

Évaluations des cours : recherche, modèles et tendances

Rapport préparé par
Pamela Gravestock et Emily Gregor-Greenleaf, l'Université de Toronto
pour le Conseil ontarien de la qualité de l'enseignement supérieur



Conseil ontarien
de la qualité de
l'enseignement supérieur

Un organisme du gouvernement de l'Ontario

Avertissement :

Les opinions exprimées dans ce rapport de recherche sont celles des auteurs et ne reflètent pas nécessairement le point de vue ou les politiques officielles du Conseil ontarien de la qualité de l'enseignement supérieur ou d'autres agences ou organismes qui ont offert leur soutien, financier ou autre, à ce projet.

Se référer au présent document comme suit :

Gravestock, P. et Gregor-Greenleaf, E. (2008) Évaluations des cours : recherche, modèles et tendances. Toronto : Conseil ontarien de la qualité de l'enseignement supérieur.

Publié par :

Le Conseil ontarien de la qualité de l'enseignement supérieur

1, rue Yonge, bureau 2402

Toronto (Ontario) Canada

M5E 1E5

Téléphone : 416 212-3893

Télécopieur : 416 212-3899

Site Web : www.heqco.ca

Courriel : info@heqco.ca

Table des matières

1. Introduction
 - A. Méthodologie
 - B. Limites

2. Contexte
 - A. Évaluation de l'enseignement supérieur
 - B. Vocabulaire lié à l'évaluation de cours par les étudiants
 - C. Perceptions des professeurs, des administrateurs et des étudiants vis-à-vis de l'évaluation des cours
 - i. Perceptions des professeurs
 - ii. Perceptions des administrateurs
 - iii. Perceptions des étudiants
 - D. Caractéristiques communes des évaluations de cours
 - i. Mesures communes de l'efficacité de l'enseignement
 - ii. Collecte et interprétation de la rétroaction qualitative
 - iii. Collecte et interprétation de la rétroaction formative
 - E. Usages communs des données d'évaluation de cours

3. Politiques et pratiques actuelles en Amérique du Nord
 - A. Introduction
 - B. Politiques relatives à l'évaluation des cours
 - i. Fréquence et emplacement des politiques
 - ii. Point central et étendue des politiques
 - C. Conception et approbation des instruments d'évaluation
 - i. Élaboration et approbation des instruments d'évaluation
 - ii. Format et contenu du questionnaire
 - iii. Examen des instruments d'évaluation
 - D. Processus de mise en place
 - i. Méthode de distribution
 - ii. Application des lignes directrices ou des politiques

- E. Analyse des résultats
 - F. Accès aux résultats
 - i. Qui a accès à quoi?
 - ii. Gestion des commentaires écrits
 - iii. Publication des résultats
 - G. Interprétation et utilisation des résultats
 - i. Résultats à des fins sommatives et formatives
 - ii. Information fournie avec les résultats d'évaluation
 - iii. Utilisation des commentaires écrits
 - iv. Titularisation, promotion et mérite
 - v. Prix d'enseignement
 - vi. Autres éléments probants de l'efficacité de l'enseignement
 - H. Lien entre les évaluations de cours et les mesures de reddition de comptes
4. Fiabilité, validité et interprétation des données provenant des évaluations de cours
- A. Introduction à la fiabilité et à la validité
 - B. Étudiants dans le rôle d'évaluateurs
 - C. Validité externe : créer l'instrument
 - i. Définition d'un enseignement efficace
 - ii. Mise au point des instruments d'évaluation
 - D. Validité externe : communication et interprétation des résultats d'évaluation
 - i. Communication des résultats d'évaluation
 - ii. Difficultés relatives à l'interprétation et à l'utilisation des données à des fins sommatives
 - E. Validité interne : influence des variables sur les résultats des évaluations
 - i. Survol des variables étudiées
 - ii. Évaluation de la validité
5. Mise en œuvre de mesures d'évaluation efficaces : recommandations tirées de la recherche
- A. Introduction
 - B. Assurance de la validité des évaluations

- C. Assurance de l'utilité des évaluations
 - i. Pour les étudiants
 - ii. Pour les enseignants
 - iii. Pour les administrateurs
 - iv. Pour les établissements

- 6. Tendances émergentes, lacunes existantes et recommandations pour les recherches futures
 - A. Tendances émergentes
 - i. Outils d'évaluation en ligne
 - ii. Association des données d'évaluation aux mesures de reddition de comptes et aux résultats de l'apprentissage axé sur les compétences
 - iii. Utilisation accrue des évaluations à des fins formatives
 - iv. Mise en contexte des données d'évaluation à des fins d'évaluation sommative de l'enseignement
 - B. Lacunes existantes et suggestions pour pousser la recherche
 - i. Définition du vocabulaire et des attentes en enseignement
 - ii. Compréhension des utilisateurs d'évaluations
 - iii. Formation des utilisateurs d'évaluations
 - iv. Évaluation des auxiliaires à l'enseignement et des enseignants chargés des étudiants de deuxième et troisième cycles

- 7. Conclusions

- 8. Travaux cités (en anglais seulement)

Note : Une annexe est disponible en anglais seulement dans un article séparé. Pour obtenir une copie de l'annexe, veuillez en faire la demande à l'adresse info@heqco.ca.

Auteures

Pamela Gravestock est codirectrice du Bureau de promotion de l'enseignement (Office of Teaching Advancement) de l'Université de Toronto, où elle travaille avec le corps professoral à différentes questions et initiatives d'enseignement et de formation. Elle est également candidate au doctorat du groupe sur l'éducation supérieure (études théoriques et politiques) de l'Institut d'études pédagogiques de l'Ontario de l'Université de Toronto. Son mémoire consiste en l'examen des pratiques et des politiques en place en matière d'évaluation de l'enseignement aux fins de titularisation et de promotion. Avec la participation d'Emily Gregor-Greenleaf, elle a rédigé deux documents à l'intention du corps professoral : *Developing Learning Outcomes: A Guide for Faculty* et *Gathering Formative Feedback with Mid-Course Evaluations*.

Emily Gregor-Greenleaf est candidate au doctorat du groupe sur l'éducation supérieure (études théoriques et politiques) de l'Institut d'études pédagogiques de l'Ontario de l'Université de Toronto. Dans son mémoire, elle présente l'historique et l'analyse du programme de premier cycle de l'Université de Toronto. Elle est également adjointe à la recherche du Bureau de promotion de l'enseignement (Office of Teaching Advancement) et du programme de formation des aides-enseignants à l'Université de Toronto. Dans le cadre de cette dernière fonction, elle élabore et réalise actuellement une vaste étude visant à évaluer les programmes de formation des assistants à l'enseignement de l'Université de Toronto.

Remerciements

Nous désirons remercier Andrew Boggs du Conseil ontarien de la qualité de l'enseignement supérieur pour son soutien et ses commentaires tout au long de ce projet. Nous remercions également le Centre d'étude sur les étudiants de niveau postsecondaire (Centre for the Study of Students in Postsecondary Education) de l'Université de Toronto. Pour terminer, nous aimerions exprimer notre reconnaissance particulière au Conseil ontarien de la qualité de l'enseignement supérieur, qui nous a permis de réaliser cette étude à grande échelle.

Section 1

Introduction

Le présent document constitue le premier examen et résumé des travaux de recherche existants sur les évaluations de cours par les étudiants selon une perspective canadienne. Les connaissances sur le sujet sont vastes et la qualité et l'étendue des recherches sont variables. Notre analyse tente de relever les principales questions et conclusions touchant la validité et l'utilité des évaluations de cours par les étudiants et d'en faire la synthèse. Notre recherche comporte sept sections :

Section 1 : Introduction – fournit un survol de l'étendue, de la méthodologie et des limites de l'étude.

Section 2 : Contexte – détermine l'état actuel des connaissances et de l'intérêt porté à l'égard de l'évaluation des cours et de l'enseignement de façon plus générale. On y trouve également une analyse des perceptions des étudiants, du corps professoral et des administrateurs vis-à-vis des systèmes d'évaluation de cours.

Section 3 : Politiques et pratiques actuelles en Amérique du Nord – présente un survol des instruments, des politiques et des processus d'évaluation en place dans 22 établissements postsecondaires du Canada et des États-Unis ainsi que des politiques se rattachant aux évaluations de cours au niveau du système et des organismes gouvernementaux.

Section 4 : Fiabilité, validité et interprétation des données des évaluations de cours – résume et analyse les résultats des études antérieures menées au cours des 40 dernières années, et plus particulièrement au cours des deux dernières décennies.

Section 5 : Mise en œuvre de mesures d'évaluation efficaces : recommandations tirées de la recherche – fait une synthèse des résultats de recherche et formule des recommandations pour améliorer l'administration et l'interprétation des évaluations de cours.

Section 6 : Tendances émergentes, lacunes existantes et suggestions pour les recherches futures – souligne les questions actuellement à l'étude dans les travaux de recherche ainsi que celles recensées comme devant être analysées plus en profondeur.

Section 7 : Conclusions – présente un bref résumé des conclusions et des recommandations les plus importantes.

Dans l'ensemble, nos résultats indiquent que, même si les instruments d'évaluation de cours fournissent généralement des données fiables et valides, il persiste néanmoins

d'importants obstacles nuisant à l'utilisation efficace de ces systèmes d'évaluation, notamment :

- des mythes et des idées fausses persistant quant aux variables qui influent sur les résultats des évaluations;
- des conceptions et des définitions vagues de ce qu'est l'enseignement efficace;
- une formation insuffisante sur les objectifs, l'utilisation et la validité des évaluations de cours chez les étudiants, le corps professoral et les administrateurs;
- une mauvaise présentation et une piètre mise en contexte des données d'évaluation;
- des politiques et des pratiques incohérentes et inéquitables en ce qui concerne la mise en place et l'administration des évaluations de cours.

Nos conclusions montrent que, peu importe la fiabilité et la validité des instruments d'évaluation mêmes, les politiques, les méthodes et les pratiques utilisées par un établissement déterminent le degré d'efficacité des évaluations comme outil de mesure de la qualité de l'enseignement.

1.A Méthodologie

Recherche documentaire

La masse d'information contenue dans le sondage présente les résultats d'un examen approfondi des travaux de recherches publiés sur les évaluations de cours et l'évaluation de l'enseignement. Notre recherche a été menée sur de nombreuses bases de données universitaires et références bibliographiques relevées dans les articles trouvés. Bien que nous ayons passé en revue la littérature remontant jusqu'aux années 1970 (période témoin de l'essor de la recherche sur les évaluations de cours), nous nous sommes principalement concentrées sur les travaux de recherche publiés au cours des 20 dernières années, car un bon nombre d'études antérieures étaient répétitives ou présentaient des résultats posant problème. De plus, les études les plus récentes comprenaient souvent un résumé des connaissances antérieures.

L'organisation de l'étude est le résultat d'un processus itératif reflétant l'évolution de notre compréhension de la matière. Nous avons tenté d'inclure tous les thèmes importants que nous avons relevés dans les travaux de recherche.

Sondage auprès d'établissements postsecondaires

La seconde partie de notre étude consistait en un sondage sur l'information accessible au public relativement aux politiques et aux pratiques en matière d'évaluation de cours auprès d'établissements et de systèmes d'enseignement postsecondaire de l'Amérique du Nord. Les établissements qui ont été sélectionnés pour ce sondage, et la justification de cette sélection, sont indiqués dans l'introduction de la *Section 3 : Politiques et pratiques actuelles en Amérique du Nord*. Nous avons tiré l'information des sites Web des établissements et des instances dirigeantes et organisationnelles, en recherchant les termes « évaluations des cours », « évaluations des enseignants », « évaluation de l'enseignement », « rétroaction des étudiants » et autres termes. Bien que ces

établissements aient été choisis parce qu'ils représentaient un groupe d'établissements de types et de mandats différents et provenant de différents territoires, nous ne pouvons prétendre être en mesure de tirer des conclusions générales sur les politiques et les pratiques en matière d'évaluation de cours des établissements qui ont été sondés ici; l'examen de nos résultats souligne plutôt les politiques et les procédures communes ou particulièrement uniques révélées au cours de ce sondage.

1.B Limites

Aucun examen de la littérature sur le sujet ne peut être exhaustif étant donné le grand nombre de travaux de recherche qui existent actuellement (et qui continuent de croître). Au moment même où nous menions notre étude, de nouvelles publications apparaissaient, amenant de nouvelles questions, soulevant à nouveau le débat sur d'anciennes questions, présentant des approches et des conclusions différentes et présentant de nouveaux résultats. Nous n'avons épargné aucun effort pour repérer le plus de sources possibles, couvrant ainsi tout l'éventail des questions pertinentes. Cependant, dans quelques cas, nous avons effectué un examen des sources figurant dans les études ou analyses de la littérature ultérieure sans toutefois en faire mention si nous estimions que ces mêmes résultats avaient été présentés avec exactitude dans les publications ultérieures. L'étendue de l'étude ne nous a pas permis d'analyser ni de réévaluer de façon approfondie les résultats des travaux de recherche antérieurs ni de mener notre propre recherche sur les questions examinées dans le présent document. Comme l'examen le montre, bon nombre de questions importantes ont déjà été suffisamment traitées, et ce, en profondeur et de façon adéquate, par les chercheurs. Toutefois, il subsiste certains points qui demandent de plus amples recherches. Ceux-ci sont mentionnés tout au long du présent document et plus particulièrement dans la *Section 6.B : Lacunes existantes et suggestions pour pousser la recherche*.

Une autre limite de cet examen est le manque de données provenant du Canada. La majeure partie de la recherche sur les évaluations de cours a été réalisée par des chercheurs américains dans des établissements situés aux États-Unis. Bien que les secteurs de l'enseignement supérieur du Canada et des États-Unis présentent des similitudes évidentes, il existe également des différences importantes en ce qui a trait à la structure, à l'organisation et aux mesures de responsabilisation, sans compter les variations d'ordre culturel et démographique. De plus, les politiques et les pratiques des établissements (plus particulièrement en matière de titularisation et de promotion) varient entre ces deux pays et à l'intérieur même de ceux-ci. À ce titre, nous sommes conscientes des limites probables quant à la généralisation des résultats de recherche entre les secteurs. Bien que nous ayons tenté d'ajouter quelques données canadiennes lors du dépouillement des établissements, notre sondage, comme nous l'avons mentionné précédemment, ne fournit pas un examen approfondi des politiques et des pratiques des établissements de l'un ou l'autre des territoires; nous fournissons plutôt un échantillon afin de démontrer une gamme d'activités en cours d'usage.

L'examen que nous avons fait de plusieurs centaines de publications traitant des évaluations de cours a révélé un nombre surprenant de divergences d'opinions entre les

chercheurs. D'une part, un consensus général, corroboré par de solides recherches, a été atteint sur certaines questions touchant à la fiabilité, à la validité et à l'utilité des évaluations. Malgré tout, des études subséquentes ramènent souvent le débat sur des questions considérées comme étant réglées depuis longtemps, ne faisant quelquefois que les embrouiller inutilement. Ainsi, même si ces questions peuvent sembler être réglées à un certain moment, leur remise en discussion soulève souvent de nouvelles questions ou recadre d'anciennes questions dans un nouveau contexte. D'autre part, certaines questions sont demeurées l'objet de constantes discussions et, vraisemblablement, ont peu de chance d'être résolues. Un bon nombre de ces discussions sont exposées en détail dans la *Section 4 : Fiabilité, validité et interprétation des données provenant des évaluations de cours*.

Comme il fallait sans doute s'y attendre, nous avons également observé que les chercheurs ayant des vues opposées sur une question donnée relèvent et choisissent des études en particulier pour les mettre en relief ou pour y référer. Même si cette situation était prévisible, nous avons été consternées et préoccupées par le manque apparent d'objectivité qui entoure ce type de « sélection ». À maintes reprises, les auteurs ne fournissent pas de précisions quant à la méthodologie, à la taille ou à l'étendue d'une étude et n'examinent pas la généralisabilité des résultats, ce qui pose problème. Par exemple, de nombreux auteurs continuent de citer des études qui ont depuis longtemps été réfutées, discréditées ou jugées non valables en regard de la méthodologie utilisée, et ce, par la majorité des chercheurs. C'est le cas de la pseudoétude « Dr Fox » par Naftulin, Ware et Donnelly (1973) qui est maintenant largement considérée comme non valide (Abrami, 2001; Ali et Sell, 1998). Certains chercheurs ont noté ce fait en y faisant référence. Toutefois, d'autres continuent de citer cette étude comme preuve selon laquelle l'enthousiasme et l'expressivité d'un enseignant engendrent possiblement des résultats plus élevés aux évaluations (Wright, 2006). De la même façon, Wright (2006) cite l'étude de Williams et Ceci (1997), considérée également par la majorité comme suivant une méthodologie non valable parce qu'elle tire des conclusions à partir de résultats obtenus dans une petite classe et au cours d'une seule session; ces résultats ne peuvent donc être généralisés.

Un défi à la généralisabilité des résultats de recherches a trait à la diversité des instruments, des politiques et des processus d'évaluation de cours et à la pluralité des contextes relatifs aux établissements et à la pédagogie, lesquels varient grandement entre les établissements et parfois à l'intérieur même de ceux-ci. Comme l'aborde la *Section 4.C.ii : Mise au point des instruments d'évaluation*, la formulation, l'ordre de présentation ou la combinaison des questions, ou même l'échelle d'appréciation utilisée dans les questionnaires, peuvent avoir une incidence importante sur les résultats obtenus; par conséquent, les études réalisées à l'aide d'un instrument de sondage donné peuvent mener à différentes conclusions comparativement à la même étude utilisant un autre instrument. Parallèlement, l'enseignement est un art complexe à multiples facettes, influencé par une telle pluralité de participants et d'influences externes que l'isolement d'une variable parmi les autres est un défi de taille dans n'importe quelle étude.

Section 2

Contexte

2.A Évaluation de l'enseignement supérieur

Moore et Kuol (2005) ont fait valoir qu'étant donné qu'il est un fait presque universel que l'activité de recherche rapporte davantage sur le plan individuel que sur le plan de l'enseignement, les efforts déployés pour mesurer les dimensions de l'enseignement quant au rendement du corps professoral et pour prêter une attention à ces mesures, dans le contexte du développement professionnel individuel, aident à considérer plus également l'enseignement et la recherche dans le rôle universitaire (p. 143). La mesurabilité et la comparabilité de la plupart des évaluations de cours font en sorte de rendre plus objectif et gérable l'art imprécis qu'est l'évaluation de l'enseignement.

Au Canada et aux États-Unis, les méthodes courantes d'évaluation de l'enseignement comprennent habituellement les évaluations de cours, les lettres d'appréciation des étudiants et des collègues, les évaluations en classe et par les pairs, les prix d'enseignement reçus, le matériel didactique et les notes de cours de même que les éléments prouvant l'utilisation de stratégies et de pratiques novatrices. Chacun de ces éléments de mesure présente des restrictions et des limites qui lui sont propres. C'est la raison pour laquelle un bon nombre d'établissements s'appuient sur plus d'une forme de mesure probante pour développer une compréhension complète des contributions pédagogiques d'un candidat. Néanmoins, les évaluations de cours, ou évaluations des étudiants, sont un des outils les plus couramment utilisés pour évaluer l'enseignement en classe (Wright, 2006; Seldin, 1999; Centra, 1979) et certains croient que ces évaluations pèsent pour la plupart lourdement (ou pèsent trop) dans les décisions touchant la gestion du personnel (Franklin, 2001). Les évaluations réalisées par les étudiants constituent également l'une des mesures les plus controversées et soulevant le plus de discussions. Elles demeurent néanmoins grandement utilisées. Beaucoup ont fait valoir qu'aucune autre option ne permet de fournir le même type de données quantifiables et comparables (Abrami, 2001).

Dans une large mesure, les évaluations de cours sont utilisées à des fins sommatives plutôt que formatives (voir *Section 2.D.iii : Collecte et interprétation de la rétroaction formative*), c'est-à-dire qu'elles servent d'outil pour prendre des décisions relatives à la gestion du personnel (comme l'embauche, la titularisation, la promotion et les évaluations annuelles) en fonction, en partie, de l'évaluation d'un étudiant concernant l'efficacité de l'enseignement d'un enseignant. Les données recueillies, plus particulièrement les réponses qualitatives, sont également utilisées par les enseignants et le personnel de soutien à l'enseignement pour fournir une rétroaction formative afin d'aider à l'amélioration de l'enseignement et à l'élaboration des cours. Wright (2008) nous met en garde contre l'utilisation d'instruments qui n'ont pas été précisément conçus pour fournir une évaluation formative à cette fin et suggère la conception d'instruments distincts destinés aux fins d'évaluation sommative et formative respectivement.

Il y a beaucoup d'écrits sur les problèmes entourant les évaluations de cours. Les chercheurs en éducation ont examiné les difficultés associées aux biais, soulevé les questions ayant trait à leur fiabilité statistique et mis en doute leur capacité à mesurer avec précision l'efficacité de l'enseignement du corps professoral. De plus, certains sont d'avis que la rétroaction que fournissent les évaluations de cours n'est pas efficace pour engendrer des changements dans les comportements des professeurs. Par contre, une majorité appréciable de chercheurs considèrent que les évaluations des étudiants sont utiles pour mesurer les comportements d'enseignement et contribuent à l'efficacité de l'enseignement (notamment Beran, Violato et Kline, 2007; Abrami, 2001; Schmelkin, Spencer et Gellman, 1997; Marsh, 1987).

2.B Vocabulaire lié à l'évaluation de cours par les étudiants

Il existe presque autant de termes qui décrivent les évaluations de cours par les étudiants que d'articles publiés sur le sujet; la terminologie la plus connue est « évaluations des étudiants », « évaluation de cours », « évaluation de l'instruction par les étudiants », « évaluation de l'enseignement par les étudiants », etc. Chaque tournure comporte une connotation quelque peu différente, selon que l'on veut mettre l'accent sur l'étudiant, le cours, l'évaluation ou autres éléments. Wright (2008) a avancé que le terme le plus approprié pour nommer l'évaluation sommative réalisée en fin de cours et que l'on utilise principalement pour prendre des décisions relatives à la gestion du personnel (et non aux fins du perfectionnement de l'enseignement) serait « évaluation de l'instruction par les étudiants », puisque ce terme illustre le mieux à quelle fin sert l'instrument d'évaluation. Pour en connaître davantage sur cette terminologie, veuillez consulter la *Section 6.B.i : Définition du vocabulaire et des attentes en enseignement*. Tout au long du présent document, nous avons utilisé différentes tournures interchangeables, mais nous avons choisi d'utiliser principalement « évaluations de cours par les étudiants », car il s'agit du terme le plus employé par les établissements d'enseignement canadiens.

2.C Perceptions des professeurs, des administrateurs et des étudiants vis-à-vis de l'évaluation des cours

« Il existe probablement plus d'idées fausses liées aux évaluations réalisées par les étudiants que de faits connus à leur sujet, même si les faits connus sont nombreux » (p. 3).

Ory, J.C. (2001). Faculty thoughts and concerns about student ratings. Dans K.G. Lewis (Ed.), *Techniques and strategies for interpreting student evaluations* [numéro spécial]. *New Directions for Teaching and Learning*, 87, 3-15.

Il existe d'innombrables mythes et perceptions erronées sur les évaluations de cours qui affectent inévitablement la perception des professeurs, des administrateurs d'universités et des étudiants vis-à-vis ce sujet. Malgré l'existence de solides recherches qui infirment ces suppositions, de telles croyances demeurent et continuent de se répandre. Il s'agit simplement de soulever la question au cours d'une réunion de département, d'un dîner entre professeurs ou d'un événement sur le campus pour déclencher un éventail d'« anecdotes probantes » de la part de différents membres du milieu universitaire. Comme Nasser et Fresko (2002) l'indiquent, quelques études approfondies ont été menées sur les attitudes et les perceptions liées aux systèmes d'évaluation de cours par les utilisateurs et les destinataires de ces évaluations (plus particulièrement les professeurs, les étudiants et les administrateurs). Les études qui se sont penchées sur cette question ont généralement été réalisées à petite échelle, recueillant les réponses d'un nombre limité d'individus; cependant, il existe une certaine uniformité dans les attitudes répandues, ce qui permet donc une certaine généralisation.

2.C.i Perceptions des professeurs

Il a été démontré que les évaluations de cours par les étudiants sont une source d'anxiété pour le corps professoral (Hodges et Stanton, 2007; Ryan, Anderson et Birchler, 1980), et certains manifestent carrément une opposition à celles-ci (Franklin et Theall, 1989). De telles attitudes sont issues de croyances ancrées selon lesquelles : les évaluations sont biaisées (Eiszler, 2002; Feldman, 1976); les étudiants ne sont pas des évaluateurs compétents (Nasser et Fresko, 2002; Ryan, Anderson et Birchler, 1980); et les évaluations sont marquées par l'influence des notes escomptées par les étudiants (Baldwin et Blattner, 2003). Ces difficultés ainsi que d'autres ont amené à la fois les professeurs et les administrateurs à remettre en question la validité globale des évaluations réalisées par les étudiants et l'utilisation de celles-ci, de même que la possibilité d'un usage indu des données (Beran, Violato et Kline, 2007; Ory, 2001), plus particulièrement en ce qui a trait aux décisions relatives à la gestion du personnel (Nasser et Fresko, 2002; Sproule, 2000; Ryan, Anderson et Birchler, 1980). Cependant, comme nous l'étudierons plus à fond à la *Section 4 : Fiabilité, validité et interprétation des données provenant des évaluations de cours*, les recherches ont montré à maintes reprises la fausseté de ces propos. Comme l'ont noté Theall et Franklin (2000), « l'inconfort qu'éprouve le corps professoral vis-à-vis des évaluations et les lacunes en matière de bonnes pratiques sont des signes d'une déconnexion persistante entre deux mondes : la recherche et la pratique » (p. 95). Ces perceptions négatives à l'égard des évaluations peuvent amener le corps professoral à minimiser leur importance et à entraver les efforts pour améliorer les cours et l'enseignement. Et comme Aleamoni (1999) et Ory (2001) ont fait valoir, tant le corps professoral que les administrateurs ont continué de générer et d'entretenir les mythes et les fausses perceptions entourant les évaluations de cours.

Les anecdotes probantes, combinées aux différentes études empiriques, démontrent clairement qu'un bon nombre de membres du corps professoral continue de s'opposer à l'utilisation des évaluations réalisées par les étudiants et à mettre en doute de tels systèmes d'évaluation. Certains ont fait valoir qu'il existe un plus grand pourcentage de professeurs montrant une attitude négative plutôt que positive ou neutre à l'égard des outils d'évaluation (Nasser et Fresko, 2002; Abrami, 2001; Theall et Franklin, 2001;

Wachtel, 1998; Centra, 1993). Les études ont également démontré que l'utilisation de systèmes d'évaluation de cours est susceptible de miner le moral du personnel enseignant (Ryan, Anderson et Birchler, 1980). Les résultats sont cependant variables et les opinions diffèrent grandement au sein du corps professoral (Wachtel, 1998; Schmelkin, Spencer et Gellman, 1997; Newport, 1996). Une étude réalisée en 2005 auprès de 357 membres du corps professoral (Beran et coll.) a révélé que la majorité des enseignants sondés ont affiché une opinion généralement positive au sujet des évaluations de cours : 63 % ont indiqué qu'ils ne les considéraient pas comme étant intrusives, 70 % ne considéraient pas qu'elles constituaient une perte de temps et 82 % ne considéraient pas qu'elles constituaient des moyens d'évaluation inappropriés. Par ailleurs, Beran et coll. (2007, 2005) ont constaté que plus de la moitié du corps professoral sondé était d'avis que les données d'évaluation sont utilisées de façon appropriée par les administrateurs de l'université.

2.C.ii Perceptions des administrateurs

La majorité des études ont démontré que les administrateurs affichent, en général, une attitude positive vis-à-vis des données d'évaluation et sont d'avis qu'elles sont une source d'information utile pour la prise de décisions relatives à la gestion du personnel (Campbell et Bozeman, 2008; Beran et coll., 2005). Sur le plan administratif, il s'agissait de l'usage le plus courant des données d'évaluation. Par exemple, Beran et coll. (2005) ont noté que 82 % des administrateurs sondés lors de l'étude utilisent les cotes des évaluations à des fins sommatives, plus particulièrement dans le but de servir aux décisions relatives à la promotion, à la titularisation et au mérite.

Bien que les administrateurs puissent convenir que ces outils sont efficaces, leurs attitudes sont également soumises aux idées fausses demeurant omniprésentes sur la validité des évaluations. Theall et Franklin (2001), Abrami (2001) et d'autres ont indiqué que ces méconnaissances perdurent en raison d'un manque de familiarisation générale avec la recherche sur la validité des évaluations ou une réticence à accepter les résultats. De plus, la littérature a également révélé que les administrateurs sont influencés par les approches et les attitudes qu'ils ont par rapport à l'enseignement et à l'enseignant. McKeachie (1997) indique que les stéréotypes propres à un administrateur sur l'enseignement influencent son jugement sur l'efficacité de l'enseignement : si un candidat ne correspond pas au stéréotype que l'administrateur entretient sur ce qu'est un bon enseignant, il sera vu défavorablement.

À la *Section 4.D.ii : Difficultés relatives à l'interprétation et à l'utilisation des données à des fins sommatives*, nous discutons de la tendance de certains administrateurs à surestimer la précision des résultats d'évaluations, plus particulièrement lorsqu'ils comparent les résultats des cours et des enseignants.

2.C.iii Perceptions des étudiants

Le nombre de recherches sur les perceptions des étudiants vis-à-vis des évaluations de cours et de l'utilisation des données qui en résultent est restreint. La majorité des études ont été réalisées à petite échelle et les échantillons ont été établis à partir d'un seul établissement (Campbell et Bozeman, 2008; Beran et coll., 2005). Selon ces études, il apparaît que les étudiants voient le processus de collecte des rétroactions des étudiants comme étant valide et utile. De plus, ils croient également que les étudiants peuvent être

et sont des évaluateurs efficaces de l'enseignement. Cependant, certaines études montrent que les étudiants ne savent pas toujours comment ces données d'évaluation sont utilisées par les établissements (Campbell et Bozeman, 2008; Beran et coll., 2005; Wachtel, 1998), ni comment elles influent sur les décisions relatives à la gestion du personnel. Certaines études révèlent que les étudiants doutent que leurs évaluations soient utilisées et examinées (Wachtel, 1998). En outre, bon nombre d'étudiants font peu usage des résultats des évaluations : dans une étude menée auprès de 1 194 étudiants et de 35 anciens étudiants provenant d'un établissement d'enseignement canadien, Beran et coll. (2005) ont constaté que 56 % des étudiants n'utilisent d'aucune façon les données des évaluations; que 43 % des étudiants ont indiqué qu'ils les avaient consultées et de ce nombre moins du tiers (31 %) les utilisent pour choisir leurs cours en fonction du contenu et de la structure (p. ex., travaux personnels, charge de travail, matière) et près des deux tiers les utilisent pour choisir leurs cours en fonction de l'enseignant (64 %).

2.D Caractéristiques communes des évaluations de cours

Au cours de leurs recherches sur les évaluations de cours, plusieurs chercheurs ont cerné les caractéristiques communes des outils d'évaluation de cours. Par exemple, Algozzine et coll. (2004) décrivent une évaluation type d'après leur recherche sur l'élaboration et l'utilisation des instruments d'évaluation de cours :

La méthode traditionnelle et historique utilisée pour évaluer l'enseignement dans les classes des universités est de demander aux étudiants de fournir leurs commentaires au sujet de l'« efficacité » des cours au moyen d'une échelle d'appréciation de type « à la carte »... Les systèmes traditionnels d'évaluation de cours de ce type possèdent des caractéristiques similaires : a) élaboration d'un instrument comprenant une série de questions ouvertes et fermées portant sur le contenu du cours et l'efficacité de l'enseignement; b) au moins une question aborde l'efficacité « générale »; c) des commentaires écrits sur le contenu du cours et l'efficacité de l'enseignant sont sollicités; d) l'anonymat des réponses est assuré et respecté; e) les réponses sont recueillies à la fin de la session en l'absence de l'enseignant; f) les réponses et les appréciations fournies sont synthétisées avec l'ensemble des enseignants, des départements et des collègues et utilisées comme preuves de l'« efficacité de l'enseignement » servant à la prise de différentes décisions relatives au développement professionnel; g) les différences qui peuvent exister entre les étudiants (ex. : moyenne pondérée cumulative, année universitaire), les cours (obligatoires, de cycle supérieur) et les enseignants (débutant, expérimenté) sont grandement laissées de côté dans l'analyse et la production des rapports des résultats exprimant l'efficacité (p. 135).

Les différentes questions posées dans les formulaires d'évaluation de cours ciblent les aspects distincts et séparables des comportements d'enseignement et du cours d'un enseignant. En général, les étudiants évaluent chaque aspect séparément, classant certains d'entre eux plus positivement que d'autres (Beran, Violato et Kline, 2007).

Comme nous le verrons à la *Section 3 : Politiques et pratiques actuelles en Amérique du Nord*, il existe des variations dans les formats et les pratiques utilisés dans les différents établissements. Cependant, certains éléments sont pratiquement universels. Les formulaires d'évaluation de cours sont le plus souvent distribués à la fin d'un cycle de formation particulier. Ils sont presque toujours anonymes (ou, plus rarement, confidentiels) et comportent très fréquemment des réponses d'ordre qualitatif et quantitatif. Pour les questions quantitatives, les étudiants doivent répondre selon l'échelle de Likert et d'autres types d'échelle, comportant le plus souvent cinq ou sept points. Certains établissements offrent un formulaire uniformisé à utiliser pour tous les cours, tandis que, dans d'autres établissements, des formulaires sont créés par les divisions et, moins fréquemment, par les départements. Dans d'autres cas, des questions types sont obligatoires pour tout le corps professoral (d'une division ou de l'ensemble de l'établissement); dans d'autres, les membres du corps professoral ou les départements peuvent ajouter des questions plus particulières à leurs programmes ou activités d'enseignement. En général, le corps professoral est retiré du processus de collecte des données d'évaluation et ne peut généralement pas avoir accès aux évaluations tant que les notes finales de tous les étudiants n'ont pas été soumises. Bien que les évaluations de cours se déroulent généralement en classe au moyen d'un formulaire papier (lu par lecteur optique), un certain nombre d'établissements se sont récemment dirigés vers l'utilisation d'outils en ligne (voir la *Section 6.A.i : Outils d'évaluation en ligne* pour un examen approfondi de cette nouvelle tendance).

Les évaluations permettent généralement de recueillir une rétroaction précise sur les mesures de l'efficacité de l'enseignement et sur certains aspects particuliers d'un cours, et posent également des questions globales avec échelle d'appréciation et souvent quelques questions ouvertes qui permettent d'obtenir des réponses écrites d'ordre qualitatif. La majorité des instruments d'évaluation sont destinés à être employés pour l'évaluation sommative de l'enseignement. Cependant, il est possible d'effectuer une évaluation formative au moyen d'autres modèles d'instruments d'évaluation d'étudiants et d'une interprétation diagnostique des résultats provenant des types d'évaluations abordés précédemment. (Voir la *Section 4.D.i : Communication des résultats d'évaluation* pour un examen approfondi des différents moyens dont les établissements peuvent se servir pour présenter efficacement les données d'évaluation à des fins sommatives et formatives.)

2.D.i Mesures communes de l'efficacité de l'enseignement

Le questionnaire des évaluations de cours permet de mettre au jour de l'information sur la conception et la présentation du cours et sur le comportement de l'enseignant. Cashin (1995) rapporte six volets qui reviennent généralement dans les évaluations : 1) des questions sur le contenu du cours; 2) des questions sur les aptitudes à communiquer de l'enseignant; 3) des questions sur l'interaction étudiant-enseignant; 4) des questions sur la difficulté du cours et la charge de travail; 5) des questions sur les pratiques d'évaluation utilisées dans le cours; et 6) des questions d'autoévaluation de l'étudiant.

Les différentes composantes des évaluations de cours proviennent également de la recherche sur l'apprentissage et sur les comportements d'enseignement efficaces. La perception quant au besoin de poser des questions sur différentes catégories de comportements d'enseignement découle de la croyance selon laquelle l'efficacité de l'enseignement est d'ordre multidimensionnel; c'est-à-dire que les enseignants peuvent exceller dans certains volets de l'enseignement et pas dans d'autres (Algozzine et coll., 2004; Marsh et Roche, 1997; Marsh 1987). En 1987, Marsh a créé un instrument d'évaluation par l'étudiant de la qualité de l'instruction (*Student Evaluations of Education Quality, ou SEEQ*), lequel comprend neuf catégories de questions sur les comportements d'enseignement qu'il considère nécessaire d'inclure afin de veiller à ce que l'efficacité de l'enseignement soit bien représentée dans l'évaluation : 1) l'apprentissage/la valeur; 2) l'intérêt de l'enseignant; 3) l'organisation; 4) le rapport individuel; 5) l'interaction avec le groupe; 6) l'étendue de la matière; 7) les examens/la notation; 8) les travaux personnels/les lectures; 9) la charge de travail/le degré de difficulté. Des mesures similaires d'évaluation de l'efficacité de l'enseignement ont été déterminées par Braskamp et Ory (1994), Centra (1993) de même que par R. Cashin avec le système d'évaluation IDEA (Individual Development and Educational Assessment) qu'il a élaboré à la Kansas State University. Ces mesures abordent l'organisation et la planification du cours; la clarté de la communication et les aptitudes à communiquer; le rapport et l'interaction enseignant-étudiant; le degré de difficulté et la charge de travail du cours; la notation et les examens, de même que l'autoévaluation par l'étudiant de son apprentissage. D'autres études, comme celles de Feldman (1989), ont déterminé jusqu'à 28 catégories de comportements d'enseignement. Les défis que posent la détermination et la définition des stratégies et des activités d'enseignement contribuant le plus à l'apprentissage de l'étudiant sont examinés à la *Section 4.C.i : Définition d'un enseignement efficace*.

Dans le contexte canadien, Harry Murray (1987) de l'Université Western Ontario a dressé une liste des comportements d'enseignement pouvant servir à recueillir des renseignements auprès des étudiants sur 60 comportements et activités d'enseignement. On mesure par ces comportements les activités d'enseignement censées améliorer l'apprentissage de l'étudiant. Ces comportements sont groupés en neuf catégories : 1) la clarté; 2) l'expression; 3) l'interaction; 4) l'organisation; 5) le rythme; 6) la prestation; 7) le discours; 8) le rapport; et 9) les aides didactiques.

Des exemples de plusieurs de ces instruments sont fournis à l'Annexe B : Échantillons d'instruments d'évaluation utilisés dans les établissements. Des versions adaptées de ces instruments en usage dans bon nombre d'établissements sont abordées à la *Section 4.C.ii : Élaboration d'instruments d'évaluation*.

2.D.ii Collecte et interprétation de la rétroaction qualitative

Les chercheurs se livrant à l'étude de l'évaluation de l'enseignement ont soutenu pendant plusieurs décennies que le mieux est de réaliser les évaluations sur l'efficacité de l'enseignement en se servant de différentes mesures qualitatives de l'efficacité de l'enseignement en supplément des évaluations des étudiants (Lattuca et Domagal-Goldman, 2007; Ory, 2001; Arreola, 1983). Un bon nombre de ces moyens d'évaluation d'ordre qualitatif (notamment les portfolios ou dossiers d'enseignement, les

autoévaluations, les évaluations par les pairs et les descriptions écrites de l'enseignement) ne s'appuient pas sur la participation des étudiants; par ailleurs, comme nous l'aborderons dans la *Section 4.B : Étudiants dans le rôle d'évaluateurs*, même si les étudiants sont efficaces pour mesurer les comportements et les activités d'enseignement en classe, ils ne le sont pas pour évaluer convenablement le contenu des cours ni les buts pédagogiques, ce qui amène la nécessité de consulter d'autres sources d'information.

Néanmoins, l'argumentation selon laquelle on doit inclure des sources d'ordre qualitatif repose également sur la valeur des rétroactions fournies par les étudiants sur les sujets abordés dans les évaluations de cours. En effet, Harper et Kuh (2007) indiquent que les moyens d'évaluation qualitatifs peuvent souvent mettre en lumière des questions qui ne peuvent être cernées au moyen des évaluations quantitatives traditionnelles. C'est pour cette raison que les rétroactions d'ordre qualitatif fournies par les étudiants sont principalement recueillies, évaluées et utilisées à des fins formatives, plutôt que sommatives (Franklin, 2001; Lewis, 2001). Elles prennent souvent la forme d'évaluations de mi-session ressemblant aux évaluations de fin de session (Lewis, 2001) ou aux évaluations non officielles qui sont réalisées en classe (Diamond, 2004). Les évaluations de mi-session sont examinées à la *Section 5.C.i : Assurance de l'utilité des évaluations pour les étudiants* et à la *Section 6.A.iii : Utilisation accrue des évaluations à des fins formatives*.

Algozine et coll. (2004) indiquent néanmoins que la majorité des formulaires d'évaluation de cours de fin de session permettent aux étudiants d'ajouter des remarques écrites. Dans ces cas, bien qu'il y ait une collecte d'information qualitative, celle-ci est souvent ni interprétée ni analysée de façon efficace et n'est pas intégrée efficacement dans les procédures d'évaluation sommative. La gestion et l'interprétation des rétroactions formulées par écrit sont examinées à la *Section 3.F.ii : Gestion des commentaires écrits* et *3.G.iii : Utilisation des commentaires écrits*. Le fait que les commentaires des étudiants peuvent induire en erreur ou être inexacts constitue un problème; Hodges et Stanton (2007) font valoir que le manque de clarté d'un étudiant à l'égard de son propre processus d'apprentissage peut donner lieu à des commentaires discordants ou créant de la confusion dans les évaluations. Un autre problème à résoudre est la perception selon laquelle l'analyse des commentaires écrits nécessite beaucoup de temps et d'efforts. Plusieurs études (Beran, Violato et Kline, 2007; Beran et coll., 2005; Wagenaar, 1995) sur l'utilisation des évaluations par les enseignants et les administrateurs indiquent que ceux-ci examinent rarement les commentaires écrits, préférant plutôt utiliser les évaluations globales qu'ils perçoivent comme prenant moins de temps. Les auteurs suggèrent d'offrir une formation aux étudiants, aux enseignants et aux administrateurs sur la valeur des commentaires écrits et, chacun en ce qui les concerne, sur les techniques efficaces d'écriture et d'analyse de ces commentaires.

Certains membres du corps professoral font davantage confiance aux réponses d'ordre qualitatif fournies par les étudiants aux questions ouvertes qu'aux évaluations quantitatives; d'autres prétendent le contraire. Cependant, les études ont démontré qu'il existe une corrélation entre les évaluations qualitatives et quantitatives (Cashin, 1995).

Abrami (2001) est d'avis que les mesures d'ordre qualitatif ne devraient pas être incorporées aux évaluations sommatives, car il est plutôt difficile d'en mesurer la fiabilité

et la validité; Harper et Kuh (2007) soutiennent que cette préoccupation, bien qu'elle ne soit pas fautive, n'est pas pertinente en ce qui concerne la façon dont l'information qualitative peut et devrait être utilisée dans les évaluations sommatives.

2.D.iii Collecte et interprétation de la rétroaction formative

Comme l'ont observé Beran, Violato et Kline (2007), Beran et coll. (2005) et Wagenaar (1995), les évaluations de l'enseignement sont principalement utilisées, en grande majorité, à des fins sommatives; c'est-à-dire par les administrateurs afin d'appuyer les décisions relatives à la gestion du personnel. L'étude de Beran, Violato et Kline (2007) démontre dans les faits que, même si les membres du corps professoral croient en l'utilité des évaluations pour fournir une appréciation de l'enseignement, ils n'emploient que très rarement les résultats des évaluations qui les concernent pour la prise de décisions relatives à leur développement professionnel ou à leurs cours.

La majorité des chercheurs attribuent cette absence relative de l'utilisation formative des évaluations de l'enseignement au manque de ressources permettant d'interpréter les résultats d'évaluation et de déterminer les stratégies d'enseignement pouvant s'attaquer aux problèmes qui surgissent (Beran et coll., 2005; Wagenaar, 1995). Cette situation peut limiter la capacité des évaluations d'améliorer l'enseignement; Ory (2001) et Marsh (2007) indiquent que les évaluations permettent d'améliorer l'enseignement seulement si leurs résultats font l'objet d'un échange de vues avec un collègue.

Quoi qu'il en soit, Lewis (2001) et Ory (2001) indiquent que, pour améliorer efficacement l'enseignement, les évaluations doivent à la fois persister dans le temps et être formatives puis être analysées en fonction des objectifs personnels d'amélioration de l'enseignant. Les auteurs font valoir que, si les ressources existent pour aider à interpréter les résultats des évaluations et à y donner suite, les évaluations pourraient être d'une grande utilité comme outil de développement professionnel.

Une rétroaction formative peut être tirée au moyen des évaluations traditionnelles de fin de session ou d'autres formes d'évaluation. Aultman (2006) et Lewis (2001) prônent l'utilisation des évaluations de début de session et de mi-session pour recueillir une rétroaction formative qui peut être mise à exécution sur-le-champ. Hodges et Stanton (2007) décrivent la façon dont les commentaires écrits des étudiants peuvent révéler de l'information sur certains aspects du processus d'apprentissage que les étudiants ne comprennent pas, pouvant ainsi servir d'outil important pour le développement du cours.

Il existe un autre type de rétroaction formative qui peut ressortir des évaluations sommatives plus communes : l'évaluation diagnostique. Si l'on utilise des évaluations de type multidimensionnel, les enseignants peuvent recevoir un rapport faisant état de leurs points forts et de ceux nécessitant une amélioration. Ce type de rapport peut faciliter le développement de l'enseignement de manière autonome et soutenue par l'établissement (Crosson et coll., 2006; Marsh et Roche, 1997). Une description plus détaillée de ce type de rapport est présentée à la *Section 4.D.i : Communication des résultats d'évaluation*.

2.E Usages communs des données d'évaluation de cours

Moore et Kuol (2005) ont constaté que les systèmes d'évaluation par les étudiants permettent de parer à l'information anecdotique véhiculée sur les comportements et l'efficacité de l'enseignement. Ils affirment également que ces outils constituent d'autres façons d'évaluer l'enseignement et permettent ainsi de réduire l'écart qui existe entre les méthodes d'évaluation de l'enseignement et la recherche.

Il existe plusieurs usages communs des données des évaluations de cours : amélioration de l'enseignement; décisions relatives à la gestion du personnel; choix de cours (par les étudiants) et, de plus en plus, compilation des dossiers de nomination pour l'attribution des prix pour l'enseignement.

Amélioration de l'enseignement

Depuis le début de l'utilisation généralisée des évaluations, les chercheurs soutiennent que les données d'évaluation de cours peuvent servir efficacement à améliorer l'enseignement et, de ce fait, l'apprentissage des étudiants (Goldschmid, 1978). Néanmoins, Marsh (2007) et Goldschmid (1978) ont constaté que les données d'évaluation de cours ne peuvent rarement, à elles seules, provoquer des changements dans les comportements d'enseignement étant donné que de nombreux membres du corps professoral ne possèdent pas la formation pour en effectuer l'analyse et, par conséquent, sont moins susceptibles de posséder les aptitudes nécessaires pour évaluer les résultats obtenus. De plus, nombre d'entre eux n'ont pas l'occasion (sur une base volontaire ou obligatoire) de discuter de leurs résultats avec le directeur ou le doyen de leur faculté et seuls certains profitent de ressources et de services offerts par les bureaux de soutien à l'enseignement et à l'apprentissage dans les campus. Par conséquent, la majorité des membres du corps professoral effectuent simplement un examen rapide des données recueillies et tentent rarement d'apporter des changements spécifiques en fonction des rétroactions fournies par les étudiants.

La recherche a fourni la preuve que, lorsque les membres du corps professoral reçoivent de la formation ou de l'aide et effectuent des consultations auprès de collègues ou de conseillers pédagogiques/formateurs en pédagogie universitaire, ils apportent des changements à leurs comportements d'enseignement (Penny et Coe, 2004). Afin d'encourager les changements et d'influer positivement sur les comportements d'enseignement, Abrami (2001) a recommandé une communication plus ouverte vis-à-vis de la collecte des données et de l'interprétation des résultats. Beran, Violato et Kline (2007) suggèrent que les évaluations soient « enrichies de sources d'information complémentaires en ce qui a trait à l'évaluation de l'efficacité de l'enseignement » et soutiennent que « tous les groupes d'utilisateurs, notamment les administrateurs, le corps professoral et les étudiants, devraient être conscients » (p. 37) qu'ils ont besoin de cette information supplémentaire lorsqu'ils utilisent les évaluations des étudiants.

Décisions relatives à la gestion du personnel

Au cours des années 1980 et 1990, les chercheurs se sont régulièrement interrogés sur l'utilisation des évaluations de cours aux fins de décisions sommatives. Leurs

préoccupations découlent en partie des croyances selon lesquelles les données d'évaluation ne sont pas utilisées de façon efficace ou équitable. Néanmoins, le débat entourant l'utilisation efficace des données d'évaluation à des fins sommatives (et également formatives) englobe également les questions qui guident les décisions relatives à la gestion du personnel. Au cours de la dernière décennie, il s'est produit un changement d'attitudes et la majorité des chercheurs, dont Abrami (2001) et Algozzine et coll. (2004), acceptent – ou attestent – de façon générale la validité des résultats provenant des évaluations de cours pour la prise de décisions relatives à la gestion du personnel, notamment l'embauche, la titularisation et la promotion des enseignants.

Il y a trente ans, la recherche faisait état du fait que, même si le corps professoral favorisait l'utilisation des évaluations d'étudiants pour prendre des décisions relatives à la promotion et à la titularisation (Rich, 1976), les administrateurs des universités n'en tenaient pas compte de façon régulière à une telle fin (McKeachie et Lin, 1975). Plus récemment, certaines études ont démontré que les administrateurs sont plus enclins à utiliser les données d'évaluation de cours que les enseignants (Beran et coll., 2005), plus particulièrement en ce qui concerne les décisions relatives à la gestion du personnel (Nasser et Fresko, 2002; Haskell, 1997; Schmelkin, Spencer et Gellman, 1997). Certaines autres études ont montré qu'il n'est pas possible de déterminer clairement si les administrateurs utilisent l'information recueillie de façon appropriée (Abrami, 2001) ou s'ils l'interprètent de façon inappropriée, car elle constitue l'unique source de données sur l'enseignement (Franklin et Theall, 1989).

Une étude récente (Beran, Violato et Kline, 2007) menée dans une université canadienne grandement axée sur la recherche a montré qu'en général les administrateurs perçoivent favorablement les évaluations des étudiants, mais émettent certaines réserves quant à l'utilisation de celles-ci. Beran et coll. (2005) ont indiqué que les administrateurs trouvent les données d'évaluation de cours utiles pour évaluer l'enseignement de façon individuelle (afin de prendre des décisions relatives à la gestion du personnel et d'émettre des recommandations quant à l'attribution de prix d'enseignement), effectuer le suivi des progrès relatifs à la correction de problèmes d'enseignement, évaluer l'enseignement au niveau de l'unité d'enseignement et pour les besoins de planification du programme d'enseignement. Dans cette étude, les administrateurs ont indiqué que les questions les plus utiles dans ces outils d'évaluation de cours étaient les questions générales qui fournissent de l'information sur la qualité globale du cours ou de l'enseignant. Ces indications correspondent aux recommandations émises dans d'autres études (d'Apollonia et Abrami, 1997) indiquant qu'il faut poser des questions générales; elles correspondent aussi aux résultats de Cashin et Downey (1992) qui mentionnent que ces questions sont les indicateurs les plus utiles de l'efficacité de l'enseignement.

Malgré leur utilité pour les évaluations sommatives et les décisions relatives à la gestion du personnel, il existe un consensus selon lequel les données des évaluations de cours ne devraient pas être utilisées de façon isolée, mais plutôt constituer l'un des nombreux indicateurs à utiliser pour évaluer l'enseignement (Beran, Violato et Kline, 2007; Ory, 2001).

Choix de cours par les étudiants

Certains établissements mettent à la disposition des étudiants les données des évaluations de cours en les publiant dans un guide spécial. Ce type de document fournit

habituellement un résumé des données d'évaluation, accompagnées de commentaires des étudiants qui ont été sélectionnés. Il est destiné aux étudiants, qui peuvent s'en servir pour faire leur choix de cours; certains éléments prouvent que l'utilisation de ce guide présente des limites (Beran et coll., 2005). Voir la *Section 3.F.iii : Publication des résultats* et *5.C.i. : Assurance de l'utilité des évaluations pour les étudiants* pour obtenir un survol des pratiques actuelles et des recommandations en ce qui concerne ce type d'utilisation des données d'évaluation.

Prix d'enseignement

Les données d'évaluation de cours sont souvent un élément demandé dans les dossiers de nomination pour l'attribution des prix d'enseignement à la fois à l'interne (au niveau du département, de la division ou de l'établissement) et à l'externe (Union des Associations des Professeurs des Universités de l'Ontario, Prix 3M pour l'excellence en enseignement). Ici, les candidats doivent faire preuve d'excellence en matière d'enseignement dans leur discipline respective, et les évaluations de cours constituent un indicateur. De plus, étant donné que ces données sont recueillies régulièrement, les candidats sont généralement en mesure de faire preuve d'une excellence soutenue en enseignement ou de fournir des données comparables pour montrer leur rendement relatif au sein de leur département, de leur division ou de leur établissement.

Section 3

Politiques et pratiques actuelles en Amérique du Nord

3.A Introduction

Comme il a été mentionné à la *Section 1.A : Méthodologie*, le présent survol des politiques et pratiques en place dans des établissements sélectionnés de l'Amérique du Nord a été réalisé afin d'apporter quelques connaissances approfondies des différents instruments, processus et activités d'évaluation. L'objectif de ce sondage n'est pas de déterminer la prévalence de certaines pratiques, mais plutôt de faire ressortir la diversité, l'étendue, les variations et les points communs quant à l'élaboration, à l'administration et à l'interprétation des évaluations de cours en usage dans les établissements, lesquelles varient selon la mission, l'importance réservée aux programmes, la taille et l'entité compétente.

Après avoir passé en revue toute l'information disponible, nous avons groupé les politiques et les pratiques selon les catégories indiquées plus bas. Les catégories tirées de l'information disponible dans chacun des établissements n'ont pas toutes été examinées; seule l'information pertinente de chaque source est donc exposée. Le fait qu'à maintes reprises l'information disponible était incomplète est un facteur important en soi : bien que certains établissements (p. ex., Harvard) gardent des renseignements sur des sites Web accessibles uniquement par les membres du corps professoral, dans bien d'autres l'information à laquelle nous avons accès était la même que celle qui aurait été disponible d'emblée aux enseignants à la recherche de renseignements sur les politiques et les pratiques d'évaluation de cours dans leur propre établissement.

Nous avons notamment observé une absence significative de politiques sur l'interprétation des résultats des évaluations de cours ou bien d'information accessible aux enseignants et aux administrateurs leur fournissant des lignes directrices quant à l'interprétation. Un petit nombre d'établissements, comme l'Université du Michigan, produisent un guide facilitant l'interprétation et la mise en contexte des résultats des évaluations de cours; cependant, la plupart des politiques et des indications que l'on trouve dans les établissements abordent uniquement la passation des évaluations et la diffusion des résultats. Il reste que l'information nécessaire à l'interprétation des évaluations est essentielle pour être en mesure d'utiliser correctement les évaluations de cours à des fins d'évaluation de l'enseignement, plus particulièrement lorsque ces évaluations servent à prendre des décisions portant sur l'embauche, la titularisation et la promotion du personnel. Compte tenu de ce qui précède, la *Section 5.C.iii : Assurance de l'utilité des évaluations pour les administrateurs* examine les recommandations pertinentes en vue de fournir des directives claires aux enseignants et aux administrateurs.

Tableau 1: Établissements sondés

Ontario	Collèges :	Collège George Brown, Collège Sheridan de technologie et d'enseignement supérieur, Collège Humber de technologie et d'enseignement supérieur, Collège Seneca d'arts appliqués et de technologie
	Universités :	Université de Guelph, Université McMaster, Université Queen's, Université Ryerson, Université Trent, Institut universitaire de technologie de l'Ontario (IUTO), Université de Toronto, Université York
Autres au Canada :		Université Brandon (Manitoba), Université Dalhousie (Nouvelle-Écosse), Université McGill (Québec), Université St. Francis Xavier (Nouvelle-Écosse), Université de l'Alberta, Université de la Colombie-Britannique (UBC)
États-Unis :		Collège Amherst (Massachusetts), Université Harvard (Massachusetts), Université du Michigan, Université du Minnesota
Organismes et entités de gouvernance :		Commission d'évaluation de la qualité de l'éducation postsecondaire (CEQEP) de l'Ontario, Commission pour l'éducation postsecondaire de la Californie, Syndicat des employées et employés de la fonction publique de l'Ontario (SEFPO, représentant le corps professoral des collèges de l'Ontario), ministère de la Formation et des Collèges et Universités de l'Ontario
Autres organismes :		Association américaine des professeurs d'université, Association canadienne des professeurs et professeurs d'université (ACPPU)

Voir l'annexe C.5 pour connaître la liste des politiques et des documents des établissements et des organismes cités en référence dans la présente section.

3.B Politiques relatives à l'évaluation des cours

3.B.i Fréquence et emplacement des politiques

La plupart des établissements maintiennent des politiques relatives à l'évaluation des cours au niveau de l'établissement, qui sont généralement complétées par des politiques ou des procédures élaborées par la division. Les politiques se trouvent généralement auprès de l'une des quatre entités administratives ou entités de gouvernance de l'établissement. Le premier emplacement correspond à la convention collective des professeurs ou au document s'y rattachant (p. ex., Brandon, Queen's, Ryerson, St. Francis Xavier, Trent). Lorsque les politiques relatives aux évaluations de cours font partie de la convention collective, elles sont généralement examinées dans un contexte d'évaluation élargi dans le but de servir à l'embauche, la promotion et la titularisation. Dans certains établissements, dont l'Université Ryerson, un exemplaire du formulaire d'évaluation à l'échelle de l'université est annexé à la convention collective. Le deuxième emplacement est les ressources humaines. C'est le cas du Collège Humber, dont le corps professoral est visé par la convention collective du SEFPO, laquelle n'aborde pas explicitement les évaluations de cours. Le troisième emplacement est les politiques du

sénat (universités Guelph, McGill, McMaster, UBC, York) ou bien du conseil de l'université (Université de l'Alberta, IUTO), lesquelles font fréquemment mention des politiques relatives à l'évaluation des cours. Pour terminer, les politiques relatives à l'évaluation des cours se trouvent sous la juridiction du bureau de l'établissement ou du centre pour le développement et le soutien de l'enseignement, comme c'est le cas des universités Dalhousie et Harvard. Michigan est unique en soi : l'évaluation de l'enseignement est réalisée par le biais de son bureau des évaluations et des examens (Office of Evaluations and Examinations), entité affectée à l'administration et à l'analyse des sondages et des examens. Dans certains établissements (p. ex., l'Université de Toronto), nous n'avons pu relever de politiques officielles à l'échelle de l'université; cependant, de l'information non officielle sur les évaluations de cours était souvent trouvée à des emplacements similaires. De plus, lorsqu'il n'y avait pas de politiques à l'échelle de l'établissement, on retrouvait des politiques à l'échelle des divisions (p. ex., la Faculté des arts et des sciences de l'Université de Toronto).

3.B.ii Point central et étendue des politiques

Les politiques fournissent principalement les directives concernant l'administration et la mise en application des évaluations (p. ex., la fréquence de passation des évaluations, les moyens par lesquels les cours à évaluer sont sélectionnés et la protection ou non de l'anonymat de l'étudiant –

le cas échéant le moyen à utiliser), ainsi que la conservation et la diffusion des résultats. Aussi, de nombreuses politiques spécifient clairement les personnes (enseignant, directeur de département, doyen) ou les entités (comité d'évaluation du département, comité de titularisation et de promotion) qui auront accès à ces données. Un certain nombre de documents de politique énoncent clairement les buts ou les objectifs de l'établissement concernant la collecte des données d'évaluation de cours (p. ex., Alberta, McGill, UBC, IUTO et York). Certaines politiques (dont celles de Brandon, Trent et St. Francis Xavier) présentent des lignes directrices sur l'élaboration ou la modification des instruments d'évaluation, tandis que d'autres précisent très clairement le type d'instrument à utiliser (Alberta, McGill, Ryerson, Queen's, IUTO), ou même le nombre de questions que doit contenir le formulaire (comme à McGill, où le nombre maximal de questions est 25 avec 4 questions obligatoires pour l'ensemble de l'établissement et jusqu'à 21 questions supplémentaires provenant des unités d'enseignement). La politique de l'Université de la Colombie-Britannique relative à l'évaluation de l'enseignement par les étudiants comprend une section sur l'attribution des responsabilités, précisant les rôles distinctifs des étudiants, des administrateurs, du corps professoral, des départements et des enseignants.

Les politiques qui sont intégrées aux conventions collectives se concentrent principalement sur la façon dont les données d'évaluation de cours peuvent être utilisées dans l'évaluation de l'enseignement dans le but de servir à la titularisation ou à la promotion. Lorsqu'il n'existe pas de politiques qui traitent directement l'évaluation, une description de l'utilisation des données d'évaluation à ces fins est souvent présentée dans d'autres documents, notamment dans les politiques et les procédures relatives aux nominations et aux promotions (p. ex., l'Université de Toronto). En général, les politiques ou l'information que l'on trouve dans les différents bureaux dédiés à l'avancement de l'enseignement, aux épreuves ou à l'apprentissage fournissent au corps professoral et aux administrateurs davantage de renseignements complets sur les données

d'évaluation et l'interprétation de celles-ci. Cette information est abordée plus longuement à la *Section 3.G.ii : Information fournie avec les résultats d'évaluation* et la *Section 5.C.iii : Assurance de l'utilité des évaluations pour les administrateurs*.

3.C Conception et approbation des instruments d'évaluation

3.C.i Élaboration et approbation des instruments d'évaluation

Le processus de conception des instruments d'évaluation des cours varie grandement au chapitre de la responsabilité relative à l'élaboration des questionnaires et de la formalité du processus d'approbation de ceux-ci. Parmi les établissements sondés, nous pouvons affirmer la présence de processus d'élaboration d'instruments d'évaluation des cours à tous les niveaux de l'administration (du niveau de la faculté, comme c'est le cas à Amherst, jusqu'au niveau du sénat/conseil de l'université, comme à l'Université de l'Alberta et à l'IUTO).

Un certain nombre d'établissements (Alberta, Dalhousie, Harvard, Humber, McGill, Michigan, Queen's, St. Francis Xavier, IUTO) ont rendu obligatoire l'utilisation d'un instrument commun d'évaluation des cours dans l'ensemble de l'établissement, tandis que d'autres délèguent cette autorité spécifiquement aux départements/divisions (p. ex., Guelph, McMaster, Trent). L'élaboration des formulaires peut être effectuée par un bureau des évaluations ou de l'enseignement (p. ex. : Michigan, Dalhousie) ou par un comité chargé des évaluations (p. ex., Harvard) ou peut être déterminée par des processus de gouvernance (p. ex., Alberta). Dans certains cas (p. ex., Alberta, Dalhousie, Michigan, Queen's), l'instrument commun est développé de façon à permettre au corps professoral d'inclure des questions choisies ou formulées par la division, le département ou un individu. Un certain nombre d'établissements fournissent une banque de questions dans laquelle le corps professoral peut puiser pour les ajouter au formulaire d'évaluation (p. ex., Queen's). En général, le corps professoral est autorisé et encouragé à mener ses propres évaluations non officielles en plus de celles élaborées au niveau de l'établissement (p. ex., Alberta, Queen's, Université de la Colombie-Britannique).

La convention collective de St. Francis Xavier stipule qu'il est nécessaire d'obtenir l'approbation préalable du sénat de la faculté pour tout changement à apporter à l'instrument d'évaluation des cours, au moyen d'un processus de requête officiel. On trouve également une stipulation semblable dans les conventions collectives de Queen's et de Ryerson. À Guelph, la politique du corps professoral confie la tâche de concevoir l'instrument d'évaluation au département; cependant, il est nécessaire d'obtenir l'approbation d'au moins les deux tiers du corps professoral au sein de l'unité d'enseignement avant la mise en application de l'instrument.

Certains établissements (comme Brandon et York) délèguent la conception des évaluations et l'approbation de celles-ci aux facultés, aux unités ou aux centres d'enseignement. Les évaluations doivent généralement se conformer aux politiques de l'ensemble de l'établissement en ce qui a trait à l'administration, à la collecte et à la

diffusion des résultats d'évaluation et doivent habituellement être approuvées par le doyen ou le directeur de la faculté, de l'unité ou du centre d'enseignement.

Amherst permet à chacun des membres du corps professoral d'élaborer ses propres instruments d'évaluation, qui doivent généralement recevoir l'approbation du directeur de département ou du doyen de la division. L'instrument peut être normalisé à l'intérieur du département ou même de la division.

3.C.ii Format et contenu du questionnaire

Nous avons effectué un échantillonnage des instruments d'évaluation venant de divers établissements. Les instruments des universités suivantes : Alberta, Dalhousie, Harvard, Humber, Michigan, Queen's, Seneca et IUTO, sont utilisés par tous les enseignants (à l'exception fréquente des auxiliaires à l'enseignement) de l'ensemble de l'établissement. La description des instruments d'évaluation suivants s'appuie sur ces échantillons pris en exemple. On doit s'attendre à ce que les instruments d'évaluation créés par la division, le département ou un individu soient significativement plus variés.

Nous avons observé que la structure et le contenu des formulaires d'évaluation des cours correspondent grandement à l'instrument d'évaluation type décrit par Algozzine et coll. (2004) à la *Section 2.D : Caractéristiques communes des évaluations de cours*. Dans la majorité des instruments, on demandait principalement une évaluation d'ordre quantitatif et on laissait dans bon nombre des formulaires une section pour l'ajout de commentaires d'ordre qualitatif par les étudiants. Nous avons relevé différentes échelles d'appréciation : des échelles contenant de quatre à sept niveaux d'appréciation, notamment l'échelle de Likert, des échelles d'évaluation qualitative et des échelles d'évaluation fréquentielle. Tous les formulaires comprenaient au moins une question demandant aux étudiants de donner une appréciation générale sur le cours ou l'enseignant. Tous les formulaires contenaient des questions sur le contenu du cours, incluant des questions portant sur les travaux personnels et, fréquemment, la pertinence de la matière vue par rapport aux autres cours ou aux activités professionnelles futures en perspective. Tous les formulaires comportaient également des questions sur les comportements d'enseignement de l'enseignant, lesquels comprenaient presque toujours des questions sur l'enthousiasme de l'enseignant pour la matière, sa disponibilité pour les étudiants, l'atmosphère de la classe et l'implication de l'enseignant. Seneca note que son formulaire a été élaboré à partir du Répertoire des comportements d'enseignement (Teaching Behaviours Inventory, 1987) présenté par Harry Murray; il s'agit d'un instrument d'évaluation de l'enseignement populaire au Canada (voir l'annexe A). La plupart des formulaires comprenaient des questions qui traitaient de l'environnement physique de la salle de classe. Plusieurs instruments posaient des questions sur l'utilisation de la technologie en classe.

À l'automne 2007, l'Université de la Colombie-Britannique a introduit un nouveau système d'évaluation utilisant six points de module universitaire pour toutes les évaluations de cours. La liste des questions peut être allongée par les divisions ou les départements, mais les six points suivants sont obligatoires et doivent être utilisés par tous les enseignants :

1. La clarté des attentes de l'enseignant par rapport à l'apprentissage;
2. La capacité de l'enseignant à communiquer efficacement le contenu du cours;

3. La capacité de l'enseignant à susciter de l'intérêt pour le sujet;
4. L'impartialité de l'enseignant dans l'évaluation de l'apprentissage (examens, essais, épreuves, etc.);
5. Le souci de l'enseignant en regard de l'apprentissage des étudiants;
6. La qualité globale de l'enseignement de l'enseignant.

(Ces questions sont évaluées sur une échelle comprenant 5 niveaux d'appréciation : 1) Très faible; 2) Faible; 3) Approprié; 4) Bien; 5) Excellent.)

Plusieurs des instruments contenaient plus d'éléments inhabituels. Le formulaire de Harvard incluait à la fois des questions avec échelle d'appréciation et des questions à réponses ouvertes permettant l'ajout de commentaires écrits pour chacun des sujets abordés. À Queen's, à Dalhousie et à Michigan, chacun utilise un instrument comportant un nombre limité de questions communes et un plus grand nombre de questions pouvant être formulées ou sélectionnées par les départements ou un membre du corps professoral. Le formulaire d'évaluation de Queen's comporte quatre questions obligatoires et permet d'ajouter jusqu'à sept questions sélectionnées par les départements, et jusqu'à un maximum de 10 (parmi une banque en contenant 200) par l'enseignant. Le formulaire de Dalhousie comprend 10 questions communes et deux sections de cinq questions auxquelles on attribue un chiffre et une échelle d'appréciation, qui peuvent néanmoins être fournies par le département et le membre du corps professoral. Celui de Michigan doit contenir quatre questions communes sur chaque formulaire d'évaluation. L'une de ces questions – « Je désirais fortement suivre ce cours » – sert à mettre en contexte les résultats obtenus dans l'évaluation, car le bureau de cette université a observé qu'une forte appréciation à ces réponses correspondait à une meilleure évaluation de cours dans l'ensemble. Les membres du corps professoral peuvent décider de mettre ou non un bloc de huit questions formulées par l'Assemblée des étudiants de l'Université du Michigan; les résultats des évaluations sont publiés chaque année dans un guide des cours à l'usage des étudiants. Les membres du corps professoral rajoutent alors 18 questions de leur choix (ou 26 s'ils décident de ne pas inclure les questions du guide) qu'ils pigent dans une banque qui en contient plus de 200 traitant de différents sujets, dont le développement de l'étudiant; l'efficacité de l'enseignant; les rédactions, les lectures, les laboratoires et autres travaux personnels; le matériel didactique, incluant la documentation audiovisuelle; l'informatique éducative; la notation et les épreuves; et les responsabilités de l'étudiant. Les enseignants peuvent également décider d'inclure jusqu'à cinq questions ouvertes portant sur le contenu du cours, les outils pédagogiques, les travaux personnels et l'enseignement.

Voir l'annexe C.2, pour obtenir une vue d'ensemble approfondie des instruments et questionnaires d'évaluation de l'enseignement de l'Université du Michigan; l'annexe C.1 en ce qui concerne l'ensemble des questions d'évaluation de McGill; et l'annexe B pour des exemples supplémentaires d'instruments d'évaluation de cours.

3.C.iii Examen des instruments d'évaluation

La récente révision des instruments d'évaluation utilisés à l'Université du Minnesota représente une étude de cas intéressante (et exemplaire) du processus d'examen des instruments d'évaluation de l'enseignement. Le document Foire aux questions (FAQ) (voir l'annexe C.4) mis à la disposition du corps professoral permet à ses membres de

répondre aux questions portant sur le processus d'examen décrit en détail les étapes ayant servi à revoir et à apporter des modifications à leur instrument d'évaluation.

La section FAQ explique que le formulaire a été révisé parce qu'il ne « s'appuyait pas sur les données de recherche en matière d'enseignement et d'apprentissage et contenait un certain nombre de questions qui ne s'avéraient pas utiles aux enseignements, aux administrateurs et aux étudiants » (p. 2), ce qui fait écho à la recherche passée en revue à la *Section 4.C.i : Définition d'un enseignement efficace*, laquelle soutient que la validité et l'utilité des instruments dépendent grandement de la capacité des établissements à élaborer des questions tenant compte des objectifs et des pratiques d'enseignement de l'établissement.

Un comité a conçu et présenté de nouvelles questions en s'appuyant sur des recherches approfondies sur l'enseignement aux études supérieures et les instruments existants, puis a procédé à l'expérimentation du nouvel instrument dans 50 cours. Le nouvel instrument a ensuite été soumis au vote, puis accepté par le sénat de la faculté. Récemment, l'Université de la Colombie-Britannique a également procédé à la révision de ses processus et instruments d'évaluation. Le nouvel instrument a été élaboré par un comité d'évaluation de l'enseignement composé d'étudiants, puis a été approuvé par le sénat de l'université. Après une première évaluation au moyen du nouvel instrument, les résultats des six questions posées dans l'ensemble de l'établissement ont été examinés afin d'en vérifier la fiabilité et la validité; en général, les questions ont été considérées comme étant valables, bien qu'il a été suggéré d'améliorer certains aspects de la formulation des questions. Les examinateurs ont également recommandé d'effectuer une évaluation continue de l'instrument et d'apporter des améliorations à la collecte et au stockage des données en ligne.

D'autres établissements, notamment Queen's et Ryerson, précisent dans leurs conventions collectives d'établir des comités mixtes administration-syndicat chargés de l'examen des formulaires et de l'approbation de toute modification ultérieure.

3.D Processus de mise en place

3.D.i Méthodes de distribution

Les établissements réalisent, analysent et diffusent les résultats des évaluations des cours, que ce soit en ligne, sur copie papier ou en combinant les deux méthodes. Bien que les établissements qui ont mené des évaluations en ligne ou ont étudié à fond la possibilité de le faire (à savoir, les formulaires d'évaluation sont remis aux étudiants par courriel ou au moyen de systèmes de gestion de cours et sont remplis à l'ordinateur) observent que les taux de réponse et les résultats d'évaluation globaux sont plus faibles (quoique la recherche officielle sur ce sujet soit variable; voir la *Section 6.A.i : Outils d'évaluation en ligne*) pour les évaluations remplies en ligne comparativement à celles remplies en classe (voir, p. ex., le rapport « Online Evaluations & Placement Examinations » rédigé par le groupe de travail de l'Université du Michigan), la conduite des évaluations en ligne demeure une perspective attrayante : les évaluations réalisées en ligne font gagner du temps considérable au personnel et, par conséquent, de l'argent. Les différences quant au taux de réponse ou aux évaluations moyennes ne présentent

pas nécessairement un problème si les évaluations sont toutes réalisées en ligne et si l'information pertinente permettant une mise en contexte est transmise au corps professoral et aux administrateurs. Pour cette raison, il n'est pas conseillé de faire passer un certain nombre d'évaluations en ligne et d'autres en classe.

Bien que le recours aux méthodes en ligne soit en hausse pour réaliser les évaluations (nous avons remarqué que certaines écoles réalisent la totalité des évaluations en ligne, notamment l'IUTO et McGill; d'autres offrent une approche par module permettant d'utiliser plusieurs moyens pour mener les évaluations, comme à Guelph, à Queen's, à Trent et à l'Université de la Colombie-Britannique; et plusieurs autres établissements ont indiqué qu'ils souhaitaient explorer la possibilité d'utiliser l'évaluation en ligne ou bien mettaient à l'essai des méthodes de distribution des évaluations en ligne, notamment à Ryerson), la principale méthode de passation des évaluations de cours demeure les formulaires papier distribués en classe. En règle générale, ces formulaires sont imprimés sur des feuilles et les réponses sont notées en noircissant des cercles. Les formulaires sont lus par lecteur optique pour en faciliter l'analyse.

La majorité des établissements font passer les évaluations sur copie papier, mais utilisent Internet ou un système informatique pour procéder à l'analyse des évaluations, à la sauvegarde des données et, moins fréquemment, à la diffusion des résultats.

3.D.ii Application des lignes directrices ou des politiques

Les lignes directrices relatives à l'administration des politiques en matière d'évaluation des cours traitent du choix des cours ou des enseignants à évaluer de même que du processus d'impression, de distribution et de collecte des formulaires d'évaluation. La fréquence et l'étendue des évaluations varient d'un établissement à l'autre. La majorité des établissements procèdent à une évaluation de chaque cours, tous les ans. Moins fréquemment, des établissements choisissent une partie des cours enseignés par chaque membre du corps professoral. Par exemple, Seneca évalue trois cours enseignés par chaque membre à temps plein; ses politiques relatives à l'évaluation énoncent que ces cours doivent faire l'objet d'une sélection menée conjointement avec le corps professoral et doivent représenter les différents types et niveaux de cours enseignés. Certains établissements ne réalisent aucune évaluation ou, dans certains cas, ne diffusent pas les résultats lorsque les classes sont très peu nombreuses (p. ex., McGill) pour des raisons de fiabilité (Cashin, 1995); dans d'autres établissements, les politiques officielles déterminent (p. ex., Alberta) d'autres méthodes d'évaluation pour les classes ayant peu d'inscriptions (moins de 10 étudiants).

La coordination des évaluations est généralement effectuée par l'entité administrative chargée de l'évaluation des cours dans l'établissement (soit à l'intérieur d'un département ou d'une division, à l'échelon du vice-recteur à l'enseignement et à la recherche, ou bien au sein des bureaux de registraire ou d'évaluation). En général, cette entité administrative se charge de l'impression des formulaires et les distribue ou les fait distribuer aux membres du corps professoral (habituellement en coordination avec le personnel administratif du département).

Dans de nombreux cas, le corps professoral reçoit la responsabilité de la coordination des évaluations de cours dans les classes. Un bon nombre d'établissements

(p. ex., McGill, Université de la Colombie-Britannique) fournissent au corps professoral l'information à communiquer aux étudiants sur le processus même et l'importance de remplir les évaluations (habituellement selon la formule d'un énoncé à lire en classe). Harvard insère un énoncé dans l'évaluation elle-même s'adressant aux étudiants et leur rappelant que les réponses demeureront anonymes, qu'elles seront lues attentivement et prises au sérieux. De plus, l'énoncé demande aux étudiants de fournir des commentaires détaillés et constructifs et d'éviter les commentaires sur l'apparence ou les traits distinctifs de l'enseignant. Des pratiques similaires sont en place à l'Université de l'Alberta et à Guelph.

Il arrive souvent que les membres du corps professoral invitent un étudiant volontaire (quoiqu'un soutien administratif soit offert dans certains cas) à distribuer et à recueillir les formulaires d'évaluation puis à les retourner au bureau administratif, habituellement du département. Ce processus sert à conserver l'anonymat des étudiants, à assurer un traitement uniforme des évaluations et à faire en sorte que les évaluations ne sont pas vues par les membres du corps professoral jusqu'à ce que les notes aient été soumises.

Tous les établissements que nous avons sondés limitent la participation du corps professoral dans la procédure d'évaluation. Les politiques de l'Université de l'Alberta et de Queen's spécifient que les enseignants ne peuvent distribuer ni recueillir les formulaires et ne doivent pas être présents dans la salle au moment où les étudiants remplissent le formulaire d'évaluation – il s'agit de pratiques communes à tous les établissements examinés.

Bien que la majorité des établissements remettent les évaluations au cours des dernières semaines d'un cours (soit en classe ou en ligne), Guelph demande que les étudiants reçoivent un exemplaire du formulaire au début de la session. Les politiques de Guelph indiquent également que les départements doivent diffuser une vue d'ensemble des politiques et des procédures connexes, pertinentes à la collecte des données d'évaluation, à tous les étudiants. Comme dans d'autres établissements, les étudiants remplissent les formulaires d'évaluation à la fin d'un cours.

L'Association canadienne des professeures et professeurs d'université (ACPPU) a préparé une Clause modèle sur l'évaluation de l'enseignement (voir l'annexe D.1) qu'elle recommande de transmettre aux facultés ayant de telles politiques au sein de leur établissement. Par exemple, l'Université de l'Alberta a ajouté un lien à cette politique dans le matériel qui accompagne les procédures et la politique d'évaluation des cours.

3.E Analyse des résultats

L'analyse des résultats des évaluations de cours consiste en un processus de rassemblement, de transfert et synthèse des réponses de tous les étudiants. Ce travail d'analyse peut être réalisé par l'entité administrative responsable des évaluations de cours au sein de l'établissement (p. ex., l'Office of Evaluations and Examinations à Michigan ou le Test Scoring and Questionnaire Services à l'Université de l'Alberta) ou bien par des experts-conseils externes (p. ex., Seneca). En règle générale, l'analyse comprend le calcul de la moyenne des résultats aux réponses de chaque question de

l'évaluation, ainsi que la moyenne des résultats aux réponses pouvant servir à des fins de comparaison. Par exemple, à Seneca, la moyenne de chaque évaluation est présentée en fonction de la moyenne au niveau du programme, de l'école, de la faculté et de l'établissement. Les commentaires écrits sont le plus souvent entrés dans un fichier informatique pour assurer l'anonymat de l'étudiant; la gestion de ce fichier est expliquée plus bas.

À l'Université de l'Alberta, la politique de l'établissement décrit en détail la manière dont les résultats d'évaluation sont analysés et présentés aux fins de diffusion. La politique générale de la faculté stipule que les résumés chiffrés, détaillant les taux des réponses pour chaque catégorie et le résultat médian arrondi à la première décimale, et ce, pour chaque question de l'évaluation, doivent être distribués à l'enseignant, aux étudiants, au directeur de département ou au doyen. De plus, tous les rapports doivent contenir les valeurs numériques qui prennent en compte et résument les données biaisées et mettent au jour les observations aberrantes, le cas échéant, de l'échantillon général.

3.F Accès aux résultats

3.F.i Qui a accès à quoi?

Dans les établissements sondés, les membres du corps professoral dont les cours font l'objet d'une évaluation ont un accès total aux données recueillies. En règle générale, cet accès est restreint tant que les notes finales de tous les étudiants inscrits au cours en question n'ont pas été soumises. Des lignes directrices et des procédures expliquent en détail le niveau d'accès autorisé aux autres membres du milieu universitaire.

Habituellement, les directeurs de département ou les chefs d'unité, les doyens et les comités de promotion et de titularisation ont le même niveau d'accès que les membres du corps professoral; c'est le cas à Guelph, à McMaster, à Queen's et à York. Dans certains cas, cet accès se limite aux questions obligatoires présentes dans tous les formulaires d'évaluation et ne comprend pas les données des questions supplémentaires et optionnelles ajoutées par l'enseignant (notamment à l'Université de l'Alberta, à Ryerson et à l'Université de la Colombie-Britannique). Le corps professoral de McGill doit donner sa permission pour rendre disponibles les résultats d'évaluation au milieu universitaire élargi, incluant les étudiants. Il en est de même à Trent, où la convention collective stipule que les évaluations demeurent accessibles uniquement aux membres du corps professoral. Aucun des établissements n'empêche l'utilisation des données d'évaluation à des fins de titularisation et de promotion.

Il existe quelques variations dans la pratique en ce qui a trait à l'accès par les étudiants aux résultats d'évaluation. Dans certains cas, les lignes directrices des établissements recommandent simplement, et n'exigent pas, que les données soient accessibles aux étudiants (p. ex., IUTO et York). Dans d'autres établissements, les étudiants ont accès à des rapports récapitulatifs produits à l'aide des données tirées de modules d'évaluation en particulier (p. ex., Université de la Colombie-Britannique). Plusieurs politiques n'abordent pas la question de l'accès aux données par les étudiants et certaines permettent seulement la visualisation des résultats sommaires aux questions obligatoires dans l'ensemble de l'établissement (p. ex., Université de l'Alberta).

À l'Université de Toronto, les associations étudiantes de certaines divisions comme la Faculté des arts et des sciences produisent un tableau récapitulatif contenant les données sommaires des évaluations de cours de premier cycle. Ici, le corps professoral peut choisir de refuser la publication de ces résultats. Harvard et Michigan ont des systèmes similaires; à Harvard, tous les résultats sont présentés, alors qu'à Michigan, les enseignants peuvent choisir d'inclure un ensemble de huit questions d'évaluation qui seront spécialement ajoutées dans un guide d'évaluation du cours pour les étudiants.

3.F.ii Gestion des commentaires écrits

Bien qu'un bon nombre de formulaires d'évaluation de cours comprennent des questions d'ordre qualitatif, demandant que les étudiants formulent des réponses écrites à des questions précises et laissant un espace pour l'ajout de commentaires généraux, les lignes directrices des établissements n'énoncent pas toujours clairement la façon dont ces données doivent être traitées. Dans certains établissements, il était simplement mentionné qu'il est d'usage de recueillir des commentaires écrits; d'autres déterminaient les personnes ayant accès à cette information; et certains autres indiquaient très clairement les processus de collecte, de production de rapports et de traitement des réponses qualitatives (p. ex., Alberta et Guelph).

La politique générale du corps professoral de l'Université de l'Alberta précise que les commentaires doivent être tapés pour assurer l'anonymat de l'étudiant. Aussi, les étudiants qui le désirent peuvent soumettre des commentaires tapés séparément du processus d'évaluation en classe ou en ligne. À Guelph, les commentaires écrits doivent être signés lisiblement par l'étudiant (comme l'indiquent la politique de l'établissement et le document protocolaire du vice-recteur principal s'y rattachant). Lorsque les commentaires ne sont pas signés par l'étudiant, ils sont uniquement communiqués à l'enseignant. Tous les formulaires d'évaluation de cours contiennent un énoncé qui explique cette politique de signature. À Queen's, le formulaire d'évaluation (University Survey of Student Assessment of Teaching, ou USAT) demeure anonyme, à l'exception d'une section qui permet d'ajouter des commentaires écrits et signés par l'étudiant. Ces politiques existent peut-être pour permettre l'ajout de commentaires écrits aux documents de titularisation et de promotion, étant donné que les conventions collectives interdisent parfois l'utilisation de données anonymes, non groupées (p. ex., Brandon).

La politique de McMaster sur l'incitation à l'excellence de l'enseignement demande aux départements de grouper les données d'évaluation pour produire un rapport de données numériques et de fournir un résumé général des commentaires écrits.

Dans certains établissements, les commentaires écrits par les étudiants ne sont révélés qu'à l'enseignant, et les exemplaires ne sont pas conservés dans un dossier centralisé (p. ex., Michigan, Queen's, Ryerson, St. Francis Xavier). La politique de McGill mentionne que les commentaires écrits sont divulgués uniquement à l'enseignant et au directeur de département. D'autres établissements conservent les commentaires écrits avec les évaluations quantitatives des étudiants (p. ex., Amherst, Harvard).

De nombreux établissements sans politique officielle sur la gestion des commentaires écrits ont néanmoins émis des recommandations sur les utilisations possibles de tels documents (voir la *Section 3.G.iii : Utilisation des commentaires écrits* plus bas).

3.F.iii Publication des résultats

Comme nous l'avons mentionné précédemment, certains établissements publient les données des évaluations de cours de manière régulière, ou bien en font un rapport (Université de l'Alberta, McGill, Queen's, Ryerson, St. Francis Xavier, Université de la Colombie-Britannique), tandis que d'autres recommandent simplement la diffusion des résultats (ex. : IUTO, York). Le mode de distribution diffère, passant de l'impression de condensés ou de rapports sommaires (Université de l'Alberta, Harvard, McMaster, Queen's, Université de la Colombie-Britannique, Faculté des arts et des sciences de l'Université de Toronto) à la diffusion en ligne de documents exposant en détail les moyennes obtenues à des questions données à l'échelle de l'établissement, des divisions et des départements (p. ex., Ryerson). La publication des résultats n'entraîne pas une utilisation à grande échelle des résultats par les étudiants, comme il a été expliqué précédemment à la *Section 2.C.iii : Perceptions des étudiants vis-à-vis de l'évaluation des cours*.

L'ACPPU a publié un « *Énoncé de principes sur l'évaluation anonyme de l'enseignement par les étudiants* » (2006) qui mentionne que « dans les cas où les associations étudiantes effectuent des enquêtes anonymes et publient les résultats pour aider les étudiants à choisir leurs cours, les membres du personnel académique devraient être libres de participer à ces enquêtes [...] » (voir l'annexe D.2 pour lire l'énoncé en entier).

3.G Interprétation et utilisation des résultats

3.G.i Résultats à des fins sommatives et formatives

Dans la plupart des établissements examinés dans le cadre de la présente étude, les évaluations de cours sont réalisées à la fin d'un cours, recueillant et fournissant ainsi des données sommatives. Cette intention est clairement indiquée dans la documentation des établissements (p. ex., Université de l'Alberta, Guelph, McMaster, McGill, Queen's, Ryerson, Trent, IUTO). Dans certains cas, les documents de politiques recommandent en outre la collecte de données formatives au moyen d'évaluations de mi-session ou par un autre moyen. C'est le cas aux universités de la Colombie-Britannique et de l'Alberta, lesquelles préconisent toutes deux une évaluation de l'enseignement par module ou à volets multiples, et ce, de façon continue et à l'aide de différents moyens d'évaluation. Il est également possible de trouver cette recommandation ailleurs que dans une politique officielle, notamment dans des lignes directrices de mise en œuvre, comme à Ryerson.

Lors du dernier examen de son instrument d'évaluation (voir la *Section 3.C.iii : Examen des instruments d'évaluation*), l'Université du Minnesota (2008) a créé parallèlement une version de mi-session de son formulaire d'évaluation afin de recueillir la rétroaction des intéressés à des fins de formation. Elle mentionne que cet instrument de mi-session comprend « les questions de base du formulaire de fin de session et un certain nombre de questions à développement en vue d'aider les enseignants à apporter des améliorations à leur enseignement » (p. 3).

3.G.ii Information fournie avec les résultats d'évaluation

La majorité des établissements sondés rendent disponibles les données brutes et les rapports sommaires au corps professoral et aux directeurs de département suivant la soumission des notes finales d'une session ou d'un semestre donné. Cependant, il existe une grande variété dans l'information supplémentaire fournie au corps professoral et aux administrateurs, lesquels devront interpréter les données. Par exemple, l'Université du Michigan présente les moyennes par division avec des sommaires d'évaluation, alors que d'autres universités incluent les moyennes de l'enseignant ou du département. À Queen's, les rapports sont envoyés aux membres du corps professoral et aux administrateurs et présentent les données groupées relatives aux questions quantitatives, en plus de la moyenne, de l'écart-type, de la fréquence de réponses et du nombre de réponses admissibles calculées. Les rapports contiennent également une représentation graphique des données.

Certains établissements ont produit des guides généraux qui décrivent la mise en application du processus d'évaluation (p. ex., Ryerson, McGill, St. Francis Xavier et IUTO). La convention collective de Queen's demande à l'université de fournir une définition claire des termes statistiques utilisés dans le processus d'évaluation à tous ceux qui sont chargés de fournir une appréciation et une évaluation du rendement de l'enseignement. Le bureau du registraire de l'Université Queen's, unité responsable de l'administration des évaluations réalisées par les étudiants, a préparé des documents à l'intention des utilisateurs, notamment une foire aux questions qui aborde le fonctionnement du système d'évaluation et une fiche de renseignements à l'intention des utilisateurs des rapports d'évaluation expliquant en détail le contenu des rapports et la façon dont les données ont été analysées.

Des établissements (McGill, Michigan) ont rédigé des guides qui précisent la façon d'interpréter les résultats d'évaluation servant à la prise de décision quant à la gestion du personnel et à l'amélioration de l'efficacité de l'enseignement. À York, le comité du sénat de l'enseignement et de l'apprentissage a rédigé un guide d'appréciation et d'évaluation de l'enseignement fournissant au corps professoral et aux administrateurs un survol des différents mécanismes d'évaluation, de leurs avantages et limites et des recommandations sur leur utilisation efficace et réfléchie. Le bureau des services de l'enseignement et de l'apprentissage de McGill a publié un guide similaire intitulé *Effective and Appropriate Use of Student Ratings of Instruction: Research based suggestions*.

À l'Université de l'Alberta, toutes les données d'évaluation sont distribuées aux directeurs, aux doyens, aux comités de titularisation et aux étudiants en fournissant une mise en garde sur les différentes formes de biais. L'établissement fournit un des documents les plus exhaustifs comme outil d'aide à l'intention des personnes qui administrent, interprètent et reçoivent les données d'évaluation. Il s'agit d'un manuel détaillé de 54 pages contenant des extraits de la politique générale de la faculté relativement à l'évaluation de l'enseignement et de l'utilisation du système USRI (Universal Student Ratings of Instruction) décrivant l'objectif des évaluations, le format et le contenu de l'instrument d'évaluation, ainsi que les procédures de mise en application. Le document présente également de l'information pratique à l'intention du corps professoral sur l'administration des évaluations réalisées par les étudiants, des

exemplaires des instruments d'évaluation et un répertoire d'un grand nombre de questions supplémentaires pouvant être rajoutées par l'enseignant. Des exemples de rapports d'enseignant et d'administrateur sont également fournis à des fins d'examen, ainsi que de brèves lignes directrices sur la lecture et l'interprétation de ces documents.

Voir l'annexe C.3 pour obtenir des exemples de rapports de données d'évaluation de l'Université de l'Alberta.

3.G.iii Utilisation des commentaires écrits

Bien que de nombreux établissements évitent de formuler des recommandations sur la façon dont les commentaires écrits doivent être gérés, certains autres (Harvard, Ryerson) soulignent que les rétroactions peuvent être utilisées à des fins de nomination pour des prix d'enseignement ou bien être ajoutées aux dossiers d'enseignement, de titularisation et de promotion. Cependant, certains établissements indiquent clairement que les commentaires écrits ne devraient pas être pris en compte au moment de prendre des décisions touchant la gestion du personnel (p. ex., McGill). Brandon interdit l'utilisation de renseignements anonymes dans la documentation servant aux décisions touchant la titularisation et la promotion, empêchant ainsi efficacement l'utilisation de commentaires provenant d'évaluations anonymes des étudiants. L'Université St. Francis Xavier autorise l'utilisation des commentaires écrits dans les documents servant à la titularisation et à la promotion uniquement dans les cas où le membre du corps professoral a accordé sa permission pour que ces documents soient ajoutés. Dans d'autres établissements, les lignes directrices à l'intention des administrateurs indiquent que ces données peuvent être utiles, mais ils mettent en garde les utilisateurs quant aux biais et limites possibles de telles données (Alberta).

3.G.iv Titularisation, promotion et mérite

Les données sommaires tirées des évaluations de cours ont été utilisées régulièrement à des fins de titularisation, de promotion et d'examen annuel du mérite. La façon d'utiliser ces données est expliquée clairement dans les politiques de l'établissement (ou du département) liées à l'évaluation des cours et dans les documents explicatifs des procédures et politiques relatives à la titularisation, à la promotion et à l'examen annuel. Tous les établissements sondés mettent en évidence l'utilisation des données d'évaluation de cours à ces fins.

Néanmoins, tous les établissements indiquent que les évaluations de cours ne peuvent servir à elles seules de fondement à l'évaluation officielle de l'enseignement, et certains exigent que les évaluations s'accompagnent de renseignements pertinents permettant une mise en contexte (prenant souvent la forme d'un dossier d'enseignement; voir la *Section 3.G.vi : Autres éléments probants de l'efficacité de l'enseignement*). À l'Université St. Francis Xavier, les évaluations de cours ne peuvent être prises en compte que dans les évaluations officielles de l'enseignement, et ce, dans les cas où les évaluations des trois années précédentes montrent une constance. Les évaluations de cours doivent être présentées comme étant une information pertinente supplémentaire sur le cours, incluant sa place dans le programme d'enseignement, le format du cours, l'information sur les outils pédagogiques, la prestation du cours et les autres tâches de l'enseignant.

3.G.v Prix d'enseignement

Bien que ce soit rarement indiqué dans les politiques officielles, la plupart des établissements sondés soulignent que les données d'évaluation pourraient éventuellement être utilisées pour établir les candidatures aux prix d'enseignement. Il s'agit d'une obligation courante pour les prix d'enseignement internes et externes.

À Harvard, les résultats des évaluations de cours servent à l'attribution d'un certificat de distinction en enseignement aux adjoints à l'enseignement, aux auxiliaires à l'enseignement, aux précepteurs et aux chargés de cours. À l'Université de Toronto, les données d'évaluation de cours font partie des éléments probants du dossier de nomination au *President's Teaching Award*.

3.G.vi Autres éléments probants de l'efficacité de l'enseignement

En règle générale, nous avons constaté que le dossier d'enseignement (ou portfolio) constitue la formule que les établissements recommandent le plus de fournir comme élément probant supplémentaire (p. ex., Guelph, McGill, Queen's, Ryerson, Trent, Université de la Colombie-Britannique). Les éléments les plus souvent présents dans ce dossier sont : un énoncé sur la philosophie d'enseignement, des renseignements sur les stratégies pédagogiques utilisées dans la classe et à l'extérieur de celle-ci, des outils pédagogiques représentatifs, des échantillons de travaux d'étudiants et des éléments probants de prix d'enseignement, de perfectionnement professionnel, de mentorat et de recherche en enseignement et en apprentissage (Seldin, 1999).

L'évaluation par les pairs est également fréquemment suggérée (p. ex., Guelph) comme autre formule d'évaluation en classe (p. ex., Alberta, Amherst), notamment l'observation effectuée par des collègues ou des formateurs du corps professoral.

3.H Lien entre les évaluations de cours et les mesures de reddition de comptes

Les résultats des évaluations de cours s'inscrivent parfois dans le cadre de mesures de reddition de comptes à l'intention des autorités. Par exemple, la Commission d'évaluation de la qualité de l'éducation postsecondaire, laquelle autorise les programmes menant à un grade dans les collèges de l'Ontario, demande l'assurance que les responsables des programmes révisent de façon régulière l'enseignement au moyen d'instruments comme les évaluations d'étudiants, mais n'examine pas les évaluations des étudiants directement dans le cadre de son programme d'évaluation. Dans son entente pluriannuelle de reddition de comptes avec le ministère de la Formation et des Collèges et Universités de l'Ontario, Sheridan s'engage à poursuivre certains niveaux de satisfaction des étudiants par rapport aux cours et au programme d'enseignement, que l'on mesure à l'aide de l'instrument de collecte des rétroactions des étudiants de l'établissement.

L'incorporation des évaluations des étudiants dans les mécanismes de reddition de comptes n'est toutefois pas une pratique universelle. Par exemple, le système de reddition de comptes de la commission de l'enseignement postsecondaire de la Californie ne demande pas les résultats des évaluations faites par les étudiants.

Section 4

Fiabilité, validité et interprétation des données provenant des évaluations de cours

4.A Introduction à la fiabilité et à la validité

Il y a depuis longtemps un consensus général dans la littérature selon lequel les instruments d'évaluation des cours peuvent être, et sont la plupart du temps, des outils fiables parce qu'ils fournissent des mesures cohérentes et stables d'éléments particuliers (p. ex., les aptitudes organisationnelles d'un enseignant ou la charge de travail relative) (voir p. ex., Abrami, 2001; Theall et Franklin, 2001; Wachtel, 1998; Goldschmid, 1978). Ce constat est particulièrement vrai lorsque l'outil a été bien construit et testé sur le plan psychométrique avant d'être utilisé (Centra, 1993; Aleamoni, 1987; Marsh, 1984). De plus, la fiabilité de ces outils se confirme davantage par le fait que les cotes représentent généralement les moyennes des évaluations recueillies auprès d'un certain nombre d'étudiants dans une classe donnée (Hoyt et Pallett, 1999). Marsh et Roche (1997) et Marsh (1987) ont étudié la fiabilité des outils d'évaluation de cours en examinant le niveau de consensus entourant des éléments particuliers auprès d'étudiants dans un même cours. Ils ont soutenu que le consensus entre évaluateurs est un indicateur de la fiabilité; toutefois, ils précisent que le facteur de fiabilité diminue légèrement dans les classes moins nombreuses. La fiabilité a aussi été examinée au moyen d'une analyse en plusieurs sections, démontrant la cohérence des résultats dans des sections de cours multiples (Ory et Ryan, 2001).

Bien que la plupart des chercheurs soient d'accord pour dire que les évaluations de l'enseignement par les étudiants constituent des outils fiables, le consensus est moins fort quant à la validité en général : le degré avec lequel l'outil mesure avec précision des éléments spécifiques (p. ex., la disponibilité de l'enseignant) ou fournit une évaluation générale du cours ou de l'enseignant. Parfois, au cours des 40 dernières années, il y avait consensus sur certains aspects liés à la validité (comme l'effet des caractéristiques particulières d'un cours, d'un étudiant ou d'un enseignant; voir la *Section 4.E.i : Survol des variables étudiées* pour en savoir davantage sur ces variables), bien que les conclusions ne soient pas demeurées constantes et que des études subséquentes ne tiennent pas souvent compte des premiers résultats basés sur des raisons méthodologiques (p. ex., la taille de l'échantillon).

En plus des variables pouvant affecter les résultats des évaluations, nous avons constaté, dans nos recherches, que bon nombre de chercheurs ont cerné d'autres menaces à la validité. Plus particulièrement, la validité est fortement déterminée par la mise au point de questions, d'échelles et de procédures d'exécution appropriées; de

plus, elle est aussi conditionnelle à une utilisation et à une interprétation justes des données d'évaluation.

Un survol historique utile de la recherche de 1975 à 1995 par Greenwald (1997) indique que la majorité des publications produites au cours de cette période montrent que les évaluations de cours sont valides. Dans un numéro spécial de la revue *American Psychologist*, paru en 1997 et portant principalement sur les évaluations de cours, les collaborateurs au volume (dont McKeachie, Greenwald ainsi que Marsh et Roche) s'entendent pour dire que les évaluations de cours par les étudiants sont la « seule source valide sur l'efficacité de l'enseignement » (McKeachie, 1997, p. 1218). Ceux qui ont trouvé que les évaluations de cours étaient valides (Abrami, 2001; Ali et Sell, 1998; Abrami, d'Apollonia et Cohen, 1990; Marsh, 1987) ont montré que les données des évaluations peuvent être mises en corrélation avec d'autres preuves de l'efficacité de l'enseignement comme les évaluations par les collègues ou par le personnel formé affecté au développement du corps professoral (Ory et Ryan, 2001; Ali et Sell, 1998; Wachtel, 1998).

Bien qu'il soit rare de trouver des recherches récentes qui rejettent catégoriquement les évaluations de cours en raison de soucis quant à la validité, le désaccord continue à persister relativement à la validité de certains aspects particuliers des évaluations en lien avec l'étendue des variables pouvant avoir un impact sur les résultats des évaluations (voir la *Section 4.E.i : Survol des variables étudiées* ci-dessous).

4.B Étudiants dans le rôle d'évaluateurs

La question fondamentale à se poser sur la validité des évaluations des cours par les étudiants est de savoir si les étudiants peuvent, en fait, évaluer correctement l'enseignement. Tel qu'indiqué à la *Section 2.C.i : Perceptions des professeurs vis-à-vis de l'évaluation des cours*, l'une des principales préoccupations relevées par le corps professoral relativement aux évaluations de cours est la crainte que les étudiants ne soient pas des évaluateurs fiables des comportements d'enseignement ou des cours. La recherche apaise et valide à la fois cette préoccupation. Le consensus quant à la compétence des étudiants dans le rôle d'évaluateurs se retrouve dans la littérature depuis les années 1970 (Goldschmid, 1978). Plusieurs études démontrent que les étudiants sont fiables et efficaces pour évaluer les comportements d'enseignement (p. ex., présentation, clarté, organisation et techniques d'apprentissage actif), le volume des connaissances acquises, la facilité ou la difficulté de leur expérience d'apprentissage dans le cours, la charge de travail du cours et la validité et la valeur de l'évaluation utilisée dans le cours (Nasser et Fresko, 2002; Theall et Franklin, 2001; Ory et Ryan, 2001, Wachtel, 1998; Wagenaar, 1995). Scriven (1995) a fait valoir que les étudiants se trouvent « dans une position unique pour coter leur propre accroissement des connaissances et de leur compréhension de même que leur changement de motivation face au sujet enseigné. En tant qu'étudiants, ils sont aussi en bonne position pour juger de telles questions, comme évaluer si les tests ont bien couvert toute la matière du cours » (p. 2).

En effet, Theall et Franklin (2001) soutiennent que, sur ces points, les étudiants peuvent de fait être plus qualifiés que des évaluateurs experts ou des pairs pour coter leur instruction; ils écrivent que « les pairs et les administrateurs s’y connaissent habituellement davantage sur le contenu et, par conséquent, ne peuvent pas nécessairement comprendre les perceptions des étudiants éprouvant des difficultés » (p. 48).

Bon nombre d’études affirment que d’autres éléments couramment présents dans les évaluations sont plus difficiles à évaluer pour les étudiants. Parmi ces éléments, notons le niveau, la quantité et la précision du contenu d’un cours de même que les connaissances ou les compétences de l’enseignant dans sa discipline (Coren, 2001; Theall et Franklin, 2001; Green, Calderon et Reider, 1998; Cashin, 1998; Ali et Sell, 1998; d’Appolonia et Abrami, 1997; Calderon et coll., 1996). De tels facteurs ne peuvent être évalués avec exactitude par les étudiants en raison de leur expérience et leurs connaissances limitées d’une discipline en particulier. Ory et Ryan (2001) avancent que « la seule dimension de l’instruction qui ne devrait pas être évaluée par les étudiants, surtout les étudiants de premier cycle, est le contenu du cours » (p. 38). Il a aussi été suggéré que les étudiants sont incapables d’évaluer les pratiques de notation et les méthodes de présentation du cours utilisées par les enseignants, de même que la pertinence des lectures choisies et ne peuvent juger si les enseignants présentent des biais dans la transmission du contenu du cours (Hoyt et Pallett, 1999; Keig et Waggoner, 1994; Cashin, 1988; Cohen et McKeachie, 1980).

Voir la *Section 4.E.ii: Évaluation de la validité* pour obtenir de l’information sur les moyens de valider les réponses des étudiants.

4.C Validité externe : créer l’instrument

Ory (2001) ainsi que Theall et Franklin (2001) soulignent que, pour que les évaluations constituent des mesures valides de l’efficacité de l’enseignement, les questions posées dans l’instrument d’évaluation doivent refléter : 1) les façons dont sont utilisées les évaluations des étudiants pour réaliser une évaluation formative ou sommative de l’enseignement; et 2) les buts pédagogiques et didactiques actuels de l’établissement. Ory et Ryan (2001) notent aussi l’importance de s’assurer que les questions d’évaluation mesurent tout ce que l’évaluation tente de mesurer, et seulement ce qu’elle tente de mesurer. Ils ont déterminé ce qui suit :

La sous-représentation du construct et la variance non pertinente du construct sont deux menaces significatives à la validité. Tel que mentionné, la sous-représentation du construct survient lorsqu’une évaluation est trop étroite ou ne réussit pas à inclure des dimensions importantes du construct; et une variance non pertinente du construct existe lorsqu’une évaluation est trop large et contient un excès de variance fiable qui affecte les réponses d’une manière non pertinente pour interpréter le construct (p. 33).

Le rôle central des questions valides pour assurer des évaluations fiables ne peut être exagéré. La tâche visant à relever et à mettre au point des questions appropriées est décrite ci-dessous. La conception de l'instrument joue aussi un rôle important pour en assurer la validité. Scriven (1995) avance que la validité des évaluations peut aussi être affectée par la longueur du formulaire, alors que Sedlmeier (2006) aborde l'effet des échelles particulières de cotation sur les résultats de l'évaluation. Ces questions sont abordées plus en détail dans la *Section 4.C.ii : Mise au point des instruments d'évaluation*.

4.C.i Définition d'un enseignement efficace

Définition

Ory et Ryan (2001) écrivent que « pour faire des déductions valides à partir des cotes accordées à l'enseignement par les étudiants, les points évalués doivent être pertinents et représentatifs des processus, des stratégies et du domaine de connaissance de la qualité de l'enseignement » (p. 32). Pour que les évaluations de cours soient des mesures valides de l'efficacité de l'enseignement, non seulement les questions doivent refléter les aspects de l'enseignement jugés comme efficaces, mais la définition même de l'enseignement efficace doit être déterminée et acceptée. Toutefois, définir ce que sont les comportements d'enseignement efficaces est difficile : Ory et Ryan font valoir qu'aucun « ensemble de caractéristiques universelles s'appliquant à des professeurs et à des cours efficaces et devant être utilisé comme cible [...] ne semble exister » (p. 32).

Plusieurs instruments d'évaluation ont tenté de cerner des comportements ou moyens spécifiques de structurer ou de présenter la matière d'un cours qui contribuent à enseigner de manière efficace. Par exemple, les concepteurs de SEEQ (Marsh, 1987), de l'inventaire des comportements d'enseignement (Murray, 1987) et d'IDEA (Cashin, 1992), ont tous examiné des recherches sur l'apprentissage des étudiants et ont procédé à des sondages auprès des étudiants et du corps professoral afin de déceler les caractéristiques d'un enseignement efficace (voir l'annexe A). Ces éléments peuvent être validés grâce à des comparaisons à d'autres mesures d'apprentissage des étudiants (p. ex., le rendement scolaire). Cependant, il existe quelques moyens au-delà de l'analyse logique (Marsh, 1987) pour garantir que ces caractéristiques de l'enseignement efficace sont représentatives, complètes ou généralisables. Wachtel (1998) note que de nombreux chercheurs ont fait une mise en garde selon laquelle d'autres moyens d'évaluer la validité des évaluations des étudiants « présument [...] un consensus qui n'existe pas encore. Autrement dit, comment pouvons-nous évaluer de manière appropriée l'efficacité de l'enseignement si nous ne nous entendons même pas sur ce que constitue l'enseignement efficace? » (p. 193).

En fait, plusieurs chercheurs ont fait une mise en garde quant à la possibilité d'effets pervers causés par la détermination des caractéristiques de l'enseignement efficace. McKeachie (1997) s'inspire de Scriven (1981) pour soutenir qu'« aucune cote portant sur le style de l'enseignement (ex. : l'enthousiasme, l'organisation, le côté chaleureux) ne devrait être utilisée, parce que l'efficacité de l'enseignement peut se concrétiser de nombreuses façons. L'utilisation de caractéristiques qui ont généralement des corrélations positives avec l'efficacité pénalise le professeur qui est efficace malgré des cotes inférieures à l'une ou à plusieurs dimensions habituellement associées à l'efficacité » (p. 1218-1219).

McKeachie constate aussi que « les membres du corps professoral et les administrateurs entretiennent des stéréotypes sur ce qu'implique un bon enseignement » et que « les enseignants qui ne se conforment pas au stéréotype [selon les éléments de l'enseignement cernés dans les évaluations] sont enclins à être jugés inefficaces malgré d'autres preuves d'efficacité » (p. 1219).

Ory et Ryan (2001) catégorisent les effets sur l'établissement des évaluations en ce qui a trait aux conséquences « intentionnelles » ou « involontaires ». Parmi les conséquences involontaires, il y a la possibilité que les enseignants utilisent des éléments du formulaire de cotation pour modéliser leur enseignement ou leurs cours. Si les éléments présents dans un formulaire d'évaluation ne reflètent pas les priorités de l'établissement ou les moyens utilisés pour un enseignement efficace, ces évaluations peuvent avoir un impact négatif sur l'enseignement dans cet établissement.

L'étude de Neumann (2001) sur les différences dans l'enseignement entre les disciplines a aussi d'importantes implications dans la définition de l'enseignement efficace pour les évaluations de cours. Neumann fait valoir que les objectifs de l'apprentissage et les styles d'enseignement varient significativement selon la discipline :

Donald (1983) a noté que les domaines purs et rigides avaient des cours bien structurés incluant des concepts et principes très reliés. Les domaines purs et souples avaient des structures de cours ouvertes et étaient organisés de façon plus flexible. En tenant compte des objectifs de formation, Braxton (1995) a trouvé que les disciplines rigides accordent une plus grande importance à la préparation de la carrière des étudiants et mettent l'accent sur les objectifs cognitifs comme l'apprentissage de faits, de principes et de concepts. Les domaines souples accordent, quant à eux, une plus grande importance à la connaissance générale élargie, au développement de la personnalité des étudiants et à la capacité de raisonner efficacement, comme développer la pensée critique. Hativa (1997) a découvert que les domaines purs et souples accordaient une plus grande importance à la créativité de la pensée et à l'expression, tant orale qu'écrite, alors que les domaines rigides et purs de même que rigides et appliqués ont fortement mis l'accent sur la capacité à appliquer les méthodes et les principes (p. 138).

Ces différences entre les disciplines peuvent grandement affecter les façons dont les évaluations sont construites et leur validité, évaluée, de même que le degré avec lequel elles reflètent avec exactitude l'efficacité selon les normes d'un domaine particulier. En général, toutefois, Neumann remarque que « dans la plupart des cas, les instruments d'évaluation employés sont génériques, ce qui signifie que l'enseignement d'une discipline à l'autre est le même » (p. 143).

À la différence de ces préoccupations, toutefois, Wachtel (1998) fait valoir que « les étudiants et le corps professoral sont généralement d'accord sur ce que sont les composantes d'un enseignement efficace et leur importance relative » (p. 192) et croit que les questions mises au point à partir de ces composantes peuvent refléter avec précision l'apprentissage des étudiants et l'efficacité de l'enseignement.

Efficacité de l'enseignement : mesures générales par rapport à multidimensionnelles

La plupart des tentatives pour relever les caractéristiques particulières de l'enseignement efficace découlent d'une croyance voulant que l'enseignement devrait se mesurer à partir d'aspects ou de catégories multiples de l'activité d'enseignement – par exemple, la disponibilité, l'enthousiasme ou l'organisation – plutôt que d'après des questions qui visent une réponse globale et vaste au cours ou à l'enseignant en général. Cet argument a été avancé de manière plus notable par Marsh (1987), puis par Marsh et Roche (1997). Ces mesures multiples proviennent de la proposition selon laquelle l'enseignement est multidimensionnel – c'est-à-dire que les enseignants peuvent exceller dans un aspect de l'enseignement, mais pas dans un autre. Ory (2001) note qu'alors qu'un ensemble particulier de comportements d'enseignement efficaces ne peut pas facilement être décelé, les résultats de recherches sur les perceptions des étudiants vis-à-vis de ce qu'est un enseignement efficace « soutiennent la notion voulant que les étudiants perçoivent la qualité de l'enseignement comme multidimensionnelle » (p. 10).

Comme ces chercheurs le soutiennent, en plus de dépeindre l'enseignement efficace avec plus de précision, les évaluations multidimensionnelles peuvent être très utiles comme outils de diagnostic pour fournir une rétroaction formative (Marsh et Roche, 1997) étant donné qu'elles peuvent aider les enseignants à connaître les domaines dans lesquels ils font preuve de force ou de faiblesse. De plus, Marsh et Roche font valoir que la rétroaction multidimensionnelle est importante pour poursuivre la recherche sur les évaluations de l'enseignement. Ils donnent plusieurs exemples de cas dans lesquels une variable pouvant avoir été repérée comme une source potentielle de biais peut, en fait, selon les évaluations des réponses des étudiants basées sur une dimension particulière de l'enseignement, être présentée comme ayant un effet positif sur l'apprentissage de l'étudiant. En d'autres mots, « un effet qui a été interprété comme un "biais" aux évaluations de l'enseignement par les étudiants peut être interprété de manière plus appropriée comme un soutien à leur validité relativement à une composante de l'enseignement efficace » (p. 1193).

McKeachie (1997) est d'accord avec le groupement en dimensions que fait Marsh des comportements d'enseignement, notant que cette façon de faire aide à la lecture et à l'interprétation des données et, par conséquent, est plus susceptible de mener à l'amélioration. Toutefois, il souligne l'incertitude relative au nombre de dimensions qui devraient être prises en compte à des fins sommatives : « une cote exprimant un résumé pondéré des facteurs devrait-elle être présentée (comme Marsh et Roche [1997] le soutiennent) ou est-ce qu'on devrait simplement utiliser les résultats d'une ou plus d'une des évaluations générales de l'efficacité de l'enseignement (comme se contentent d'avancer d'Apollonia et Abrami, 1997)? » (p. 1218).

Cependant, de nombreux chercheurs avancent que les questions générales sont tout aussi, sinon plus, valables pour évaluer de manière sommative l'enseignement que les mesures multidimensionnelles (Algozine et coll., 2004; Abrami, 2001; Cashin, 1995, 1992). Dans leur étude sur les façons dont les évaluations sont utilisées par le corps professoral et les administrateurs, Beran, Violato et Kline (2007) indiquent que les mesures générales sont fréquemment les seules à être évaluées dans les processus officiels d'évaluation (voir la *Section 2.E : Usages communs des données d'évaluation de cours*). Abrami (2001) et Cashin (1995, 1992) font remarquer qu'il y a une forte

corrélation entre les questions générales et les autres mesures de l'efficacité de l'enseignement. Lorsqu'un nombre limité de résultats sont examinés, les questions générales peuvent servir avec précision d'indicateurs pour des mesures plus complexes et dans ces cas, donc, une véritable question générale, plutôt qu'une sélection limitée de mesures multidimensionnelles, est plus représentative de l'efficacité de l'enseignement. McKeachie (1997) est d'avis que de telles cotes générales sont valables pour l'évaluation sommative de l'enseignement, même si elles ne sont pas particulièrement valables pour fournir une rétroaction formative. Marsh et Roche (1997) ne sont toutefois pas d'accord avec cette perspective, soutenant que « des cotes globales ou générales ne peuvent pas représenter de manière appropriée l'aspect multidimensionnel de l'enseignement. Elles peuvent aussi être plus sensibles au contexte, à l'humeur et à d'autres biais potentiels que ne le sont d'autres éléments spécifiques plus intimement liés aux comportements d'enseignement réels » (p. 1188).

Dans une révision récente de son instrument d'évaluation (voir la *Section 3.C.iii : Examen des instruments d'évaluation*), l'Université du Minnesota a décidé d'éliminer sa question générale : « Comment coteriez-vous la capacité d'enseignement générale de l'enseignant? » Le comité chargé de réviser l'instrument a fait valoir que cet élément était trop souvent la seule cote évaluée dans une évaluation sommative de l'enseignement, que les étudiants avaient de la difficulté à répondre à cette question, que l'élément n'est pas un diagnostic et que les questions générales comme celle-là ne sont pas en corrélation avec les cotes aux questions qui portent sur des caractéristiques spécifiques de l'enseignement.

Il y a peu de discussion dans la littérature actuelle sur l'énoncé particulier des questions générales. Scriven (1995) a indiqué que de nombreuses évaluations sommatives posent de mauvaises questions générales. Il cite les erreurs courantes suivantes : des questions requérant des comparaisons entre les enseignants; des questions où les étudiants doivent indiquer s'ils recommanderaient le cours aux autres; et des questions qui demandent aux étudiants si un cours est le « meilleur » qu'ils ont eu à ce jour. Malheureusement, Scriven ne précise pas pourquoi une telle formulation n'est pas souhaitable.

4.C.ii Mise au point des instruments d'évaluation

Sélection des questions

Comme Marsh et Roche (1997) font valoir, la sélection des questions d'évaluation est un facteur essentiel pour assurer que les évaluations sont des mesures valides de l'efficacité de l'enseignement et que :

La validité et l'utilisation de l'information tirée de l'évaluation de l'enseignement par les étudiants dépendent du contenu et de la couverture des éléments. Des énoncés mal formulés ou inappropriés ne fourniront pas de l'information utile, tandis que des cotes dont on établit la moyenne à partir d'un assortiment d'énoncés mal définis n'offrent aucune base pour savoir ce qui est mesuré (p. 1187).

Et ce, pour des raisons liées aux façons dont les étudiants répondent aux questions (voir la *Section 4.C.ii : Mise au point des instruments d'évaluation*), à la relation entre les questions d'évaluation et celles sur les caractéristiques de l'enseignement jugées importantes ou efficaces dans un contexte particulier (voir la *Section 4.C.i : Définition d'un enseignement efficace*) et à l'éventail de questions auxquelles les étudiants peuvent répondre avec exactitude (voir la *Section 4.B : Étudiants dans le rôle d'évaluateurs*). Malgré ces considérations importantes, les éléments de l'évaluation sont souvent choisis avec moins d'attention que ce à quoi on devrait s'attendre. Marsh et Roche (1997) écrivent que « en pratique, la plupart des instruments reposent sur une combinaison de considérations logiques et pragmatiques, incluant occasionnellement des observations psychométriques comme la fiabilité ou l'analyse de facteurs » (p. 1187). Ory et Ryan (2001) notent que « bon nombre des formulaires d'évaluation de cours utilisés aujourd'hui ont été mis au point à partir d'autres formulaires existants sans mûre réflexion à la théorie ni aux domaines du construct » (p. 32). Comme le démontre la *Section 3.C : Conception et approbation des instruments d'évaluation*, la mise au point d'une évaluation et les politiques et pratiques d'approbation varient significativement d'un établissement à l'autre. L'imprécision de la sélection des questions et du développement d'instruments demeure donc un obstacle significatif à la validité de l'évaluation.

Test psychométrique

La formulation, l'ordre et l'échelle utilisés pour les questions peuvent eux-mêmes avoir un effet significatif sur les évaluations. Par conséquent, un élément important pour assurer la validité des formulaires d'évaluation est le test psychométrique. Tel qu'indiqué ci-dessus, Ory et Ryan (2001) soutiennent que de nombreux établissements mettent au point des évaluations en se servant de questions provenant de formulaires existants qu'ils ont simplement adaptées. Bien que ces formulaires originaux – par exemple l'ensemble de questions mis au point par l'Université du Michigan ou l'évaluation du système d'éducation IDEA par les étudiants – aient été soumis à un test psychométrique poussé, les évaluations adaptées décrites par Ory et Ryan n'ont pas conservé, et peuvent ne pas conserver, la validité des formulaires originaux. Comme Marsh et Roche (1997) le soulignent, « les sondages maison conçus par des chargés de cours ou des comités (pour l'évaluation de l'enseignement par les étudiants) sont rarement évalués relativement à des considérations psychométriques rigoureuses et révisés en conséquence » (p. 1188). Franklin (2001) dégage des problèmes couramment observés relativement à de tels sondages maison, dont les questions à double-détente, « énoncés trop complexes ou ambigus » ou « choix de réponses mal gradués » (p. 89).

Ory et Ryan écrivent aussi qu'il y a peu de connaissances sur le processus par lequel les étudiants répondent aux questions d'évaluation et sur la question de savoir si les étudiants répondent à l'échelle de cotation de manière constante. Ils soulignent, par exemple, qu'aucune recherche ne détermine si les « étudiants répondent aux énoncés en comparant le rendement de l'enseignant à celui des autres enseignants ou à une norme idéalisée » (p. 33). De la même façon, il existe peu de recherches démontrant comment les étudiants interprètent les points individuels sur les échelles de cotation et que « nous avons besoin de déterminer s'il y a correspondance entre la signification de l'échelle pour les étudiants et son sens voulu » (p. 34). Finalement, ces auteurs notent que les façons qu'ont les étudiants de répondre aux échelles de cotation peuvent varier en fonction des facteurs démographiques, dont l'âge, le niveau de scolarité et le bagage culturel. D'autres études démontrent des menaces similaires à la validité des

évaluations : Greenwald (1997) observe que selon la façon dont le formulaire est construit, les étudiants peuvent fournir la même cote, ou des cotes similaires, à tous les énoncés; Sedlmeier (2006) démontre que l'ordre et l'échelle utilisée dans les évaluations quantitatives des étudiants affectent les résultats de ces évaluations. Coren (2001) aborde l'« effet de halo », c'est-à-dire que lorsque l'on perçoit certains aspects d'une personne sous un angle positif, il y a une tendance à percevoir tout ce que la personne dit ou fait sous ce même angle, ce qui offre donc moins de certitude quant à savoir si les cotes accordées aux énoncés individuels reflètent des forces et des faiblesses spécifiques. L'effet de halo amplifie aussi les perceptions négatives.

Examen de l'instrument

Déterminer la fréquence optimale avec laquelle les évaluations sont examinées est une question d'équilibre à trouver entre s'assurer que les énoncés de l'évaluation reflètent la pratique et les priorités pédagogiques actuelles de l'établissement et s'assurer que les énoncés de l'évaluation sont sélectionnés et évalués avec suffisamment d'attention pour qu'ils répondent au construct et aux critères de validité psychométrique décrits ci-dessus. Ory et Ryan (2001) font une mise en garde contre le recours à des questions d'évaluation périmées. Ils soutiennent que :

Par exemple, de nombreux collèges et universités encouragent maintenant le corps professoral à utiliser l'informatique pour enseigner. Les formulaires d'évaluation utilisés sur le campus ont-ils été modifiés pour inclure des questions sur cette technologie? Les implications liées à la valeur des évaluations des étudiants, qu'elles soient intentionnelles ou non, peuvent être que le contenu de l'évaluation définisse les dimensions de l'enseignement qui sont valorisées et encouragées par l'établissement (p. 38).

Theall et Franklin (2001) recommandent que, « lorsque des changements aux programmes ou des changements liés à l'établissement sont apportés, l'établissement devrait revoir le système d'évaluation et l'adapter selon les besoins », mais ils mettent l'accent sur le fait que les établissements devraient « demander l'avis et l'aide d'experts lorsque c'est nécessaire » (p. 53) afin de respecter une autre de leurs recommandations : « adhérer à des principes et des pratiques rigoureuses en matière de psychométrie et de mesures » (p. 52).

4.D Validité externe : communication et interprétation des résultats d'évaluation

d'Apollonia et Abrami (1997) observent que « de nombreux experts en évaluation du corps professoral considèrent que la validité des évaluations sommatives basées sur des évaluations réalisées par les étudiants est menacée par une réalisation inappropriée de la collecte, de l'analyse, de la communication et de l'interprétation des données » (p. 1203, citant les travaux d'Arreola, 1995; de Theall, 1994; de Franklin et Theall, 1990).

Dans le même ordre d'idées, ils soutiennent que « les questions particulières posées, l'utilisation des cotes générales par rapport à des cotes factorielles, l'introduction possible de biais par les variables, etc. causent des problèmes relativement mineurs. Le plus grand problème relié à la validité réside dans l'utilisation qu'en font les comités de gestion de personnel et les administrateurs » (p. 1222). Franklin et Theall (1989) en viennent à des conclusions semblables. La présentation appropriée des données d'évaluation de même qu'une interprétation juste de celles-ci par des gens formés sont des éléments essentiels, voire fondamentaux, pour assurer la validité des évaluations.

4.D.i Communication des résultats d'évaluation

Sélection des résultats et des synthèses de résultats pour la production des rapports

Comme l'expliquent la *Section 2.D : Caractéristiques communes des évaluations de cours* et la *Section 3.C.ii : Format et contenu du questionnaire*, un bon nombre d'instruments d'évaluation incorporent une série de questions portant principalement sur des aspects du cours (contenu, matière, charge de travail) et de l'enseignant (disponibilité, rapidité du retour des travaux corrigés et notés, clarté des attentes), doublée d'un ensemble de questions d'ordre général (les étudiants devant fournir une appréciation d'ensemble du cours et de l'efficacité de l'enseignant en général). Lorsqu'un formulaire est bien structuré, les résultats obtenus à chacune des questions permettent de fournir de précieuses données au corps professoral, aux administrateurs ainsi qu'aux étudiants. Cependant, la façon dont les résultats aux questions générales et multidimensionnelles sont organisés dans les rapports destinés au corps professoral et aux administrateurs peut avoir une incidence sur la validité des interprétations faites à partir de ces données.

Abrami (2001) indique que les données d'évaluation des questions individuelles posées dans les formulaires d'évaluation peuvent et devraient servir à des fins formatives dans le but d'améliorer l'efficacité de l'enseignement. Toutefois, fournir au corps professoral des rapports pour chaque question individuelle pourrait alourdir et rendre plus difficile la tâche d'évaluation. Pour y remédier, Algozzine et coll. (2004) ont suggéré de grouper les questions par catégorie de comportement d'enseignement ou d'éléments de cours et de ne fournir au corps professoral que les cotes des catégories, plutôt que les cotes de chaque question.

Comme l'aborde la *Section 4.C.i : Définition d'un enseignement efficace*, l'enseignement est une activité à multiples facettes et certains chercheurs (p. ex., Marsh) soulignent la nécessité d'utiliser un questionnaire d'évaluation multidimensionnel pour être en mesure d'obtenir une évaluation approfondie de l'efficacité de l'enseignement. Cependant, Abrami (2001) a fait valoir que bien que l'enseignement soit une activité à multiples facettes, la prise de décisions sommatives concernant l'efficacité de l'enseignement ne l'est pas, et qu'il est plus utile de fournir aux administrateurs un seul résultat représentant une définition étendue et exhaustive de l'efficacité de l'enseignement. Cependant, les chercheurs ne s'entendent pas sur la provenance de ce résultat unique, à savoir s'il devrait être issu a) d'une question générale demandant aux étudiants d'évaluer le cours ou l'enseignant (position défendue par Abrami, 2001) ou b) d'un résultat calculé en faisant la moyenne des données de plusieurs dimensions d'enseignement, pondéré en fonction des priorités de l'établissement (proposé par d'Apollonia et Abrami, 1997).

Cashin et Downey (1992) conviennent du fait que les cotes attribuées aux dimensions prises individuellement ne sont pas très utiles aux administrateurs aux fins d'évaluation sommative. L'étude qu'ils ont réalisée sur les données tirées de 17 183 cours dans 105 établissements a permis d'observer que les formulaires de type court et économique (p. ex., comportant des questions générales) ont permis de saisir la majorité de l'information nécessaire aux fins sommatives. Cependant, des questions et des résultats plus détaillés peuvent aider aux évaluations formatives.

Présentation des données

Des chercheurs ont été préoccupés par la façon de présenter des résumés statistiques des données d'évaluation au corps professoral et ont indiqué que bon nombre des responsables de l'interprétation des données ne sont pas munis de l'information ni des compétences nécessaires pour faire le travail de façon efficace. Dans de nombreux établissements, le corps professoral et les administrateurs reçoivent des rapports sommaires pouvant ou non présenter de l'information sur les écarts statistiques, des détails sur la façon dont les données ont été préparées ou un guide d'interprétation ou d'utilisation des données. Comme l'ont fait valoir Abrami (2001) ainsi que Theall et Franklin (2001), sans ce type de renseignements, les administrateurs peuvent prendre des décisions de gestion de personnel qui ne sont pas justes et possiblement inappropriées. Pour surmonter certaines de ces difficultés, Abrami (2001) recommande de présenter les données normatives sous forme visuelle (diagramme, graphique) – particulièrement pour les personnes possédant peu ou pas du tout d'expérience en statistiques.

4.D.ii Difficultés relatives à l'interprétation et à l'utilisation des données à des fins sommatives

Valeur statistique ou données d'évaluation

Scriven (1995) met en lumière une série d'erreurs communes liées à l'utilisation des données d'évaluation de cours, notamment l'utilisation des cotes sans tenir compte de la distribution statistique; le fait d'accorder de l'importance aux petites différences; et l'utilisation des données d'évaluation comme outil principal dans les évaluations sommatives ou formatives. Abrami (2001) a soulevé des inquiétudes quant à l'interprétation erronée et à la mauvaise utilisation des données d'évaluation au moment de la prise de décisions relatives à la gestion du personnel. Il met en garde les administrateurs de ne pas trop insister sur les petites différences dans les évaluations, particulièrement s'ils ne sont pas très expérimentés en analyse statistique. Parallèlement, McKeachie (1997), d'Appolonia et Abrami (1997), ainsi que Wagenaar (1995) préviennent les administrateurs de ne pas surestimer le degré de précision des résultats d'évaluation et recommandent qu'au lieu de se servir des résultats bruts présentés à une ou deux décimales à des fins d'interprétation, les administrateurs devraient classer les résultats en trois catégories générales d'appréciation : exceptionnels, satisfaisants et inacceptables. Ces grandes catégories permettent d'atténuer les variations ou les biais qu'introduisent les caractéristiques du cours ou de la discipline afin de réaliser une comparaison équitable entre les cours et les enseignants. De plus, McKeachie souligne que ces grandes catégories reflèteraient davantage les façons dont les évaluations de l'enseignement sont utilisées à des fins sommatives.

Même si de telles catégories n'étaient pas mises en place, les administrateurs devraient veiller à pouvoir faire une distinction significative entre les niveaux possibles de cotes examinées (p. ex., entre un 3,5 et un 3,6 sur une échelle de 5 points) avant d'employer ces cotes à des fins d'évaluation officielle. McKeachie (1997) fait également valoir que « la présentation des moyennes ou des médianes (souvent à deux décimales près) conduit à des décisions qui s'appuient sur des différences minimales entre les nombres – différences qui ne permettent pas de départager les enseignants compétents et incompetents » (p. 1223). On ne devrait pas fournir aux administrateurs des résultats qui sont plus précis que probants.

Connaissances des administrateurs en matière de recherche et de statistiques

Theall et Franklin (2001) laissent entendre que la validité des évaluations des étudiants se bute en grande partie au peu d'aisance avec laquelle de nombreux administrateurs interprètent les résultats fournis et le manque de formation offerte pour leur permettre d'améliorer ces compétences. Menges (2000) est du même avis; il écrit qu'« un grand nombre de personnes du domaine de l'évaluation affirmeraient que peu importe le degré de validité et de fiabilité de l'instrument d'évaluation, les utilisateurs peuvent et effectuent une mauvaise utilisation des résultats qu'il fournit » (p. 8). Franklin (2001) émet une mise en garde aux personnes qui emploient les évaluations de cours :

[...] sans assumer que ceux qui examineront ces résultats possèdent les compétences et les connaissances nécessaires pour les utiliser dans le cadre des principes directeurs recommandés par les experts en évaluation. [...] Dans une étude multi-établissements, plus de la moitié des membres du corps professoral utilisant les évaluations des collègues ne pouvaient répondre aux questions de base portant sur les statistiques qui sont habituellement présentées sur les rapports types d'évaluation, notamment les moyennes et les écarts-types (p. 86).

Wachtel (1998) fait une mise en garde similaire, affirmant que « le corps professoral et les administrateurs possèdent peu de connaissances sur les travaux de recherche existants dans ce domaine et, par conséquent, peuvent [...] faire en quelque sorte des abus (p. ex., accorder trop d'importance à la signification de la dernière décimale des moyennes calculées pour une classe) » (p. 193).

Franklin et Theall (1989) soulignent, toutefois, qu'il n'incombe pas nécessairement aux administrateurs et aux membres du corps professoral de développer ces compétences en statistiques. Ils affirment que « parce que les évaluations existent dans des systèmes plus grands, nous ne pouvons raisonnablement nous attendre à ce que chaque utilisateur final soit statisticien ou possède les compétences psychométriques pour évaluer sa propre capacité à interpréter les évaluations » (p. 21). Ils recommandent plutôt qu'on fournisse aux utilisateurs « des lignes directrices, des mises en garde, des interprétations et des commentaires » pour mettre en contexte les évaluations et aider à les interpréter (p. 21). L'examen que nous avons fait des pratiques courantes dans les établissements de l'Amérique du Nord examinés a révélé qu'il semble y avoir peu d'information fournie aux administrateurs des universités et collèges pour les aider à interpréter les données d'évaluation. Nous avons trouvé très peu de modèles d'établissements qui créent ou offrent du matériel de formation ou des guides à cette fin.

Theall et Franklin (2000) de même qu'Abrami (2001) ont soulevé certaines préoccupations concernant l'expertise des personnes qui examinent et interprètent les données d'évaluation et qui s'appuient sur celles-ci pour prendre des décisions, observant que les administrateurs des universités sont rarement des experts en matière de recherche et ne sont pas formés pour interpréter de façon efficace les données d'évaluation de leur propre établissement : un état de fait peut influencer négativement sur les décisions relatives à la gestion du personnel. Par conséquent, Abrami (2001) souligne l'importance d'accroître « l'expertise des personnes qui interviennent dans la prise de décision » en modifiant « le système de production de rapports et en orientant le processus de prise de décision » (p. 64). De plus, il fait des recommandations aux établissements sur les façons d'améliorer le jugement porté sur l'efficacité de l'enseignement au moyen des données d'évaluation.

McKeachie (1997) et d'autres ont fortement recommandé que les établissements augmentent leurs efforts pour aider les étudiants à devenir de meilleurs évaluateurs de l'enseignement et pour mieux former les administrateurs relativement à l'interprétation et à l'utilisation des données d'évaluation dans la prise de décision touchant la gestion du personnel. Dans le même ordre d'idées, Abrami (2001) avance que la méfiance et les inquiétudes du corps professoral vis-à-vis l'emploi des données d'évaluation dans les décisions relatives à la promotion et à la titularisation peuvent être réglées ou atténuées en réformant les structures hiérarchiques des établissements et en assurant une transparence dans le processus de prise de décisions. Si les processus d'évaluation ne sont pas normalisés dans l'ensemble de l'établissement, de la division ou du département, il n'est pas possible de comparer les enseignants entre eux de façon valable (Hoyt et Pallett, 1999).

Utilisation des évaluations à des fins de comparaison

Pour améliorer le processus de prise de décision, Abrami (2001) recommande que les établissements déterminent à l'avance leur stratégie d'évaluation, suggérant soit l'évaluation normative ou critériée. En ce qui a trait à l'évaluation normative, chaque membre du corps professoral est comparé à un groupe approprié du corps professoral (p. ex., en fonction du type de cours ou de la discipline). Si les données d'évaluation doivent servir à des fins formatives, l'enseignant peut trouver utile de savoir où il se situe par rapport à son département ou de comprendre comment se comparent ses cotes par rapport à des sections de cours enseignées antérieurement. L'autre évaluation possible, soit l'évaluation critériée, établit une norme de rendement normalisée (avec ou sans mesure de triangulation de l'efficacité de l'enseignement) et compare le rendement de l'enseignant par rapport à cette norme (Abrami, 2001).

Les chercheurs divergent d'opinions quant à savoir si les évaluations de cours devraient faire l'objet de comparaisons. Bien que McKeachie (1997) croie en la validité et l'utilité des résultats des évaluations à des fins sommatives, il est davantage préoccupé par l'utilisation de ceux-ci à des fins de comparaison. Il soutient que les administrateurs emploient incorrectement les données d'évaluation pour établir des comparaisons en se servant des moyennes et des médianes. Il indique que :

La comparaison des évaluations de différentes classes pose problème non seulement parce que les étudiants diffèrent d'une classe à l'autre,

mais aussi parce qu'il y a des différences sur le plan des buts, des méthodes d'enseignement, du contenu et de bien d'autres variables. En outre, comme je l'ai proposé précédemment, il n'est pas nécessaire de procéder à une comparaison pour prendre des décisions touchant la gestion du personnel. Dans la mesure où l'on tient compte des évaluations des étudiants au moment de la prise de telles décisions, il est possible de classer de manière fiable les membres du corps professoral dans trois ou quatre catégories en jetant un regard rapide sur la distribution des résultats des évaluations par les étudiants : combien d'étudiants ont évalué l'enseignant comme étant très bon ou excellent? Combien d'étudiants ont été insatisfaits? (p. 1222)

Cette préoccupation est reprise par Algozine et coll. (2004) et Zabaleta (2007). McKeachie poursuit en suggérant néanmoins que de telles comparaisons sont possibles seulement lorsqu'on a recours à des évaluations élargies (voir la *Section 4.D.i : Communication des résultats d'évaluation*).

Cashin (1990) est en désaccord, soutenant que « sans données de comparaison, il n'est pas possible d'interpréter valablement les données des évaluations des étudiants » (p. 2). L'argument de Cashin favorable aux comparaisons laisse toutefois entendre que les comparaisons agissent souvent comme une procuration ou un substitut pour une analyse ou un examen soigneux de l'instrument d'évaluation ou des échelles d'évaluation. Si les établissements connaissaient davantage les questions posées dans les évaluations ou la façon dont les étudiants répondent aux évaluations, il ne serait peut-être pas nécessaire d'effectuer de telles comparaisons.

Pour les établissements qui comparent les résultats d'évaluation entre les enseignants, Abrami (2001) propose plusieurs moyens pouvant atténuer les erreurs statistiques et conceptuelles découlant des évaluations normatives, notamment l'ajout de marges d'erreur et la présentation visuelle des résultats d'évaluation (voir la *Section 4.D.i : Communication des résultats d'évaluation*).

Comparaison des cours

Plusieurs chercheurs soutiennent que les résultats d'évaluation représenteront mieux l'efficacité de l'enseignement d'un enseignant si l'on atténue les variations possibles causées par les caractéristiques relatives au cours en établissant une note moyenne dans les différents cours. Abrami (2001) suggère que chaque enseignant doit déterminer une série de cours assurant un équilibre entre les cours de niveau inférieur et supérieur, de même que les cours obligatoires et à option. Abrami fait valoir également que :

[...] puisque les décisions d'ordre sommatif s'appuient souvent sur un ensemble d'évaluations de cours, on devrait combiner la moyenne, la variance et la taille de l'échantillon pour un membre du corps professoral provenant de différents cours [...] Les résultats individuels des cours peuvent être plus utiles à des fins formatives, tandis que les résultats des cours combinés sont plus utiles à des fins sommatives (p. 72).

Franklin (2001) est d'avis que « les résultats dont on a fait la moyenne et qui proviennent de cours comparables suivis sur plusieurs sessions sont plus susceptibles d'être

beaucoup plus fiables pour les besoins de comparaison que les résultats provenant d'un seul cours » (p. 92). Franklin observe de plus que « le nombre de cours requis pour constituer des résultats "moyens" augmente à mesure que la taille de la classe diminue. En règle générale, on recommande cinq cours ou plus dans la majorité des cas, bien que les classes très peu nombreuses en requièrent davantage. Par exemple, les cours composés d'aussi peu que cinq étudiants peuvent nécessiter 20 sessions pour les besoins de comparaison » (p. 92).

Abrami indique toutefois qu'il faudrait élaborer des politiques claires indiquant les cours à inclure et à exclure des évaluations générales qui devraient être appliquées de façon équitable et cohérente.

Comparaison des enseignants

Selon la recherche, étant donné les différences de styles et de buts de l'enseignement selon les disciplines, les évaluations de l'enseignement (si on les compare) devraient être comparées seulement entre les enseignants d'une même discipline ou de disciplines similaires. De plus, toute comparaison effectuée entre les enseignants devrait fournir suffisamment d'occasions de mettre en contexte les données et devrait assurer que les cours comparés ont des caractéristiques similaires (voir la *Section 4.E.i : Survol des variables étudiées* pour obtenir une description des caractéristiques des cours pouvant influencer sur les résultats des évaluations).

Neumann (2001) met en lumière les différentes définitions de l'efficacité de l'enseignement dans différentes disciplines. Pour mener des évaluations génériques de manière différente auprès des différents départements, elle note le travail d'autres chercheurs proposant de « concevoir des instruments d'évaluation de l'enseignement spécifiques à la discipline » ou de « concevoir un certain nombre d'instruments reflétant les différentes philosophies d'enseignement convenant à la diversité des disciplines » (p. 143). Elle se réfère également à son travail précédent qui « souligne comment les résultats d'évaluation provenant d'instruments génériques peuvent être utilisés par les universités d'une manière qui tient compte des variations entre les disciplines » (p. 143).

Même au niveau d'un département, les variations qui existent entre les enseignants peuvent affecter la validité des données de comparaison (Addison, Best et Warrington, 2006). Par exemple, Theall et Franklin (2001) abordent un exemple particulier où un biais lié au sexe pourrait s'expliquer par le fait que la plupart des cours obligatoires de niveau inférieur étaient enseignés par des femmes dans le département. Ce qui a semblé être un biais lié au sexe dans les évaluations était en réalité un reflet du fait que des enseignants particuliers ont donné des cours selon des caractéristiques distinctives. Cet exemple renforce le besoin de veiller à ce que les données soient présentées à partir d'une sélection de cours qui est équitable ou représentative.

4.E Validité interne : influence des variables sur les résultats des évaluations

4.E.i Survol des variables étudiées

Les recherches ont porté une grande attention au large éventail de facteurs pouvant ou non avoir un impact sur la validité des données des évaluations réalisées par les étudiants (Zabaleta, 2007; Addison et coll., 2006; Algozzine et coll., 2004; Ory et Ryan, 2001; Theall et Franklin, 2001; Ali et Sell, 1998; Wachtel, 1998; Cashin, 1995, 1988). L'information de la présente section est principalement tirée de ces travaux de recherche (voir l'annexe E pour des exemples de résumés de recherche sur les variables pouvant introduire un biais dans ces travaux).

Les variables abordées dans la littérature sont classées en quatre catégories : les conditions administratives, les caractéristiques du cours, les caractéristiques de l'enseignant et les caractéristiques de l'étudiant. Le tableau ci-dessous énumère les facteurs particuliers associés à chacune de ces catégories.

Il est important de souligner que tout effet sur les évaluations générales causé par n'importe laquelle de ces variables particulières, même statistiquement significatives, est presque toujours très faible – faisant varier les évaluations de moins d'un dixième de 1 % dans les évaluations. C'est pour cette raison que, même si ces variables ont un effet sur les résultats des évaluations, la validité des résultats peut presque toujours être maintenue en rapportant les cotes à une seule décimale ou en les présentant à l'intérieur d'une grande catégorie, tel que décrit à la *Section 4.D.i : Communication des résultats d'évaluation*.

D'autres stratégies pour gérer les variables et veiller à ce qu'elles n'aient pas d'impact sur la validité générale sont présentées à la *Section 5.C : Assurance de l'utilité des évaluations*.

Tableau 2 : Variables examinées

Catégorie	Variables
Conditions administratives	Moment des évaluations
	Instructions aux étudiants
	Anonymat
	Présence de l'enseignant
	Objet
Caractéristiques/variables du cours (sur lesquelles l'enseignant ne peut agir)	Taille de la classe
	Moment de la journée
	Cours obligatoire/cours à option
	Charge de travail/difficulté
	Niveau du cours
	Discipline

Caractéristiques/variables de l'enseignant	Âge
	Productivité en recherche
	Race
	Personnalité/popularité
	Expressivité
	Rang et expérience
	Sexe
Caractéristiques/variables de l'étudiant	Âge
	Sexe
	Année d'étude
	Moyenne pondérée cumulative
	Personnalité
	Sexe
	Motivation
	Présence au cours
Notes	

Dans l'ensemble, la recherche sur ces variables est impressionnante et inégale sur les plans de la qualité et de l'étendue. Un certain nombre d'études exhaustives récentes dans la recherche fournissent un résumé juste des études précédentes, notamment celles d'Algozzine et coll. (2004), d'Ory et Ryan (2001), de Wachtel (1998), d'Ali et Sell (1998) et de Marsh et Roche (1997). En règle générale, aucune variable n'a révélé avoir un effet important sur les évaluations (p. ex., un élément qui fausserait les évaluations au-delà de la deuxième décimale), à l'exception des notes attendues. Des études (cf William et Ceci, 1997 et l'étude « D' Fox ») ont déterminé des facteurs qui semblaient révéler des biais (p. ex., aptitudes à la présentation, enthousiasme ou personnalité de l'enseignant); cependant, ces études ont été en grande partie écartées pour des raisons méthodologiques ou bien parce que ces facteurs peuvent en fait servir de mesure de l'amélioration de l'enseignement. Dans leur l'examen de la longue liste de variables qui se sont révélées influentes, à différents degrés, sur les évaluations des étudiants, Algozzine et coll. (2004) ont fait valoir que ces variables ne peuvent être perçues comme sources de biais à moins qu'elles modifient les évaluations sans qu'il n'y ait de mesure des écarts dans l'efficacité de l'enseignement. De la même manière, d'Apollonia et Abrami (1997) concluent que même si les caractéristiques liées à l'administration, au cours et à l'enseignement peuvent avoir une incidence sur les évaluations, elles ne correspondent pas à la définition de biais et, par conséquent, ne traduisent pas une invalidité des résultats. Cashin (1988) et Marsh (1984) ont également soutenu que les seules variables pouvant introduire un biais sont celles qui « ne représentent pas une fonction de l'efficacité de l'enseignement de l'enseignant » (Cashin, 1988, p. 3) – par exemple, la taille de la classe. Cashin poursuit en faisant remarquer que ces variables « peuvent avoir une incidence sur l'efficacité de l'enseignement, mais les enseignants ne devraient pas être considérés comme fautifs si leur enseignement est moins efficace dans de grandes classes d'étudiants non motivés comparativement à leurs collègues

dans de petites classes avec des étudiants motivés » et que ce type de facteurs devrait « être pris en compte en recourant à des données de comparaison appropriées » (p. 3).

Comme nous l'avons mentionné précédemment, le large éventail de variables présentées dans la littérature a été examiné et réexaminé de façon exhaustive. L'étendue de la présente étude ne nous permet pas de faire une revue approfondie de toutes les variables étudiées; plutôt, ce qui suit met en lumière certaines de ces variables qui sont soit plus litigieuses et font l'objet de vifs débats dans la littérature ou qui ont mené à des résultats présentant un intérêt particulier.

Conditions administratives

Moment des évaluations : En règle générale, il a été démontré que le moment où les évaluations sont tenues n'a pas d'impact significatif sur les résultats obtenus (Wachtel, 1998). Il existe une certaine preuve montrant que lorsque les évaluations sont réalisées au cours de la période des examens finaux, les résultats sont plus faibles (Ory, 2001); par conséquent, la majorité des chercheurs recommandent que les évaluations aient lieu avant les examens finaux et la remise des notes finales (d'Apollonia et Abrami, 1997).

Explication des objectifs d'évaluation et communication des instructions aux étudiants : Expliquer le but des évaluations (p. ex., en soulignant que celles-ci servent à des décisions qui touchent la gestion du personnel) peut avoir un effet positif sur les résultats (Algozzine et coll., 2004; Cashin, 1995); cependant, les résultats liés à cette variable ont été irréguliers. Cashin (1995) suggère de remédier à cette situation en utilisant des instructions normalisées. Fox (2006) a souligné que ces évaluations peuvent être améliorées si les enseignants demandent une rétroaction plus critique de la part de leurs étudiants.

Anonymat : Les étudiants semblent se préoccuper davantage des répercussions possibles sur leur dossier universitaire lorsqu'on leur demande de signer les questionnaires d'évaluation; par conséquent, les cotes des évaluations signées tendaient à être plus élevées. Ainsi, la majorité des chercheurs recommandent que les évaluations demeurent anonymes (Cashin, 1995), tandis que certains ont suggéré qu'elles soient plutôt confidentielles (l'établissement d'enseignement, et non l'enseignant, serait en mesure d'identifier la personne qui a réalisé l'évaluation), et ce, dans le but d'encourager les étudiants à remettre des évaluations responsables, de s'assurer qu'ils le fassent, de même que de permettre un suivi ultérieur (Wright, 2006).

Présence de l'enseignant pendant le processus d'évaluation : Les résultats semblent être supérieurs lorsque l'enseignant est présent pendant les évaluations; néanmoins, on peut remédier à la situation en veillant à ce que l'enseignant quitte la salle au moment où les étudiants remplissent le questionnaire (Cashin, 1995). Algozzine et coll. (2004) observent que la présence de l'enseignant n'a pas d'impact significatif sur la validité, sauf si cette pratique est combinée à des évaluations non anonymes.

Caractéristiques du cours

Taille de la classe : Bien qu'il existe des études montrant que les classes de plus petite taille produisent des résultats légèrement plus élevés aux évaluations (Algozzine et coll., 2004; Williams et Ory, 1992; Centra et Creech, 1976), la corrélation entre la taille de la classe et les résultats n'est pas statistiquement significative et n'est donc pas considérée comme ayant une incidence sur la validité (Marsh et Roche, 1997; d'Apollonia et Abrami, 1997; Aleamoni, 1997; McKeachie, 1997; Cashin, 1995, 1988; Marsh, 1987). McKeachie (1997) indique qu'il existe des éléments probants suggérant que les membres du corps professoral dispensent un meilleur enseignement dans des classes de plus petite taille, ce qui se solderait par un signe, non un biais, voulant que les résultats soient le reflet précis de l'efficacité de l'enseignement. Cependant, étant donné que les enseignants ne peuvent agir énormément sur la taille de la classe, il faut prendre soin de tenir compte de celle-ci dans les rapports sur les données d'évaluation ou de s'assurer que les enseignants dont les résultats seront comparés aient aussi des classes de taille comparable en moyenne.

Cours obligatoire/cours à option : Les étudiants évaluent souvent les cours à option un peu plus positivement que les cours obligatoires; cependant, aucun impact significatif n'a été observé sur les résultats (Algozzine et coll., 2004; Cashin, 1995, 1988). Le statut d'un cours, qu'il soit obligatoire ou à option, devrait être géré de la même manière que la taille de la classe dans les rapports sommatifs présentant les résultats des évaluations.

Charge de travail/difficulté du cours : Bien que de nombreux membres du corps professoral croient que les cours plus difficiles ou présentant une charge de travail plus élevée entraînent des résultats plus faibles aux évaluations, cette croyance n'a cependant pas été corroborée par les travaux de recherche, lesquels ont montré des résultats irréguliers (Marsh, 1987). Les cours dits « faciles » ne sont pas garants de résultats plus élevés aux évaluations. De plus, des études ont indiqué que les cours dits difficiles ou dont la charge de travail est plus élevée sont évalués plus favorablement (Cashin, 1988).

Niveau du cours : Les conclusions de recherches indiquent que le niveau du cours peut avoir une incidence sur les résultats (Algozzine et coll., 2004; Marsh, 1997; Cashin, 1988, 1995; Aleamoni et Hexner, 1980) et montrent certaines évidences que les cours de niveau plus élevé peuvent obtenir des résultats plus élevés. Encore une fois, cette information doit être mise en contexte lors de la production des rapports de données d'évaluation.

Discipline : Certaines études ont montré que certaines disciplines obtiennent de meilleurs résultats (les meilleurs étant obtenus par les lettres et sciences humaines, suivies des sciences sociales, puis des sciences naturelles (Johnson, 2003; Neumann, 2001; Ory, 2001; Wachtel, 1998; Cashin, 1990). Cette situation est le reflet des différences en lien avec la discipline qui existent dans les styles et les objectifs d'enseignement plutôt que d'une source de biais. Neumann (2001) et Cashin (1995, 1988) indiquent qu'il faut faire preuve de prudence, car les comparaisons entre les disciplines peuvent ne pas être exactes.

Caractéristiques de l'enseignant

Personnalité/popularité : La revue de la littérature par Ali et Sell (1998) sur la popularité et la personnalité d'un enseignant a abouti à la conclusion qu'il y a un accord commun disant que ces facteurs ont un impact non significatif sur les résultats des évaluations. Deux études, une publiée par Naftulin, Ware et Donnelly (1973), connue également sous le nom de l'étude « D^r Fox », et une autre par Williams et Ceci (1997) ont conclu que l'enthousiasme de l'enseignement pouvait avoir une incidence sur les évaluations. Cependant, ces deux conclusions ont été largement réfutées par la majorité des chercheurs des dernières années pour des raisons liées à la méthodologie (Abrami, 2001; Kulik, 2001; Marsh et Dunkin, 1992). Abrami (2001) ainsi que Theall et Franklin (2001) ont soutenu qu'il n'existe pas de recherches pour corroborer cette allégation voulant que la popularité ou la personnalité de l'enseignant amène des résultats plus élevés, et Ory (2001) fait valoir que la « personnalité » peut effectivement mesurer les comportements d'enseignement, comme l'enthousiasme, pouvant en fait influencer sur l'efficacité de l'enseignement.

Expressivité : La recherche sur l'expressivité de l'enseignant, comme celle entourant la personnalité et la popularité, est complexe et parfois illogique. Certaines études ont établi clairement que l'expressivité est susceptible d'améliorer l'apprentissage et, par conséquent, ne peut être prise comme étant un facteur de biais (Cashin, 1995).

Rang et expérience : d'Apollonia et Abrami (1997) puis Arreola (2000) indiquent que ces variables n'affectent pas de façon significative les résultats des évaluations. Marsh (2001) a souligné que l'expérience ne mène pas à de meilleures évaluations et peut en fait avoir une relation négative vis-à-vis de l'efficacité de l'enseignement.

Sexe : En règle générale, les études relatives au sexe des personnes ont mené à des résultats non concluants, mais la majorité d'entre elles ont montré que cette variable a peu ou pas du tout d'incidence sur les évaluations (Algozzine et coll., 2004; Theall et Franklin, 2001; Marsh et Roche, 1997; Cashin, 1995; Arreola, 2000; Aleamoni et Hexner, 1980).

Caractéristiques de l'étudiant

Sexe : Il y a certains éléments de preuve montrant que les étudiants sont enclins à noter légèrement à la hausse les enseignants du même sexe qu'eux (Ory, 2001). Cette situation est significative uniquement dans les disciplines où il y a un déséquilibre important entre le nombre d'hommes et de femmes, auquel cas ce facteur peut être mis en contexte dans la présentation des données.

Motivation : La motivation de l'étudiant ou son intérêt *a priori* pour le cours peut avoir une incidence sur les évaluations, donnant lieu à des résultats supérieurs (Cashin, 1988/95). L'Université du Michigan se sert d'une question sur la motivation pour mettre en contexte les données d'évaluation.

Présence au cours : Une étude récente auprès de plus de 9 000 étudiants d'un collège en Israël a révélé qu'il existe un lien positif entre un taux de présence élevé et une évaluation positive du cours. En général, cette situation n'était pas perçue comme étant une variable introduisant un biais, car une plus grande présence au cours améliore l'apprentissage (Davidovitch et Soen, 2006). Notons qu'il s'agit de la seule étude exhaustive que nous avons repérée sur cette question.

Notes : attentes, inflation et indulgence – Mythe ou réalité?

La variable la plus controversée examinée dans les travaux de recherches est sans doute la relation entre les notes et les évaluations. Les attentes des étudiants quant à leur note finale ont-elles un impact sur leur évaluation de l'efficacité de l'enseignement d'un enseignant? Une étude récente réalisée par Baldwin et Blattner (2003) a révélé que 40 % du corps professoral croit que c'est le cas. Cette question a fait l'objet d'un examen soigneux dans la recherche et demeure l'objet de nombreux débats. Aleamoni (1999) a repéré 37 études ayant révélé des corrélations entre les notes attendues/reçues et une évaluation positive et 24 études n'ayant trouvé aucun lien significatif. (Pour voir une revue récente de la littérature sur l'hypothèse d'une relation entre les notes et l'indulgence, consulter l'étude de Gump, 2007).

Quelques études ont montré une relation entre les évaluations positives et les notes. Cette corrélation a été interprétée par certains comme une indication claire que l'indulgence dans la notation peut aboutir à de meilleures évaluations (Wachtel, 1998). Greenwald et Gillmore (1997) ont fait valoir qu'étant donné que les attentes des étudiants à l'égard de leurs notes finales ont une incidence sur l'évaluation de l'enseignant, les évaluations devraient faire l'objet d'un ajustement statistique pour corriger ce facteur. Abrami (2001) et d'autres ont réfuté cette allégation, faisant valoir que l'impact n'était pas important. Abrami a soutenu que ni l'indulgence ni la sévérité dans les pratiques de notation n'ont une incidence sur les évaluations de cours sur le plan statistique. De la même manière, Marsh (1987) ainsi que Marsh et Roche (1997) ont fait valoir que bien que les attentes vis-à-vis des notes puissent révéler un niveau de biais, l'incidence sur les évaluations est faible et relativement peu substantielle.

McKeachie (1997) et d'autres se montrent préoccupés par les conclusions de Greenwald et Gillmore suivant leur étude de 1997 sur l'indulgence dans la notation, proposant que les arguments utilisés sont faibles. Le réexamen des ensembles de données de Greenwald et Gillmore par Marsh et Roche (2000) a permis d'observer que les cours et les enseignants présentant des charges de travail plus élevées recevaient des évaluations plus élevées.

Heckert et coll. (2006) ont examiné certaines études sur la relation entre les notes et les évaluations, soulignant les opinions opposées dans la littérature. Ils ont procédé à une étude particulière pour vérifier l'hypothèse de l'indulgence dans la notation, laquelle a été menée auprès de 463 étudiants et a examiné l'incidence de deux variables : la difficulté du cours et l'effort de l'étudiant. Heckert et coll. ont constaté que les évaluations les plus fortes avaient été données aux cours dont le degré de difficulté correspondait à ce à quoi s'attendaient les étudiants. De plus, les évaluations étaient positives lorsque les étudiants indiquaient qu'ils avaient fourni davantage d'efforts qu'ils ne l'avaient prévu. Dans l'ensemble, cette étude a conclu que les enseignants les plus exigeants recevaient des

évaluations supérieures; par conséquent, elle a réfuté l'hypothèse de l'indulgence dans la notation et la notion voulant que le corps professoral puisse « acheter » de meilleures évaluations par une notation plus généreuse.

Wachtel (1998) et d'autres (Marsh et Dunkin, 1992; Murray, 1987) ont proposé qu'une corrélation positive entre les notes attendues et les évaluations de l'enseignant peut simplement faire foi de l'apprentissage de l'étudiant : les étudiants s'attendent à des notes supérieures et évaluent l'enseignant plus favorablement lorsqu'ils ont profité d'une expérience positive en classe. En outre, Chambers et Schmitt (2002) postulent un modèle de processus comparatif pour expliquer la relation entre les attentes vis-à-vis des notes et les évaluations. Dans cette théorie, les étudiants basent leurs attentes quant aux notes sur l'expérience acquise dans les autres cours (charge de travail, effort et note finale). Si leur comparaison est positive, ils font une évaluation positive; cependant, si elle est négative, les étudiants le refléteront dans leur évaluation de l'enseignant. Addison et coll. (2006) réfutent cette hypothèse en montrant les résultats de leur étude à petite échelle, laquelle conclut que les attentes vis-à-vis des notes sont également tributaires des idées préconçues sur le niveau de difficulté que présentera le cours, c'est-à-dire s'il sera difficile ou facile. Leur sondage auprès des étudiants a indiqué que ceux ayant trouvé le cours plus difficile qu'ils ne s'y attendaient au départ ont évalué moins favorablement le cours, tandis que ceux l'ayant trouvé plus facile que ce à quoi ils s'attendaient au départ l'ont évalué plus favorablement. Addison et coll. sont arrivés à la conclusion que l'incidence du degré de difficulté perçu n'était pas liée aux notes obtenues par l'étudiant dans un cours, ce qui montre que les pratiques en matière de notation ont une incidence limitée sur les résultats d'évaluation.

En examinant la recherche qui se concentre sur l'étude de l'hypothèse de l'indulgence dans la notation, Gump (2007) met en doute la généralisabilité des résultats de ces études, souvent contradictoires. Il relève plus particulièrement les questions relatives à la méthodologie de l'étude, l'applicabilité des résultats au-delà d'un établissement en particulier (c.-à-d. la capacité de reproduire les résultats) et les différences dans les définitions et les utilisations des principaux termes utilisés dans la recherche (p. ex., biais, charge de travail).

4.E.ii Évaluation de la validité

Ory et Ryan (2001) présentent cinq méthodes de base qui ont été utilisées pour déterminer la validité des évaluations de cours : l'étude multisections, l'étude multitraits-multiméthodes, l'analyse des biais, l'étude de laboratoire et l'étude multidimensionnelle. Ory et Ryan soutiennent que seules les trois premières méthodes ont contribué à la compréhension de la validité des évaluations de cours. Ils ont écarté les études de laboratoire comme méthode d'évaluation appropriée de la validité en raison de l'environnement artificiel dans lequel elles sont menées. Ils indiquent également que les évaluations de dimensions – études qui tentent de « déterminer un ensemble de facteurs "communs" sous-jacents au construct mesuré dans les évaluations de l'enseignement par les étudiants » – n'ont pas permis de « déterminer un simple ensemble de dimensions et confortent tout simplement la notion selon laquelle les étudiants ont une vue multidimensionnelle de la qualité de l'enseignement » (p. 31) (tel qu'abordé à la *Section 4.C.i : Définition d'un enseignement efficace*).

Fait important, Ory et Ryan (2001) constatent que la majorité des outils utilisés pour évaluer la validité des cotes attribuées par les étudiants se sont concentrés sur le degré avec lequel les évaluations correspondent aux autres méthodes d'évaluation de l'efficacité de l'enseignement et sur la détermination des influences externes exercées sur les évaluations. Cependant, les études ne tendent pas à évaluer les façons dont les cotes sont interprétées et mises à la disposition des étudiants, du corps professoral et des administrateurs. Ces éléments peuvent influencer significativement sur la validité des instruments d'évaluation même si les questions elles-mêmes ont été soigneusement mises à l'épreuve comme l'explique la *Section 4.C.ii : Mise au point des instruments d'évaluation*.

Études multisections sur la validité

Comme le décrivent Greenwald (1997) ainsi qu'Ory et Ryan (2001), une méthode d'évaluation couramment utilisée pour évaluer la validité des évaluations est l'étude multisections. Ce type d'études compare le rendement scolaire des étudiants par rapport à différentes sections d'un même cours et aux résultats d'évaluation. En règle général, Ory et Ryan soulignent que « les études multisections sur la validité des évaluations montrent des corrélations importantes avec le rendement scolaire des étudiants comme le démontre l'examen du rendement » (p. 30). Cependant, ces études ont été critiquées pour deux raisons. En premier lieu, parce qu'elles doivent évaluer les cours comportant de multiples sections, elles évaluent en général seulement les cours de niveau inférieur. Ory et Ryan (2001) ont fait valoir que les objectifs d'apprentissage de ces cours sont différents de ceux des cours de niveau supérieur et que les conclusions tirées de ces études ne peuvent être généralisées aux évaluations des cours de niveau supérieur. La deuxième raison est que l'évaluation réalisée dans ces cours donnés à de grands groupes et adaptée à ces études se limite souvent à des tests à choix multiples qui peuvent ne pas mesurer un large éventail d'objectifs d'apprentissage, ce qui suggère encore une fois que la généralisabilité de ces études peut être limitée.

Étude multitraits-multiméthodes

Ces études comparent des évaluations d'étudiants aux autres moyens d'évaluation de l'enseignement, notamment les sondages auprès des anciens étudiants, les évaluations par les collègues et les autoévaluations. Elles peuvent également utiliser différents moyens d'évaluation (p. ex., l'analyse du contenu d'un dossier d'enseignement ou l'évaluation par des groupes de discussion). Ory et Ryan (2001) soutiennent que ces études ont montré en général une corrélation importante entre les évaluations obtenues par ces moyens d'évaluation multiples.

Étude par analyse des biais

Ces études utilisent une analyse des facteurs pour déterminer les influences externes ou environnementales sur les cotes attribuées par les étudiants. Ory et Ryan (2001) indiquent que « de nombreuses études ont été menées dans le but de déterminer les relations (ou l'absence de relation) entre les cotes et un grand éventail d'influences possibles », mais que « les comptes rendus de recherche révèlent qu'il y a peu d'influences, s'il en est, introduisant un biais possible dans le processus d'évaluation » (p. 31). Ils constatent également que les résultats de ces études ne sont pas toujours valides ni concluants.

Méthodes de validation des réponses des étudiants

La précision des cotes des étudiants est généralement évaluée en comparant les cotes des étudiants avec d'autres mesures de l'efficacité de l'enseignement, plus particulièrement au moyen des résultats scolaires (Ory et Ryan, 2001; Theall et Franklin, 2001; Wachtel 1998). De tels travaux mettent souvent en corrélation les notes finales (en tant qu'indicateur d'apprentissage de l'étudiant) avec les résultats des évaluations. Certains chercheurs ont fait valoir que des évaluations aux résultats élevés sont un indicateur de l'apprentissage de l'étudiant; cependant, comme il est expliqué à la *Section 4.E.i : Survol des variables étudiées*, d'autres ont avancé que des résultats élevés peuvent être le résultat d'autres facteurs (p. ex., indulgence dans la notation de la part de l'enseignant). Ce modèle d'évaluation de la précision des évaluations des étudiants a été critiqué par un petit nombre de chercheurs, car ces études se sont seulement concentrées, pour des raisons de méthodologie, sur des cours de niveau inférieur qui reposent sur des évaluations uniformisées (Ory et Ryan, 2001). Abrami (2001) a réalisé un examen de la recherche et conclut qu'il existe une preuve empirique solide démontrant que les données d'évaluation de cours peuvent être et sont un indicateur de l'apprentissage.

Le corps professoral exprime souvent son inquiétude quant à la possibilité que les étudiants soient facilement manipulés en vue de faire des évaluations plus élevées par le biais de l'inflation des notes ou d'enseignants particulièrement charismatiques (Theall et Franklin, 2001). Un grand nombre d'études sur cette question ont été réalisées et sont expliquées à la *Section 4.E.i : Survol des variables étudiées*. En règle générale, cette recherche montre que même si les résultats des évaluations peuvent sembler démontrer que les étudiants se montrent généreux envers les enseignants indulgents et sympathiques, la relation réelle entre ces facteurs et les résultats d'évaluation est beaucoup plus complexe et qu'en général l'exactitude des évaluations des étudiants est maintenue.

Les réponses des étudiants sont également vérifiées en les comparant aux résultats d'autres évaluateurs : Ory et Ryan (2001) décrivent comment « la recherche a relevé de grandes corrélations positives entre les évaluations des étudiants et celles des anciens étudiants [...] et ont trouvé des corrélations moyennement positives entre les évaluations générales que font les étudiants et les autoévaluations [...] et les évaluations par les pairs » (p. 36). De la même manière, Murray (1987) a constaté que les évaluations de l'enseignement réalisées par les étudiants sont comparables à celles des observateurs qualifiés. Arreola (2000), Aleamoni (1987) et d'autres ont montré que les évaluations des étudiants sont cohérentes, stables et sont en corrélation avec les évaluations par les collègues et les observations par les pairs.

Selon Nasser et Fresko (2002), Ory (2001) de même que Remedios et Lieberman (2008), ce qui inquiète couramment le corps professoral en ce qui a trait à la validité des évaluations par les étudiants, c'est la capacité des étudiants à évaluer avec précision la valeur d'une expérience en enseignement avant d'avoir pu utiliser leurs connaissances acquises dans le cours dans d'autres cours ou dans leur carrière. Cependant, plusieurs études ayant comparé les évaluations d'anciens étudiants avec celles d'étudiants en formation montrent que l'évaluation qu'un étudiant fait d'un cours ne change pas considérablement au fil du temps (Ory et Ryan, 2001; Theall et Franklin, 2001).

Validité du contexte

Theall et Franklin (2000) ont introduit un groupe de variables liées au contexte susceptibles d'influer sur la validité, mais qui n'avaient pas encore été explorées à fond dans la littérature jusqu'à maintenant. Ces variables comprennent les pratiques d'enseignement variables, les populations étudiantes variables, les besoins variables du corps professoral, les priorités variables de l'établissement, les technologies changeantes et les exigences relatives aux données de même que la formation professorale en évolution et les pratiques d'évaluation variables. La plupart des formulaires d'évaluation ont été créés lorsque l'enseignement magistral constituait la norme. Toutefois, au cours des dernières années, les pratiques d'enseignement ont changé afin d'inclure des techniques d'apprentissage coopératif, d'apprentissage par l'action et par résolution de problèmes et utilisent davantage la technologie éducative. Les instruments d'évaluation existants peuvent ne plus être en mesure d'évaluer avec précision ou pertinence ces nouvelles situations d'enseignement et d'apprentissage (voir la *Section 3.C.ii : Format et contenu du questionnaire*). Certains établissements ont peut-être déjà abordé cette difficulté au moyen de formulaires personnalisables permettant au corps professoral de choisir les questions pertinentes tandis que d'autres ont réagi en procédant à un examen approfondi de leurs instruments d'évaluation.

Une grande partie des travaux de recherche ont été menés sur une population étudiante qui ne représente plus les étudiants bacheliers actuels. L'évolution sociodémographique (âge, appartenance ethnique et statut socio-économique) peut avoir un impact sur les attitudes des étudiants à l'égard de l'efficacité de l'enseignement et, par conséquent, de leurs évaluations des enseignants et des cours (Theall et Franklin, 2000). Compte tenu de ces changements contextuels et du fait que la grande partie de la recherche réalisée à ce jour n'a pas tenu compte correctement de ces changements, Theall et Franklin (2000) soulèvent des préoccupations quant aux généralisations des systèmes d'évaluation de cours s'appuyant sur les recherches actuelles.

Sources susceptibles d'établir la validité des évaluations de cours par les étudiants :

- La corrélation positive et statistiquement significative entre les évaluations et l'apprentissage des étudiants;
- La position privilégiée et les compétences de l'étudiant pour évaluer sa propre compréhension et acquisition de connaissances;
- La position privilégiée des étudiants pour évaluer le changement dans leur degré de motivation face à la matière enseignée et le changement dans leur attitude générale envers leur apprentissage ultérieur de la discipline;
- La position privilégiée de l'étudiant pour évaluer des éléments observables se rapportant à l'enseignement compétent (p. ex. : la ponctualité de l'enseignant);
- La position privilégiée de l'étudiant pour déterminer la présence régulière des indicateurs de style d'enseignement (p. ex. : être enthousiaste, encourager les étudiants);
- Les étudiants sont dans la meilleure position pour juger si les épreuves couvrent la matière du cours.

[adaptation de Scriven, M. (1995). Student ratings offer useful input to teacher evaluations. *Practical Assessment, Research & Evaluation*, 4(7), 4-5.]

Section 5

Mise en œuvre de mesures d'évaluation efficaces : recommandations tirées de la recherche

5.A Introduction

Comme l'aborde la *Section 4 : Fiabilité, validité et interprétation des données provenant des évaluations de cours*, les politiques et les pratiques entourant la création, l'administration et l'interprétation des évaluations forment un volet important de la validité des évaluations. Les recommandations présentées ci-dessous exposent en détail les actions et les politiques qu'un établissement, une division ou un département peut mettre en œuvre pour assurer la validité et l'utilité des évaluations. Les recommandations sont un complément et sont tirées de plusieurs articles présentant aux établissements un ensemble de recommandations pour la mise en œuvre des évaluations valables et équitables des cours. Les auteurs comprennent Moore et Kuol (2005), Franklin (2001), Ory et Ryan (2001), Theall et Franklin (2001) de même que Cashin (1990).

L'annexe F contient un certain nombre de lignes directrices en matière de bonnes pratiques d'évaluation découlant de la recherche actuelle.

5.B Assurance de la validité des évaluations

La recherche sur les évaluations des étudiants a permis de faire plusieurs recommandations pour veiller à ce que les évaluations de cours puissent générer des données valides à des fins d'évaluation sommative et formative de l'efficacité de l'enseignement :

Établir des objectifs d'évaluation clairs, notamment déterminer clairement en quoi consiste un enseignement efficace au sein de l'établissement et veiller à ce que les questions posées reflètent les objectifs établis.

La *Section 4.C.i : Définition d'un enseignement efficace* explique l'importance de veiller à ce que les questions d'évaluation concordent avec les priorités de l'enseignement de l'établissement et fournissent l'information pertinente permettant de réaliser le type d'évaluations sommatives aux fins desquelles l'instrument est utilisé. L'exercice d'identification des mesures d'enseignement à évaluer de même que d'élaboration des questions doit être perçu comme une occasion favorisant les échanges à la grandeur de l'établissement sur les objectifs de l'enseignement et les façons d'utiliser les évaluations. Pour que les questions fournissent une rétroaction significative aux enseignants et qu'elles puissent être utilisées pour l'évaluation

sommative de l'enseignement, les questions retenues devraient pouvoir mesurer les aspects de l'enseignement qui tiennent compte des conclusions tirées.

Concevoir et mettre à l'épreuve des instruments d'évaluation en fonction de normes théoriques et psychométriques rigoureuses

L'élaboration d'instruments servant à l'évaluation devrait faire l'objet d'un processus sérieux et substantiel comptant sur la participation de nombreux membres de l'établissement. Le choix des questions devrait se faire soigneusement en fonction de constructs théoriques bien élaborés et basés sur la recherche. Les échelles d'appréciation doivent être logiques et clairement expliquées. Les instruments devraient également être approuvés par une instance dirigeante ou un comité approprié au moyen d'un processus de consultation transparent. Les instruments approuvés devraient être évalués par des experts en réalisation de sondages et continuellement faire l'objet d'un examen par le biais de recherches menées par l'établissement (voir la *Section 5.C.iv : Assurance de l'utilité des évaluations pour les établissements*). Dans le cas où un établissement ne peut consacrer le temps et les fonds nécessaires à l'élaboration d'un instrument interne rigoureux, il pourrait alors prendre en considération le recours à une licence d'un instrument validé provenant d'un autre établissement.

Établir des politiques et processus pertinents et normalisés pour l'administration des évaluations de cours

Des politiques et des processus clairs et cohérents doivent être élaborés pour ne pas mettre en péril les évaluations recueillies. Pour ce faire, on doit s'assurer que :

Les pratiques et les politiques relatives à l'administration des évaluations sont normalisées sur le plan administratif à partir duquel se fait la comparaison entre les enseignants ou entre les cours (le cas échéant)

Un bon nombre de menaces à la validité découlent de l'administration incohérente des évaluations. Pensons notamment à la présence de l'enseignant pendant les évaluations, à l'utilisation de formulaires incohérents ou à la passation des évaluations au moyen de méthodes différentes, par exemple l'utilisation concomitante de formulaires en ligne et en format papier, et ainsi de suite. En veillant à ce que les politiques d'administration et de communication des évaluations soient équitables et appliquées de façon uniforme, les établissements peuvent faire de grands pas vers l'amélioration de la validité des évaluations.

Chaque cours atteint un taux de réponse approprié

Cashin (1990) recommande de recueillir les rétroactions d'au moins 10 étudiants ou d'au moins les deux tiers de la classe, selon le nombre le plus élevé. Comme l'explique la *Section 4.D.i : Communication des résultats d'évaluation*, pour faire en sorte que les résultats d'évaluation soient davantage représentatifs, plusieurs chercheurs proposent de faire la moyenne de quelques-unes ou de la totalité des évaluations d'un enseignant pour voir à ce que les réponses recueillies représentent fidèlement son enseignement.

L'anonymat des réponses des étudiants est protégé

Il y a peu de données montrant que les réponses anonymes sont plus précises ou valides que les réponses non anonymes ou bien si elles le sont moins.

Wright (2006) a fait valoir que les évaluations anonymes soustraient les étudiants de leur responsabilité envers leurs commentaires et opinions et que « sans possibilité de suivi, les étudiants ne réfléchissent pas nécessairement à leurs opinions » (p. 419). Wright souligne que le but des évaluations anonymes est de faire en sorte que les étudiants ne reçoivent pas de blâme de la part du corps professoral en raison de commentaires négatifs. Cependant, il fait valoir que même si les intentions sous-jacentes visant à protéger l'anonymat peuvent être positives, un tel système fait davantage confiance aux étudiants qu'au corps professoral. Il soulève également certaines préoccupations quant à la possibilité que les étudiants utilisent les évaluations pour extérioriser leur colère ou leur déception découlant de faibles notes (il est à remarquer que Wright ne réfère à aucune étude en particulier pour corroborer cette théorie).

Wright fait également ressortir la possibilité que l'anonymat incite à des abus des instruments d'évaluation et du processus d'administration, émettant l'hypothèque selon laquelle « des étudiants pourraient entrer dans la salle et remplir l'évaluation sans même qu'ils soient inscrits au cours » (p. 419). Pour remédier à cette situation, Wright recommande que les évaluations demeurent confidentielles et que les noms soient retirés des données avant de les présenter au corps professoral, ce qui permet ainsi à l'administration de faire un suivi auprès des élèves (p. ex., en vue d'approfondir une évaluation extrêmement élevée ou faible ou pour déterminer les variables qui concourent à l'attribution d'évaluations élevées ou faibles).

La recherche montre néanmoins que les étudiants peuvent ne pas se sentir à l'aise de fournir des renseignements non anonymes et que ces réponses non anonymes produisent des évaluations un peu plus élevées (Wachtel, 1998). Par conséquent, les politiques visant à protéger l'anonymat devraient être appliquées de manière cohérente et uniforme, car il y a beaucoup à perdre à mettre en péril la confiance, déjà faible, des étudiants vis-à-vis du système d'évaluation. Les pratiques doivent également veiller à ce que les étudiants comprennent que leur anonymat sera protégé et qu'ils connaissent la manière dont il sera protégé.

Une quantité appropriée de données sont distribuées aux personnes appropriées et des politiques cohérentes et pertinentes sont élaborées pour conserver ces données ou pour y accéder.

Les étudiants, la faculté et les enseignants ont besoin et tirent avantage des différentes données découlant des évaluations de cours. Wachtel (1998) soutient que les étudiants sont en droit d'avoir accès aux résultats de leur appréciation sous la forme de résultats d'évaluation distribués publiquement. Un bon nombre d'établissements qui font connaître publiquement les résultats des évaluations (voir la *Section 3.F.iii : Publication des résultats*) choisissent de faire ressortir un petit nombre de questions générales qu'ils diffusent aux étudiants pour les aider à choisir leurs cours. Un certain nombre d'établissements publient les résultats d'évaluation dans le principal but de fournir de l'information aux étudiants afin de

les aider dans le processus de choix de cours. Voir la *Section 5.C.i : Assurance de l'utilité des évaluations pour les étudiants* pour obtenir d'autres recommandations sur la communication des résultats aux étudiants.

Les administrateurs devraient recevoir les données individuelles et comparatives appropriées qui concordent avec l'utilisation qu'ils feront des données d'évaluation. Les administrateurs qui ne fournissent pas de diagnostics ni de rétroactions formatives pourraient recevoir seulement les données provenant des questions générales aux fins sommatives (voir la *Section 4.D.ii : Difficultés relatives à l'interprétation et à l'utilisation des données à des fins sommatives*). Les enseignants pourraient recevoir les autres résultats servant à des fins formatives. Les données que reçoivent les administrateurs devraient correspondre à leur habileté à effectuer des analyses avec les données et les statistiques. Les résultats d'évaluation devraient être accompagnés de renseignements complémentaires nécessaires afin de pouvoir mettre en contexte les données de manière appropriée (p. ex., des guides explicatifs, des moyennes comparatives ou des notes circonstanciées de la part du corps professoral; voir la *Section 5.C.iii : Assurance de l'utilité des évaluations pour les administrateurs*).

Les membres du corps professoral devraient pouvoir accéder aux données d'évaluation de cours recueillies au sujet de leur enseignement, y compris les commentaires écrits par les étudiants dont l'anonymat a été préservé. Les enseignants devraient également recevoir les rapports sommaires des données (voir la *Section 4.D.i : Communication des résultats d'évaluation*), lesquels aident à mettre en contexte les données qui leur sont fournies.

Les établissements devraient conserver les évaluations de l'enseignement dans des dossiers centraux (voir la *Section 3.F.i Qui a accès à quoi?* pour consulter des exemples sur la façon dont les données d'évaluation sont maintenues dans différents établissements). Les formulaires originaux devraient être conservés pendant une période de temps limitée mais suffisamment longue pour pouvoir vérifier tout résultat contesté. Les données traitées devraient être gardées confidentielles par les départements, les divisions ou dans une base de données centralisée aussi longtemps qu'elles peuvent être utilisées par les enseignants et les établissements.

5.C Assurance de l'utilité des évaluations

5.C.i Pour les étudiants

Pour que les évaluations soient précises, les étudiants doivent recevoir suffisamment d'information pour être en mesure de fournir correctement des réponses qui sont utiles et appropriées. Par conséquent, les politiques et les pratiques relatives aux évaluations de cours doivent aborder les moyens par lesquels un établissement peut arriver à :

Fournir suffisamment d'information aux étudiants sur l'administration et l'utilisation des évaluations

Ory (2001) cite des études montrant que les étudiants fournissent des évaluations plus constructives, exhaustives, précises et positives lorsqu'ils ont été préalablement renseignés sur les buts et les utilisations des évaluations de cours (bien que Wachtel [1998] soutienne que les études sont variables et non concluantes). Cette situation s'explique par le fait que les étudiants remplissent normalement les évaluations uniquement à la fin du cours et ne sont pas en mesure de constater les effets de leurs efforts. Beran et coll. (2007, 2005) de même que Wachtel (1998) soulignent que les étudiants ont donc souvent l'impression que les résultats de leur évaluation ne sont pas examinés et que leurs suggestions ne sont pas mises en pratique. Les étudiants ont également parfois l'impression que leur anonymat n'est pas protégé lorsque le processus de collecte et de conservation des données n'est pas expliqué convenablement, plus particulièrement lorsqu'il s'agit d'une évaluation en ligne (demandant souvent une forme d'authentification même lorsque les résultats sont uniquement enregistrés sous forme de données groupées). Tel qu'indiqué à la *Section 2.C.iii : Perceptions des étudiants*, les résultats d'évaluation en sont affectés.

Les étudiants devraient être bien informés sur l'utilisation des évaluations dans le but de perfectionner et d'évaluer l'enseignement et sur le rôle que ces évaluations jouent dans le processus de cheminement de carrière, d'embauche et de titularisation, de même que sur la conservation et la consultation des données. Les enseignants peuvent également discuter des changements qu'ils ont apportés à leurs cours ou à leur enseignement en fonction des évaluations antérieures.

Svinicki (2001) propose plusieurs voies possibles que les enseignants peuvent suivre pour aborder les évaluations et aider les étudiants à comprendre les types de réponses les plus utiles aux enseignants et aux administrateurs.

Permettre aux étudiants de consulter les résultats appropriés des évaluations

La question de savoir si les résultats groupés des évaluations devraient être communiqués aux étudiants est étonnamment complexe. Comme le montre la *Section 4.D.i : Communication des résultats d'évaluation*, Wachtel (1998) soutient que les étudiants, en ayant participé au processus d'évaluation de l'enseignement, méritent d'accéder aux résultats découlant de leurs commentaires. Cependant, plusieurs études (telles qu'examinées par Wachtel, 1998) ont laissé entendre que la « réputation » d'un enseignant (pouvant découler de la publication de résultats d'évaluation) peut avoir une incidence sur les réponses données par les étudiants faisant leur évaluation par la suite. Toutefois, aucune de ces incidences n'a néanmoins montré que la validité des évaluations des étudiants est compromise par la communication des résultats aux étudiants; elles indiquent plutôt que les résultats d'évaluation doivent être diffusés uniformément de manière à ce que toute influence sur les résultats d'évaluation soit constante dans le cas où ces résultats sont comparés entre les cours ou entre les enseignants.

Plusieurs écoles semblent pallier ces préoccupations en fournissant un accès aux données groupées à un nombre limité de questions d'évaluation (voir la *Section 3.F.iii : Publication des résultats*). Il s'agit habituellement de questions vastes et

générales qui influent peu sur les attentes des étudiants relativement aux traits de caractère particuliers d'un enseignant.

Au cours d'une étude menée sur 1 229 étudiants, Beran et coll. (2005) ont observé que 52 % des étudiants n'avaient jamais consulté ni utilisé les cotes d'évaluation de cours (principalement parce qu'ils en ignoraient l'existence), tandis que 47 % ont affirmé les utiliser pour leur choix de cours ou d'enseignants. Ces résultats laissent entendre qu'il peut être nécessaire d'améliorer la publication des résultats d'évaluation ainsi que leur accès, et ce, même dans les établissements où les résultats sont mis à la disposition des étudiants.

Offrir aux étudiants d'autres moyens de fournir une rétroaction

Étant donné la valeur et l'importance d'aider les étudiants à comprendre le processus d'évaluation et l'incidence de leurs commentaires et afin de pallier le scepticisme des étudiants à l'égard des évaluations, le recours aux évaluations de mi-session peut améliorer de façon significative la confiance des étudiants (Wachtel, 1998) de même que la capacité de ceux-ci à fournir des commentaires utiles. Les évaluations de mi-session, particulièrement lorsque les enseignants discutent des résultats des évaluations avec leurs étudiants, aident les étudiants à comprendre les façons dont leur rétroaction est interprétée et incluse dans les changements apportés au cours ou à l'enseignement de l'enseignant, améliorant ainsi leur perception de la valeur et de l'utilité des évaluations pour les inciter à fournir une rétroaction plus constructive (Svinicki, 2001). Lewis (2001) montre également que la conduite d'évaluations de mi-session peut améliorer les cotes des évaluations de fin de cours, les étudiants étant devenus des évaluateurs plus habiles et davantage investis dans le cours.

Un certain nombre d'auteurs fournissent des lignes directrices sur la conduite des évaluations de mi-session, notamment Lewis (2001) et Felder (1993).

5.C.ii Pour les enseignants

Plusieurs études arrivent à la conclusion selon laquelle la majorité des membres du corps professoral ont une vue négative de l'évaluation de l'enseignement par les étudiants et affichent même une résistance (Nasser et Fresko, 2002; Abrami, 2001; Theall et Franklin, 2001; Centra, 1993). Cependant, Beran et coll. (2005) ont observé que presque tous les membres du corps professoral avaient une opinion positive des systèmes d'évaluation, mais peu avaient réellement utilisé les résultats pour apporter des changements à leurs cours ou à leur enseignement. Cet état de fait est corroboré par les conclusions tirées par d'autres chercheurs dont les études ont montré que les données d'évaluation avaient souvent une incidence minimale sur l'efficacité ou le rendement de l'enseignement (Campbell et Bozeman, 2008, Marsh, 2007, Centra, 1998), plus particulièrement lorsqu'elles sont fournies sans le bénéfice des consultations. Pour cette raison, les établissements devraient donc :

Demander au corps professoral de joindre une note circonstanciée

Comme nous l'avons vu à la *Section 3.G.iv : Titularisation, promotion et mérite*, le corps professoral est souvent invité à fournir des données sommaires (habituellement dans les dossiers d'enseignement) pour les dossiers de promotion et de titularisation de même que pour les examens annuels du mérite. Cependant, Ory

(2000) indique qu'idéalement « l'évaluation va au-delà du dénombrement, de la mesure, de l'enregistrement ou de la compilation des données. Il préconise d'utiliser les évaluations de l'enseignement non pas comme un exercice scientifique ayant comme objectif la recherche de la vérité absolue, mais plutôt comme une forme d'argument que le corps professoral utilise pour établir le bien-fondé de son enseignement » (p. 17). Franklin (2001) avance que cette compréhension de l'utilisation des données d'évaluation peut être favorisée si le corps professoral a l'occasion de mettre en contexte ses résultats d'évaluation en joignant une note circonstanciée qui souligne les aspects particuliers du cours (p. ex., les techniques d'évaluation expérimentale) pouvant clarifier certains résultats d'évaluation. Le corps professoral pourra également mettre en contexte les résultats dans le cadre de l'amélioration continue de son enseignement, soulignant les domaines à améliorer ou les changements apportés au cours ou aux méthodes d'enseignement par suite des évaluations antérieures. Franklin soutient que de telles notes circonstanciées permettront d'« améliorer les chances que les examinateurs tiennent compte des opinions des étudiants en prenant en considération l'ensemble du contexte des facteurs complexes qui les composent » (p. 85), ainsi que d'aider les examinateurs à éviter d'interpréter ou d'utiliser incorrectement les données.

Utiliser les données d'évaluation pour fournir une rétroaction formative

Les résultats des évaluations, particulièrement ceux découlant des instruments mesurant les comportements d'enseignement ou les éléments de cours particuliers, peuvent fournir une rétroaction diagnostique précieuse des forces et faiblesses d'un enseignant ou d'un volet particulier d'un cours. La rétroaction qualitative, sous la forme de réponses écrites à des questions ouvertes, peut également fournir des renseignements précis et utiles (voir la *Section 2.D.ii : Collecte et interprétation de la rétroaction qualitative*).

Néanmoins, il existe suffisamment de preuves (comme il a été abordé à la *Section 2.E : Usages communs des données d'évaluation de cours*) démontrant que le simple examen des résultats d'évaluation n'est pas suffisant pour mener à une amélioration de l'enseignement. Pour y arriver, il est nécessaire de soumettre les résultats d'évaluation à des consultations (voir les explications ci-dessous).

Voir la *Section 6.A.iii : Utilisation accrue des évaluations à des fins formatives* qui propose des façons d'adapter les instruments d'évaluation afin de pouvoir les utiliser à des fins d'évaluation formative de l'enseignement.

Favoriser et fournir une infrastructure de consultation des évaluations d'enseignement

Comme nous l'avons vu, il existe des preuves qui indiquent que l'accès à des données diagnostiques a beaucoup plus d'impact lorsqu'elles sont combinées à des consultations (auprès du personnel affecté au développement du corps professoral ou aux directeurs de département). Lang et Kersting (2007) de même que Marsh (2007) ont étudié l'incidence de la rétroaction des étudiants sur les efforts d'amélioration de l'enseignement et ont conclu que, lorsque les données d'évaluation ne sont pas accompagnées d'une certaine forme de consultation, elles produisent peu d'effet à long terme. Leurs études ont été réalisées sur une période de quatre semestres et de 13 ans, respectivement, et ont démontré que, même si les données

d'évaluation prises seules ont un impact positif et immédiat sur les enseignants, l'effet ne perdure pas et, en fait, s'amenuise assez rapidement.

Hodges et Stanton (2007) soulignent que, lorsque les membres du corps professoral reçoivent de l'aide pour l'analyse des résultats d'évaluation, ils sont plus susceptibles de considérer les évaluations plus positivement et comme « faisant partie d'une approche de recherche en enseignement » pouvant en retour « servir de base pour des changements efficaces dans notre approche d'enseignement et pouvant également alimenter notre réflexion sur les questions touchant la matière » (p. 280). Moore et Kuol (2005) proposent différentes stratégies pratiques destinées aux membres du corps professoral pour analyser des données d'évaluation visant à les aider à gérer leurs réactions et à concentrer leurs efforts à utiliser les évaluations pour améliorer leur rendement en enseignement (voir le tableau à la fin de la présente section).

Penny et Coe (2006) ont déterminé un certain nombre de stratégies que les formateurs en pédagogie universitaire ou les collègues peuvent utiliser pour assurer des consultations efficaces. Elles comprennent l'implication active des membres du corps professoral dans le processus; l'utilisation de multiples sources d'information (évaluations, observations en classe); la création d'occasions où les membres du corps professoral peuvent interagir avec leurs pairs; l'attribution de temps suffisant consacré au dialogue et à l'interaction (entre les conseillers et les membres du corps professoral); l'utilisation des autoévaluations des enseignants; l'utilisation de rétroactions de grande qualité; l'examen et la compréhension de l'approche des membres du corps professoral vis-à-vis de l'enseignement (p. ex., philosophie et stratégies pédagogiques); et l'établissement d'objectifs d'amélioration pour les membres du corps professoral.

Fournir aux enseignants l'occasion de recevoir des évaluations individualisées

Wright (2006) soutient qu'« il est fréquent de voir que tous les membres du corps professoral sont évalués de la même manière, qu'ils comptent une ou quinze années d'enseignement » (p. 420). Il suggère que les systèmes d'évaluation soient plutôt adaptés afin de tenir compte du rang occupé au sein du corps professoral. Les enseignants débutants pourraient recevoir des rétroactions plus approfondies et être évalués en fonction d'un certain nombre de mesures de l'enseignement, tandis que les enseignants titulaires et chevronnés pourraient profiter davantage de rétroactions plus ciblées en fonction de leurs objectifs individuels d'enseignement.

Hoyt et Pallett (1999) ont dressé un calendrier d'évaluation exhaustif pour les établissements, comprenant des propositions de procédures s'adressant aux enseignants de première année et à des groupes particuliers d'enseignants (comme les titulaires et les non-titulaires). En ce qui concerne les membres du corps professoral qui en sont à leur première affectation à un poste, ils recommandent que les évaluations des étudiants soient menées dans tous les cours et soient combinées à au moins une analyse formative réalisée par un collègue, permettant ainsi aux directeurs de département d'établir rapidement les améliorations à apporter. Pour ceux qui aspirent à la titularisation, ils recommandent que les évaluations des étudiants soient réalisées dans tous les cours, au moins deux fois pendant une période de cinq ans (une première fois au début de leur affectation et une deuxième

fois au cours de la plus récente activité d'enseignement). De plus, il serait souhaitable que des évaluations formatives soient réalisées chaque année dans un ou deux cours, et ce, jusqu'à l'année de titularisation.

Fournir au corps professoral l'information relative à la collecte et à l'utilisation des données d'évaluation

Il y a clairement lieu d'indiquer que les établissements ne prennent pas suffisamment de mesures pour renseigner et former le corps professoral sur les politiques et les procédures relatives à la collecte des données d'évaluation. Plus particulièrement, les efforts pour s'assurer que les membres du corps professoral comprennent les méthodes de collecte, d'analyse et de communication des données sont inconstants et souvent limités. L'examen du matériel utilisé par les établissements de même que des résultats des sondages auprès des administrateurs des collèges et des universités montre que les responsables des décisions touchant la gestion du personnel (que ce soit pour le mérite annuel, la promotion ou la titularisation) n'assurent pas systématiquement la transparence des processus liés à l'administration des évaluations de cours. Comme l'ont démontré Abrami (2001), Kulik (2001) et d'autres, la prestation d'une formation destinée au corps professoral sur les évaluations de cours contribue à dissiper les mythes et les idées fausses tenaces et à apaiser les craintes concernant la façon dont les données peuvent être utilisées par les administrateurs. De plus, les membres du corps professoral possédant une meilleure compréhension des attentes de l'établissement sont plus susceptibles d'aller chercher l'information et l'aide nécessaires pour améliorer l'efficacité de leur enseignement.

5.C.iii Pour les administrateurs

Les administrateurs sont les principaux utilisateurs de données d'évaluation et requièrent une formation et un appui substantiels afin d'instaurer et d'interpréter les évaluations avec efficacité. Pour aider les administrateurs à réaliser ces tâches, les établissements devraient :

Utiliser les données d'évaluation à des fins sommatives

Beran et coll. (2005) ont observé qu'une grande majorité d'administrateurs (84 % dans une étude menée auprès de 52 administrateurs) trouvent que les évaluations de cours constituent une source de renseignements utile (bien que leur étude subséquente, en 2007, a montré que seulement 31 % des administrateurs croyaient que les évaluations étaient un indicateur valide de la qualité de l'enseignement), particulièrement en ce qui a trait aux décisions touchant la gestion du personnel. Beran, Violato et Kline (2007) ont constaté que les administrateurs s'accordent en général pour dire que les données d'évaluation pourraient être utilisées efficacement pour déterminer la qualité de l'enseignement de même que pour accorder les mérites et les prix d'excellence en enseignement. Les données d'évaluation sont suffisamment valables pour que les administrateurs puissent les utiliser avec confiance lorsqu'ils effectuent des évaluations sommatives de l'efficacité de l'enseignement, tout en tenant compte de plusieurs mises en garde.

Comme il a été indiqué précédemment, d'Apollonia et Abrami (1997) de même que McKeachie (1997) ont soutenu que, même si les évaluations peuvent fournir des

renseignements utiles sur l'enseignement, elles devraient être utilisées par les administrateurs uniquement pour porter un « jugement global ». Ils s'entendent pour dire qu'à des fins sommatives, les comités de promotion et de titularisation n'ont pas à aller au-delà de trois catégories de rendement : exceptionnel, satisfaisant et inacceptable.

Dans l'étude menée par Beran, Violato et Kline (2007), une part importante des administrateurs (23 %) estimaient que les évaluations devraient être mises en contexte et accompagnées de renseignements complémentaires. L'utilisation des données d'évaluation à des fins sommatives devrait également tenir compte des recommandations citées ci-après en ce qui a trait à la présentation des données d'évaluation et à la formation des utilisateurs de ces données.

Renseigner et former les administrateurs

Plusieurs études citées en référence dans la présente analyse indiquent que, même si les administrateurs utilisent possiblement les données d'évaluation à différentes fins (décisions quant à la gestion du personnel), ils ne sont pas très familiers avec la recherche qui a été menée sur la validité des évaluations et les pratiques exemplaires en ce sens. Abrami (2001), Theall et Franklin (2001), Beran, Violato et Kline (2007), Beran et coll. (2005) et d'autres ont montré que les administrateurs manquent souvent de connaissances générales sur la meilleure façon d'interpréter et d'utiliser les données obtenues dans les évaluations. Cette situation est la cause de préoccupations étant donné que les données d'évaluation servent régulièrement pour la prise de décisions touchant la gestion du personnel. Bien qu'il soit déraisonnable de s'attendre à ce que les administrateurs atteignent des connaissances approfondies de ce vaste champ de recherche de l'enseignement supérieur, ils profiteraient des connaissances élémentaires des principales questions à examiner étant donné qu'elles sont en lien avec les fins particulières pour lesquelles les données sont utilisées. L'utilisation des données ne requiert pas nécessairement une connaissance poussée de l'analyse statistique, mais elle requiert une connaissance élémentaire de la façon dont les outils fonctionnent et des mesures qu'ils peuvent ou non fournir. Comprendre les limites d'un instrument donné est un élément clé. La formation sur la valeur statistique des données d'évaluation, les influences externes possibles sur les résultats d'évaluation et les moyens efficaces pour gérer et interpréter les données, notamment les mesures comparatives appropriées, contribueraient à ce que, lorsque les données sont utilisées à des fins sommatives, les décisions prises soient justes et équitables. Voir la *Section 4.E.i : Survol des variables étudiées* pour connaître la liste des questions qui devraient être abordées pendant la formation des administrateurs sur l'utilisation des données d'évaluation.

Présenter des données permettant une interprétation aisée et juste

De nombreux chercheurs ont émis des commentaires sur le besoin de fournir des renseignements concis, mais utiles, aux administrateurs sur la meilleure façon possible de relayer les données d'évaluation à d'autres. Abrami (2001) soutient qu'il ne faut pas sous-estimer la puissance de la présentation des données. Il fait remarquer que les présentations visuelles, comme les tableaux ou les graphiques, peuvent influencer positivement sur la capacité d'un examinateur à interpréter l'information. Pour certains, la simple présentation des moyennes n'est pas

suffisante : un tableau ou un graphique avec des données de comparaison (p. ex., les moyennes des départements) peut aider à clarifier les notes obtenues par chacun et à les situer par rapport aux autres collègues ou à les comparer aux années d'enseignement précédentes (p. ex., faire un suivi dans le temps). Voir la *Section 4.D.ii : Difficultés relatives à l'interprétation et à l'utilisation des données à des fins sommatives* pour davantage d'information sur la comparaison des données d'évaluation entre les cours ou entre les enseignants. Les rapports ne devraient pas présenter plus de données, ou plus de détails, que ce dont les administrateurs ont besoin pour réaliser les tâches d'évaluation qui leur sont propres. Ory et Ryan (2001) de même que Theall et Franklin (2001) ont indiqué que les administrateurs ne devraient recevoir que les données qui correspondent au niveau de spécificité (p. ex., nombre de décimales) leur permettant d'établir des distinctions statistiques et conceptuelles significatives.

Pour aider les administrateurs à prendre des décisions efficaces au moyen des données d'évaluation, Hoyt et Pallett (1999) recommandent de présenter toutes les données disponibles de tous les cours enseignés par un membre du corps professoral et d'utiliser les données cumulatives enregistrées pour réaliser l'évaluation de l'efficacité de l'enseignement de l'enseignant (six cours au minimum).

Ajouter des preuves supplémentaires et appropriées aux données d'évaluation

Pour faire une appréciation valable de l'efficacité de l'enseignement, Cashin (1988) recommande d'utiliser des données provenant de sources multiples. Celles-ci peuvent comprendre des autoévaluations ou des commentaires de réflexion, des renseignements sur les objectifs du cours, des échantillons du matériel pédagogique, des modèles de barèmes d'évaluation, des détails sur les changements apportés au cours, des recherches probantes sur l'enseignement et des preuves d'activités de perfectionnement professionnel (lesquelles constituent généralement la base d'un portfolio ou d'un dossier d'enseignement) (Hoyt et Pallett, 1999; Seldin, 1993, 1999). Ces données peuvent être triangulées et elles peuvent mettre en contexte les évaluations des étudiants de même que couvrir des éléments du cours ou des stratégies d'enseignement qui ne ressortent pas dans les données d'évaluation.

Certains experts ne sont pas d'accord quant à la valeur de ces autres mesures de l'efficacité de l'enseignement à des fins d'évaluations sommatives. Dans sa revue de littérature sur l'évaluation de l'enseignement par les pairs, Bernstein (2008) souligne que, même si plusieurs défendent l'utilisation des observations de l'enseignement faites en classe à des fins formatives (en vue d'améliorer l'efficacité de l'enseignement), la plupart des chercheurs font une mise en garde contre l'utilisation d'observations informelles à des fins d'examen sommatif en raison de leur manque relatif de validité. Afin d'améliorer cette forme d'évaluation, DeZure (1999) a recommandé de multiples observations par plus d'une personne qualifiée et l'utilisation d'un formulaire d'évaluation valide.

Bien qu'il soit d'accord avec l'utilisation de mesures d'évaluation supplémentaires (p. ex., évaluations par les pairs, matériel didactique, etc.) afin de compléter ou de compléter les données d'évaluation, Abrami (2001) est d'avis que ces mesures sont « moins valides sur le plan psychométrique » que les instruments d'évaluation et ne

devraient pas remplacer les formulaires d'évaluation officiels de fin de session (p. 65).

5.C.iv Pour les établissements

Mettre à l'épreuve et réexaminer les instruments d'évaluation lorsque changent les priorités de l'établissement ou les pratiques d'enseignement

Les instruments d'évaluation devraient être régulièrement mis à l'épreuve et réexaminés par les établissements qui les utilisent. Étant donné que les méthodes d'enseignement se modifient, que les étudiants, le corps professoral et les administrateurs varient et que les politiques de l'établissement sont révisées, les formulaires d'évaluation peuvent nécessiter une révision et une mise à jour. Les établissements doivent s'assurer que les instruments mesurent efficacement les points spécifiques qui représentent un intérêt pour eux, pour le corps professoral ainsi que pour les étudiants. Les chercheurs en évaluation ne recommandent pas une révision générale annuelle des formulaires d'évaluation, ni même une révision mineure annuelle de l'outil. Cette pratique peut avoir une incidence négative sur la capacité de l'instrument à fournir des examens longitudinaux au niveau de l'établissement ou de la faculté. (Par exemple, elle peut avoir un impact négatif sur un membre du corps professoral dans le cas où les questions du formulaire d'évaluation ou l'échelle d'appréciation sont modifiées à plusieurs reprises au cours des années menant à la titularisation ou à une promotion. Si les données ne sont pas clairement présentées pour indiquer ces changements, il est possible que les examinateurs n'interprètent pas correctement les résultats.)

En outre, étant donné que l'enseignement supérieur évolue, les instruments d'évaluation devraient être réexaminés, et ajustés s'il y a lieu, pour répondre à l'évolution des contextes de l'enseignement postsecondaire. Ces changements peuvent inclure des transformations quant aux pratiques pédagogiques et aux données démographiques des étudiants ou des changements dans les besoins en matière d'évaluation des membres du corps professoral (à des fins formatives), les mesures de responsabilisation des établissements, les technologies, les pratiques de perfectionnement du corps professoral et la recherche en évaluation de l'enseignement (Theall et Franklin, 2000).

Mener des auto-analyses et faire de la recherche à l'interne

Un établissement devrait envisager de procéder à une recherche à l'interne sur son système d'évaluation. Cette recherche pourrait comprendre l'examen de l'instrument ou la conduite d'un sondage auprès de la communauté en ce qui a trait à l'utilité de l'outil ou aux attitudes à l'égard de celui-ci. Ory (2001) observe que l'incidence découlant des influences qu'exercent le cours, l'étudiant et l'enseignant peut varier d'un établissement à l'autre et que différents établissements ou même différentes divisions peuvent conclure qu'ils ont besoin de contrôler certaines variables en vue de produire des données d'évaluation permettant d'être comparées de façon juste entre les cours et les enseignants. La recherche sur un établissement et un instrument n'est pas nécessairement généralisable dans sa totalité et doit être validée par une recherche interne menée par l'établissement.

Établir des politiques-cadres en matière de collecte, d'administration et d'utilisation des systèmes d'évaluation de cours par les étudiants

À la *Section 3 : Politiques et pratiques actuelles en Amérique du Nord*, nous avons examiné différentes politiques et pratiques utilisées dans plusieurs dizaines d'établissements postsecondaires démontrant des variations entre elles dans l'ensemble de l'Amérique du Nord. En règle générale, nous avons constaté que la plupart des établissements ont élaboré des politiques qui traitent de la collecte, de l'administration et de l'utilisation des systèmes d'évaluation de cours par les étudiants. Cependant, à notre étonnement, plusieurs établissements utilisent régulièrement des systèmes d'évaluation, mais, semble-t-il, sans politiques-cadres officielles abordant ces questions clés. Pour assurer une cohérence, une transparence et une compréhensibilité, de telles politiques-cadres devraient être adoptées.

L'ACPPU recommande que les membres du corps professoral et les associations qui les représentent participent à l'élaboration de ces politiques. L'« Énoncé de principes sur l'évaluation anonyme de l'enseignement par les étudiants » (2006) de l'ACPPU mentionne que :

Toute méthode établie par l'administration ou l'instance supérieure de l'établissement pour évaluer la qualité de l'enseignement, y compris tout projet d'utiliser des questionnaires anonymes pour les étudiants, devrait recevoir l'accord de l'association de personnel académique ou avoir été négociée avec elle, et devrait être ajoutée à la convention collective ou au guide du personnel académique (voir l'annexe D.2).

Des pensées similaires sont partagées par l'American Association of University Professors (AAUP) dans son énoncé intitulé « Statement on Teaching Evaluation » (voir l'annexe D.3).

Établir des pratiques administratives claires

Un certain nombre d'études, notamment celles réalisées auprès des utilisateurs des évaluations, ont recommandé que les établissements améliorent les processus et les pratiques relatives à l'administration des systèmes d'évaluation de cours.

Pour les établissements qui envisagent l'utilisation de systèmes d'évaluation en ligne (voir la *Section 6.A.i : Outils d'évaluation en ligne* pour un exposé de cette nouvelle tendance), Sorenson et Reiner (2003) présentent une liste utile de points à considérer, notamment la meilleure façon d'amener des changements organisationnels, d'anticiper et d'aborder les objections, d'évaluer la réceptivité, de former les utilisateurs, de créer un système pratique et sécurisé et de promouvoir l'esprit de collaboration et d'appartenance.

Articuler des buts et des objectifs d'évaluation

Soulignant que leur étude a mis en lumière certaines ambiguïtés entourant les objectifs d'évaluation, Campbell et Bozeman (2008) recommandent que les établissements déterminent et formulent clairement l'objectif de la conduite des évaluations et affinent leurs procédures administratives pour refléter les buts établis.

Élaborer du matériel pédagogique et soutenir des réseaux destinés aux utilisateurs

Franklin et Theall (1989) ont montré que moins les connaissances d'une personne à l'égard des évaluations sont grandes, plus elle est susceptible de remettre en question l'utilité de celles-ci comme indicateurs de l'efficacité de l'enseignement. Ces auteurs et d'autres ont également démontré que la sensibilisation quant aux évaluations des étudiants est faible et très variable.

Comme il a été observé précédemment, les étudiants, le corps professoral et les administrateurs pourraient grandement profiter de connaissances et de formation sur l'utilisation des systèmes d'évaluation – une responsabilité qui devrait incomber à l'établissement (ou à une autorité déléguée).

Il est grandement recommandé que les établissements travaillent dans le but d'améliorer la formation de ceux qui utilisent et interprètent les systèmes d'évaluation (Theall et Franklin, 2000). De plus, une grande partie de la littérature demande à ce que les établissements fassent plus que fournir simplement des rapports sommaires des résultats (Theall et Franklin, 2000).

Theall et Franklin (2000) soutiennent que les pratiques d'évaluation et de développement du corps professoral sont « inextricablement liées », car une « bonne évaluation nécessite de déterminer les caractéristiques et le rendement à évaluer de même que le degré d'engagement des établissements et des individus qui y évoluent afin de pouvoir utiliser les éléments probants le plus justement et équitablement possible au moment de la prise des décisions » (p. 103).

Franklin et Theall (1989) ont noté que les guides ou les manuels sur les systèmes d'évaluation de cours peuvent constituer des sources importantes de renseignements pour les personnes ayant à examiner, à recevoir, à communiquer et à interpréter les données d'évaluation, de même qu'à prendre des décisions à partir de ces données. Ces sources peuvent comprendre une description des instruments en vigueur; des recommandations sur les procédures administratives, notamment la mise en œuvre des pratiques et des politiques; et des méthodes d'analyse, de communication et d'interprétation des données.

**Stratégies individuelles à appliquer pour l'analyse de la
rétroaction fournie par les étudiants**

1. Maîtriser les mécanismes de défense.
2. Analyser la source ayant généré les réactions des étudiants d'une façon qui fait la lumière sur les préoccupations et les problèmes soulevés.
3. Tâcher de ne pas avoir une réaction exagérée ou insuffisante vis-à-vis de l'information reçue par le biais des commentaires contenus dans les évaluations.
4. Classer les préoccupations soulevées par les étudiants en deux catégories : celles donnant lieu à une action réalisable et à une action non réalisable.
5. Échanger avec les étudiants avant et après leur rétroaction.
6. Ne pas émettre l'hypothèse simpliste selon laquelle toutes les réponses positives sont attribuables à un bon enseignement et toutes les réponses négatives, à un mauvais enseignement.
7. Garder en tête que les changements aussi minimes soient-ils peuvent avoir de grandes incidences.
8. Élaborer une stratégie d'amélioration de l'enseignement qui tient compte des rétroactions des évaluations (145-6).

Moore, S., et Kuol, N. (2005). A punitive tool or a valuable resource? Using student evaluations to enhance your teaching. Dans G. O'Neill, S. Moore et B. McMulline (Éd.), *Emerging issues in the practice of university learning and teaching* (pp. 141-148), Dublin, All Ireland Society for Higher Education.

Section 6

Tendances émergentes, lacunes existantes et recommandations pour les recherches futures

6.A Tendances émergentes

La recherche et les pratiques en cours dans les établissements nord-américains révèlent quelques nouvelles orientations. Nous avons choisi les sujets qui sont le plus à même de conduire à des changements dans l'administration et l'utilisation des évaluations, et ce, dans un délai assez rapproché.

6.A.i Outils d'évaluation en ligne

La tendance à offrir des évaluations de cours informatisées ou en ligne s'est amorcée il y a environ 20 ans et s'est répandue davantage au cours de la dernière décennie. La recherche dans ce domaine est encore nouvelle et le débat sur certaines questions liées à la mise en place de ces outils fait toujours rage.

Comme nous l'avons souligné à la *Section 3.D.i : Méthodes de distribution*, un bon nombre d'établissements de l'Amérique du Nord ont commencé à faire remplir les évaluations en ligne. Certains ont totalement adopté cette méthode, tandis que d'autres offrent une alternative entre la version papier et en ligne et quelques-uns continuent d'utiliser uniquement un questionnaire en version papier. Une étude menée en 2003 auprès de 256 établissements américains a révélé que 10 % des établissements ont indiqué qu'ils utilisent les outils en ligne comme principaux moyens de mener les évaluations de cours, tandis que 78 % ont indiqué qu'ils ont recours aux formulaires papier pouvant être lus par lecteur optique. Les 12 % d'établissements restants utilisent des formulaires papier sans lecture optique (Hoffman, 2003).

Le processus de passation des évaluations en ligne varie d'un établissement à l'autre, certains laissant du temps en classe pour remplir le questionnaire de sondage (p. ex., dans un laboratoire informatique) et d'autres demandant aux étudiants de le remplir en dehors des heures de cours. En règle générale, une adresse Web est fournie aux étudiants pour leur permettre d'accéder à l'instrument d'évaluation du ou des cours qu'ils suivent. Les étudiants peuvent avoir accès au lien Web en classe ou le recevoir par le biais d'une plate-forme LMS (système de gestion de l'apprentissage, comme le Blackboard or WebCT). Les étudiants utilisent normalement un code d'ouverture de session unique et sécurisé qui leur est fourni ou se servent de leur propre code d'identification d'étudiant.

Bien qu'un bon nombre confirment l'éventail d'avantages de l'utilisation d'un système en ligne, d'autres en sont moins convaincus. De plus, certains désavantages ont été cernés (p. ex., faibles taux de réponse) et n'ont pu être éliminés jusqu'à maintenant. De

nombreux chercheurs ont proposé les meilleures façons de s'attaquer aux problèmes et les établissements ont eux-mêmes essayé diverses méthodes; aucune n'a cependant été largement acceptée ni mise à l'épreuve.

Voici un résumé des principaux sujets abordés dans la recherche :

Rentabilité et efficacité

La réalisation des évaluations en ligne peut procurer des avantages importants, particulièrement en ce qui a trait aux coûts tant en argent (coût d'impression des formulaires) qu'en temps de la part du personnel (temps consacré à la distribution, à la collecte, à la lecture optique, à la transcription des commentaires et au stockage des données) (Donmeyer et coll., 2004; Bothell et Henderson, 2003; Johnson, 2003; Sorenson et Reiner, 2003).

Étant donné qu'elles ne sont habituellement pas réalisées durant les heures de classe, les évaluations en ligne ne grugent pas le temps qui pourrait être utilisé pour l'enseignement (Donmeyer et coll., 2004; Sorenson et Reiner, 2003).

De plus, grâce à la collecte des données en ligne, ces dernières peuvent être traitées plus rapidement que si elles étaient sur des formulaires papier (Donmeyer et coll., 2004; Sorenson et Reiner, 2003) et permettent généralement de produire des rapports personnalisés de plus grande qualité et plus exhaustifs (Llewellyn, 2003).

Anonymat de l'étudiant

Étant donné que l'écriture des étudiants ne paraît pas dans les évaluations en ligne, certains ont fait valoir que l'anonymat est davantage assuré (Donmeyer et coll., 2004). Cependant, de nombreux chercheurs ont souligné que le facteur de l'anonymat demeure une préoccupation auprès des étudiants même lorsque ceux-ci utilisent des systèmes d'évaluation en ligne (Avery et coll., 2006; Reid, 2001). Cette situation peut, en partie, être en lien avec une croyance voulant que les codes d'accès puissent être appariés aux étudiants individuellement, une crainte qui s'accroît lorsque les étudiants accèdent aux questionnaires de sondage au moyen de leur code d'identification unique. Comme l'aborde précédemment la *Section 4.E.i : Survol des variables étudiées*, l'anonymat n'est pas un facteur qui a une énorme incidence sur les résultats d'évaluation; cependant, il est encore généralement recommandé que les établissements fassent de leur mieux pour assurer l'anonymat des étudiants, ce qui s'applique également à l'environnement en ligne. Afin de respecter cette recommandation, quelques universités et collèges ont recours à des entreprises externes sous contrat pour faire la collecte et l'analyse des données de même que la préparation des rapports.

Variables liées à l'enseignant : influence du corps professoral

La littérature se concentre également sur certaines variables abordées dans le cadre des évaluations traditionnelles sur papier. Donmeyer et coll. (2004) soutiennent que « les évaluations en ligne sont moins susceptibles de subir l'influence du corps professoral que les évaluations réalisées en classe », suggérant qu'avec les évaluations sur papier, les enseignants peuvent intervenir

de façon quelconque la journée où sont passées les évaluations, ce qui résulterait en des cotes plus élevées (p. 612). Ils allèguent en outre que la simple présence de l'enseignant peut avoir un impact sur les cotes d'évaluation. Cependant, comme nous l'avons vu lors de l'examen des variables à la *Section 4.E.i : Survol des variables étudiées*, la présence de l'enseignant n'influe pas de façon importante sur les résultats d'évaluation.

Administration des formulaires d'évaluation : amélioration des réponses des étudiants

Lorsque les évaluations se déroulent en ligne et sont remplies à l'extérieur de la classe, les étudiants ne sont pas limités par le temps attribué à la fin d'une rencontre en classe. De nombreux outils permettent aux étudiants de revenir au sondage pour ajouter de l'information supplémentaire ou modifier leurs commentaires avant de le soumettre définitivement. Certaines études ont révélé que les réponses qualitatives fournies dans les formulaires en ligne sont plus étayées (Donmeyer et coll., 2004; Sorenson et Reiner, 2003).

Adaptabilité du questionnaire de sondage

Donmeyer et coll. (2004) de même que Sorenson et Reiner (2003) ont fait valoir que les évaluations en ligne procurent aux enseignants plus de souplesse, car ceux-ci peuvent personnaliser les questionnaires en y ajoutant des questions qui se rapportent à leur cours en particulier et à leur style d'enseignement. Néanmoins, cette option est également offerte aux membres du corps professoral utilisant les formulaires papier.

Taux de réponse

Une des principales préoccupations abordées dans la littérature concerne le taux de réponse (voir Avery et coll., 2006, pour une analyse de la recherche connexe). Un bon nombre d'établissements ayant recours au système d'évaluation en ligne ont constaté une diminution importante de la participation, souvent de l'ordre de la moitié (ou plus), par rapport au taux de participation lorsque les évaluations se font à l'aide de formulaires papier. Avery et coll. (2006) indiquent que certaines études ont démontré des taux de réponse aussi faibles que 43 % (comparativement à des taux variant de 61 à 82 % pour les évaluations sur formulaires papier); leur propre étude a révélé des résultats similaires. La récente étude pilote en ligne de Ryerson a affiché un taux de réponse de 38 %, comparativement à l'intervalle normal de 50 à 60 % (Faculty Course Survey, FAQ). Les membres du corps professoral ont eux-mêmes soulevé la question, ce qui a affecté leur volonté d'adopter l'évaluation en ligne (Donmeyer et coll., 2004). Les chercheurs ont avancé que le faible taux de réponse traduit, en partie, une préoccupation concernant l'anonymat (Avery et coll., 2006; Donmeyer et coll., 2004), mais qu'il peut être le résultat de l'obligation des étudiants de remplir les évaluations en dehors des heures de classe (Sorenson et Reiner, 2003). D'autres causes peuvent se rattacher à des problèmes d'ordre technique (Sorenson et Reiner, 2003).

Certains établissements ont eu recours à des incitatifs pour encourager les réponses des étudiants. Ces mesures passent d'une légère bonification des notes, à la publication hâtive des notes finales et au tirage au sort de prix. Dans

son Faculty Course Survey FAQ, Ryerson formule une mise en garde aux membres du corps professoral contre l'utilisation de telles mesures incitatives, précisant que les points de bonification constituent une forme de coercition, ne sont pas appropriés dans le cadre du travail non universitaire et exigent que les répondants soient identifiés.

Donmeyer et coll. (2004) ont indiqué qu'en offrant une légère bonification des notes, le taux de réponse pouvait s'accroître pour atteindre celui obtenu par l'administration des évaluations en classe. Les résultats de cette étude ont révélé que l'utilisation des points de bonification comme mesure incitative ne produisait pas un biais important dans les résultats; cependant, étant donné que cette étude était menée à une échelle plutôt petite, il est difficile de procéder à la généralisation des résultats. D'autres recherches devraient être réalisées pour vérifier les impacts des mesures incitatives de toutes sortes.

Au lieu de recourir à des mesures incitatives pour motiver les étudiants, certains établissements ont misé sur un processus de messages répétitifs pour garantir un taux de réponse acceptable. Les éléments de preuve suggèrent qu'il s'agit d'un moyen efficace pour améliorer la participation (Donmeyer et coll., 2004).

Fiabilité et validité des résultats

Comme le signalent Avery et coll. (2006), « il y a peu de connaissances sur la comparabilité des résultats d'évaluation obtenus à l'aide des mécanismes de collecte sur le Web par rapport à ceux obtenus à l'aide du processus traditionnel des formulaires papier », particulièrement en ce qui a trait à l'incidence sur les cotes moyennes des évaluations de cours (p. 22). En se basant sur ce qui précède, ils mettent en garde les administrateurs considérant de mettre en place un système en ligne, du fait que les données d'évaluation sont utilisées à des fins sommatives. Des études ont examiné la validité et la fiabilité des systèmes d'évaluation en ligne, en vérifiant si la méthode de passation des évaluations avait une incidence ou non sur les résultats dans leur ensemble. Certaines conclusions avancent que les résultats sont généralement cohérents et que les écarts ne sont pas statistiquement importants (Avery et coll., 2006; Donmeyer et coll., 2004). Avery et coll. (2006) ont observé que les questions de sondage obtenaient de meilleurs résultats sur les formulaires en ligne que sur les formulaires papier. En revanche, Hardy (2003) soutient que la recherche n'est pas concluante et que les résultats des évaluations en ligne « peuvent être inférieurs, supérieurs ou égaux » (p. 33).

D'autres recherches sur certains aspects de la réalisation des évaluations de cours en ligne sont encore nécessaires. Davantage d'études plus approfondies sur les taux de réponse se révéleraient utiles, particulièrement pour les établissements envisageant la possibilité de migrer vers un outil en ligne. Une telle recherche devrait s'attaquer à la comparaison des taux de réponse des évaluations en ligne avec celles sur papier et devrait examiner les différents types de mesures incitatives qui sont offerts, particulièrement en ce qui a trait à l'introduction de biais. Les données non scientifiques laissent entendre que le corps professoral et les administrateurs croient que les évaluations en ligne sont uniquement remplies par les étudiants qui « aiment » ou « détestent » un enseignant. Il peut s'agir d'une idée fautive, mais elle nécessite un

examen plus approfondi. Avery et coll. (2006) ont proposé qu'il faudrait également faire l'étude de l'impact des facteurs environnementaux, notamment l'influence des pairs sur les réponses ou les distractions possibles lorsque le questionnaire est rempli (à la maison, dans un lieu public). De plus, d'autres enquêtes sur le contenu et la structure des formulaires d'évaluation sont aussi nécessaires afin de déterminer si la prestation en ligne requiert ou non des changements.

Pour obtenir plus de renseignements sur les établissements utilisant actuellement les systèmes d'évaluation en ligne, consultez le site <http://onset.byu.edu> (en anglais).

6.A.ii Association des données d'évaluation aux mesures de reddition de comptes et aux résultats de l'apprentissage axé sur les compétences

L'évaluation des programmes et l'évaluation par les étudiants aux cycles supérieurs se font de plus en plus, depuis plusieurs années, à l'aide d'un modèle d'évaluation axée sur les compétences. Ce mode d'évaluation se concentre sur la mesure de résultats ou d'objectifs prédéterminés et la définition de moyens authentiques d'évaluation dans la mesure du possible (c'est-à-dire, des mesures d'évaluation qui demandent aux étudiants de faire directement la preuve de leurs résultats, plutôt que d'utiliser des indicateurs comme les notes obtenues au cours) (l'Association of American Colleges and Universities est un fervent adepte de ce modèle d'évaluation des programmes et d'évaluation par les étudiants comme l'explique leur rapport de 2007 intitulé *College Learning for the New Global Century*).

Comme il a été abordé à la *Section 3.H : Lien entre les évaluations de cours et les mesures de reddition de comptes*, les évaluations de cours et les autres moyens de fournir une appréciation de l'apprentissage et de l'expérience des étudiants commencent à apparaître dans les processus d'évaluation, de reddition de compte et de planification des systèmes et des établissements. Ces exercices sont encore au stade embryonnaire, surtout parce que les évaluations de cours ne sont pas nécessairement conçues pour mesurer des résultats particuliers d'un programme et qu'en général elles ne constituent pas des moyens d'évaluation authentiques.

Pour que la tendance vers l'incorporation des évaluations de cours dans le processus de reddition de compte des systèmes et des établissements se poursuive efficacement, les instruments d'évaluation de cours devront être modifiés pour refléter plus directement les résultats des programmes. Ce constat correspond à l'argumentation présentée par Ory et Ryan (2001) de même que Theall et Franklin (2001) selon laquelle les questions des évaluations devraient refléter les priorités de l'établissement. Hoyt et Pallet (1999) soutiennent de façon similaire que lorsque les évaluations servent principalement à des fins sommatives, les instruments devraient viser à mesurer les éléments déterminés (p. ex., dans quelle mesure les objectifs du cours sont-ils pris en compte?).

6.A.iii Utilisation accrue des évaluations à des fins formatives

Comme Ory l'a signalé en 2000, l'évaluation de l'enseignement continue d'être une tâche à multiples facettes et d'ordre formatif. L'étude de Marsh (2007) démontre que la simple communication des résultats des évaluations sommatives aux enseignants se traduit par

peu d'améliorations à l'enseignement. Il est plutôt nécessaire de procéder à des consultations intensives pour voir apparaître des améliorations substantielles et durables dans l'enseignement.

Plusieurs chercheurs (Abrami, 2001; McKeachie, 1997) ont néanmoins indiqué que les évaluations de cours ne sont pas conçues principalement à des fins d'évaluation formative, bien que certains instruments puissent comporter une certaine utilité diagnostique. En revanche, les évaluations de mi-session permettent une rétroaction individualisée et souvent qualitative pouvant être plus profitable aux enseignants souhaitant recevoir de l'information sur les façons d'améliorer leur enseignement. Ainsi, les évaluations de mi-session sont en train de devenir des outils de plus en plus populaires (Aultman, 2006), quoiqu'en raison de leur caractère relativement nouveau, peu de recherches ont été menées jusqu'ici sur les pratiques exemplaires en matière de développement et d'administration ou sur les meilleures méthodes pour intégrer le plus efficacement possible les résultats de ces évaluations aux activités de perfectionnement et d'amélioration de l'enseignement.

6.A.iv Mise en contexte des données d'évaluation à des fins d'évaluation sommative de l'enseignement

Depuis environ les 20 dernières années, la recherche démontre un intérêt grandissant pour les questions traitant de l'amélioration des pratiques d'évaluation. D'une part, cet intérêt a introduit et exploré de nouveaux outils d'évaluation comme les dossiers ou les portfolios d'enseignement de même que le recours aux évaluations par les pairs ou en classe par des professionnels; d'autre part, une plus grande attention a ainsi été portée aux pratiques et aux politiques des établissements.

En ce qui a trait au domaine du perfectionnement du corps professoral, un mouvement est né pour répondre à certaines préoccupations à l'égard de l'utilisation et de l'interprétation des données d'évaluation de cours à des fins de décisions touchant la gestion du personnel. Dans le cadre de leur pratique régulière, les formateurs pédagogiques travaillent avec chaque membre du corps professoral pour « décoder » les données numériques et qualitatives, et ce, tant à des fins formatives que sommatives. Par le biais de consultations, ils aident également le corps professoral à apporter des changements appropriés à leurs stratégies d'enseignement de même qu'à la conception de leurs cours (p. ex., barème d'évaluation, conception des travaux pratiques) en réponse aux commentaires et aux évaluations des étudiants. De plus, certains, comme Franklin (2001), ont recommandé que le corps professoral rédige des notes circonstanciées pour compléter et mettre en contexte les données d'évaluation à l'intention des évaluateurs de l'établissement, notamment les comités de titularisation et de promotion. Ce processus aide les administrateurs dans l'analyse des données d'évaluation et profite aux membres du corps professoral. En outre, elle fait valoir que « les notes circonstanciées peuvent aider vos examinateurs à se rendre davantage compte que les résultats sont des mesures précieuses, bien qu'imparfaites, de l'efficacité de l'enseignement et les aider, par conséquent, à éviter les mauvaises interprétations et utilisations des données courantes pouvant porter préjudice à l'évaluation de votre enseignement » (p. 85).

Comme Franklin (2000) le souligne, les chiffres ne donnent qu'une partie des renseignements; une note de mise en contexte permet de présenter les multiples dimensions de l'enseignement et d'aborder des aspects particuliers du cours ou du style d'enseignement de l'enseignant, ce qu'un condensé ou un résumé des résultats ne permet pas. Elle recommande au corps professoral d'examiner la recherche sur les variables (voir la *Section 4.E.i : Survol des variables étudiées*) pouvant possiblement influencer sur les résultats à divers degrés et de traiter ces variables dans les notes circonstanciées, s'il y a lieu. Par exemple, certains éléments probants indiquent que les cours à option donnent lieu à des évaluations plus élevées; par conséquent, les enseignants dont tous les cours sont qualifiés d'obligatoires voudront peut-être ajouter une note circonstanciée à cet effet aux évaluateurs.

6.B Lacunes existantes et suggestions pour pousser la recherche

6.B.i Définition du vocabulaire et des attentes en enseignement

Theall et Franklin (2000) soulignent le besoin d'établir un « vocabulaire commun fiable et étendu pour décrire des phénomènes postsecondaires importants » (p. 104), notamment le vocabulaire pour évaluer l'enseignement, de même que la nature changeante de l'enseignement postsecondaire (p. ex., les nouvelles pratiques pédagogiques). Ils affirment que c'est une démarche « essentielle pour toute généralisation des résultats des évaluations » (p. 104). Sans une compréhension universelle des termes essentiels utilisés pour parler de l'enseignement et de l'apprentissage des étudiants, notre incapacité à obtenir un consensus sur ce qui constitue un enseignement efficace persistera.

Comme Wright (2008) le suggère, le vocabulaire employé pour aborder les évaluations de cours doit exprimer avec justesse les façons dont les instruments sont utilisés. Bien que ce travail reflète la pratique courante, Wright propose le remodelage de la pratique courante pour que les étudiants, le corps professoral et les établissements comprennent pleinement leurs rôles respectifs dans le processus d'évaluation de l'enseignement : les étudiants évaluent l'instruction, alors que les administrateurs, les établissements et le corps professoral participent à l'évaluation de l'enseignement, en partie selon les résultats des évaluations des étudiants. Plus particulièrement, si les évaluations demandent strictement aux étudiants d'évaluer l'instruction, par opposition au contenu du cours, le nom choisi pour l'instrument devrait indiquer ce thème central et on devrait reconnaître que le titre employé pour faire référence à l'instrument peut influencer sur la façon dont il est utilisé. La pratique courante sur ce point n'est pas constante et de futures recherches peuvent être orientées, de façon plus utile, vers la détermination d'une terminologie commune et significative pour un usage dans tous les secteurs postsecondaires au Canada.

6.B.ii Compréhension des utilisateurs d'évaluations

Peu d'études ont été menées sur les utilisateurs d'évaluations : les administrateurs, le corps professoral et les étudiants. Celles qui existent ont habituellement une envergure limitée et sont restreintes à l'établissement visé. Étant donné les différences de mandat, de secteurs disciplinaires et de culture des établissements, il est difficile de reporter ces résultats sur les sphères élargies de l'enseignement supérieur. Ces facteurs peuvent avoir une incidence sur les perspectives individuelles et les attitudes de même que sur leur utilisation des données d'évaluation. Des travaux de recherche à plus grande échelle contribueraient certainement à comprendre les utilisateurs d'évaluations; toutefois, il ne faut pas exclure la poursuite d'études dans les établissements qui permettront de continuer à informer les administrateurs, le corps professoral et les étudiants des universités et des collèges dans le contexte de leur propre culture d'établissement. De plus, de telles études peuvent aussi englober les outils d'évaluations spécifiques utilisés dans l'établissement. D'autres recherches sur les utilisateurs d'évaluations devraient également être menées dans les deux contextes.

Theall et Franklin (2000) demandent d'autres recherches sur les besoins des différents utilisateurs de données d'évaluations de cours, indiquant que la tâche d'interpréter les résultats varie selon le but. Comme ils le soulignent, l'interprétation va au-delà de la simple capacité d'analyser les nombres, mais elle implique fréquemment la capacité d'offrir une consultation complémentaire pour traduire les évaluations en amélioration réelle des méthodes d'enseignement. De la même façon, Menges (2000) recommande d'autres recherches sur la façon dont les administrateurs utilisent les données d'évaluation des étudiants pour prendre des décisions relatives à la gestion du personnel. McKeachie (1997) avance que des recherches basées sur l'observation et portant sur le processus de prise de décision aideraient à avoir une compréhension plus profonde de la façon dont sont utilisées les évaluations. Pour ce faire, il suggère que les chercheurs assistent aux rencontres des comités de titularisation et de promotion afin d'observer comment les données d'évaluation sont réellement utilisées par les réviseurs.

Comme l'ont noté Schmelkin, Spencer et Gellman (1997), nous avons également besoin de mieux comprendre comment le corps professoral perçoit et utilise les données d'évaluation. À ce jour, les résultats ont été mitigés. L'étude de Schmelkin et coll. auprès de 400 membres du corps professoral a montré que contrairement à quelques cas isolés, le corps professoral n'a pas été trop résistant à l'utilisation des données d'évaluation de cours à des fins formatives ou sommatives, alors que des articles d'opinion et des discussions sur des sites Web et sur des questionnaires de listes de diffusion suggèrent le contraire.

Parmi tous ceux qui ont participé à la collecte et à l'utilisation des évaluations, les moins étudiés demeurent les étudiants. McKeachie (1997) a recommandé que l'attention soit portée sur la façon dont les étudiants comprennent le processus d'évaluation et la manière dont ils remplissent les formulaires d'évaluation. De là, il fait valoir que nous pouvons aller vers une meilleure formation des étudiants pour qu'ils deviennent des évaluateurs plus avertis.

6.B.iii Formation des utilisateurs d'évaluations

Il y a un besoin général et maintes fois répété d'améliorer la formation des utilisateurs de formulaires d'évaluation de cours et des données qui y sont liées. Pour les étudiants, ce besoin signifie s'assurer qu'ils comprennent le but et l'utilisation subséquente des évaluations. Pour les membres du corps professoral, cela veut dire de s'attaquer aux mythes persistants qui mettent en péril la considération et l'utilité conséquente des évaluations pour le perfectionnement de l'enseignement. Cela signifie aussi les aider à trouver des données de mise en contexte utiles pouvant être employées dans leur propre interprétation des résultats et fournies aux administrateurs pour les aider dans le cas des évaluations sommatives. Pour les administrateurs, ce constat signifie de mettre au point et d'offrir une formation pour devenir de meilleurs évaluateurs, ce qui comprend l'information sur la communication et l'interprétation des données statistiques. Pour les établissements, une telle formation assure que des politiques détaillées sont conçues et appliquées équitablement.

Pour atteindre ces objectifs, d'autres travaux de recherche sont nécessaires. Par ailleurs, pour améliorer la formation et le matériel pédagogique, nous devons mieux comprendre les utilisateurs, notamment leurs attitudes envers les outils d'évaluation et leurs perceptions de ces outils. De plus, nous réclamons davantage d'information sur les connaissances qu'ont les utilisateurs de la recherche sur les évaluations de cours ainsi que sur les processus et pratiques de l'établissement. Évidemment, il faut aussi une compréhension plus profonde de la façon dont les données d'évaluation sont utilisées à des fins formatives ou sommatives. La recherche dans ces domaines est limitée. Une enquête plus poussée profiterait aux chercheurs actuels, mais de façon plus importante, elle permettrait la formulation future de recommandations fondées et plus efficaces sur la création d'instruments d'évaluation, l'administration et la mise en place de systèmes d'évaluation et l'utilisation et l'interprétation des données d'évaluation.

6.B.iv Évaluation des auxiliaires à l'enseignement et des enseignants chargés des étudiants de deuxième et troisième cycles

Dans le milieu de l'éducation postsecondaire, il est de plus en plus courant d'offrir des occasions de perfectionnement professionnel et de formation aux étudiants de deuxième et troisième cycles. De nombreuses ressources de perfectionnement professionnel ont été mises au point pour les auxiliaires à l'enseignement, et les bureaux destinés à offrir de la formation et du soutien ont proliféré sur les campus (p. ex., le programme de formation des auxiliaires à l'enseignement de l'Université de Toronto ou le programme de préparation de la relève du corps professoral aux États-Unis). Étant donné que les auxiliaires à l'enseignement de deuxième et troisième cycles ont un contact assez prolongé et direct avec les étudiants de premier cycle (par le tutorat, les laboratoires, la disponibilité au bureau et les responsabilités reliées à la notation), la qualité de leur enseignement devrait être évaluée. Bien que certains établissements aient conçu des mécanismes pour évaluer les auxiliaires à l'enseignement de deuxième et troisième cycles, de nombreux établissements ne s'engagent pas du tout dans cette pratique. Dans les quelques établissements qui le font, les procédures varient d'un département à l'autre et d'une division à l'autre. De plus, les évaluations des étudiants peuvent contribuer au perfectionnement professionnel et à la recherche d'emploi dans le milieu universitaire. Un

nombre croissant d'établissements exigent des dossiers d'enseignement (ou portfolios) des candidats à un poste; des preuves d'expérience et d'expertise en enseignement constituent une part essentielle de ce document.

Actuellement, il existe peu de recherches relatives à l'évaluation des étudiants de deuxième et troisième cycles comme auxiliaires à l'enseignement. L'un des aspects sur lesquels il faudrait se pencher de façon particulière est le type de formulaires d'évaluation et l'étendue des questions à poser. Bien que les formulaires d'évaluation existants puissent fournir des occasions d'interroger les étudiants sur leur expérience avec les auxiliaires à l'enseignement, des questions devraient être formulées pour aborder spécifiquement l'éventail d'activités et de comportements d'enseignement particuliers aux auxiliaires à l'enseignement. Des instruments adaptés au corps professoral, comportant fréquemment des questions sur la structure du cours, les lectures choisies et les examens, ne conviennent généralement pas à l'évaluation des auxiliaires à l'enseignement, étant donné que ces derniers ne participent habituellement pas à ces aspects d'un cours.

Consultez aussi l'annexe G pour avoir un résumé des recommandations pour les recherches futures tirées de Greenwald (1997).

Section 7

Conclusions

Malgré le fait qu'il y a maintenant des milliers d'articles consacrés au thème des évaluations de cours par les étudiants, beaucoup de recherche reste encore à faire. Dans l'étendue des travaux, une attention particulière a été portée aux questions en lien avec la validité et la fiabilité : en fait, une grande majorité d'études et de revues de littérature ont principalement porté sur ces domaines et continuent de le faire. Toutefois, comme nous l'avons démontré dans cette revue, la fiabilité et la validité interne des évaluations de cours sont maintenant largement acceptées par de nombreux chercheurs, comme l'ont prouvé les cotes d'un élément probant fondé. Il est peut-être maintenant temps de porter notre attention vers quelques-unes des autres questions, dont on réclame largement un examen approfondi. Celles-ci comprennent :

- Améliorer l'information et la formation destinées aux utilisateurs des évaluations et des résultats analysés;
- Élaborer et tester des outils servant à l'interprétation ainsi que des moyens efficaces de communiquer les résultats (en relation avec les besoins des utilisateurs);
- S'assurer de l'engagement du corps professoral et des étudiants envers le processus d'évaluation;
- Examiner régulièrement des instruments d'évaluation selon les besoins et les objectifs de l'établissement et en relation avec les résultats de recherche courants.

Alors que les chercheurs doivent réorienter leurs travaux sur la validité des évaluations de cours et s'éloigner de la recherche sur des éléments de sondage individuels pour se tourner vers ces questions plus vastes sur la conception, l'application et l'interprétation des sondages, les établissements doivent aussi adapter leur vision et leur utilisation des évaluations. Ces dernières doivent être conçues de manière à correspondre soigneusement au contexte de l'établissement et accompagnées d'un soutien substantiel de la part de l'établissement. Les politiques doivent être détaillées et équitables. La formation des utilisateurs des évaluations – les étudiants, le corps professoral et les administrateurs – doit chasser les mythes et les perceptions erronées et améliorer le savoir-faire et la transparence. Les évaluations doivent être accompagnées d'un dialogue en continu et de mécanismes de soutien, notamment la consultation, afin d'assurer qu'elles contribuent au soutien et à l'amélioration de l'enseignement dans l'établissement.

Notre recherche a clairement démontré que les évaluations sont des outils valables et importants pour évaluer l'enseignement, à condition qu'elles soient mises au point et soutenues tout en gardant à l'esprit que la validité est déterminée par beaucoup plus que simplement la façon dont les étudiants répondent aux questions individuelles d'un sondage.

Travaux cités (en anglais seulement)

- Abrami, P.C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. In M. Theall, P.C. Abrami, and L.A. Mets (Eds.). The student ratings debate: Are they valid? How can we best use them? [Special issue]. *New Directions for Institutional Research* 109, 59-87.
- Abrami, P.C., d'Apollonia, S., & Cohen, P.A. (1990). Validity of student ratings of instruction: What we know and what we do not know. *Journal of Educational Psychology*, 82(2), 219-231.
- Addison, W.E., Best, J., & Warrington, J.D. (2006). Students' perceptions of course difficulty and their ratings of the instructor. *College Student Journal*, 40(2), 409-416.
- Aleamoni, L.M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, 13(2), 153-166.
- Aleamoni, L.M. (1997). Issues in linking instructional improvement research to faculty development in higher education, *Journal of Personnel Evaluation in Education*, 11(1), 31-37.
- Aleamoni, L. M. (1987). Typical faculty concerns about student evaluation of teaching. *New Directions for Teaching and Learning*, 31, 25-31.
- Aleamoni, L. M. & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science*, 9(1), 67-84.
- Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L., Mohanty, G., & Spooner, F. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching*, 52(4), 134-141.
- Ali, D.L & Sell Y. (1998) *Issues regarding the reliability, validity and utility of student ratings of instruction: A survey of research findings*. Calgary: University of Calgary APC Implementation Task Force on Student Ratings of Instruction.
- Arreola, R.A. (2000). *Developing a comprehensive faculty evaluation system* (2nd ed.). Bolton, MA: Anker.
- Arreola, R.A. (1995). *Developing a comprehensive faculty evaluation system*. Bolton, MA: Anker.
- Arreola, R. A. (1983). Establishing successful faculty evaluation and development programs. In A. Smith (Ed.). Evaluating faculty and staff [Special issue]. *New Directions for Community Colleges*, 41, 83-93.

- Aultman, L.P. (2006). An unexpected benefit of formative student evaluations. *College Teaching*, 54(3), 251.
- Avery, R.J., Bryan, W.K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations? *Journal of Economic Education*, 37(1), 21-37.
- Baldwin, T. & Blattner, N. (2003). Guarding against potential bias in student evaluations: What every faculty member needs to know. *College Teaching*, 51(1), 27-32.
- Beran, T., Violato, C., & Kline, D. (2007). What's the 'use' of student ratings of instruction for administrators? One university's experience. *Canadian Journal of Higher Education*, 17(1), 27-43.
- Beran, T., Violato, C., Kline, D., & Frideres, J. (2005). The utility of student ratings of instruction for students, faculty, and administrators: A "consequential validity" study. *Canadian Journal of Higher Education*, 35(2), 49-70.
- Bernstein, D.J. (2008). Peer review and evaluation of the intellectual work of teaching, *Change*, 40(2), 48-51.
- Bothell, T.W., & Henderson, T. (2003). Do online ratings of instruction make sense? In D.L. Sorenson & T.D. Johnson (Eds.), *Online student ratings of instruction* [Special issue]. *New Directions for Teaching and Learning*, 96, 69-80.
- Braskamp, L.A. & Ory, J.C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco: Jossey-Bass.
- Calderon, T.G., Gabbin, A.L., & Green, B.P. (1996). *Report of the committee on promoting evaluating effective teaching*. Harrisonburg, VA: James Madison University.
- Campbell, J.P. & Bozeman, W.C. (2008). The value of student ratings: Perceptions of students, teachers and administrators. *Community College Journal of Research and Practice*, 32(1), 13-24.
- Cashin, W.E. (1995). *Student ratings of teaching: The research revisited* (IDEA Paper No. 32). Manhattan, KS: Kansas State University Center for Faculty Evaluation and Development.
- Cashin, W.E. (1992). Student ratings: The need for comparative data. *Instructional Evaluation and Faculty Development*, 12(2), 1-6.
- Cashin, W.E. (1990). Students do rate different academic fields differently. In Theall, M. & Franklin, J. (Eds.), *Student ratings of instruction: Issues for improving practice* [Special issue]. *New Directions for Teaching and Learning*, 43, 113-121.

- Cashin, W.E. (1989). *Defining and evaluating college teaching* (IDEA Paper No. 21). Manhattan, KS: Kansas State University Center for Faculty Evaluation and Development.
- Cashin, W.E. (1988). *Student ratings of teaching: A summary of the research* (IDEA Paper No. 20). Manhattan, KS: Kansas State University Center for Faculty Evaluation and Development.
- Cashin, W.E., & Downey, R.G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology*, 84(4), 563-572.
- Centra, J.A. & Gaubatz, N.B. (1998, April). *Is there a gender bias in student evaluation of teaching?* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.
- Centra, J.A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco: Jossey-Bass.
- Centra, J.A. (1979). *Determining faculty effectiveness: Assessing teaching, research, and service for personnel decisions and improvement*. San Francisco: Jossey-Bass.
- Centra, J.A., & Creech, R.F. (1976). *The relationship between student, teacher and course characteristics and student ratings of teacher effectiveness*. Princeton, N.J.: Educational Testing Service.
- Chambers, B.A., & Schmitt, N. (2002). Inequity in the performance evaluation process: How you rate me affects how I rate you. *Journal of Personnel Evaluation in Education*, 16(2), 103-112.
- Cohen, P.A., & McKeachie, W.J. (1980). The role of colleagues in the evaluation of teaching. *Improving College and University Teaching*, 28(4), 147-154.
- Coren, S. (2001). Are course evaluations a threat to academic freedom? In S.E. Kahn & D. Pavlich (Eds.), *Academic Freedom and the Inclusive University* (pp. 104-117). Vancouver: University of British Columbia Press.
- Crosson, A.C., Boston, M., Levison, A., Matsumura, L.C., Matsumura, L.C., Resnick, L.B., et al. (2006). *Beyond summative evaluation: The instructional quality assessment as a professional development tool* (CSE Technical Report 691). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- d'Apollonia, S., & Abrami, P.C. (1997). Navigating student ratings of instruction. *American Psychologist*, 52(11), 1198-1208.
- Davidovitch, N., & Soen, D. (2006). Class attendance and students' evaluation of their college instructors. *College Student Journal*, 40(3), 691-703.

- DeZure, D. (1999). Evaluating reaching through peer classroom observation. In P. Seldin (Ed.), *Changing practices in evaluating teaching. A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 70-96). Bolton, MA: Anker.
- Diamond, M. R. (2004). The usefulness of structured mid-term feedback as a catalyst for change in higher education classes. *Active Learning in Higher Education* 5(3), 217-231.
- Donmeyer, C.J., Baum, P., Hanna, R.W. & Chapman, K.S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education*, 29(5), 611-623.
- Eiszler, C.F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education*, 43(4), 483-501.
- Felder, R. (1993). What do they know anyway? 2: Making evaluations effective. *Chemical Engineering Education*, 27(1), 28-29.
- Feldman, K.A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education*, 30(2), 137-194.
- Feldman, K.A. (1976). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education*, 4(1), 69-111.
- Fox, C.R. (2006). The availability heuristic in the classroom: How soliciting more criticism can boost your course ratings. *Judgment and Decision Making*, 1(1), 86-90.
- Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. In K.G. Lewis (Ed.), *Techniques and strategies for interpreting student evaluations* [Special issue]. *New Directions for Teaching and Learning*, 87, 85-100.
- Franklin, J., & Theall, M. (1990). Communicating student ratings to decision makers: Design for good practice. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice* [Special issue]. *New Directions for Teaching and Learning*, 43, 75-93.
- Franklin, J., & Theall, M. (1989). *Who reads ratings: Knowledge, attitude and practice of users of student ratings of instruction*. Paper presented at the Annual meeting of the American Educational Research Association, San Francisco.
- Goldschmid, M.L. (1978). The evaluation and improvement of teaching in higher education. *Higher Education*, 7(2), 221-245.

- Green, B.P., Calderon, T.G., & Reider, B.P. (1998). A content analysis of teaching evaluation instruments used in accounting departments. *Issues in Accounting Education, 13*(1), 15-30.
- Greenwald, A.G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*(11), 1182-1186.
- Greenwald, A.G., & Gillmore, G.M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209-1217.
- Gump, S.E. (2007). Student evaluations of teaching effectiveness and the leniency hypothesis: A literature review. *Educational Research Quarterly, 30*(3), 55-68
- Hardy, N. (2003). Online ratings: Fact and fiction. In D.L. Sorenson & T.D. Johnson (Eds.), Online student ratings of instruction [Special issue]. *New Directions for Teaching and Learning, 96*, 31-38.
- Harper, S.R., & Kuh, G. (2007). Myths and misconceptions about using qualitative methods in assessment. In S.R. Harper & S.D. Museus (Eds.), Using qualitative methods in institutional assessment [Special issue]. *New Directions for Institutional Research, 136*, 5-14.
- Haskell, R.E. (1997). Academic freedom, tenure, and the student evaluation of faculty: Galloping polls in the 21st century. *Education Policy Analysis, 5*(6).
- Heckert, T.M., Latier, A., Ringwald-Burton, A., & Drazen, C. (2006). Relations among student effort, perceived class difficulty appropriateness, and student evaluations of teaching: Is it possible to “buy” better evaluations through lenient grading? *College Student Journal, 40*(3), 588-596.
- Hodges, L.C., & Stanton, K. (2007). Translating comments on student evaluations into the language of learning. *Innovative Higher Education, 31*, 279-286
- Hoffman, K.M. (2003). Online course evaluations and reporting in higher education. In D.L. Sorenson & T.D. Johnson (Eds.), Online student ratings of instruction [Special issue]. *New Directions for Teaching and Learning, 96*, 25-30.
- Hoyt, D.P., & Pallett, W.H. (1999). *Appraising teaching effectiveness: Beyond student ratings* (IDEA Paper No.36). Manhattan, KS: Kansas State University Center for Faculty Evaluation and Development.
- Johnson, T.D. (2003). Online student ratings: Will students respond? In D.L. Sorenson & T.D. Johnson (Eds.), Online student ratings of instruction [Special issue]. *New Directions for Teaching and Learning, 96*, 49-60.
- Keig, L., & Waggoner, M.D. (1994). Collaborative peer review: The role of faculty in improving college teaching. *ASHE-ERIC Higher Education Report No. 2*. Washington, DC: George Washington University.

- Kulik, J.A. (2001). Student ratings: Validity, utility, and controversy. In M. Theall, P.C. Abrami, & L.A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* [Special issue]. *New Directions for Institutional Research*, 109, 9-25.
- Lang, J.W.B., & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instructional Science*, 35, 187-205.
- Lattuca, L.R., & Domagal-Goldman, J.M. (2007). Using qualitative methods to assess teaching effectiveness. In S.R. Harper & S.D. Museus (Eds.), *Using qualitative methods in institutional assessment* [Special issue]. *New Directions for Institutional Research*, 136, 81-93.
- Lewis, K.G. (2001). Using midsemester student feedback and responding to it. In K.G. Lewis (Ed.), *Techniques and strategies for interpreting student evaluations* [Special issue]. *New Directions for Teaching and Learning*, 87, 33-44.
- Llewellyn, D.C. (2003). Online reporting of results of online student ratings. In D.L. Sorenson & T.D. Johnson (Eds.), *Online student ratings of instruction* [Special issue]. *New Directions for Teaching and Learning*, 96, 61-68.
- Marsh, H.W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology*, 99(4), 775-790.
- Marsh, H.W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11(3), 253-388.
- Marsh, H.W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, 76(5), 707-754.
- Marsh, H.W., & Dunkin, M.J. (1992). Students' evaluations of university teaching: A multidimensional approach. In J.C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 8) (pp. 143-223). New York: Agathon Press.
- Marsh, H.W., & Roche, L. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology*, 92(1), 202-228.
- Marsh, H.W., & Roche, L. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias and utility. *American Psychologist*, 52(11), 1187-1197.
- McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist*, 52(11), 1218-1225.

- McKeachie, W.J., & Lin, W.G. (1975). Multiple discriminant analysis of student ratings of college teachers. *Journal of Educational Research*, 68(8), 300-305.
- Menges, R.J. (2000). Shortcomings of research on evaluating and improving teaching in higher education. In K.E. Ryan (Ed.), *Evaluating teaching in higher education: A vision for the future* [Special issue]. *New Directions for Teaching and Learning*, 83, 5-11
- Moore, S., & Kuol, N. (2005). A punitive tool or a valuable resource? Using student evaluations to enhance your teaching. In G. O'Neill, S. Moore, & B. McMulline (Eds.), *Emerging issues in the practice of university learning and teaching* (pp. 141-148). Dublin: All Ireland Society for Higher Education.
- Murray, H.G. (1987). Acquiring student feedback that improves instruction. In Weimer, M.G. (Ed.), *Teaching large classes well* [Special issue]. *New Directions for Teaching and Learning*, 32, 85-96.
- Naftulin, D.H., Ware, J.E., & Donnelly, F.A. (1973), The Dr. Fox lecture: A paradigm of educational seduction. *Journal of Medical Education*, 48, 630-635.
- Nasser, F. & Fresko, B. (2002). Faculty view of student evaluation of college teaching. *Assessment & Evaluation in Higher Education*, 27(2), 187-198.
- Neumann, R. (2001). Disciplinary differences and university teaching. *Studies in Higher Education*, 26(2), 135-146.
- Newport, J.F. (1996). Rating teaching in the USA: Probing the qualifications of student raters and novice teachers. *Assessment & Evaluation in Higher Education*, 21(1), 17-21.
- Ory, J.C. (2001). Faculty thoughts and concerns about student ratings. In K.G. Lewis (Ed.), *Techniques and strategies for interpreting student evaluations* [Special issue]. *New Directions for Teaching and Learning*, 87, 3-15.
- Ory, J.C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In M. Theall, P.C Abrami, & L.A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* [Special issue]. *New Directions for Institutional Research*, 109, 27-44.
- Penny, A.R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research*, 74(2), 215-253.
- Reid, I.C. (2001). Reflections on using the Internet for the evaluation of course delivery. *The Internet and Higher Education*, 4(1), 61-75.
- Remedios, R., & Lieberman, D.A. (2008). I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal*, 34(1), 91-115.

- Rich, H.E (1976). Attitudes of college and university faculty toward the use of student evaluations. *Educational Research Quarterly*, 1, 17-28
- Ryan, J.J., Anderson, J.A., & Birchler, A.B. (1980). Student evaluation: The faculty responds. *Research in Higher Education*, 12(4), 317-333.
- Schmelkin, L.P., Spencer, K.J. & Gellman, E.S. (1997). Faculty perspectives on course and teacher evaluations. *Research in Higher Education*, 38(5), 575-592.
- Scriven, M. (1981). *Evaluation thesaurus* (3rd ed.). Pt. Reyes, CA: Edgepress.
- Scriven, M. (1995). Student ratings offer useful input to teacher evaluations. *Practical Assessment, Research & Evaluation*, 4(7).
- Sedlmeier, P. (2006). The role of scales in student ratings. *Learning and Instruction*, 16(5), 401-415.
- Seldin, P. (1999). *Changing practices in evaluating teaching. A practical guide to improved faculty performance for promotion/tenure decisions*. Bolton, MA: Anker.
- Seldin, P. (1993). *Successful use of teaching portfolios*. Bolton, MA: Anker.
- Sorenson, D.L. & Reiner, C. (2003). Charting the uncharted seas of online student ratings of instruction. In D.L. Sorenson & T.D. Johnson (Eds.), *Online student ratings of instruction* [Special issue]. *New Directions for Teaching and Learning*, 96, 1-24.
- Sproule, R. (2000). Student evaluation of teaching: A methodological critique of evaluation practices. *Education Policy Analysis*, 8(50).
- Theall, M. (2002). *Student ratings: Myths vs. research evidence*. Retrieved June 26, 2008, from <https://studentratings.byu.edu/info/faculty/myths.asp>.
- Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? In M. Theall, P.C Abrami, & L.A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* [Special issue]. *New Directions for Institutional Research*, 109, 45-56.
- Theall, M., & Franklin, J. (2000). Creating responsive student ratings systems to improve evaluation practice. In K.E. Ryan (Ed.), *Evaluating teaching in higher education: A vision for the future* [Special issue]. *New Directions for Teaching and Learning*, 83, 95-107.
- Theall, M. (1994). What's wrong with faculty evaluation: A debate on the state of practice. *Instructional Evaluation and Faculty Development*, 14, 27-34
- Wagenaar, T.C. (1995). Student evaluation of teaching: Some cautions and suggestions. *Teaching Sociology*, 23(1), 64-68.

- Wachtel, H.K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, 29(2), 191-121.
- Williams, W.M. & Ceci, S.J. (1997). How'm I doing? Problems with student ratings of instructors and courses. *Change*, 29(5), 13-23.
- Williams, R and Ory, J.C. (1992). *A further look at class size, discipline differences and student ratings*. Urbana-Champaign: Office of Instructional Resources, University of Illinois at Urbana-Champaign.
- Wright, R.E. (2006). Student evaluations of faculty: Concerns raised in the literature, and possible solutions. *College Student Journal*, 40(2), 417-422.
- Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education*, 12(1), 55-76.

