REVUE FRANÇAISE DE PÉDAGOGIE

Revue française de pédagogie
Recherches en éducation

165 | octobre-décembre 2008
L'éducation et les politiques de la mémoire

L'Évaluation des enseignements par les étudiants : état de la recherche et perspectives

Assessment of teachers by students: progress of research and its current standpoints

Evaluación de las docencias por los estudiantes : estado de la investigatión y perspectiva

Die Evaluierung des Unterrichtstoffs durch die Studenten: Zustand der Forschung und Perspektiven

Pascal Detroz

revues.org

Éditeur
ENS Éditions

Édition électronique
URL : http://rfp.revues.org/1165
DOI : 10.4000/rfp.1165
ISSN : 2105-2913

Édition imprimée
Date de publication : 1 octobre 2008
Pagination : 117-135
ISBN : 978-2-7342-1143-3
ISSN : 0556-7807

Référence électronique
Pascal Detroz, « L'Évaluation des enseignements par les étudiants : état de la recherche et perspectives », *Revue française de pédagogie* [En ligne], 165 | octobre-décembre 2008, mis en ligne le 01 octobre 2012, consulté le 01 février 2017. URL : http://rfp.revues.org/1165 ; DOI : 10.4000/rfp.1165

NOTE DE SYNTHÈSE

L'Évaluation des enseignements par les étudiants : état de la recherche et perspectives

Pascal Detroz

L'Évaluation des enseignements par les étudiants (EEE) est une pratique en extension à travers le monde. Depuis cinquante ans, la littérature, essentiellement anglo-saxonne, s'interroge sur la fidélité et la validité de la mesure recueillie à travers ce dispositif. Le présent article se propose de synthétiser les débats en cours. Ils amènent à la conclusion que les divers chercheurs n'ont pu établir, de manière univoque, décontextualisée et formelle, la preuve que ces critères de qualité sont atteints. Toutefois, de nombreuses présomptions allant dans ce sens ont pu être mises à jour. En conséquence, cet article présente un certain nombre de précautions qui doivent être prises lorsque l'on souhaite mettre en place une EEE de qualité.

Descripteurs (TEE) : évaluation des enseignements, enseignement supérieur, fidélité, validité, validité conséquentielle.

INTRODUCTION

Historiquement, l'évaluation des enseignements (1) est une pratique ancienne. Selon J. J. O. Doyle (1983), on en trouve déjà des traces en 350 avant Jésus-Christ. Si certaines études sur ce sujet ont été publiées à la fin du XIX^e^ siècle (2), c'est surtout dans le courant du XX^e^ siècle que ce champ théorique s'est étoffé.

Les pratiques d'évaluation des enseignements, centrées sur l'avis des étudiants ont, quant à elles, été décrites plus tardivement. Si, en 1924, un groupe d'étudiants de l'université de Harvard publie le *Confidential Guide to Course*, probablement le premier répertoire de cours évalués (Bernard, 1992), il faudra attendre le début des années soixante pour que la pratique d'EEE se généralise. Cette généralisation se fera d'abord lentement, puis de manière plus spectaculaire à la fin du XX^e^ siècle. Ainsi, une étude longitudinale de P. Seldin (1993) portant sur

600 collèges américains montre que 29 % de ceux-ci utilisaient l'EEE en 1973, qu'ils étaient 68 % à la pratiquer en 1983 et que ce chiffre atteignait les 86 % en 1993.

En Europe, et plus particulièrement dans les pays francophones, l'évaluation des enseignements par les étudiants (EEE), que nous définissons comme le fait de recueillir l'avis des étudiants sur la qualité des enseignements qu'ils ont suivis afin de permettre un jugement menant à des régulations, se diffuse plus lentement. Par exemple, en Belgique francophone, c'est seulement dans le milieu des années quatre-vingt que ce type de méthodologie a été mis en œuvre au niveau institutionnel (Remy, 1994). Avant cela, l'utilisation de ces méthodes était l'apanage de certains cercles d'étudiants ou d'enseignants isolés. En France, J. Dejean (2002) conclut, dans son rapport sur l'évaluation de l'enseignement dans les universités, que cette forme d'évaluation est encore peu développée et a du mal à s'imposer. De manière générale, le rythme de diffusion de cette méthode s'est très largement accéléré partout en Europe grâce au processus de Bologne, c'est-à-dire à partir des années deux mille.

Parallèlement à ces pratiques de terrain, de nombreux articles ont été publiés depuis les années soixante. Ainsi, W. E. Cashin (1995) a répertorié plus de 1 500 références scientifiques portant sur l'évaluation de l'enseignement par les étudiants (*student ratings).* Cet article vise à présenter l'état de la recherche sur ces deux aspects. Nous sommes conscients que l'efficacité d'un dispositif EEE ne peut pas être réduite uniquement à ses aspects psychométriques. Toutefois, lorsque le dispositif d'EEE est remis en cause par des porteurs d'enjeux réfractaires, c'est souvent au sujet de la qualité de la mesure que les critiques se font les plus virulentes.

Très tôt, des recherches crédibles démontrent la validité de l'EEE. C'est ce qui a permis à Cohen de conclure, dès 1981, à la validité de l'EEE après une méta-analyse portant sur 41 études indépendantes. Pourtant, malgré le sérieux de ces travaux qui démontrent la validité et la fidélité de la procédure, le scepticisme reste souvent de mise parmi les enseignants évalués. Ce qui fait dire à P. A. Cohen (1990) que « les attitudes négatives au sujet de l'EEE sont spécialement résistantes au changement et qu'il semble que les enseignants et les administrateurs renforcent leurs croyances en certains mythes relatifs à l'EEE par des éléments personnels et anecdotiques qui, pour eux, ont plus de poids que les preuves apportées par la recherche empirique (3) ». En conséquence, les chercheurs-praticiens en EEE ont dû, et doivent encore, régulièrement convaincre et faire la preuve de la qualité des mesures effectuées à l'aide de cette méthode. Ceci explique le très large champ théorique consacré à cet aspect. À ce titre, il est d'ailleurs symptomatique de remarquer que L. M. Aleamoni a publié en 1987 un article portant le titre *Student rating myths versus research facts* et qu'un peu plus tard M. Theall (2002) a produit une note interne dans son université s'intitulant *Student Ratings: Myths vs. Research Evidence*.

Ce combat pour convaincre n'est toujours pas gagné. Le sondage de W. A. Wright et M. C. O'Neil (1995), auprès de différentes institutions universitaires (canadiennes, américaines, australiennes et européennes) sur les stratégies permettant l'amélioration de l'enseignement, démontre que, parmi 36 stratégies proposées, l'EEE est classée par les administrateurs en 34^{e} position en termes de préférence.

À première vue, donc, il semblerait que nous soyons face à une situation antagoniste et paradoxale avec, d'une part, une recherche validant pleinement l'EEE et, d'autre part, des enseignants rejetant nettement cette pratique.

Une analyse plus fine de la documentation scientifique radicalise cependant moins ces positions. La preuve en est, notamment, les études francophones approfondies sur les représentations/attitudes des enseignants par rapport à l'EEE (Bernard, 2000 ; Younes, 2006) qui concluent que ces dernières ne sont pas univoques et sont souvent ambivalentes sur certains aspects. Dans l'étude de H. Bernard, par exemple, un peu plus de la moitié des enseignants accepte le fait que « l'évaluation faite par les étudiants constitue une opération utile qui aide à identifier les points forts et les aspects à améliorer de son enseignement ». Par ailleurs, la littérature n'est pas non plus univoque quant à la qualité de la mesure récoltée par l'EEE. À la lumière de dizaines d'articles traitant du sujet, il semble que certaines critiques méritent réflexion. Selon nous, ces critiques ne devraient pas mener au rejet pur et simple de la méthode, mais plutôt inciter à la prudence et, à tout le moins, à adopter quelques principes de précaution.

Cet article vise à dresser un tableau de la littérature portant sur la validité et la fidélité de l'EEE. Cette centration est liée à l'état de la recherche : la littérature anglo-saxonne, majoritaire dans ce domaine, se concentre essentiellement sur les caractéristiques psychométriques du recueil d'informations et s'intéresse plus rarement aux jugements et à la régulation souhaitée, que celle-ci soit à visée formative ou à visée administrative. Autrement dit, les recherches que nous soumettons ici à une discussion critique, portent sur la qualité de la mesure effectuée lorsqu'on recueille l'avis des étudiants, le plus souvent par questionnaires, sur les enseignements qu'ils ont reçus. L'objectif de la présente note de synthèse est de mettre en lumière le débat d'idées et de faire comprendre au lecteur les difficultés méthodologiques liées à ce type d'études. Nous proposerons ensuite certaines précautions qui nous semblent devoir découler des divers constats que nous dresserons et nous aborderons quelques perspectives de recherche qui nous semblent fécondes.

LA FIDÉLITÉ DE LA MESURE

Pour être fidèle, un questionnaire doit recueillir des résultats strictement identiques s'il est appliqué à un groupe de sujets similaires, dans un contexte similaire. La fidélité est donc définie par certains auteurs comme un critère de qualité important en termes d'EEE. La littérature décrit trois principales fidélités : la stabilité, l'équivalence et la consistance interne (Cohen *et al*., 2007).

Comme son nom l'indique, l'indice de stabilité vérifie que l'information récoltée est stable à travers le temps, mais aussi qu'elle sera identique pour des échantillons semblables. Le concept d'équivalence peut prendre deux formes distinctes. Il peut s'agir de mesurer l'équivalence entre deux formes parallèles d'un même questionnaire, ou alors, de vérifier la concordance intercorrecteurs. Dans ce deuxième cas de figure, un test sera valide si deux correcteurs différents (ou un même correcteur à deux moments différents) obtiennent les mêmes résultats finaux. Enfin, la notion de consistance interne, aussi appelée homogénéité, analyse dans quelle mesure l'ensemble des items d'un questionnaire (ou d'une de ses parties) a bien trait à la même entité sous-jacente. Dans le cadre de l'EEE, il est important de s'assurer que l'ensemble des items est bien en phase avec le *construct* sous-jacent que l'on ambitionne de mesurer : la qualité de l'enseignement. La fidélité est donc importante en EEE, en ce sens qu'elle est le garant de la consistance et de la réplicabilité des résultats.

En termes de stabilité de l'EEE, H. W. Marsh (1987), relatant de nombreuses études, conclut que la fidélité, quelle que soit la manière dont on la mesure (dans la classe, entre les classes, à des moments différents ou de toute autre façon) est remarquablement constante. La fidélité intercorrecteurs de l'EEE par questionnaire a également été abordée par une série d'auteurs. H. W Marsh (1987) a, par exemple, constaté que celle-ci augmentait avec le nombre de juges. Selon son étude, cette fidélité était de .60 pour 5 étudiants et montait jusqu'à .95 pour 50 étudiants. Suite à cette étude, W. E. Cashin (1995) recommande de n'utiliser l'EEE par questionnaire que dans les groupes supérieurs à 10 étudiants. Enfin, en ce qui concerne la consistance interne du questionnaire, elle dépend des items qui le constituent et varie donc d'un questionnaire à l'autre. Elle doit donc être mesurée au cas par cas. L. M. Aleamoni (1987) publie toutefois une liste de questionnaires obtenant une fidélité supérieure à .90. Il s'agit de l'*Educational Testing Service Instrument (ETS)*, du *Student Instructional Report*, de l'*Instructional Developement and Effectiveness Assessment* et du *Arizona Teachers Course Evaluation Questionnaire*. F. Costin, W. T. Greenough et R. J. Menges (1971) précisent que, d'après leur expérience, lorsque les questionnaires ont été construits de manière rigoureuse, la fidélité est de .90.

L'importance accordée par les différents auteurs au concept de fidélité, associé au dispositif EEE, s'explique par le fait que ceux-ci – anglo-saxons pour la plupart – ont été très influencés par des dispositifs d'évaluation visant la certification à l'aide de références normatives. Cet éclairage explicatif est important. Sinon, comment expliquer que le critère de fidélité, pour ces auteurs, rende compte de la qualité de la mesure de l'efficacité de l'enseignement qui est une action multidimensionnelle par essence ? Et comment expliquer qu'une certaine stabilité dans la mesure soit recherchée alors qu'idéalement ces différents indicateurs devraient être sensibles à des micro-variations (dans le temps ou au sein d'un groupe) ?

LA VALIDITÉ DE LA MESURE

L'essentiel de la littérature du XX^{e} siècle a porté sur la validité de la mesure récoltée via l'EEE. P. C. Abrami (2001), influencé par les théories classiques de la mesure, donne la définition suivante de la validité : « Un score valide est une valeur hypothétique qui représente le mieux la dextérité, l'habilité ou l'attribut d'une personne. Une valeur issue de la production de connaissances consistante sur les différences individuelles sans être influencée par des [...] facteurs qui contribuent à des résultats imprécis et instables ». Cette notion est extrêmement importante puisque, selon J. C. Ory et K. Ryan (2001), elle s'intéresse à la relation entre la/les mesure(s) issue(s) de l'EEE et le *construct* sous-jacent évalué, à savoir « la qualité » de l'enseignement. Il s'agit là d'une définition classique de la validité, basée sur le point de vue qu'il est essentiel de démontrer qu'un instrument donné mesure effectivement ce qu'il se propose de mesurer.

Le champ théorique se concentrant sur la validité dans son acception classique comprend plusieurs centaines d'articles. Et leurs conclusions ne sont pas toujours concordantes. Ainsi A. G. Greenwald et G. M. Gillmore (1997), qui ont catégorisé les recherches sur la validité des EEE, dressent, d'un côté, la liste de celles qui plaident en faveur de sa validité (Cashin, 1995 ; Cohen, 1981 ; Feldman, 1988 ; Howard, Conway et Maxwell, 1985 ; Howard et Maxwell, 1980, 1982 ; Marsh, 1980, 1982, 1984 ; Marsh et Dunkin, 1992 ; McKeachie, 1979), de l'autre côté, ils listent celles qui mettent en cause un manque de validité de cette

procédure (Dowell et Neal, 1982 ; Holmes, 1972 ; Powell, 1977 ; Snyder et Clair, 1976 ; Vasta et Sarmiento, 1979 ; Worthington et Wong, 1979).

Comme A. G. Greenwald et G. M. Gillmore (1997) le mettent en évidence, il semble que la littérature concernant la validité de l'EEE soit au moins partiellement contradictoire. Ceci s'explique par la difficulté méthodologique d'évaluer la validité de l'EEE. En effet, sans une définition univoque de ce qu'est un enseignement de qualité, les nombreux auteurs ont dû recourir à des méthodologies plus ou moins complexes, dont certaines ont été vivement critiquées. Cette complexité est soulignée par P. C. Abrami, S. d'Apollonia et P. Cohen (1990) qui ne répertorient pas moins de 5 grandes méthodologies de recherche portant sur la validité de l'EEE qui sont [a] l'approche multitrait-multiméthode, [b] l'approche multisection, [c] l'analyse des biais, [d] les méthodes de laboratoire et [e] l'approche multidimensionnelle. Passons-les en revue rapidement.

L'approche « multitrait-multiméthode »

Elle consiste à corréler les résultats obtenus par un enseignement lors de l'EEE à d'autres mesures critériées de la qualité de cet enseignement (multitrait) ou à des résultats obtenus en utilisant d'autres méthodes de recueil de données (multiméthode). L'aspect multitrait consiste typiquement à comparer le score d'un enseignement à l'EEE avec, par exemple, le résultat obtenu à l'examen par les étudiants qui l'ont évalué (nous verrons plus loin que ce type d'indicateur doit être remis en question). Cette comparaison peut également être faite avec l'évaluation d'anciens étudiants, l'évaluation par les observateurs ou l'évaluation par les pairs. Pour l'aspect multiméthode, il s'agit d'utiliser différentes méthodes de recueil de données et de les comparer. Par exemple, un enseignant aura-t-il une évaluation aux conclusions identiques à celles fournies par l'EEE, quel que soit le mode de recueil de données (questionnaire, *focus group*,…) ? Les études les plus sérieuses qui se sont penchées sur ce sujet concluent en faveur d'une validité convergente et discriminante de l'EEE (Howard, Conway et Maxwell, 1985 ; Murray, 1984 ; Marsh, 1982 ; Aleamoni et Hexner, 1980).

La problématique inhérente à ce type de recherche est toutefois la validité des indicateurs comparés. Le plus souvent, il s'agit du score obtenu par les étudiants à l'examen ou de l'évaluation par les pairs. L'un et l'autre indicateurs soulèvent beaucoup de questions, car l'un et l'autre n'offrent que des mesures incomplètes et imparfaites de ce que l'on souhaite appréhender : « la qualité » de l'enseignement.

Ainsi, le score obtenu par les étudiants à l'examen ne peut être considéré comme le simple reflet de l'enseignement reçu par ceux-ci. Plusieurs remarques s'imposent à cet égard :

- cet indicateur ne tient, en général, pas compte des compétences des étudiants à l'entrée dans le cursus d'apprentissage. De manière caricaturale, si un étudiant connaît préalablement la matière, il peut afficher un excellent score à l'examen sans avoir rien appris au cours. Les pédagogues peuvent contourner ce problème en recourant au gain relatif (GR = Gain effectué/Gain possible, Mc Guigan, 1967). Mais cette méthode exige un pré-test et un post-test exactement de la même difficulté, ce qui nécessite souvent le recours à des techniques difficilement applicables comme l'*item Response Théory* ;
- ce gain relatif peut être influencé par des variables externes à l'enseignement. Ainsi, P. Ramsden et N. J. Entwistle (1981) ont démontré que l'ambiance de la section ou du département influence les apprentissages ;

- on sait, depuis H. Piéron (1963), à quel point le résultat des étudiants aux examens est soumis à un manque de fidélité inter-juges et intra-juge. Le score obtenu ne reflète donc pas uniquement la performance de l'étudiant, il est contaminé par des erreurs de mesure liées au manque de fidélité de la correction ;
- souvent, l'examen manque de validité de contenu et/ou de validité de *construct*. En d'autres mots, les questions de l'examen ne sont souvent qu'un échantillon peu représentatif de l'ensemble du contenu ou des processus mentaux abordés pendant le cours ;
- les examens, la plupart du temps, ont lieu juste après les enseignements et ne présagent en rien de l'apprentissage à long terme, par exemple, du transfert qui pourra être réalisé dans un contexte professionnel.

Le score obtenu par les étudiants à l'examen est donc une mesure imparfaite, souvent incomplète qui n'est affectée que partiellement par l'enseignement dispensé (4). En conclusion, l'approche multitrait-multiméthode révèle une bonne corrélation entre l'EEE et d'autres mesures critériées de la qualité de l'enseignement. On ne peut toutefois pas s'appuyer sur ce résultat pour conclure que l'EEE mesure complètement, parfaitement et exclusivement la qualité de l'enseignement.

Renversant en quelque sorte le raisonnement, les auteurs ayant contribué au numéro thématique de l'*American psychologist* (Greenwald, 1997 ; Marsh et Roche, 1997 ; d'Apollonia et Abrami, 1997 ; Greenwald et Gillmore, 1997 ; McKeachie, 1997) soutiennent que l'EEE est la source d'informations la plus valide concernant la qualité de l'enseignement puisque, contrairement aux autres sources de données, pour lesquelles il y a peu de preuves de la validité, les études sur l'EEE apportent certains arguments allant dans le sens de la validité de cette méthode.

Les études multisection

Elles sont utilisées pour mesurer la relation entre l'évaluation des enseignements et le score des étudiants dans des chapitres différents d'un même cours donné par des enseignants différents. Cette méthodologie présente l'avantage de contrôler les variables inhérentes au contexte et celles inhérentes à l'étudiant, puisque ce sont les mêmes cours suivis par les mêmes étudiants. Au total, ces études montrent une corrélation entre les résultats et le jugement des étudiants qui va dans le sens de la validité de l'EEE, puisque les cours où les étudiants ont les meilleurs scores sont aussi ceux qu'ils évaluent le plus favorablement (Abrami, d'Appolonia et Cohen, 1990 ; d'Appolonia et Abrami, 1997). Cette méthodologie n'échappe toutefois pas à l'analyse que nous avons effectuée précédemment concernant l'évaluation des étudiants au moyen de tests.

Les études portant sur les biais

Elles partent du principe que, pour être valide, l'évaluation des enseignements par les étudiants doit évaluer la qualité de l'enseignement, sans être influencée par d'autres variables exogènes. Par exemple, si le score obtenu par un enseignant était influencé par certains traits de sa personnalité qui, par ailleurs, seraient sans effet sur les apprentissages, on pourrait en conclure que ce score ne reflète pas uniquement la qualité de l'enseignement. Ce serait la preuve d'un manque de validité de l'EEE. Les conditions d'une invalidité sont donc que, premièrement,

les variables exogènes soient corrélées avec l'EEE et que, deuxièmement, ces variables exogènes ne soient pas corrélées avec la qualité de l'enseignement.

Beaucoup de recherches, mettant en œuvre des méthodes très variées (de la simple corrélation à la méta-analyse), ont tenté de mettre à jour de telles influences. J. C. Ory (2001) souligne que ces recherches s'intéressent à trois grands types de biais qui sont les biais liés aux caractéristiques de l'enseignant, les biais liés aux caractéristiques des étudiants et les biais liés au contexte de l'EEE.

En ce qui concerne *les biais liés à l'enseignant*, divers chercheurs se sont demandé si certaines caractéristiques biographiques ou de personnalité de l'enseignant n'influençaient pas systématiquement les résultats obtenus à l'EEE. Il apparaît que certaines de ces variables biographiques sont légèrement corrélées avec les scores EEE. Ainsi, les enseignants *expérimentés* obtiennent un meilleur score que les enseignants *novices* (Feldman, 1983) et les *enseignants nommés* un meilleur score que les *assistants* (Brandenburg, Slinde et Batista, 1977 ; Centra et Creech, 1976). D'autres variables, telles que l'*âge de l'enseignant*, n'ont pas d'effets (Ory, 2001).

L'effet de la variable *genre* sur l'EEE a été longuement discuté dans la littérature. À ce propos, il semblerait que les auditoires masculins évaluent plus favorablement les hommes et que les auditoires féminins évaluent plus favorablement les femmes. Cet effet est cependant marginal (Bennett, 1982 ; Bernard, Keefauver Elsworth et Maylor, 1981 ; Feldman, 1992). Il semblerait également que les enseignantes obtiennent un meilleur score que les enseignants sur certaines variables, comme par exemple *l'attention portée aux progrès des étudiants* (Feldman, 1983). Par ailleurs, S. K. Bennet (1982), cité par N. Younes (2006), a mis en évidence que les étudiants seraient plus exigeants par rapport aux comportements, étiquetés par eux comme « étant féminins », (ex. : disponibilité envers les étudiants) chez les femmes que chez les hommes. En d'autres termes, un même comportement pourrait être perçu différemment selon le genre de l'enseignant. H. W. Marsh et L. A. Roche (1997) concluent toutefois que le genre de l'enseignant n'a que peu ou pas d'effet sur le score à l'EEE. C'est aussi la conclusion de plusieurs auteurs qui ont effectué un relevé soigné de la littérature sur ce sujet et qui concluent qu'il n'y a pas de relation forte ou régulière entre sexe et EEE (Centra et Gaubatz, 2000 ; Feldman, 1992).

L'ethnie de l'enseignant, quant à elle, a très peu été investiguée. Des études récentes (Hamermesh et Parker, 2005 ; Glascock et Ruggerio, 2006) montrent que, toutes choses égales par ailleurs, les enseignants faisant partie de minorités ethniques ont tendance à être crédités de scores inférieurs à ceux obtenus par les enseignants issus des majorités ethniques.

Selon la littérature, il semblerait donc que, parmi les variables biographiques de l'enseignant, seuls l'ancienneté, le statut et l'ethnie soient légèrement corrélés au score à l'EEE. De nombreux auteurs se sont également penchés sur les traits de personnalité en lien avec la *popularité* de l'enseignant et de son rapport avec l'EEE. La plupart de ces études montrent que la popularité influence, à des degrés divers, le jugement des étudiants par rapport à l'enseignement dispensé (Aleamoni, 1987 ; Feldman, 1978 et Theall et Franklin, 1990).

Au-delà des corrélations, la plupart du temps très faibles, entre ces variables et l'EEE, il est nécessaire de s'interroger sur la signification de ces résultats de recherche et, en particulier, de se demander s'il s'agit de biais. Tendanciellement, un enseignant expérimenté, nommé et populaire, recevrait une meilleure évaluation qu'un jeune assistant impopulaire. Cet effet ne montre-t-il pas tout simplement

qu'un enseignant expérimenté prodigue un meilleur enseignement qu'un novice, qu'un enseignant nommé est plus efficace que celui qui ne l'est pas encore et que la popularité est étroitement liée à la capacité d'enseigner aux étudiants ?

Certains auteurs vont dans ce sens. Ceux-ci appuient leur argumentation sur le fait que la recherche montre que les étudiants apprécient les enseignants compétents, chaleureux, extravertis et enthousiastes, ces caractéristiques améliorant également l'enseignement (Murray, Rushton et Paunonen, 1990). En 1983, H. G. Murray écrivait d'ailleurs à ce sujet que : « *expressive teaching behaviors serve to communicate the lecturer's enthusiasm for the subject matter and, thereby elicit and maintain student attention to lecture material* ».

Certaines études prêtent plus à la controverse. C. R. Emery (1995) (5), par exemple, dans une étude non publiée, semble avoir démontré que les enseignants qui amenaient de la nourriture en classe recevaient le meilleur score à l'EEE. Plus sérieusement, P. C. Abrami, L. Leventhal et R. P. Perry (1982) ont montré que certaines caractéristiques de l'expressivité de l'enseignant avaient un effet sur son score à l'EEE, mais n'avait pas d'influence sur la qualité des travaux réalisés. Les études en laboratoire se sont également largement penchées sur l'effet potentiel du style de l'enseignant sur le score à l'EEE. Nous y reviendrons dans la section suivante.

Parmi les *biais liés aux étudiants*, ont été étudiés les scores obtenus ou espérés à l'examen par les étudiants, leur motivation pour la matière et leur personnalité. Les liens entre le *score obtenu* ou espéré par les étudiants *à l'examen* et leur jugement sur la qualité de l'enseignement reçu ont fait l'objet d'une très large littérature. Il semble judicieux de s'y attarder et d'adopter à leur propos un point de vue historique. En effet, cette perspective historique permettra de mettre en évidence la difficulté méthodologique inhérente à l'étude des biais et la difficulté à interpréter les résultats obtenus.

La première étude sur ce thème qui mérite d'être mentionnée est celle de M. Rodin et B. Rodin (1972), non pas pour sa qualité scientifique, mais surtout pour sa signification historique. Publiée dans la célèbre revue *Science*, cette étude conclut de façon interpellante : les auteurs y présentent une corrélation négative de –.75 entre le score à l'examen (*grading*) et le score obtenu à l'EEE (*rating*). À ce jour, c'est la seule étude à montrer une corrélation négative de cette ampleur. Si elle a eu le mérite d'ouvrir un nouveau champ de recherche, cette étude a subi depuis un lot important de critiques. Ainsi, K. O. Doyle (1975) écrit à son sujet une réplique cinglante : « L'attention accordée à l'étude de M. Rodin et B. Rodin semble proportionnelle à son manque de rigueur » (6) (p. 59).

Plus sérieuse, la méta-analyse menée par P. A. Cohen (1981) sur le sujet conclut à une corrélation moyenne de .43 entre le score à l'examen et les réponses à l'EEE. Ceci dit, toutes les études recensées ne vont pas dans le même sens. Ainsi, certaines de celles-ci aboutissent à une corrélation légèrement négative. P. A. Cohen (1981) analyse les raisons de ce manque de constance et met en avant deux facteurs explicatifs. Le premier est lié aux items du questionnaire EEE. En effet, P. A. Cohen montre que les items liés à la capacité de l'enseignant et à l'organisation du cours sont très corrélés avec le score à l'examen, alors que ceux liés au rapport élève-enseignant le sont faiblement et ceux liés à la difficulté du cours ne le sont pas du tout. Le second est lié à certaines variables contextuelles. Ainsi, les variables « professeur à temps plein », « résultat à l'examen obtenu ou non lors de l'EEE » et « évaluation interne ou externe » influencent significativement la corrélation.

Les études de A. G. Greenwal et G. M. Gillmore (1997) vont dans le même sens : ces chercheurs concluent que, toutes choses égales par ailleurs, plus le score à l'examen augmente plus le score à l'EEE augmente. Ils y voient la preuve que l'EEE n'est pas valide. Selon eux, il suffirait de diminuer les exigences à l'examen pour obtenir une bonne évaluation. H. W. Marsh et L. A. Roche (1997) ne partagent pas cette analyse : ils soulignent que ce n'est pas parce qu'il y a une corrélation entre score à l'examen et réponse à l'EEE qu'il y a une causalité. Ils argumentent qu'il est probable qu'une variable causale, la qualité de l'enseignement, influence simultanément le score à l'examen et les réponses à l'EEE. L'étude de J. Franklin, M. Theall et L. Ludlow (1991) constitue également un argument majeur contre l'hypothèse énoncée par A. G. Greenwald et G. M. Gillmore (1997). En effet, ils analysent des milliers de cours pendant les cinq premières années de l'introduction de l'EEE. Ils mettent en évidence une augmentation faible, mais significative des notes et, dans un même temps, une diminution des scores à l'EEE. Augmenter le score à l'examen ne garantit donc pas une meilleure évaluation à l'EEE.

Il importe de rappeler que l'analyse des corrélations entre scores à l'examen et résultats obtenus lors d'une EEE est au cœur des études multitrait-multiméthode. Dans ce contexte, si un lien est avéré, cela renforce la validité de l'EEE. Les chercheurs qui s'inscrivent dans le courant de l'analyse des biais, font l'interprétation inverse : un tel lien diminue la validité de l'EEE car il peut s'interpréter comme le signe que les enseignants cotant largement les examens sont aussi ceux qui obtiennent les meilleurs scores à l'EEE.

Ces interprétations divergentes ont été discutées par P. C. Abrami (2001). Il explique que cette opposition de points de vue est inhérente au score à l'examen qui est le reflet de deux facteurs combinés. Le premier est l'apprentissage à l'intérieur du cours. Le second est la sévérité de l'enseignant. Il note aussi que le score à l'examen reflète également les compétences transversales des étudiants. J. A. Kulik (2001), pour sa part, met en avant les problèmes de validité inhérents aux examens. Comme on ne peut maîtriser ces différentes sources, P. C. Abrami (2001) préconise d'avoir recours à l'expérimentation en laboratoire ou aux études de terrain. P. C. Abrami, W. J. Dickens, R. P. Perry, L. Leventhal (1980) ont mené des expériences de ce type. Ils y ont démontré, dans plusieurs études de laboratoire que, si la sévérité de l'enseignant influençait le score obtenu à l'EEE, le degré et la direction de cette influence dépendaient de variables contextuelles. Par exemple, une politique de score clémente envers les étudiants augmentait le score à l'EEE pour les enseignants très expressifs, mais le diminuait pour les enseignants peu expressifs. De même, la sévérité de l'enseignant avait une influence plus grande pour les EEE où il y avait peu d'enjeu. À l'inverse, lorsque l'EEE avait un enjeu fort, par exemple, dans le cadre de nomination d'enseignants, l'effet de la sévérité était très faible. En conclusion, ils spécifient que l'effet de la sévérité est, dans des conditions idéales de passation, faible et inconsistant.

La relation entre la *motivation* initiale des étudiants à suivre le cours et leur jugement lors de l'EEE a été analysée par diverses études. Celles-ci montrent que les étudiants ayant un intérêt prononcé pour une matière enseignée évaluent légèrement plus favorablement les cours s'y rapportant que ceux n'ayant pas cet intérêt (Marsh et Cooper, 1981 ; Ory, 1980 ; Perry, Abrami, Leventhal et Check, 1979). H. W Marsh (1983) signale toutefois que la variable « intérêt pour la matière » influence à la fois les apprentissages et le score à l'EEE. Un effet de colinéarité est donc à craindre. Il serait dès lors intéressant de considérer l'effet de la variable de la motivation, à la lumière d'une étude longitudinale. Comment la

motivation de l'étudiant pour la matière varie-t-elle tout au long du cours et quel est, au final, l'impact de cette variation sur l'EEE ?

En ce qui concerne la *personnalité de l'étudiant*, les études de P. C. Abrami, R. P. Perry et L. Leventhal (1982) concluent qu'il n'y a pas de liens consistants et porteurs de sens entre des traits de personnalité de l'étudiant et ses évaluations.

Enfin, un certain nombre de *biais liés à des variables contextuelles* ont été investigués. Ces dernières sont relatives au caractère obligatoire du cours, au niveau d'enseignement auquel il s'adresse, à la taille de la classe, à la discipline enseignée, à la charge de travail donnée aux étudiants et aux méthodes de passation de l'EEE.

Les liens entre EEE et le *caractère optionnel ou obligatoire du cours* ont été questionnés par de nombreux auteurs (Costin, Greenough et Menges, 1971 ; Brandenburg, Slinde et Batista, 1977 ; Feldman, 1978 ; McKeachie, 1979 ; Marsh, 1984 ; Aleamoni, 1989). Tous concluent que les cours obligatoires obtiennent une évaluation inférieure à celle des cours optionnels. Même si la part de la variance expliquée par cette variable est minime, D. C. Brandenburg, J. A. Slinde et E. E. Batista (1977) recommandent de rédiger des normes différentes pour les cours obligatoires et optionnels.

L'impact du *niveau d'enseignement,* dans lequel est donné le cours, sur l'EEE a été exploré par F. Costin, W. T. Greenough et R. J. Menges (1971), L. M. Aleamoni et N. H. Graham (1974), R. B. Bausell et C. R. Bausell (1979), K. A. Feldman (1978), J. A. Kulik et W. J. McKeachie, (1975), J. J. O. Doyle (1983) et par L. M. Aleamoni (1989). Même si les corrélations sont peu élevées, tous signalent, qu'en règle générale, un enseignant donnant cours dans les niveaux supérieurs aura une évaluation plus élevée qu'un enseignant dans un niveau inférieur.

La *taille de l'auditoire* a également été étudiée. Les enseignants pensent souvent qu'enseigner à un petit groupe leur permet d'être plus performants que face à de grands groupes. Les études ne démontrent pourtant pas ce fait. Ainsi, K. A. Feldman (1978), dans une revue portant sur 52 études réalisées dans des auditoires de taille différente trouve une corrélation de –,09 entre la taille de l'auditoire et le score à l'EEE. La corrélation ne semble cependant pas être l'indicateur qui convient le mieux pour ce type d'analyse. En effet, certaines études parmi celles que K. A. Feldman a examinées montrent une relation curvilinéaire, pour laquelle les évaluations sont en tendance plus positives pour les très petits groupes (< 15 étudiants) ou pour les très grands groupes (> 100 étudiants). Plus récemment, W. E. Cashin (1992, 1993) a conclu que lorsqu'elle est « isolée » la variable « taille de l'auditoire » ne constitue pas une source sérieuse de biais. S. Chiu (1999), cité par J. C. Ory (2001), en appliquant une analyse de variance ANOVA sur des groupes de taille différente (*Unbalanced Nested Anova*), arrive à la même conclusion : la variable « taille du groupe » n'explique que 1,19 % de la variance totale du score EEE. H. W. Marsh (1987) remarque, quant à lui, que la taille de l'auditoire n'influence que certains indicateurs de l'EEE comme, par exemple, les interactions dans le groupe et l'interaction avec l'enseignant.

En ce qui concerne la *discipline enseignée*, K. A. Feldman (1978) avait déjà démontré que les mathématiques et les sciences obtenaient, en moyenne, un score inférieur aux autres disciplines. Ce qui l'amenait à dire qu'il fallait tenir compte de cette variable, soit en créant des normes spécifiques pour les enseignants de ces matières, soit en contrôlant statistiquement cette différence. Plus récemment, W. E. Cashin (1990, 1992) et J. A. Centra (1993) sont arrivés à des

conclusions proches en établissant une classification. En ordre décroissant, les arts et sciences humaines, les langues, la biologie et sciences sociales, l'administration des affaires, l'informatique, les mathématiques, les sciences appliquées et physiques obtiennent tendanciellement des évaluations différentes, les premières étant les plus largement rétribuées que les dernières.

Le rapport entre la *charge de travail* et le score à l'EEE est assez étonnant. La plupart des auteurs (Marsh, 1980, 1982, 1983, 1984 ; Cashin, 1988) trouvent une relation positive entre ces deux variables : plus la charge de travail inhérente à un enseignement est grande, plus le score obtenu lors de l'EEE pour cet enseignement augmente.

Enfin, toujours dans les variables contextuelles, les *méthodes de passation* de l'EEE peuvent influencer, même si ce n'est que marginalement, les résultats à l'EEE. Ainsi les EEE *recueillies pendant un examen final* (Frey, 1976) sont plus sévères que celles recueillies en cours d'année.

Les évaluations des enseignements par les étudiants réalisées de manière *anonyme* sont en règle générale supérieures à celles pour lesquelles l'étudiant doit s'identifier (Argulewiz et O'keefe, 1978 ; Feldman, 1979 ; Hartnett et Seligsohn, 1967 ; Stone, Spool et Rabinowitz, 1977).

D'autres auteurs spécifient que certaines conditions influencent positivement les résultats de l'EEE. C'est le cas si les résultats sont utilisés à des fins de promotion (Centra, 1976 ; Feldman, 1979 ; Overall et Marsh, 1979), si l'enseignant reste dans la classe pendant l'évaluation et si un administrateur fait un « court speech » sur l'importance de l'évaluation (Frey, 1976).

La conclusion relative aux biais liés au contexte revient à M. Theall et J. Franklin (2001). Ils citent l'étude de H. W. Marsh (1987) qui conclut que les variables contextuelles n'influencent pas grandement les EEE, mais précisent toutefois qu'un jeune enseignant d'un cours de premier cycle et obligatoire en physique aura, en moyenne, un score plus faible qu'un enseignant confirmé de second cycle, donnant un cours optionnel. Ils précisent, toutefois, que cette affirmation ne constitue pas nécessairement un biais, dans la mesure où il est sans doute plus difficile de donner un enseignement de qualité sous certaines conditions.

Les méthodes de laboratoire, quant à elles, présentent le défaut d'être très peu authentiques. Elles consistent à recréer une situation d'enseignement en laboratoire. L'idée sous-jacente est de garder la variable « qualité de l'enseignement » sous contrôle et de faire varier expérimentalement un certain nombre d'autres variables pour voir si le score à l'EEE se modifie. Si c'est le cas, cela signifierait que la variable enseignement n'est pas la seule qui influence le score à l'EEE ce qui, dès lors, pourrait l'invalider en tant que mesure de l'efficacité de l'enseignement.

La plus connue des études de laboratoire est probablement celle dite du « Doctor Fox ». Dans cette étude, D. H. Naftulin, J. E. Ware et F. A. Donnelly (1973) ont demandé à un acteur (nommé « Doctor Fox ») de donner une leçon de mathématique à des médecins en formation. Cette leçon était truffée d'erreurs en termes de contenu (néologisme inexistant présenté comme concept clé,...) mais avec un maximum d'emphase. À la fin de cette leçon, un questionnaire d'EEE fut rempli. L'évaluation était très positive, ce qui aux yeux des auteurs signifiait que l'EEE mesurait le style de l'enseignant plutôt que la qualité de son enseignement. Cette étude a très vite été critiquée dans la littérature, notamment sur la base de son caractère peu authentique. Ainsi, P. C. Abrami, L. Leventhal et R. P. Perry (1982),

P. W. Frey (1979) critiquèrent le fait que les étudiants aient dû porter leur jugement après une seule séance de cours, dans une matière qu'ils ne connaissaient absolument pas (ils étaient en début de premier cycle) et sans avoir pu fonder leur jugement sur d'autres aspects de l'enseignement (par exemple les notes de cours, les supports d'apprentissage, les examens) ce qui est très loin des conditions réelles d'EEE. De plus, H. W. Marsh et J. E. Ware (1982) ont découvert que, lorsque les étudiants reçoivent une stimulation à apprendre (par exemple une évaluation certificative à la fin du cours), l'effet « Doctor Fox » est annihilé.

Dans la lignée de l'expérience du « Doctor Fox », W. M. William et S. J. Ceci (1997) ont tenté de démontrer que l'EEE était plus influencée par le style de l'enseignant que par le contenu du cours. À partir d'une étude multisection, ils ont en effet comparé les scores EEE lorsqu'un même enseignant variait son style d'enseignement (intonation de la voix, comportement non verbal, enthousiasme,…) dans deux groupes différents sans toutefois modifier ni le matériel, ni le contenu du cours. Ils arrivent à la conclusion que les performances des étudiants en fin de semestre à l'examen ne varient pas entre les deux groupes, mais que le score EEE, lui, varie de manière significative (il passe de 3.08 à 3.92 sur 5). Ils en concluent que l'EEE ne dépend pas du contenu, mais bien du style de l'enseignant. S. d'Apollonia et P. C. Abrami (1997) ont sévèrement critiqué cette recherche d'un point de vue méthodologique, la qualifiant de recherche pré-expérimentale. Ils soulignent qu'eux-mêmes (Abrami, Leventhal et Perry, 1982) ont publié une revue des études quantitatives sur le sujet. Ils y concluent que l'expressivité de l'enseignant a un plus grand impact sur son score EEE que sur l'apprentissage des étudiants. Dans cette même étude, ils concluent également que le contenu du cours a un plus grand impact sur l'apprentissage des étudiants que sur le score à l'EEE. Toutefois, ils interprètent ces résultats très différemment des détracteurs de l'EEE. Pour eux, ces résultats ne sont en rien la preuve de la non-validité de l'EEE. Selon eux, plutôt que de la remettre en cause, ces résultats posent la question des raisons de la validité de l'EEE. Ils soutiennent que des études comme celles du « Doctor Fox » et de W. M. William et S. J. Ceci (1997) sont utiles pour comprendre ce qui peut influencer la relation entre le score à l'EEE et l'apprentissage des étudiants, mais ils ne voient pas en quoi celles-ci permettent de mettre en doute la validité de l'EEE.

Une autre étude de laboratoire, est celle menée par N. Ambady et R. Rosenthal (1992). Dans cette recherche, ils ont soumis à des observateurs des séquences vidéo de trente secondes, qu'ils ont appelées « fines tranches d'expressivité (*Thin slices of expressive behavior*) ». Ces séquences ne comprenaient pas de son. Ils ont demandé aux observateurs de prédire quel serait le score de l'EEE pour ces enseignements. Ils observent une corrélation positive élevée (76) entre le jugement des étudiants et le pronostic des observateurs. Ils en concluent que le score attribué à un enseignement dépend essentiellement de comportements non verbaux, indépendamment de la qualité de l'enseignement. J. A. Kulik, en 2001, a commenté cette étude. Il signale que le faible échantillonnage de cours vidéoscopés rend l'erreur de mesure très importante. Il fait également référence aux travaux de K. A. Feldman (1989), qu'il juge plus sérieux sur les liens entre score EEE et score attribué par les observateurs. Les études revues par K. A. Feldman comprenaient l'observation de longues séquences d'enseignement (visuelles et auditives). La corrélation moyenne relevée par K. A. Feldman est, dans ce cadre, de .50.

Enfin, certaines études sur la validité se sont penchées sur la *structure conceptuelle de l'EEE*. Ainsi, comme le spécifient J. C. Ory et K. Ryan (2001), beaucoup d'études (Kulik et McKeachie, 1975 ; Feldman, 1976 ; Marsh, 1987) ont été conduites, revues ou méta-analysées pour tenter de repérer un ensemble commun de

facteurs sous-tendant le *construct* qui est mesuré par le score à l'EEE. Bien qu'il y ait quelques éléments communs à toutes ces études, la recherche n'est pas parvenue à isoler un seul ensemble de dimensions constant à travers toutes ces études, ce qui plaide pour le fait de considérer la qualité de l'enseignement comme étant multidimensionnelle. Cette difficulté à isoler un *construct* homogène plaide pour le fait que les auteurs travaillant dans le domaine de l'EEE définissent exactement ce que les procédures en place dans leur institution mesurent et, *a contrario*, ce qu'elles ne mesurent pas. Il est probable qu'une partie des débats contradictoires présents dans la littérature découle de ce manque de définition préalable.

CONCLUSION

La question de la validité de l'EEE, apparue dans les années soixante, reste polémique, comme on a pu le constater tout au long de cette revue de littérature. Les études multitrait-multiméthode et les études multi-section montrent une corrélation de l'EEE avec d'autres indicateurs – imparfaits – de la qualité de l'enseignement. Les chercheurs qui s'inscrivent dans ce courant considèrent que les résultats engrangés plaident en faveur de la validité de l'EEE sans être décisifs. Quant aux études axées sur les biais et les études en laboratoire, elles manquent d'uniformité. Ainsi, on y trouve une littérature analysant l'interaction d'une variable isolée (score à l'examen, sexe, âge, taille de l'auditoire, etc.) avec l'EEE. Certaines interactions sont ainsi relevées. Répondant à ce type de littérature, des méta-analyses mettant en œuvre des modèles statistiques performants relativisent la plupart du temps l'impact des résultats de ces études de laboratoire, soit en mettant en évidence l'inconstance des interactions mises à jour, soit en minimisant leur impact sur le score EEE. Les auteurs de ces études reconnaissent toutefois que, bien que faibles, certaines interactions peuvent influencer les résultats à l'EEE.

Selon nous, la littérature scientifique a échoué dans sa tentative de faire la preuve formelle, indiscutable et décontextualisée de la validité de l'EEE. Si un nombre important d'arguments en faveur de l'EEE a été mis à jour par la recherche, d'autres, plus défavorables, ont également été mis en évidence. Que faut-il en conclure ? En accord avec J. A. Kulik (2001), nous pensons que les résultats de cette méthode ont généralement montré des preuves partielles de validité convergente et discriminante, mais pas de manière parfaite. On peut, tout au plus, parler d'un faisceau d'arguments qui plaident pour la validité de l'EEE et ce, même si les travaux de P. A. Cohen (1981) et les méta-analyses de S. d'Appolonia et P. C. Abrami (1997) soutiennent l'idée qu'il existe une liaison établie, signifiante et porteuse de sens entre l'EEE et l'apprentissage.

Ces conclusions sont d'ailleurs peu étonnantes dans la mesure où il nous semble que la littérature en EEE a hérité de l'incapacité actuelle des auteurs en sciences de l'éducation à s'entendre sur une définition de ce qu'est un enseignement de qualité (Shmanske, 1988 ; Scriven, 1989).

Discussion

L'enseignement est une tâche complexe. Dans l'enseignement supérieur, elle se situe dans un contexte en profonde mutation. Massification, nouvelles technologies, démocratisation, professionnalisation, démarche de qualité, autant de

facteurs qui peuvent amener un enseignant à s'interroger sur son enseignement. Une approche cybernétique de l'enseignement avec des boucles régulatrices doit, dès lors, être envisagée, ce qui implique que des informations de rétroaction soient disponibles afin de permettre à l'enseignant de réguler ses pratiques d'enseignement. Selon Romainville (2004), cette information peut être recueillie de diverses façons et auprès de différentes sources :

– l'enseignant, en tant que « praticien réflexif » peut en quelque sorte s'auto-observer ;
– les collègues peuvent renvoyer des informations utiles ;
– les étudiants, enfin, constituent une source d'informations valide et pertinente à propos de l'évaluation de certains aspects de leur formation.

Parmi les sources possibles, rappelons que les auteurs ayant co-publié dans le numéro thématique de la revue *American Psychologist* (Greenwald, 1997 ; Marsh et Roche, 1997 ; d'Apollonia et Abrami, 1997 ; Greenwald et Gillmore, 1997 ; McKeachie, 1997) soutiennent que l'EEE est la source d'information la plus valide concernant la qualité de l'enseignement.

Même si sa validité n'a pas pu être mise en évidence une fois pour toutes par la recherche, nous remarquons également que les nombreuses études l'ayant questionnée n'ont pas pu prouver son caractère non-valide. Mieux, un faisceau d'évidences plaide pour sa validité.

De notre point de vue, l'EEE peut et doit être utilisée à des fins de régulation formative. Même si certains auteurs, comme P. C. Abrami (2001), plaident pour l'utilisation prioritaire de l'EEE à des fins administratives (c'est-à-dire dans une perspective qui influence la carrière des enseignants), nous sommes de ceux qui pensent que l'EEE ne peut être utilisée seule qu'à des fins formatives. En effet, au vu de la littérature, il est probable que les scores à l'EEE soient influencés, même faiblement, par d'autres caractéristiques que la qualité de l'enseignement *stricto sensu*. Au vu de l'intérêt de l'information fournie par l'EEE et sachant qu'il n'existe pas de recueils d'informations qui présentent des caractéristiques plus robustes, nous pensons que ces légers biais sont acceptables dans le cadre d'une évaluation formative. Dans le cadre d'une évaluation certificative, toutefois, ceux-ci pourraient mener à des problèmes d'équité, surtout dans le cadre d'une évaluation normative, ce qui semble inacceptable.

Par ailleurs, nous pensons que l'usage de l'EEE à des fins de régulation formative doit être soumis à des conditions liées à un principe de précaution. Sans prétention à l'exhaustivité, nous en mentionnons quatre :

- ne pas utiliser l'EEE par questionnaire si le groupe est inférieur à 10 étudiants. En effet, W. E. Cashin a démontré que, dans ce cadre, l'évaluation manque de fidélité ;
- si l'évaluation est à référence normative, comparer ce qui est comparable. Nous avons pu voir que certaines variables contextuelles pouvaient interférer, la plupart du temps légèrement, avec le score à l'EEE. Un moyen de contrôler ces biais est de ne comparer entre eux que les enseignements d'une même section, c'est-à-dire proposés aux mêmes étudiants, dans les mêmes conditions ;
- mettre en place des évaluations des enseignements par les étudiants conformes aux *standards* de la littérature afin de s'assurer de la validité du questionnaire dans le cadre d'un enseignement donné ;
- obtenir un consensus sur ce que recouvre la procédure d'EEE dans l'institution et sur ce qu'elle ne recouvre pas.

Comme le relevé de la littérature présenté ci-dessus le laisse apparaître, beaucoup d'études portent sur la validité de l'EEE. Paradoxalement, assez peu d'études portent sur l'effet qu'a eu l'EEE sur la régulation des enseignements, sur ce que P. Lather (1986) a appelé la « validité catalytique ». Or l'EEE n'est qu'un outil et, comme tout outil, il doit servir. Nous sommes donc en phase avec W. J. McKeachie (1997) qui plaide pour que les recherches se concentrent désormais sur la validité conséquentielle (Miller et Linn, 2000) de l'EEE. Cet auteur spécifie que les recherches doivent porter sur les représentations véhiculées par le dispositif mais aussi sur son effet sur les divers acteurs.

Quelques études de ce type ont été effectuées et mettent en avant des résultats paradoxaux. Ainsi, si de nombreux effets positifs des EEE sur les pratiques d'enseignement ont été mis en évidence dans les recherches de H. W. Marsh et de ses collaborateurs, des effets négatifs (McKeachie, 1979) ou une absence d'effets (Bernard *et al*., 2000) ont aussi été rapportés. Ces effets contradictoires ont aussi été retrouvés dans la recherche conduite dans une université française sur les effets de l'EEE par N. Younes (2006). Pourquoi de tels effets contradictoires et quelles sont les conditions d'une utilisation de l'EEE à des fins de régulation formative de l'enseignement ? Il est probable que les réponses résident dans les caractéristiques et les conditions de l'implémentation de l'EEE dans les universités. En effet, les auteurs les plus rassurants sur la validité et la fidélité de l'EEE précisent, qu'avant toute chose, ce sont les conditions d'implémentation de l'EEE qui sont les gages de qualité. Ainsi, J. C. Ory et K. Ryan (2001) adoptent-t-ils l'idée de R. L. Linn (1998) qui précise que la validité de l'EEE est sous la responsabilité de l'ensemble des groupes d'utilisateurs.

La qualité de la conception et de la mise en œuvre du processus d'EEE est ainsi cruciale. Or, selon R. A. Arreola (1995), il arrive encore trop souvent que l'EEE repose sur un ensemble non cohérent d'outils, rassemblés à la « va-vite » par une administration, un groupe d'enseignants ou d'étudiants. La procédure de recueil d'informations ressemble alors plus à du bricolage qu'à une méthode structurée. Il est donc urgent que la littérature aborde en profondeur les conditions d'émergence d'une EEE menant réellement à la régulation des enseignements. Or, si une série de *guidelines* a été publiée dans la littérature (Cashin, 1999 ; Menges, 1990 ; Brinko, 1991 ; Centra, 1993 ; Ory, 2001 ; Arreola, 1994 ; Theall et Franklin, 2001 ; Emery *et al.*, 2003 ; Younes, 2006), il semble que cette littérature n'ait encore jamais été unifiée. De plus, à notre connaissance, aucune étude n'a porté sur les effets de l'application de l'une ou l'autre de ces recommandations.

À nos yeux, il s'agit pourtant d'un domaine d'investigation indispensable. Décrire les pratiques en matière d'EEE ne nous semble pas suffisant. Nous pensons urgent de constituer un savoir stratégique, au sens de J.-M. Van der Maren (2004), à savoir une critique évolutive des pratiques à la lumière des résultats scientifiques engrangés dans ce domaine. Pour ce faire, nous pensons que les démarches d'EEE doivent reposer sur les conseils issus de la littérature en EEE, s'appuyer sur les connaissances en pédagogie et en docimologie, tenir compte du système de valeurs des divers acteurs, mais également s'inscrire dans une démarche cybernétique dans laquelle le contexte, le dispositif EEE et ses effets seraient soigneusement documentés et analysés. De cette manière, nous pourrons nourrir une approche *Evidence Based* en termes d'EEE.

Pascal Detroz
Système méthodologique d'aide
à la réalisation de tests,
université de Liège,
www.smart.ulg.ac.be

NOTES

(1) Notons d'emblée que la présente note de synthèse traite des enseignements et non des enseignants.

(2) Rice (1898) publie une étude comparative de la performance en épellation de 33 000 étudiants, étude dans laquelle il fit une critique sévère des procédés d'enseignement jusqu'alors utilisés (cité par Nadeau, 1990).

(3) "Negative attitude towards student ratings are especially resistant to change, and it seems that faculty and administrators support their belief in student-rating myths with personal and anecdotical evidence wich [for them] outweighs empirically based research evidence".

(4) Conclusion qui peut d'ailleurs s'étendre au portfolio de l'enseignant (voir les travaux de Kane, Krooks et Cohen 1999; Richlin et Manning, 1996), et à l'évaluation par les anciens étudiants (Kulik, 2001).

(5) Cité par Emery, Kramer and Tian (2003).

(6) "The attention received by the Rodin and Rodin study seems disproportionate to its rigor".

BIBLIOGRAPHIE

ABRAMI P. C. (2001). « Improving judgements about teaching effectiveness using teacher ratings forms ». In M. THEALL, P. C. ABRAMI & L. A. METS (éd.), *The student ratings debate: Are they valid? How can we best use them*. San Francisco : Jossey-Bass, p. 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 ». *Journal of Educational Psychology*, n° 82, p. 219-231.

ABRAMI P. C., DICKENS W. J., PERRY R. P. & LEVENTHAL L. (1980). « Do teacher standards for assigning grades affect student evaluations of instruction? ». *Journal of Educational Psychology*, n° 72, p. 107-118.

ABRAMI P. C., PERRY R. P. & LEVENTHAL L. (1982). « Educational seduction ». *Review of Educational Research*, n° 52, p. 446-464.

ALEAMONI L M. (1987). « Student rating myths versus research facts ». *Journal of Personnel Evaluation in Education*, vol. 1, n° 1, p. 111-119.

ALEAMONI L. M. (1987). « Typical faculty concerns about student evaluation of teaching ». In L. M. Aleamoni (éd.), *Techniques for Evaluation and improving Instruction*. San Francisco : Jossey-Bass.

ALEAMONI L. M. (1989). « Typical faculty concerns about evauation of theaching ». In L. M. ALEAMONI (éd.), *Techniques for evaluating and improving Instruction*. San Francisco : Jossey-Bass.

ALEAMONI L. M. & GRAHAM N. H. (1974). « The relationship between CEQ ratings and instructor's rank, class size, and course level ». *Journal of Educational Measurement*, n° 11, p. 189-201.

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*, vol. 9, n° 1, p. 67-84.

AMBADY N. & ROSENTHAL R. (1992). « Half a minute: predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness ». *Journal of Personality and Social Psychology*, n° 64, p. 431-441.

ARGULEWIZ E. & O'KEEFE T. (1978). « An investigation of signed versus anonymously completed ratings of high school student teachers ». *Educational Research Journal*, n° 3, p. 39-44.

ARREOLA R. A. (1995). *Developing a comprehensive faculty Evaluation System: A handbook for College Faculty and Administrators on designing and operating a comprehensive Faculty Evaluation System*. Boston : Anker Publishing Co.

BAUSELL R. B. & BAUSELL C. R. (1979). « Student rating and various instructional variables from a within-intructor perspective ». *Research in Higher Education*, n° 11, p. 167-177.

BENNETT S. K. (1982). « Student perceptions of and expectations for male and female instructors: evidence relating to the question of gender bias in teaching evaluation ». *Journal of Educational Psychology*, vol. 74, n° 2, p. 170-179.

BERNARD H. (1992). *Processus d'évaluation de l'enseignement supérieur. Théorie et pratique.* Laval : Éditions études vivantes.

BERNARD H., POSTIAUX N. & SALCIN A. (2000). « Les paradoxes de l'évaluation de l'enseignement universitaire ». *Revue des sciences de l'éducation*, vol. 26, n° 3, p. 625-650.

BERNARD M. E., KEEFAUVER L. W., ELSWORTH G., & MAYLOR F. D. (1981). « Sex role behavior and gender in teachers-student evaluations ». *Journal of Educational Psychology*, n° 73, p. 681-696.

BRANDENBURG D. C., SLINDE J. A., & BATISTA E. E. (1977). « Student ratings of instruction: validity and normative interpretations ». *Journal of Research in Higher Education*, n° 7, p. 67-98.

BRINKO K. T. (1991). « The interactions of teaching improvement. Effective practices for improving teaching ». *Directions for Teaching and Learning*, n° 48, p. 39-49.

CASHIN W. E. (1988). *Students ratings of Teaching: A Summery of the Research.* Kansas State University : Center for Faculty Evaluation and Development.

CASHIN W. E. (1990). « Students do rate different academic fields differently ». In M. THEALL & J. FRANKLIN (éd.), *Student ratings of Instruction: Issues for improving Practice*. San Francisco : Jossey-Bass.

CASHIN W. E. (1992). « Student ratings: the need for comparative data ». *Instructionnal Evaluation and Faculty Developpement*, vol. 12, p. 1-6.

CASHIN W. E. (1983). « Concerns about using student ratings in community colleges ». In A. SMITH (éd.), *Evaluating Facultv and Staff:New directions for communitv Colleges* San Francisco : Jossey-Bass, p. 57-68.

CASHIN W. E. (1995). « Student ratings of teaching: the research revisited ». *IDEA Paper*, n° 32.

CASHIN W. E. (1999). « Student ratings of teaching: Uses and misuses ». In P. SELDIN (éd.), *Changing practices in evaluating teaching. A practical guide to improved faculty performance and Promotion/Tenure decisions.* Bolton, MA : Anker Publishing, p. 25-44

CENTRA J. A. (1976). « The influence of different directions on student ratings of instruction ». *Journal of Educational Measurement*, n° 13, p. 277-282.

CENTRA J. A. (1993). *Reflexive faculty evaluation effectiveness. Enhancing teaching and determining faculty effectiveness.* San Francisco : Jossey-Bass.

CENTRA J. A. & CREECH F. R. (1976). *The relationship between Students, Teachers, and course Characteristics and student Ratings of Teacher effectiveness.* Princeton, N. J. : Educational Testing Service Ed.

CENTRA J. A. & GAUBATZ N. B. (2000). « Is there gender bias in student evaluation of teaching ». *The Journal of Higher Education*, vol. 70, n° 1, p. 17-33.

CHIU S. (1999). *Use of the unbalanced nested ANOVA to exam Factors influencing student Ratings of instructional Quality.* Thèse de doctorat, University of Illinois et Urbana-Champaign (non publiée).

COHEN P. A. (1981). « Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies ». *Review of Educational Research Fall,* vol. 51, n° 3, p. 281.

COHEN P. A. (1990). « Bring research into practice ». In M. THEALL & J. FRANKLIN (éd.), *Student ratings of instruction: Issues for improving practice: New directions for teaching and learning.* San Francisco : Jossey-Bass Publishers, p. 123-132.

COHEN L., MANION L., MORRION K. (1997). *Research Methods in Education*. Londres, New York : Routledge Falmer.

COSTIN F., GREENOUGH W. T. & MENGES R. J. (1971). « Student ratings of college teaching: Reliability, validity, and usefulness ». *Review of Educational Research*, n° 41, p. 511-535.

D'APOLLONIA S. & ABRAMI P. C. (1997). « Navigating student ratings of instruction ». *American Psychologist*, vol. 52, n° 1198, p. 1208.

DEJEAN J. (2002). *L'évaluation de l'enseignement dans les universités françaises : rapport suivi de l'avis du Haut conseil de l'évaluation de l'école*. France : Haut conseil de l'évaluation de l'école.

DOWELL D. A. & NEAL J. A. (1982). « A selective review of the validity of student ratings of teachings ». *The Journal of Higher Education*, vol. *53*, n° 1, p. 51-62.

DOYLE J. J. O. (1983). *Evaluating teaching*. Lexington, Mass. : Lexington Books. Ed.

DOYLE K. O. (1975). « Student evaluation of instruction ». *Student Evaluation of Instruction.* Lexington, MA : D. C. Heath and Co.

EMERY C. R. (1995). *Student Evaluations of faculty Performance* (non publié). Cité in P. Detroz, *Évaluation des enseignements : de la contrainte administrative à l'amélioration des pratiques.* Disponible sur Internet : <http://www.smart.ulg.ac.be/smartweb/documents/fribourg2007/fribourg240907.pdf> (consulté le 25 février 2009).

EMERY C. R., KRAMER R. & TIAN R. G. (2003). « Return to academic standards: a critique of student evaluations of teaching effectiveness ». *Quality Assurance in Education*, vol. 11, n° 1, p. 37-46.

FELDMAN K. A. (1976). « The superior college teacher from the student's view ». *Research in Higher Education*, n° 5, p. 223-274.

FELDMAN K. A. (1978). « Course characteristics and college students' ratings of their teachers: what we know and what we don't ». *Research in Higher Education*, n° 9, p. 199-242.

FELDMAN K. A. (1979). « The signifiance of circumstances for college students' ratings of their teachers and courses: A review and analysis ». *Research in Higher Education*, n° 10, p. 149-172.

FELDMAN K. A. (1983). « Seniority and experience of college teachers as related to evaluations they receive from their students ». *Research in Higher Education*, n° 18, p. 3-124.

FELDMAN K. A. (1987). « Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: a review and exploration ». *Research in Higher Education*, n° 26, p. 227-298.

FELDMAN K. A. (1988). « Effective college teaching from the students' and faculty's view: matched or mismatched priorities? ». *Research in Higher Education*, n° 28, p. 291-344.

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*, n° 30, p. 137-194.

FELDMAN K. A. (1992). « College student's views of male and female college teachers: part I – evidence from the social laboratory and experiments ». *Research in Higher Education*, n° 33, p. 317-375.

FRANKLIN J., THEALL M., LUDLOW L. (1991) « Grade inflation and student ratings: a closer look ». Paper presented at the American Education Research Association.

FREY P. W. (1976). « Validity of student instructional rating as a function of their timing ». *Journal of Higher Education*, n° 47, p. 327-336.

FREY P. W. (1979). « The Dr. Fox effect and its implications ». *Instructional Evaluation*, n° 3, p. 1-5.

GLASCOCK J. & RUGGIERO T. (2006). « The relationship of ethnicity and sex to professor credibility at a culturally diverse university ». *Communication Education*, n° 55, p. 197-207.

GREENWALD A. G. (1997). « Validity concerns and usefulness of student ratings instruction ». *American Psychologist*, n° 52, p. 1182-1186.

GREENWALD A. G. & GILLMORE G. M. (1997). « Grading leniency is a removable contaminant of student ratings ». *American Psychologist,* n° *52*, p. 1209-1217.

HAMERMESH D. S. & PARKER A. (2005). « Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity ». *Economics of Education Review*, vol. 24, n° 4, p. 369-376.

HARTNETT R. T. & SELIGSOHN H. C. (1967). « The effects of varying degrees of anonymity on response to different

types of psychological questionnaires ». *Journal of Educational Measurement*, n° 4, p. 95-103.

HOLMES D. S. (1972). « Effects of grades and disconfirmed grade expectancies on students' evaluation of their instructor ». *Journal of Educational Psychology*, n° 63, p. 130-133.

HOWARD G. S., CONWAY C. G. & MAXWELL S. E. (1985). « Construct validity of measures of college teaching effectiveness ». *Journal of Educational Psychology*, n° 77, p. 187-196.

HOWARD G. S. & MAXWELL S. E. (1980). « Correlation between student satisfaction and grades: a case of mistaken causation ? ». *Journal of Educational Psychology*, n° 72, p. 810-820.

HOWARD G. S. & MAXWELL S. E. (1982). « Do grades contaminate students' evaluation of college teaching: a multi-trait multimethod analysis ». *Journal of Educational Psychology*, n° 74, p. 264-279.

KANE M., CROOKS T. & COHEN A. (1999). « Validating measures of performance ». *Educational Measurement: Issues and Practice*, vol. 18, n° 2, p. 5-17.

KULIK J. A. (2001). « Student rating: validity, utility, and controversy ». In M. THEALL, P. C. ABRAMI & L. A. METS (éd.), *The student ratings debate: Are they valid? How can we best use them?* San Francisco : Jossey-Bass (New Directions for Institutional Research Ed.).

KULIK J. A. & MCKEACHIE W. J. (1975). « The evaluation of teachers in higher education ». In F. N. KERLINGER (éd.), *Review of research in education*. Itasca, Ill. : Peacock Ed.

LATHER P. (1986). « Research as praxis ». *Harvard Educational Review*, n° 56, p. 257-277.

LINN R. L. (1998). « Partitioning responsability for the evaluation of the consequences of assessment program ». *Educational Measurement: Issues and Practice*, vol. 17, n° 2, p. 28-36.

MARSH H. W. (1980). « The influence of student, course, and instructor characteristics in evaluations of university teaching ». *American Educational Research Journal*, vol. 17, n° 2, p. 219-237.

MARSH H. W. (1982). « Validity of students' evaluations of college teaching: a multitrait-multimethod analysis ». *Journal of Educational Psychology,* vol. 74, n° 2, p. 264-279.

MARSH H. W. (1983). « Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics ». *Journal of Educational Psychology*, vol. 75, n° 1, p. 150-166.

MARSH H. W. (1984). « Students evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility ». *Journal of Educational Psychology,* vol. 76*,* n° 5, p. 707-754.

MARSH H. W. (1987). « Students' evaluations of university teaching: research findings, methodological issues, and directions for future research ». *International Journal of Educational Research*, vol. 11, n° 3, p. 255-388.

MARSH H. W. & COOPER T. (1981). « Prior subject interest, students' evaluations, and instructional effectiveness ». *Multivariate Behavioral Research*, n° 16, p. 82-104.

MARSH H. W. & DUNKIN M. (1992). « Students' evaluations of university teaching: a multidimensional perspective ». In J. C. SMART (éd.), *Higher Education: Handbook on Theory and Research*. Edison : Agathon Press, p. 143-234.

MARSH H. W. & ROCHE L. A. (1997). « Making students' evaluations of teaching effectiveness effective ». *American Psychologist*, n° 52, p. 1187-1197.

MARSH H. W. & WARE J. E. (1982). « Effects of expressiveness, content coverage, and incentive on multidimensionnal student rating scale: new interpretations of the Doctor Fox effect ». *Journal of Educational Psychology*, n° 74, p. 126-134.

MCGUIGAN F. J. (1967). « The G. statistics, an index of amount learned ». *National Society for Programmed Instruction*, n° 6, p. 14-16.

MCKEACHIE W. J. (1979). « Student rating of faculty: a reprise ». *Academe*, n° 65, p. 384-397.

MCKEACHIE W. J. (1997). « Student ratings: the validity of use ». *American Psychologist*, n° 52, p. 1218-1225.

MENGES R. J. (1990). « Using evaluative information to improve instruction ». In P. SELDIN (éd.), *How administrators can improve teaching. Moving from talk to action in higher education*. San Francisco : Jossey-Bass, p. 104-121.

MILLER D. M. & LINN R. L. (2000). « Validation of performance-based assessments ». *Applied Psychological Measurement*, vol. 24, n° 4, p. 367-378.

MURRAY H. G. (1983). « Low-inference classroom teaching behaviors and students' ratings of college teaching effectiveness ». *Journal of Educational Psychology*, n° 75, p. 138-149.

MURRAY H. G. (1984). « The impact of formative and summative evaluation of teaching in north american universities ». *Assessment and Evaluation in Higher Education,* vol. 9, n° 2, p. 117-132.

MURRAY H. G., RUSHTON J. P. & PAUNOMEN S. V. (1990). « Teacher personality traits and student instructional ratings in six types of university courses ». *Journal of Educational Psychology*, vol. 82, n° 2, p. 250-261.

NADEAU M. A. (1990). *L'évaluation de programme : théorie et pratique*. Laval : Les presses de l'université Laval.

NAFTULIN D. H., WARE J. E. & DONNELLY F. A. (1973). « The Doctor Fox lecture: a paradigm of educational seduction ». *Journal of Medical Education*, n° 48, p. 630-635.

ORY J. C. (1980). « The influence of students' affective entry on instructor and course evaluations ». *Review of Higher Education*, n° 4, p. 13-24.

ORY J. C. (2001). « Faculty thoughts and concerns about student ratings ». In K. G. LEWIS (éd.), *Techniques and Strategies for interpreting student Evaluations*. San Francisco : Jossey-Bass, p. 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 (éd.), *The student Ratings Debate: Are they valid? How can we best use them?* San Francisco : Jossey-Bass, p. 27-44.

OVERALL J. U. & MARSH H. W. (1979). « Midterm feedback from student: its relationship to instructional improve-

ment and students' cognitive and affective outcomes ». *Journal of Educational Psychology*, n° 71, p. 856-865.

PERRY R. P., ABRAMI P. C., LEVENTHAL L. & CHECK J. (1979). « Instructor reputation: an expectancy relationship involving student ratings and achievement ». *Journal of Educational Psychology*, n° 71, p. 776-787.

PIÉRON H. (1963). *Examens et docimologie*. Paris : Presses universitaires de France.

POWELL R. W. (1977). « Grades, learning, and student evaluation of instruction ». *Research in Higher Education*, n° 7, p. 193-205.

RAMSDEN P. & ENTWISTLE N. J. (1981). « Effects of academics departments on students' approaches to studying ». *British Journal of Educational Psychology*, vol. 51, p. 368-83.

REMY S. (1995). *Contribution à l'élaboration d'un système d'évaluation de l'enseignement à la FAPSE*. Mémoire de fin d'étude (non publié), faculté de psychologie et sciences de l'éducation, université de Liège.

RICE J. M. (1898). *The Rational Spelling Book.* New York : American Book.

RICHLIN L. & MANNING B. (1996). « Using portfolios to document teaching excellence ». In M. D. SVINICKI & R. J. MENGES (éd.), *Honoring exemplary teaching*. San Francisco : Jossey-Bass, p. 65-70.

RODIN M. & RODIN B. (1972). « Student evaluations of teachers ». *Science*, vol. 177, n° 4055, p. 1164-1166.

ROMAINVILLE M. (2004). « Esquisse d'une didactique universitaire ». *Revue francophone de gestion*, numéro spécial consacré au deuxième prix de l'innovation pédagogique en sciences de gestion, La Sorbonne/ CIDEGEF.

SCRIVEN M. (1989). « The design and use of forms for the student evaluation of teaching ». *Instructional Evaluation*, n° 10, p. 1-13.

SELDIN P. (1993). « The use and abuse of student ratings of instruction ». *The Chronicle of Higher Education,* A-40.

SHMANSKE S. (1988). « On the measurement of teacher effectiveness ». *Journal of economic*, vol. 19, n° 4, p. 307-314.

SNYDER C. R. & CLAIR M. (1976). « Effects of expected and obtained grades on teacher evaluation and attribution of performance ». *Journal of Educational Psychology*, n° 68, p. 75-82.

STONE E. F., SPOOL M. D. & RABINOWITZ S. (1977). « Effects of anonymity and retaliatority potential on student evaluations of faculty performance ». *Research in Higher Education*, n° 6, p. 313-325.

THEALL M. (2002). « Student rating: myths vs research evidence ». *Brigham Young University's Focus on Faculty Newsletter*, vol. 10, n° 3, p. 2.

THEALL M. & FRANKLIN J. (1990). « Student ratings of instruction: issues for improving practice ». In M. THEALL & J. FRANKLIN (éd.), *New directions for teaching and learning*. San Francisco : Jossey-Bass.

THEALL M. & FRANKLIN J. (2001). « Looking for bias in all the wrong places: A search for truth or a with hunt in student ratings of instruction ». In M. THEALL, P. C. ABRAMI & L. A. METS (éd.), *The student ratings debate: Are they valid? How can best use them?* San Francisco : Jossey-Bass, p. 45-56.

THIVIERGE A. & BERNARD H. (1996). « Les croyances des étudiants à l'égard de l'évaluation de l'enseignement ». *Mesure et évaluation en éducation,* vol. 18, n° 3, p. 59-84.

VASTA R. & SARMIENTO R. F. (1979). « Liberal grading improves evaluations but not performance ». *Journal of Educational Psychology*, n° 71, p. 207-211.

VAN DER MAREN J.-M. (2004). *Méthode de recherche pour l'éducation* [2e édition]. Bruxelles : De Boeck.

WILLIAMS W. M. & CECI S. J. (1997). « How'm I doing? Problems with student ratings of instructors and courses ». *Change*, vol. 29, n° 5, p. 13-23.

WORTHINGTON A. G. & WONG P. T. P. (1979). « Effects of earned and assigned grades on student evaluation of an instructor ». *Journal of Educational Psychology*, n° 71, p. 764-775.

WRIGHT W. A. & O'NEIL M. C. (1995). « Teaching improvement practices: international perspectives ». In W. A. WRIGHT (éd.), *Teaching improvement practices. Successful strategies for higher education*. Bolton : Anker Publishing, p. 1-57.

YOUNES N. (2006). *L'effet évaluation de l'enseignement supérieur par les étudiants.* Thèse de doctorat (non publiée), sciences de l'éducation, université de Grenoble.