « collège invisible » qui assurerait une cohérence aux critères d'évaluation des enseignants-chercheurs d'une même discipline : *« les experts savent quelque part ce qu'est un niveau de licence ou de maîtrise en droit ou en économie »,* ce qui justifie alors l'absence de définition d'objectifs de formation et donc de critères d'évaluation. Dans l'étude de cas évoquée ci-dessus, des enseignants rapportent qu'ils ont abandonné le recours à des grilles de correction parce que *« avec l'expérience, une pesée globale suffit ; avec l'habitude, on sait ce que vaut telle ou telle copie ».* Et ce que l'enseignant expérimenté cherche dans la copie, ce sont justement des signes de l'affiliation de l'étudiant à la communauté de recherche auquel il tente de s'affilier : *« un étudiant qui pose des questions*

sans trop y répondre de manière définitive, on sait que c'est quelqu'un qui a compris ce qu'est l'histoire ».

En conclusion de son étude sur les pratiques d'évaluation dans les universités anglaises, Warren Piper (1994) regrette que l'évaluation ne soit pas considérée comme une des composantes de la « professionnalité » de l'enseignant-chercheur. Les enseignants du supérieur n'y sont pas formés. Ils apprennent « sur le tas », se définissent progressivement des règles pragmatiques et se rassurent en postulant une capacité à identifier le niveau « d'excellence », signe distinctif de l'université. Peu de réglementations explicites encadrent cet aspect de leur travail. Et quand elles existent, elles ne sont pas toujours connues et respectées. Dans l'enquête de Blais *et al.* (1997), 57 % des enseignants interrogés estiment qu'aucune contrainte ne pèse sur leur façon d'évaluer les étudiants : ils se sentent donc libres de les évaluer comme bon leur semble. Mais, la même enquête montre qu'une part importante des enseignants interrogés contreviennent à une série de dispositions du règlement pédagogique de l'Université de Montréal. Ainsi, 62 % des enseignants déclarent qu'il n'existe pas, dans leur unité, de jury d'examen, pourtant expressément prévu dans le règlement pour l'attribution définitive des notes. Le système littéral de communication des résultats (de A à E) ne semble pas adopté par les 60 % des enseignants qui continuent à communiquer aux étudiants les résultats en pourcentage. La conclusion ne laisse pas d'inquiéter : *« Force est de constater que ceux qui ont répondu au questionnaire ou participé aux entrevues ne semblent pas se formaliser outre mesure de certaines dispositions du règlement pédagogique. »* (Blais et al, 1997, p. 116).

De plus, les activités didactiques et, en particulier, le choix de ce que chaque enseignant décide de mesurer lors de ses examens restent considérés comme relevant de la sphère privée (Dejean, 2002). L'enquête de Blais *et al.* (1997) illustre aussi l'aspect individuel de l'activité d'évaluation : les enseignants disent montrer rarement leurs questions d'examen à leurs collègues. Ces derniers interviennent peu dans le choix des moyens d'évaluation et encore moins dans la correction : il est, par exemple, exceptionnel qu'un professeur sollicite l'avis d'un collègue en cas d'indécision quant à la note à attribuer à une copie.

3.3 Une évaluation à fonction sélective et non d'inventaire

En définitive, on pourrait risquer l'hypothèse selon laquelle l'évaluation des acquis est hétérogène et peu explicite quant à ses critères parce qu'elle remplit prioritairement des fonctions de sélection, de tri et de certification de réussite globale, ces fonctions ne nécessitant pas de disposer d'inventaires détaillés des compétences et connaissances acquises. Conçue comme une série d'obstacles permettant d'éliminer progressivement les étudiants les plus faibles, l'évaluation des étudiants, surtout ceux de premier cycle, ne chercherait pas prioritairement à valider des acquis précis. Comme

le rappelle Hutmacher (2001), le diplôme universitaire possède une fonction de signal à l'égard du monde extérieur, lui indiquant que le détenteur de ce diplôme a réussi son passage par l'enseignement supérieur, ce qui garantit un certain niveau de compétences et l'acquisition de la culture du champ social correspondant. On se souviendra d'ailleurs que, dès 1964, Bourdieu et Passeron, dans leurs travaux sur les héritiers, soutenaient que l'absence d'explicitation des critères d'évaluation qui caractérise l'examen universitaire, « *asile privilégié de l'irrationalité* », favorise les étudiants des classes cultivées *« puisqu'ils détiennent, implicitement, le moyen d'y satisfaire »* (Bourdieu et Passeron, 1964, p. 113).

4. UN MANQUE DE VALIDITÉ, DE FIDÉLITÉ ET DE FIABILITÉ

Un quatrième thème de recherche sur les pratiques d'évaluation au supérieur a trait à la qualité docimologique des examens qui y ont cours. Certes, la critique docimologique n'est guère spécifique au supérieur. Cependant, l'importante hétérogénéité des pratiques et l'absence d'explicitation des critères posent, avec plus d'acuité encore qu'aux autres niveaux scolaires, la question de la fidélité, de la validité et de la fiabilité de l'évaluation.

Des études de docimologie, parfois anciennes, ont aussi été réalisées dans l'enseignement universitaire, avec des résultats similaires à ceux enregistrés aux autres paliers de la scolarité (Ashcroft & Palacio, 1996 ; Beattie, 1995 ; Brown & Glasner, 1999 ; Heywood, 2000 ; Warren Piper, 1994).

Ainsi, dès les années 1930, une série d'études ont été consacrées à la **fidélité** des examens universitaires : des correcteurs ont été invités à noter à nouveau des copies qu'ils avaient déjà corrigées. Les notations d'une centaine de professeurs de mathématiques ont été comparées à propos de cinq mêmes copies. Les résultats sont sans appel : dans la détermination de la note, la part du correcteur se révèle finalement plus décisive que celle de la performance réalisée par l'étudiant ! Ces études ont été confirmées par un programme de recherche de la Society for Research into Higher Éducation réalisé dans les années 1960, puis par des études portant plus spécifiquement sur des disciplines particulières, notamment les sciences médicales et les sciences de l'ingénieur.

Une des explications du manque de fidélité réside dans le fait que les critères d'évaluation varient considérablement d'un examinateur à l'autre, parce que ces examinateurs sont guidés par des cadres de référence différents pour juger leurs étudiants (Warren Piper, 1994 ; Webster, Pepper & Jenkins, 2000). Par exemple, certains correcteurs tiennent compte de la performance des autres étudiants, d'autres s'attachent à repérer les signes d'un progrès depuis le début du cours. Le jugement porté sur une dissertation est dépendant de ce que le correcteur valorise dans un tel travail : certains privilégient l'apport de données empiriques, d'autres l'aspect théorique. Pour une même

filière, les jurys de différents établissements n'attribuent pas les mentions selon les mêmes critères et sont ignorants des critères appliqués ailleurs. Par exemple, la mention la plus élevée est accordée par certains examinateurs si le candidat se distingue par son originalité, par d'autres si le candidat fait montre de compétences de communication remarquables. Cette observation est à mettre en parallèle avec le fait que peu de discussion et d'explicitation semble être consacré, au sein des unités de formation, à l'établissement collectif des critères d'évaluation.

Même quand les correcteurs se réfèrent à des critères identiques, ils ne les font pas intervenir selon la même pondération et ils ne leur attribuent pas la même signification. Ainsi, si les enseignants sont assez unanimes pour identifier les critères pertinents pour classer les étudiants (l'analyse critique, par exemple), on observe que des comportements précis très différents sont retenus comme indicateurs de ce critère selon la discipline enseignée et même parfois selon les théories privilégiées au sein de la même discipline (Ecclestone, 2001). Dans le même sens, Webster *et al.* (2000) ont demandé à 24 enseignants de préciser ce qu'ils entendaient quand ils annonçaient à leurs étudiants qu'un de leurs critères d'évaluation résiderait dans la « qualité de l'analyse ». Pour un premier groupe d'enseignants, l'analyse est définie comme la décomposition d'un tout en ses différentes parties, de manière à mieux comprendre les relations qui les unissent. Pour un second groupe, il s'agit de l'application d'outils méthodologiques aux données recueillies. Neuf autres des 24 enseignants interrogés fournissent chacun une définition qui ne se rapproche ni de celle du premier groupe ni de celle du second. La même étude révèle que les critères effectivement pris en compte lors de la correction diffèrent parfois considérablement des critères annoncés. Ainsi, tous les critères annoncés ne sont pas nécessairement appliqués et certains critères effectifs, induits à partir des annotations des copies, n'avaient pas été annoncés. Dans le même sens, Dejean et Magoga (2001) ont observé que les qualités strictement linguistiques des réponses écrites des étudiants (de l'orthographe à la construction discursive) interviennent dans la notation à des degrés très divers selon le correcteur.

Plusieurs études récentes montrent que des problèmes de **fidélité** se rencontrent aussi pour l'obtention du diplôme de **doctorat**. Johnston (1997) a ainsi réalisé une analyse détaillée de rapports de thèse. Sur les 16 thèses analysées et jugées chacune par quatre professeurs, trois d'entre elles font l'objet de désaccord entre les enseignants à propos de la décision d'échec ou de réussite. Pour l'une des thèses analysées, un rapport relève la qualité et l'originalité de la revue de la littérature (« éclectique, organisée et cohérente »), alors que le rapport de l'expert prônant l'échec contient sept pages de critiques sur cette partie de la thèse, qui se terminent en regrettant que le candidat n'ait pas eu recours à la structure classique et académique d'un relevé de la littérature. Dans un autre cas, c'est manifestement une incompatibilité idéologique qui explique le désaccord. Des commentaires identiques n'ont, de plus, pas le même impact chez tous les rapporteurs : des

suggestions de modifications de forme (écriture, références…) peuvent justifier un avis d'acceptation sous condition de modification, tout comme elles peuvent être simplement mentionnées après un avis favorable sans condition. En ce qui concerne le type de critères utilisés pour juger de la qualité de la thèse, les rapporteurs se réfèrent tous à des critères de qualité formelle (écriture et présentation), mais divergent quant aux autres critères (gestion de la complexité, originalité, pensée critique, produit publiable…), au point que peu de thèmes sont communs à l'ensemble des rapports. Enfin, un même critère peut revêtir des sens différents : ainsi, quand les rapporteurs se réfèrent au caractère « publiable » du travail de recherche, cela signifie, pour certains, que la thèse est publiable telle quelle ; pour d'autres, que le travail mérite publication, mais après avoir subi une série de révisions parfois importantes. Dans le même sens, Morley et ses collaborateurs (2002) ont décrit l'importante hétérogénéité des pratiques d'évaluation de la thèse de doctorat selon les institutions, les facultés et les départements : poids relatif de l'oral par rapport à l'écrit, modalités de rapports de thèse et de constitution du jury.

La **validité** des examens universitaires a aussi fait l'objet de questionnements critiques. On s'est ainsi interrogé sur la *validité prédictive* de ces examens : permettent-ils de prédire les performances ultérieures, voire la réussite professionnelle (Goldschmid, 1992) ? On a aussi mis en évidence l'existence de *biais idéologiques* dans la correction de copies (Husbands, 1976). Cette étude a porté sur un département de sciences sociales au sein duquel deux écoles de pensée coexistaient, l'une centrée sur l'action sociale et l'autre sur des approches sociales cliniques. Six enseignants, dont on connaissait le « penchant » épistémologique, ont été invités à noter des copies d'étudiants, la moitié de ces copies manifestant une certaine sympathie pour la première école de pensée et l'autre moitié pour la seconde. L'existence de biais idéologiques n'est pas systématique : certains correcteurs semblent faire fi de leurs options méthodologiques personnelles. Mais d'autres surévaluent assez systématiquement les copies qui valorisent les cadres d'analyse qu'ils privilégient eux-mêmes en tant que chercheurs. L'auteur conclut son étude en montrant comment l'attitude stratégique de certains étudiants qui adaptent leur discours aux positions épistémologiques qu'ils pensent être celles de leur correcteur est finalement assez rationnelle, alors qu'elle est parfois sévèrement punie, notamment par les évaluateurs externes qui y voient une sorte de jeu de séduction inutile et peu « universitaire ».

L'existence de *biais culturels* dans les procédures d'évaluation des étudiants a aussi fait l'objet d'investigation, dans le contexte d'un enseignement supérieur mondialisé. On s'est notamment interrogé sur l'adéquation de certaines formes d'examen aux différents publics d'étudiants, en particulier les étudiants étrangers et ceux qui appartiennent aux minorités ethniques ou culturelles. Ainsi, De Vita (2002) a montré que les différentes formes d'examen n'offrent pas aux étudiants d'origine culturelle différente les mêmes opportunités de manifester leur maîtrise du cours. Par exemple, les étudiants d'origine chinoise inscrits dans les universités anglaises semblent, par certains

côtés, désavantagés par les QCM car ils répugnent à choisir au hasard une réponse en cas de doute. Par d'autres aspects cependant, les examens écrits à réponse longue les désavantagent également, comme tous les étudiants non anglophones d'ailleurs, tant ce type d'examen mesure autant (sinon davantage) les compétences langagières que la stricte maîtrise des connaissances.

Enfin, des interrogations critiques portent aussi sur la **fiabilité** de l'évaluation des acquis, c'est-à-dire sur la capacité qu'a l'université de masse à assurer que les examens se déroulent sans incidents. À l'occasion de plusieurs entretiens, des doutes ont été émis à propos des garanties qu'offre l'université quant à la fiabilité des procédures et des dispositifs d'évaluation, surtout dans les premiers cycles massifiés. Des problèmes particuliers ont été évoqués : les mauvaises conditions de passation des épreuves écrites qui aboutissent parfois à l'impossibilité de s'assurer du caractère individuel de l'examen ; la répétition des mêmes questions d'année en année ; la présence aux examens d'étudiants qui ont disposé à l'avance des questions ; les fraudes en cours d'examen, notamment à l'aide d'outils technologiques de plus en plus pointus.

Comme pour confirmer ces interrogations critiques, on trouve dans la littérature contemporaine à la massification une série d'articles sur la fraude aux examens (Franklyn-Stokes et Newstead, 1995 ; Wankat et Oreovicz, 2002). Les comportements frauduleux prennent des formes très variées dans le supérieur. Certains, parmi les moins graves, touchent un étudiant sur deux dans le premier cycle : permettre à un autre étudiant de copier un devoir ou une préparation ; paraphraser sans citer explicitement ; modifier, voire inventer des données ; surévaluer le travail d'un pair lors de l'évaluation par les pairs ; recopier un travail antérieur ; inventer des références. Les comportements de fraudes durant les examens seraient plus rares, bien qu'ils puissent toucher plus d'un étudiant sur dix : utiliser du matériel non autorisé, passer un examen à la place de quelqu'un d'autre, obtenir à l'avance des informations sur l'examen, copier sur un voisin. En tout cas, la fraude serait plus présente que ne le pensent les enseignants-chercheurs et se développerait corrélativement à l'évolution actuelle des procédures d'évaluation. En effet, pressée d'évaluer, plus souvent suite à la modularisation et à la semestrialisation, un nombre d'étudiants plus important, l'université de masse se tournerait davantage vers des formules de questionnaires à choix multiple et à correction automatique, forme d'examen qui semble la plus propice à la fraude. Par ailleurs, les comportements frauduleux sont moins fréquents quand l'évaluation a la réputation d'être correctement et justement conduite et quand elle vise à mesurer des compétences qui dépassent la restitution, par exemple via des examens à livre ouvert.

En conclusion de son relevé de la littérature sur la docimologie dans le supérieur, Beattie note que « *l'impression qui se dégage des diverses études est que les méthodes d'évaluation généralement utilisées pour s'assurer que des connaissances ont bien été acquises ont grand besoin d'être examinées. Elles n'ont guère de chance, sous leur forme actuelle, d'être considérées comme acceptables par ceux qui militent en faveur d'une*

approche plus rigoureuse de l'évaluation des résultats des études » (Beattie, 1995, p. 316).

5. CONCLUSION : UNE ÉVALUATION QUI « PILOTE » LES ÉTUDIANTS

Ce tableau des pratiques d'évaluation est d'autant plus inquiétant que les étudiants sont véritablement « pilotés » par les exigences de l'évaluation. Ils sont à l'affût de ce qui leur sera demandé aux examens et ont tendance à y adapter leurs manières d'étudier (Biggs, 1999a ; Romainville, 2000). Si l'évaluation apparaît aux enseignants comme la toute fin d'un processus, elle est, par contre, le début de toute chose pour les étudiants. Ce sont les examens qui définissent le curriculum réel.

5.1 L'évaluation par la restitution appelle l'étude en surface

Les méthodes d'évaluation ont parfois tendance, au premier cycle surtout, à privilégier des mesures de restitution des matières enseignées, notamment parce que ces méthodes sont les plus aisées à mettre en œuvre au sein des grands groupes de l'université de masse. On constate dès lors que bon nombre de questions d'examen n'exigent pas beaucoup plus qu'une reproduction de parties de polycopié. Ces pratiques d'évaluation inciteraient les étudiants à privilégier une étude en surface des matières. Plusieurs observations vont dans ce sens. Des questions portant sur la restitution de faits pointus engendrent une étude mécanique, « par cœur », morcelée et superficielle (Montgomery, 1995). Une accumulation d'évaluations sommatives partielles encourage l'étude par la seule mémorisation (Tan, 1992). L'opacité des méthodes d'évaluation est aussi responsable du développement de l'approche en surface : l'étudiant qui ne sait pas très bien ce qui lui sera demandé à l'examen se réfugie dans une reproduction stricte des contenus du cours (Edwards & Knight, 1995).

À l'inverse, les étudiants adoptent plus volontiers une approche en profondeur quand l'évaluation se réalise par une production écrite ouverte plutôt que par des examens à questions fermées, le questionnaire à choix multiples en particulier (Fallows & Chandramohan, 2001). D'ailleurs, les étudiants qui réussissent aux QCM sont ceux qui ont bien compris que cette forme d'évaluation mesure la maîtrise de connaissances ponctuelles (faits, dates, formules…) : ils s'y préparent en conséquence en négligeant d'exercer des opérations intellectuelles plus englobantes, comme la comparaison (Scouller, 1998).

En adaptant leurs stratégies à l'évaluation, les étudiants réalisent en définitive des apprentissages de nature différente. Ainsi, les productions écrites d'étudiants engagés dans une pédagogie active manifestent davantage de traces d'opérations intellectuelles de haut niveau (comparaison, généralisa-

tion, classification...) que les écrits d'étudiants qui ont assisté, pour la même matière, à un cours magistral suivi d'un examen traditionnel (Tynjälä, 1998). Dans le même sens, l'introduction d'évaluations formatives intermédiaires permet aux étudiants de réaliser des apprentissages de qualité supérieure (Greer, 2001). Bref, quand on déplore que les étudiants ne travaillent qu'en vue de la réussite d'un test, c'est surtout la mauvaise qualité de ce test qui est en cause : s'ils étudient superficiellement, c'est qu'ils savent qu'ils seront interrogés superficiellement. Il semble, cependant et heureusement, exister une certaine immunité des meilleurs étudiants aux effets des mauvaises pratiques d'évaluation des acquis : les meilleurs étudiants en médecine, même s'ils sont confrontés à des tests qui n'exigent que de la mémorisation, continuent à étudier dans une perspective principale de compréhension et d'application de leurs connaissances (LindBlom-Ylänne & Lonka, 2001).

Cet impact des procédures d'évaluation sur les apprentissages étudiants est aussi évoqué dans certains rapports du CNE. Ainsi, le rapport transversal portant sur les formations de pharmacie signale que *« dans trop de cas encore, des examens multiples et parcellaires sont imposés par les professeurs, dont chacun vérifie que son cours est appris, sans réflexion d'ensemble sur ce qui doit être appris, et non enseigné (toutes les classes de médicaments), et sur le type d'entraînement qui favoriserait l'acquisition de concepts et de mécanismes intellectuels rigoureux. Trop souvent, les examens confortent une démarche encyclopédiste et morcellent davantage encore les connaissances »* (CNE, 1998a, p. 39). L'évaluation des différentes UFR se réfère également à ce critère : il est régulièrement reproché à telle ou telle UFR de recourir à des formes d'examen peu appropriées à développer les compétences de haut niveau attendues, comme l'esprit de synthèse. De l'avis des étudiants, *« pas de temps, pas de place à la réflexion : il faut redonner le cours mot à mot et de préférence avec les virgules, ce qui fait qu'on ne retient pas »* (CNE, 1998b, p. 62). C'est notamment au nom de leur impact positif sur la qualité des apprentissages des étudiants que le CNE recommande que des UFR accordent plus de place aux oraux et aux questions ouvertes.

5.2 L'évaluation par la restitution est peu compatible avec les plus hautes finalités de l'enseignement universitaire

Ce questionnement critique par rapport aux effets délétères des pratiques d'évaluation sur l'apprentissage des étudiants s'est exacerbé quand l'enseignement universitaire a été prié de participer plus directement au développement économique et social des pays et de « rendre compte » de l'utilisation de la part des finances publiques qui lui était consacrée (« accountability »). Les critères retenus par l'Université pour évaluer les étudiants semblaient fort éloignés de ceux qui ont cours dans les environnements de travail actuels pour évaluer l'efficacité professionnelle. La remise en cause des pratiques d'évaluation est ici significative d'une volonté d'assigner de nou-

veaux objectifs à l'enseignement tertiaire et de créer des synergies plus étroites entre le système éducatif et les politiques économique et sociale de chaque pays (Neave, 1996). Par exemple, Edwards et Knight (1995) déplorent que l'évaluation académique des acquis privilégie les épreuves qui portent sur les connaissances déclaratives alors que le monde du travail souhaite des étudiants qui maîtrisent des connaissances procédurales, mobilisables pour l'action. Dans la même veine, ils critiquent le recours trop exclusif à l'essai personnel, alors que ce serait les compétences de communication orale et de travail de groupe qui seraient prioritairement recherchées par les employeurs. Au total, les compétences les plus estimées par les employeurs, comme la capacité de résoudre des problèmes réels, ne seraient acquises par les étudiants que fortuitement, comme des conséquences « collatérales » d'un curriculum universitaire formel qui ne s'en préoccupe pas directement (Montgomery, 1995).

Dans le même sens, Beattie (1995) estime que les pratiques actuelles d'évaluation restent avant tout conçues pour tester l'acquisition de connaissances factuelles. Les détracteurs de l'Université ont alors beau jeu de prétendre que cette dernière remplit la tête des étudiants de faits arides et sans intérêt qui, même s'ils sont retenus pour les examens, sont ensuite rapidement oubliés. De plus, ce qui est privilégié dans les pratiques d'évaluation reste la maîtrise de savoirs déclaratifs formels et les problèmes issus de situations pratiques ne sont considérés qu'à titre d'exemples. Les étudiants ne sont incités à établir des liens entre ces connaissances formelles apprises, ni avec leur savoir naturel, ni avec la résolution de problèmes réels. Ils développent dès lors des « savoirs morts », c'est-à-dire des connaissances déconnectées, apprises par cœur, peu mobilisables dans l'action. Ramsden (1988) décrit ainsi comment des étudiants qui ont pourtant réussi des examens universitaires de connaissance n'ont pas modifié durablement leur manière d'appréhender les phénomènes de la vie quotidienne en rapport avec leur discipline de base. Or, ce sont surtout des « savoirs vivants », des compétences, des savoirs pour penser et pour agir que l'ensemble du système éducatif, du primaire au supérieur, est actuellement invité à développer chez les jeunes, sous la pression du monde économique et social.

Enfin, comme l'ensemble du processus évaluatif est aux mains de l'enseignant (de la définition des objectifs du cours jusqu'à la détermination des critères), les pratiques actuelles ne semblent pas en mesure de faire acquérir aux étudiants certaines compétences qui sont pourtant au cœur même des finalités de l'enseignement universitaire, comme l'autonomie, l'esprit d'initiative, la responsabilité et l'esprit critique. Il conviendrait dès lors de modifier les pratiques d'évaluation, de manière à maintenir la poursuite d'objectifs de haut niveau, comme les capacités de réflexion et de synthèse, par exemple en proposant des épreuves synthétiques portant sur plusieurs enseignements (Boud, 1990 ; Poirier, 2001 ; Reynolds & Trehan, 2000).