Qualités métriques des résultats académiques universitaires

Léon Harvey
Université du Québec à Rimouski

Marie-Hélène Hébert
Université du Québec à Rimouski

Catherine Simard
Université du Québec à Rimouski

Résumé

Cet article présente une synthèse des écrits qui attestent que les résultats obtenus par les étudiants dans le cadre d'une formation universitaire, et rapportés au dossier académique, sont valides, fiables et en lien avec les objectifs des programmes. Sont passés en revue la validité et la fiabilité des résultats académiques, les indices utilisés, les principales procédures, ainsi que les principaux constats au niveau universitaire.

Mots clés : résultats académiques, validité, fiabilité, analyse factorielle, assurance qualité

Abstract

This paper reviews studies that claim that results obtained by students during university schooling, as reported in their academic files, are valid, reliable and related to the program objectives. The themes examined include the validity and reliability of grades, the indices used, the main procedures, and some important aspects concerning university education.

Keywords: grades, validity, reliability, factorial analysis, quality assurance

Introduction

Il existe, dans les écrits, un intérêt grandissant pour les mécanismes complexes qui régissent l'évaluation et la transparence dans l'obtention des diplômes afin d'assurer la qualité en enseignement. En particulier, la croissance des curriculums de formation professionnalisants génère des pressions accrues de la communauté scientifique, du public, des médias, des gestionnaires et des acteurs eux-mêmes (étudiants, enseignants), afin de déterminer comment ces programmes permettent d'attester du développement des compétences (Ryan & Cousins, 2009). Dans certaines professions, la certification professionnelle est réalisée *a posteriori* par des organismes externes autres que l'institution de formation. Ces derniers sont responsables de la mise en place d'épreuves valides qui attesteront de la compétence des futurs professionnels. Dans d'autres professions, dont l'éducation, l'admission à l'exercice de la profession repose sur l'obtention par le candidat des crédits prévus au programme de formation (Louis, Jutras, & Hensler, 1996).

En parallèle, différents organismes d'accréditation sont mis en place tant en Europe, aux États-Unis qu'au Canada afin de procéder à l'audit des programmes universitaires de formation en fonction de critères de qualité (Conseil Supérieur de l'Éducation, 2012). Ces organismes ont des attentes claires face à l'évaluation des retombées de l'apprentissage. Aux États-Unis, le *Council of Higher Education Accreditation* (CHEA) (2003) considère que les institutions et les programmes sont responsables d'établir des énoncés clairs relatifs aux acquis et aux aptitudes des étudiants en fin de cursus et, spécifie que des données probantes à cet effet doivent être recueillies et communiquées (voir Lewis, 2011). En Europe, l'*European Association for Quality Assurance in Higher Education* (2009) a

publié des standards généraux destinés à l'enseignement supérieur ; le standard 1.3 stipule que les étudiants doivent être évalués en fonction de critères, de règles et de procédures qui sont appliqués de façon cohérente. Au Québec, le *Comité d'Agrément des Programmes de Formation à l'Enseignement* (CAPFE, 2010) procède spécifiquement à l'audit des programmes de formation des maîtres et « reconnaît que l'évaluation des compétences comporte un niveau certain de difficulté, mais il s'attend [. . .] à ce que les universités recueillent des données et des observations qui témoignent du développement des douze compétences professionnelles chez leurs étudiantes et étudiants » (p. 4). Si certaines politiques institutionnelles explicitent que les évaluations doivent être équitables, équivalentes, valides et fiables, et en cohérence avec les objectifs d'un programme (p. ex., Collège d'enseignement général et professionnel Marie-Victorin, 2005), des preuves empiriques sont nécessaires afin de démontrer que ces critères sont respectés.

Les résultats académiques ont trois fonctions dans un système éducatif (Lekholm & Cliffordson, 2008). La fonction première est de donner de l'information à l'étudiant quant à l'atteinte des objectifs d'une formation. Ces objectifs peuvent être variés. Ils sont généralement clairement spécifiés par ledit programme qui se doit de décliner les produits d'apprentissage attendus au terme de la formation. Ces produits peuvent notamment se décliner en connaissances liées à des contenus considérés essentiels par les experts d'un domaine, ou en des compétences, attitudes et comportements particuliers à une profession. Les résultats académiques ont également comme fonction de réguler les apprentissages et permettre de sélectionner les étudiants pour les formations et même les emplois ultérieurs. Finalement, ils sont une base d'information utilisée pour évaluer la qualité d'un système éducatif. Considérant ces trois fonctions essentielles, un système éducatif qui évalue d'une manière inadéquate les apprentissages réalisés dans le cadre d'une formation ne peut pas être considéré comme un système de qualité. L'ensemble des fonctions se trouve alors compromis.

Or, régulièrement depuis 30 ans, les écrits (Agazzi, 1967; *Assessment and Learning Research Synthesis Group*, 2004; Pfister, 1975; Smith, 1992) soulignent que les évaluations réalisées par les enseignants ont la réputation d'être peu fiables et sujettes à des biais. De plus, les notes « A » et « B » sont obtenues plus facilement que par le passé (Jewell & McPherson, 2012; Rojstaczer & Healy, 2012). De ce fait, les résultats académiques n'ont plus la même signification et l'enseignement supérieur semble perdre la confiance du public (Caruth & Caruth, 2013).

Dans cette perspective, les programmes de formation doivent démontrer qu'ils prévoient des mesures valides et fiables des connaissances et des compétences. Considérant que dans plusieurs systèmes éducatifs (Lekholm & Cliffordson, 2008), c'est le dossier académique qui permet d'attester formellement de l'atteinte de l'objectif visé grâce au cumul des différents crédits obtenus, la question fondamentale qui se pose alors est de savoir si ce dossier académique permet de porter un jugement valide et fiable afin d'attester du développement des compétences et des connaissances d'un domaine.

Objectif de la recherche

Notre objectif est de présenter une synthèse d'études qui permettent de vérifier si les résultats obtenus dans le cadre d'une formation universitaire et rapportés au dossier académique sont valides, fiables et en lien avec les objectifs des programmes de formation. Sont passés en revue la validité et la fiabilité des résultats académiques, les indices utilisés, les procédures qui émergent, ainsi que les constats au niveau universitaire.

La validité et la fiabilité des résultats académiques universitaires

L'évaluation en contexte de classe est le processus qui consiste à porter un jugement sur les apprentissages, à partir de données recueillies, analysées et interprétées, en vue de décisions pédagogiques et administratives (Ministère de l'Éducation du Québec (MEQ), 2003). Le jugement de l'enseignant est au centre de ce processus de cueillette, d'analyse et d'interprétation. En ce sens, l'évaluation en contexte de classe n'est pas un processus standard; les conditions de cueillettes, d'analyse et d'interprétation sont sujettes à des changements multiples en fonction des enseignants, des cours, des années, des étudiants et des institutions. Cette absence de conditions standards lors de l'évaluation ne rend pas obsolètes les questions relatives à la validité et a la fiabilité des résultats académiques. Au contraire, ces études sont d'autant plus pertinentes que de multiples sources de variations existent; elles permettent de distinguer ce qui, dans l'évaluation, est cohérent et lié à la présence de facteurs pertinents à la formation, de ce qui est attribuable à des fluctuations aléatoires ou non pertinentes à la formation.

Validité

La validité est considérée comme un jugement global porté sur l'ensemble des preuves empiriques et théoriques qui attestent que les interprétations et les actions basées sur les résultats à un test ou à une autre modalité d'évaluation sont adéquates (Messick, 1995). Dans cette définition, la validité est un concept unifié, intégrant des aspects liés au contenu et à la représentativité des évaluations, à la structure et au construit, à des critères internes et externes, ainsi qu'à la généralisabilité et aux conséquences de l'évaluation. De plus, les preuves doivent permettre d'éviter deux menaces importantes à la validité. La première est la présence de facteurs non pertinents à l'évaluation. La seconde est la sous-représentation du construit qui indique que ce qui est évalué ne couvre pas l'ensemble du concept qui est attendu et, qu'ainsi, la mesure est incomplète. Cette conception de la validité concilie l'apport de considérations théoriques et empiriques, ainsi que les conséquences sociales et éthiques de l'utilisation de l'outil sous validation. Cette définition inclut les différentes formes de validité dont la validité de construit et la validité de critère qui sont principalement utilisées pour la validation des résultats académiques. La validité de construit réfère à l'adéquation entre la structuration attendue des évaluations en lien avec celle empiriquement obtenue.

La validation à partir de critères externes est également largement utilisée. Les critères externes de référence sont variés. Il peut s'agir d'antécédents académiques, de succès concomitants ou conséquents à une formation universitaire. Généralement, la relation avec les antécédents est étudiée à partir de la capacité des études antérieures ou de tests variés d'aptitudes à prédire le succès à l'université. Les succès concomitants sont obtenus à partir de résultats à des tests externes qui sont mis en relation avec la moyenne cumulative. Les tests externes sont la plupart du temps des tests de connaissances d'une discipline spécifique (p. ex., mathématiques, langue, psychologie, biologie). Dans certains pays tels que la Suède (Lekholm & Cliffordson, 2008) et les États-Unis, le succès à des tests nationaux concomitants est également rapporté. Les conséquents aux études universitaires peuvent se mesurer à partir de l'accomplissement dans la carrière ultérieure en termes de prestige occupationnel, de salaire ou par l'obtention de diplômes ultérieurs (Baird, 1985; Pattison, Grodsky, & Muller, 2013).

Borsboom, Mellenbergh et van Heerden (2004) proposent une autre conception de la validité. Ils considèrent que les questions liées à la validité sont d'ordre ontologique et liées à l'existence même d'un attribut et de son lien de causalité avec les résultats des

évaluations. Ils considèrent qu'un instrument procure une mesure valide d'un attribut lorsqu'il est possible de démontrer que cet attribut existe et que des fluctuations de celui-ci provoquent (dans le sens de causalité) des changements dans les résultats mesurés. Cette conception s'avère utile lorsqu'il est possible de manipuler expérimentalement certains aspects du curriculum à valider. En ce sens, Goova, Hollett, Tesfay, Gala, Puzziferri, Kehdy et Scott (2008) ont utilisé une démarche quasi-expérimentale pour valider leur construit mesuré lors d'un curriculum en médecine dédié spécifiquement au développement de la compétence à suturer des points. À partir de dispositifs technologiques de simulation de tâches complexes, la démarche permet d'effectuer un suivi de la progression des étudiants et de la comparer à celles d'experts sur le même ensemble de tâches standardisées.

Fiabilité

La fiabilité réfère à la qualité de la mesure (Saupe & Eimers, 2012). Une mesure est fiable lorsqu'elle génère une erreur de mesure faible par rapport à ce qui est mesuré. Comme il a été mentionné auparavant, pour certains acteurs, les résultats académiques sont peu fiables, car ils sont constitués d'une erreur de mesure forte (voir Smith, 1992). Dans cette perspective, le défi est de taille pour les systèmes d'assurance de la qualité en enseignement; ils doivent ainsi faire preuve de transparence et démontrer que les évaluations qui sont réalisées dans le cadre de leurs formations sont bel et bien fiables.

Les procédures disponibles pour établir la fiabilité des évaluations sont diverses mais ne sont pas toutes applicables à l'analyse des résultats académiques (Saupe & Eimers, 2012). Certaines requièrent l'utilisation de deux mesures telles que l'utilisation de formes parallèles et équivalentes d'un même test (p. ex, version A vs version B); cette condition n'a pas réellement de sens dans l'étude du résultat moyen (GPA). La procédure test-retest nécessite également que deux mesures identiques soient réalisées avec le même instrument; cette procédure ne s'applique que si on considère que les résultats moyens d'un trimestre sont comparables à ceux d'un second trimestre. La corrélation entre deux trimestres successifs peut alors être utilisée.

Des procédures alternatives existent et ne nécessitent qu'une prise de mesure du test. Elles peuvent ainsi s'appliquer sur les résultats académiques où il n'y a qu'une prise de mesure (une note unique) à chacun des cours. Ces procédures sont basées sur le calcul de la consistance interne (p. ex., le coefficient alpha de Cronbach), sur la fiabilité fractionnée de l'échelle (corrélation d'une moitié de l'échelle avec l'autre) ou sur l'analyse de

la variance (p. ex., étude de généralisabilité). Les procédés psychométriques permettent ainsi de distinguer ce qui dans une note est attribuable aux fluctuations individuelles entre les élèves, de ce qui est attribuable aux différences entre les enseignants, les classes, les institutions ou autres facteurs (Brown, 2006; O'Connell & McCoach, 2008).

Les indices considérés

Le dossier académique est le dossier faisant état des progrès d'un étudiant tout au long de sa formation (Ministère de l'Éducation de l'Ontario, 2000). Il répertorie différents indices de la réussite académique; il n'y a pas cependant de consensus sur celui qui reflète le mieux la réussite des objectifs d'une formation. Les indices les plus souvent utilisés sont les résultats aux cours et la moyenne cumulative. Cependant, les résultats dans certains travaux (Mason & Dragovich, 2010), les crédits obtenus (Smith, 1992), la graduation, l'attrition ou la persistance dans les études, ainsi que l'obtention de mentions spéciales sont également considérés (Camara & Echternacht, 2000). Dans le cadre de la présente recension, seuls les principaux indices de la réussite académique seront considérés, soit les résultats à certains travaux, la note obtenue dans les cours et la moyenne cumulative.

La moyenne académique cumulée sur l'ensemble des crédits obtenus (*grade point average,* GPA) est largement utilisée dans la documentation pour des fins de validation (Luthy, 1996; Saupe & Eimers, 2012). Le GPA peut être calculé sur l'ensemble des années (p. ex., 4 ans) ou sur des périodes de temps plus limitées (un trimestre, la première année ou l'année la plus récente, etc.). Cependant, cette note unique ne donne que peu d'information en lien avec les trois fonctions visées par les résultats académiques. Ce résultat unique informe peu sur l'atteinte des objectifs d'une formation, ne permet pas de réguler précisément les apprentissages, et ne donne que peu de rétroaction aux concepteurs (ou évaluateurs) de programmes.

Ainsi, Mason et Dragovich (2010) considèrent que la moyenne cumulative et les notes aux cours sont des scores composés et que ceux-ci ne peuvent pas être utilisés directement pour attester l'atteinte d'un objectif de formation parmi un ensemble. Ces auteurs mentionnent, à titre d'exemple, qu'un score de 80% obtenu dans un cours qui poursuit deux objectifs peut soit attester de l'atteinte de ceux-ci à 80% chez un étudiant ou alternativement, être obtenu par un étudiant qui n'atteint que difficilement l'un des deux objectifs, mais qui excelle dans l'autre.

Les limites de la moyenne cumulative ne sont pas nouvelles. Dans les années 70, Pfister (1975) fait également valoir qu'il n'est pas possible d'utiliser directement les notes scolaires telles qu'elles sont données par les enseignants. Ainsi, la note obtenue dans une classe n'est pas comparable à la même note obtenue dans une autre classe, car les critères diffèrent d'un enseignant à un autre. Cet auteur, suivant la recommandation de la commission d'étude suisse de l'époque, a utilisé l'écart à la moyenne de la classe pour l'étude des résultats scolaires au primaire. Cet écart exprime la distance entre le résultat d'un élève et celui de sa classe. La démarche de Pfister illustre que les résultats académiques peuvent être utilisés comme substrat à des analyses mais nécessitent un traitement psychométrique particulier. À notre connaissance cependant, l'écart à la moyenne est peu usité en psychométrie. Les dossiers académiques rapportent parfois le rang cinquième (ou autre rang) et la moyenne du groupe; le seul indice utilisé qui effectue une correction de la note individuelle en fonction de l'écart au groupe est la cote de rendement au collège (Conférence des recteurs et des principaux des universités du Québec, 2013) au Québec. Il n'existe cependant pas d'équivalent au niveau universitaire.

Ainsi, les procédures utilisées dans les écrits (Harvey, 2012; Lekholm & Cliffordson, 2008; Rexwinkel, Haenen, & Pilot, 2013; Thorsen & Cliffordson, 2012) ne sont pas basées sur la comparaison directe d'un résultat unique d'un contexte à un autre ou sur l'utilisation de l'écart à la moyenne, mais plutôt sur l'extraction des facteurs qui sous-tendent les évaluations. L'extraction des facteurs permet de constituer une mesure formée des composantes communes aux évaluations réalisées et permet également de corriger cette mesure en tenant compte de l'erreur aléatoire ou de celle spécifique au contexte (cours, enseignant, etc.).

Les différentes procédures

Plus spécifiquement, l'analyse factorielle (Bourque, Poulin, & Cleaver, 2006; Brown, 2006) est fréquemment utilisée afin de valider un construit (Muis & Winne, 2012). L'analyse permet d'inférer la présence d'états latents. Un état latent est défini comme un construit non directement observable, mais dont les valeurs peuvent être estimées à partir de données observables. En éducation et en psychologie, la notion d'état latent réfère très largement à l'état interne d'un individu. Griffin (2007) précise qu'il n'y a pas de restriction quant à la nature des variables latentes mesurées que celles-ci soient des

connaissances acquises, des attitudes ou des compétences (p. ex., Morlaix, 2009). Toujours selon Griffin (2007), il n'y a pas non plus de restriction en ce qui a trait à la nature des tâches considérées. Il peut s'agir de tests standardisés, mais également des performances observées lors de tâches en milieu professionnel (p. ex., Harvey, 2009), des folios, des variables associées à la production langagière, etc.

L'analyse factorielle a deux variantes : les procédures de validation exploratoire et confirmatoire. Par définition, une procédure exploratoire compare un modèle de formation obtenu *a posteriori* à des critères d'acceptabilité. Le modèle obtenu doit notamment être interprétable à partir du curriculum étudié. En ce sens, la procédure proposée par Rexwinkel et al. (2013) est exploratoire et se décline en cinq étapes : l'examen de l'ensemble des données, l'analyse du construit, la fiabilité des échelles de mesure, l'inspection de la matrice de corrélations et la vérification auprès d'acteurs concernés afin d'assurer la validité apparente (*face validity*). Lors de l'inspection des résultats scolaires, sur une échelle à 10 échelons (de 1, vraiment faible à 10, excellent), ces auteurs considèrent que des cours avec un écart-type d'environ 0,70 sont acceptables. À l'opposé, des écarts-types inférieurs à 0,34 et supérieurs à 1,40 sont respectivement considérés trop faibles et trop accentués. L'analyse factorielle exploratoire est utilisée pour déterminer la nature des construits mesurés; une structure factorielle forte, i.e. où les valeurs propres (*eigenvalues*) supérieures à 1 expliquent plus de 50% de la variance, est recherchée. La fiabilité des résultats scolaires doit également être supérieure à 0,60. La matrice de corrélation doit présenter des relations positives et significatives entre les cours. Des corrélations positives indiquent que des construits communs sont mesurés; des corrélations nulles sont acceptables dans l'éventualité où il est prévu, dans le curriculum, que des cours ne mesurent pas de construits communs; finalement, la présence de corrélations négatives est problématique, car elles indiquent que ce qui est valorisé et mesuré dans un cours est dévalorisé et négativement évalué dans un autre. Toute corrélation négative devrait faire l'objet d'explications de la part des responsables de programme.

Finalement, des études utilisant des questionnaires permettent de vérifier l'opinion des acteurs (étudiants, professeurs) quant à la qualité des résultats académiques. L'ensemble des opinions doit être favorable et dépasser un taux d'acceptation de 50%.

L'approche proposée par Rexwinkel et al. (2013) est intéressante dans la perspective où elle peut être appliquée à tous les types de programmes. Elle possède également une valeur diagnostique puissante en ce qu'elle fournit une information précieuse

aux concepteurs de programme; elle permet de porter un jugement sur la contribution de chacun des cours à l'évaluation de construits sous-jacents à la formation. Les cours qui ne contribuent pas tel qu'attendu peuvent ainsi être revus afin d'apporter des correctifs dans les procédures d'évaluation et d'enseignement.

Un des désavantages de la procédure de Rexwinkel et al. (2013) est que le jugement qui est porté sur la structure factorielle demeure très relatif. Ainsi, ce n'est pas parce qu'il existe une structure factorielle exploratoire forte que celle-ci correspond à l'intention initiale du programme de formation. Les structures théoriques et empiriques peuvent ne pas être en adéquation. Cette procédure permet de valider la cohérence des résultats académiques, mais constitue un test ambigu de l'adéquation avec ce qui est prévu au programme de formation.

En complément à une approche exploratoire, il est judicieux de recourir à une approche confirmatoire. Une procédure est confirmatoire lorsqu'un modèle de formation existe *a priori,* que ce modèle spécifie explicitement les liens qui existent entre les objectifs du programme, les cours et les évaluations réalisées et que ce modèle est mis à profit lors des étapes de la validation. Une approche confirmatoire vérifie l'adéquation entre la structuration des évaluations et l'intention du programme. Ce type de procédure est plus restrictif, plus contraignant, et permet d'apporter des preuves supplémentaires quant au construit mesuré dans le cadre d'une formation. Le modèle de formation se doit également d'être fertile et doit permettre de générer de nouvelles hypothèses, qui une fois confirmées, constitueront des preuves supplémentaires de la qualité des résultats associés à une formation. Mason et Dragovich (2010), Harvey (2012), Lekholm et Cliffordson (2008) utilisent des procédures confirmatoires.

Principaux constats

Les études recensées couvrent des secteurs variés et proviennent principalement de secteurs où il existe des ordres professionnels ou des organismes externes d'accréditation. Ces études couvrent alors des secteurs tels que le génie (Mason & Dragovich, 2010), la psychologie (Smith, 1992), la médecine vétérinaire et la physiothérapie (Rexwinkel et al., 2013), les études en marketing (Bacon & Bean, 2006) ainsi que l'éducation (Harvey, 2012). Smith (1992) a également considéré des disciplines telles que la biologie et les sciences biologiques, la chimie et l'anglais. Quelques études (Kuncel, Credé, & Thomas, 2007; Kuncel, Wee, Serafin, & Hezlett, 2010; Luthy, 1996) se sont intéressées aux résultats académiques aux études avancées.

Validation du construit

Rexwinkel et al. (2013) ont utilisé une procédure exploratoire afin d'analyser les évaluations réalisées dans les cours en médecine vétérinaire et en physiothérapie. En médecine, des corrélations positives de faibles (0,09) à moyennes (0,48) ont été observées entre les cours du programme avec une bonne fiabilité de l'ensemble (alpha de Cronbach de 0,75). De plus, les cours saturent fortement dans trois facteurs qui expliquent plus de 50% de la variance totale. À l'opposé, en physiothérapie, des corrélations négatives (-0,20) à faibles (0,17) ont été observées entre les cours avec un indice de fiabilité insatisfaisant (alpha de Cronbach = 0,20). De plus, les cours saturent faiblement dans trois facteurs et expliquent moins de 50% de la variance totale. À partir des critères psychométriques établis, les évaluations réalisées dans le cadre du programme de médecine vétérinaire ont été validées, tandis que celles réalisées en physiothérapie ne le furent pas.

En éducation, une étude (Harvey, 2012) suggère qu'un programme d'enseignement secondaire d'une université québécoise possède une bonne validité de construit curriculaire, et ce, basée sur une analyse confirmatoire incluant 14 cours de psychopédagogie, dont des stages en milieu professionnel. Ainsi, la matrice de corrélation entre les cours révèle que 99% (90 sur 91) des coefficients sont soit positivement corrélés (64%, 58 sur 91) ou non corrélés (35%, 32 sur 91) entre eux. La fiabilité (alpha de Cronbach) de l'ensemble des cours est de 0,69, supérieure au critère de 0,60 fixé par Rexwinkel et al. (2013). Dans l'ensemble, ce curriculum atteint les critères de qualité fixés par la procédure de Rexwinkel et al. (2013). L'atteinte des objectifs de formation a fait l'objet d'une attention spéciale. Ainsi, les résultats académiques ont été soumis à une analyse factorielle confirmatoire, où la mesure de quatre groupes de compétences est confirmée. Ces groupes sont liés, tel qu'attendu dans le plan de formation, aux fondements de l'éducation, à l'acte d'enseigner, aux aspects sociaux et au développement de l'identité professionnelle. L'analyse révèle également une continuité entre ce qui est évalué dans les cours en institution et les cours de stages en milieu professionnel. Globalement, l'étude suggère que ce curriculum atteint 75% des objectifs visés par le programme. Pour arriver à cet estimé, les objectifs visés par chacun des cours en lien avec chacun des groupes de compétences attendues ont été précisés puis, utilisés comme prédicteurs dans une analyse factorielle confirmatoire. De cette analyse, 75% des saturations factorielles attendues se sont avérées significatives.

Cependant, l'un des quatre groupes de compétences, celui lié au développement identitaire, contient des saturations polarisées (positives et négatives) qui suggèrent des incohérences dans le curriculum quant à la mesure de cette dimension. L'origine de cette polarisation est attribuée à des conceptions différentes de cette composante par différents acteurs de la formation (Chevrier, Gohier, Anadon, & Godbout, 2007). De plus, plusieurs saturations dans les facteurs sont faibles et laissent entrevoir une fiabilité faible de ces dimensions.

L'étude de Mason et Dragovich (2010) permet de tracer l'adéquation entre l'évaluation de certains travaux, les objectifs poursuivis dans le cadre des cours, ainsi qu'avec les objectifs de la formation. Le personnel enseignant doit cependant procéder à la spécification des liens entre ces trois entités (travaux, objectifs de cours, objectifs de programme). La spécification de telles valeurs initiales est complexe et difficile à implanter lorsqu'un programme mobilise un grand nombre de travaux, d'objectifs de cours et de programmes et qu'un grand nombre de ressources professorales interviennent dans le cursus. Elle nécessite de plus l'implantation et le maintien de bases de données nouvelles et concurrentes avec celles actuellement existantes; il peut alors exister des résistances importantes à l'implantation, au maintien et à l'utilisation de ces bases de données.

Validation de critère

Les critères considérés sont soit liés aux réalisations antérieures, présentes, ou futures des étudiants. Ainsi, le succès dans les études universitaires peut être prédit à partir des résultats antérieurs et à partir de tests standardisés d'aptitudes. Le GPA au secondaire et les tests standardisés d'aptitudes (Camara & Echternacht, 2000; DiPerna, 2004; Geiser & Santelices, 2007; Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008; Komarraju, Ramsey & Rinella, 2013; Pattison et al., 2013; Pike & Saupe, 2002) s'avèrent de bons prédicteurs du succès (GPA) des études universitaires. Aux études supérieures, des métaanalyses confirment la robustesse d'un test d'aptitude et de la moyenne cumulative au premier cycle à prédire la moyenne cumulative en gestion des affaires (Kuncel et al., 2007) et d'un test d'admission aux études graduées à prédire le succès lors d'études de maîtrise et de doctorat (Kuncel et al., 2010).

La validité concomitante a été investiguée par Smith (1992) à partir de cinq programmes de premier cycle, par Bacon et Bean (2006) dans un programme de baccalauréat en marketing et par Luthy (1996) aux études avancées dans neuf secteurs. Des

tests externes ont été utilisés et corrélés avec le GPA. Dans l'ensemble de ces études, des relations positives ont été observées entre les tests externes et le GPA universitaire.

Le lien entre les résultats académiques universitaires et les conséquences ultérieures en termes de réalisations professionnelles est controversé. Si dans certains secteurs, il est primordial que les aptitudes à réussir les tâches sollicitées en contexte universitaire préparent directement à la réalisation des mêmes tâches en milieu professionnel, dans d'autres secteurs, cette relation est beaucoup moins importante. La littérature rapporte régulièrement une relation (p. ex., Jones & Jackson, 1990), mais tout secteur confondu, elle est généralement considérée faible (Baird, 1985; Pattison et al., 2013; Wingard & Williamson, 1973).

Il existe par ailleurs une limite importante à tout pronostic à long terme du succès académique. Une étude, celle de Bacon et Bean (2006), rapporte une baisse progressive des corrélations entre les facteurs latents (équivalents annuels du GPA) du baccalauréat en marketing et ce, de la première année universitaire versus les années subséquentes. Ainsi, la capacité prédictive des résultats antérieurs diminue avec les années. Une diminution progressive des corrélations a déjà été observée ailleurs (Pfister, 1975). Sa provenance reste encore indéterminée. Elle peut provenir soit du transfert partiel des apprentissages d'une année à l'autre, de la transformation des compétences, de l'impact des conditions créées par le milieu ou de l'ensemble de ces facteurs.

Cette diminution des corrélations entre les résultats obtenus en fonction des années peut constituer un obstacle à la validation d'un construit et plus d'attention devrait y être accordée. Ainsi, deux cours qui mesurent la même compétence pourraient ne pas être en corrélation et ne pas saturer dans un facteur commun simplement parce qu'ils se situent respectivement en début et en fin de curriculum. Un curriculum universitaire de premier cycle est généralement offert sur une période de trois à quatre années. Durant cette période, plusieurs changements cognitifs, affectifs, motivationnels, situationnels ou autres peuvent intervenir et ce, qu'ils soient liés ou non aux cours offerts. Notamment, sur le plan cognitif, les compétences développées par les étudiants peuvent évoluer. Cette évolution est d'ailleurs spécifiée dans la progression des apprentissages dans certains curriculums, mais n'est cependant pas prise en compte explicitement dans les procédures exposées jusqu'à présent. Ces transformations des facteurs cognitifs et non cognitifs peuvent générer une variance non expliquée et affecter les preuves de la validité.

Les corrélations peuvent également être affectées par l'attrition du groupe d'étudiants entre le début et la fin du programme. Lorsque les dossiers académiques des étudiants diplômés sont utilisés, seuls les résultats des étudiants qui ont terminé sont analysés. Cette attrition contribue à deux phénomènes réduisant la force des corrélations : perte de sujets et restriction de l'étendue des cotes (les meilleurs demeurant dans le programme).

Une étude de la validité doit donc prendre en compte cette possible évolution des résultats académiques lorsqu'une formation s'échelonne sur plusieurs années ainsi que la possible attrition des étudiants.

Fiabilité

Finalement, la documentation (Bacon & Bean, 2006; Saupe & Eimers, 2012) confirme que la fiabilité de la moyenne cumulative (GPA) est excellente et est généralement supérieure à 0,80. Cependant, lorsqu'elle est calculée sur des intervalles plus restreints (un ou deux trimestres), elle peut être inférieure à 0,70 et s'avère alors moins fiable (Saupe & Eimers, 2012).

Tel qu'il a été mentionné dans l'introduction, un phénomène d'*inflation des notes* est également parfois perçu comme une menace à la fiabilité des résultats académiques. L'inflation des notes est définie comme une augmentation des résultats académiques qui n'est pas justifiée par un accroissement de la compétence des étudiants. Ce phénomène a été observé dans différents collèges et universités (Jewell & McPherson, 2012; Rojstaczer & Healy, 2012) et a fait passer les résultats académiques moyens de la note *C* à *B* au fil des ans. Ce phénomène d'inflation des notes soulève un problème d'équité entre les générations; il est une indication que certains diplômes sont plus faciles à obtenir que par le passé. Cependant, sur le plan psychométrique, il n'y a pas en soi d'effet de l'inflation des notes sur la fiabilité de la moyenne cumulative (Millman, Slovacek, Kulick, & Mitchell, 1983). Un changement dans les résultats moyens ne signifie pas qu'une formation n'atteint pas ses objectifs en termes d'évaluation. Selon Pattison et al. (2013), il faut plutôt s'interroger sur la valeur informative des diplômes. La valeur informative est la puissance de la moyenne cumulative à envoyer un signal adéquat (*signaling power of grades*) quant à la qualité des étudiants qui sont diplômés. Une augmentation de la moyenne ne change pas en soi cette valeur informative. Il faut plutôt vérifier les changements dans la variance des évaluations; changements qui informent que la distance qui sépare

les excellents étudiants de ceux qui sont soit très bons, moyens ou médiocres a changé avec le temps.

Ce qui est encore plus déterminant selon Pattison et al. (2013), ce sont des changements dans les covariations qui existent entre les antécédents des étudiants en termes de réussite et d'efforts et les conséquents des évaluations en termes de niveaux d'études atteints et de conditions de travail (prestige occupationnel, salaire, etc.). En ce sens, Pattison et al. (2013) n'ont trouvé aucune preuve qu'il y a eu une perte dans la valeur informative de la moyenne cumulative (GPA) entre 1972 et 1992 dans les universités américaines. Quoi qu'il en soit, l'inflation des notes au fil des années reste un phénomène préoccupant pour les programmes universitaires de formation, car cette inflation concerne directement la valeur des diplômes et qu'elle est largement médiatisée. Elle doit faire l'objet d'un suivi par les programmes et les institutions dans une optique de maintien de la qualité (Caruth & Caruth, 2013).

La présence de facteurs non pertinents

La littérature rapporte fréquemment des différences dans les résultats académiques attribuables à des facteurs liées au genre (Lekholm & Cliffordson, 2008; Luthy, 1996; Pfister, 1975), à l'âge (Luthy, 1996) des étudiants, à des facteurs socio-économiques tels que l'éducation des parents et le revenu familial (Geiser & Santelices, 2007) ou l'ethnie (Fletcher & Tienda, 2010).

Des différences dans l'attribution des notes existent également entre les institutions qui offrent des programmes semblables qui sont liées soit au contexte ou au climat de l'institution (Ma, Ma, & Bradley, 2008). Des différences attribuables à la localisation (nord ou sud) de l'institution, à son caractère public ou privé, ainsi qu'à sa vocation (technique vs non technique) sont aussi rapportées et ne s'expliquent pas par des différences dans les aptitudes des candidats (Rojstacker & Healy, 2012). Devant la multitude de facteurs en présence, ainsi que les ressources en jeu, le défi est de taille pour tout système d'éducation. Tel qu'il a été mentionné précédemment, le critère de qualité fixé par Rexwinkel et al. (2013) est qu'un programme se doit d'expliquer au moins 50% de la variance dans les résultats académiques à partir de facteurs cohérents. En contrepartie, 50% de la variance reste inexpliquée, ce qui laisse amplement de latitude aux facteurs responsables d'iniquités dans les évaluations.

Discussion et conclusion

Avec le foisonnement des programmes par compétences et l'instauration à l'échelle mondiale des politiques d'assurance de la qualité, les pratiques d'évaluation subissent des transformations importantes et, dans l'ensemble d'un curriculum, il devient nécessaire de documenter les mécanismes d'évaluation des connaissances et des compétences professionnelles. En ce sens, cette recension a présenté une synthèse des études qui permettent de vérifier si les résultats obtenus dans le cadre d'une formation et rapportés au dossier académique sont valides et fiables et en lien avec les objectifs du programme.

Cette recension a permis de mettre en évidence que les résultats académiques peuvent s'avérer valides et fiables tout au long du parcours des études universitaires. Dans une perspective d'assurance qualité, il est de la responsabilité de chaque programme et de chaque institution d'en apporter les preuves. Certaines preuves sont associées à la moyenne cumulative (GPA) qui s'avère généralement valide et fiable. Cette moyenne cumulative a cependant une valeur informative limitée et plus d'informations sont nécessaires afin de remplir les trois visées d'un dossier académique : information auprès des étudiants, régulation des apprentissages et rétroaction auprès des concepteurs de programmes.

Dans cette perspective, des preuves supplémentaires doivent provenir des résultats obtenus dans les différents cours d'un programme en lien avec les objectifs visés. À ce titre, les corrélations qui existent entre les résultats académiques obtenus dans les cours entre eux sont une preuve essentielle. L'ensemble de ces corrélations permet d'attester que les évaluations ne sont pas aléatoires et qu'elles sont en continuité. En ce sens, des corrélations qui varient entre faiblement négatives (-0,20) à moyennement positives (environ 0,40) sont recensées.

L'identification des facteurs à partir d'analyses factorielles ajoute à ces preuves et permet de confirmer la mesure de construits communs entre les cours d'un curriculum. Dans la plupart des études recensées, trois ou quatre facteurs émergent. Cependant, les facteurs obtenus par analyse exploratoire n'émergent qu'*a posteriori* et peuvent être très différents de ceux projetés par le modèle de formation qui lui est déterminé *a priori* lors de la conception du programme. L'analyse confirmatoire permet de confronter un modèle théorique de formation avec les évaluations réalisées. Or, cet exercice est très contraignant et à ce stade-ci de l'avancement des connaissances, peu d'études permettent de con-

firmer que les évaluations reportées au dossier académique sont clairement en adéquation avec le plan initial de formation.

L'utilisation d'épreuves externes à la formation permet de clarifier partiellement cet enjeu. Ainsi, la validation des résultats académiques à partir de critères externes (Bacon & Bean, 2006; Luthy, 1996; Smith, 1992) constitue une preuve concomitante forte du construit mesuré par les enseignants et attendu au terme des études.

Par ailleurs, il est difficile de déterminer si les réformes des curriculums ont permis d'augmenter la qualité des évaluations qui sont réalisées. Aucune étude ne compare la qualité des évaluations réalisées avant les réformes (obtenues avant 1990) avec celles obtenues plus récemment et des recherches en ce sens devraient également être entreprises afin d'évaluer l'impact des réformes.

L'évolution des construits mesurés dans le temps devrait également recevoir plus d'attention dans une perspective d'assurance qualité. Une baisse des corrélations entre les résultats de la première année et des années subséquentes est parfois rapportée et représente une limite à la capacité prédictive des résultats académiques (Bacon & Bean, 2006; Pfister, 1975). Une baisse prononcée et non anticipée par les concepteurs d'un programme, constitue une variance non expliquée. Cette variance non expliquée peut alors affecter la validité et la fiabilité des résultats.

Par ailleurs, une limite des études recensées est liée à la nature du produit de l'apprentissage évalué. Cette documentation n'effectue pas de distinction entre les notions de connaissances ou de compétences. Or, à notre connaissance, au moins un organisme d'accréditation, le CAPFE, exige des preuves en ce sens. En fait, cet aspect est une limite des dossiers académiques, car les produits évalués n'y sont pas distingués. Par conséquent, les procédés de validation empiriques ne peuvent suppléer adéquatement à cette lacune. Mason et Dragovich (2010) suggèrent de mettre en place un système de suivi des résultats des travaux réalisés dans le cadre des cours et de les mettre directement en lien avec les objectifs des cours et des programmes. D'autres (voir Guédé, 2009) proposent des systèmes de gestion et de suivi des compétences qui pourraient faciliter la validation des procédés utilisés à partir de procédures psychométriques. Cependant, comme il a été mentionné précédemment, de tels systèmes de gestion et de suivi nécessitent l'implantation et l'utilisation de nouveaux systèmes d'information incompatibles avec ceux actuellement en place et nécessitent un changement important de culture.

Par souci de représentativité, il apparaît nécessaire de préciser que le présent article est influencé par un biais inévitable de sélection et de publication; les articles sélectionnés sont ceux dont les résultats sont publiés et projettent une image positive des formations à l'étude. Les preuves moins positives sont rarement diffusées. L'étude de la validité de construit curriculaire est une démarche relativement nouvelle qui émerge en réponse aux besoins d'assurer la qualité des produits de l'apprentissage des formations universitaires; les écrits sur cette thématique sont encore peu nombreux, et de ce fait, nos recommandations restent basées sur un nombre relativement limité d'études.

Néanmoins, dans la perspective où les organismes accorderont encore plus de place à l'évaluation des résultats et aux retombées de l'apprentissage tel que le suggèrent certains écrits (Conseil Supérieur de l'Éducation (CSE), 2012; Lewis, 2011), cet article peut servir de guide aux acteurs concernés dans l'élaboration d'un ensemble de preuves qui attestent que le dossier académique de leur programme permet de porter un jugement valide et fiable quant au développement des compétences et des connaissances attendues au terme d'une formation.

Références

Agazzi, A. (1967). *Les aspects pédagogiques des examens*. Strasbourg, France : Conseil de l'Europe.

Assessment and Learning Research Synthesis Group. (2004). *A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes*. University of London, England: EPPI-Centre.

Bacon, D. R., & Bean, B. (2006). GPA in research studies: An invaluable but neglected opportunity. *Journal of Marketing Education*, *28*(1), 35–42. doi: 10.1177/0273475305284638

Baird, L. L. (1985). Do grades and tests predict adult accomplishment? *Research in Higher Education*, *23*(1), 3–85. doi: 10.1007/BF00974070

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Bourque, J., Poulin, N., & Cleaver, A. F. (2006). Évaluation de l'utilisation et de la présentation des résultats d'analyses factorielles et d'analyses en composantes principales en éducation. *Revue des sciences de l'éducation*, *32*(2), 325–344. doi: 10.7202/014411ar

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Camara, W. J., & Echternacht, G. (2000). *The SAT I and high school grades: Utility in predicting success in college* (Research Notes RN-10). New York, NY: The College Board, Office of Research and Development.

Caruth, D. L., & Caruth, G. D. (2013). Grade inflation: An issue for higher education? *Turkish Online Journal of Distance Education*, *14*(1), 102–110.

Chevrier, J., Gohier, C., Anadon, M., & Godbout, S. (2007). Construction de l'identité professionnelle des futures enseignantes : dispositifs de formation présents et souhaités selon les acteurs responsables de la formation des maîtres au préscolaire et au primaire. Dans C. Gohier (Éd.), *Identités professionnelles d'acteurs de l'enseignement : regards croisés* (pp.137–168). Québec, QC : Presses de l'Université du Québec.

Collège d'enseignement général et professionnel Marie-Victorin. (2005). *Politique no-12, Politique institutionnelle d'évaluation des apprentissages*. Repéré à http://www.collegemv.qc.ca/CMS/Media/2226_294_fr-CA_0_pol_12_PIEA.pdf

Comité d'Agrément des Programmes de Formation à l'Enseignement (CAPFE). (2010). *La visite de suivi de l'agrément d'un programme de formation à l'enseignement : cadre de référence et modalités d'application pour les visites de suivi de l'agrément qui seront effectuées entre 2010 et 2013*. Québec, QC : Ministère de l'Éducation, du Loisir et du Sport.

Conférence des recteurs et des principaux des universités du Québec. (2013). *La cote de rendement au collégial : aperçu de son rôle et de son utilisation.* Document approuvé le 30 novembre 2000 par le Comité de gestion des bulletins d'études collégiales. Mis à jour le 4 mars 2013. Repéré à http://www.crepuq.qc.ca/spip.php?article227&lang=fr.

Conseil Supérieur de l'Éducation (CSE) (2012). *L'assurance qualité à l'enseignement universitaire : une conception à promouvoir et à mettre en œuvre. Avis à la ministre de l'Éducation, du Loisir et du Sport.* Québec, QC : Gouvernement du Québec.

Council for Higher Education Accreditation (CHEA). (2003). *Statement of mutual responsibilities for student learning outcomes: Accreditation, institutions, and programs*. Washington, DC: Council for Higher Education Accreditation.

DiPerna, J. C. (2004). Structural and concurrent validity evidence for the Academic Competence Evaluation Scales-College edition. *Journal of College Counseling*, *7*(1), 64–72. doi: 10.1002/j.2161-1882.2004.tb00260.x

European Association for Quality Assurance in Higher Education. (2009). *Standards and guidelines for quality assurance in the european higher education aera.* Helsinki, Finland: ENQA.

Fletcher, J., & Tienda, M. (2010). Race and ethnic differences in college achievement: Does high school attended matter? *The Annals of the American Academy of Political and Social Science*, *627*(1), 144–166. doi: 10.1177/0002716209348749

Geiser, S., & Santelices, M. V. (2007). *Validity of high-school grades in predicting student success beyong the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes.* Research & Occasional Paper Series: CSHE.6.07. University of California, Berkeley. Repéré à http://cshe.berkeley.edu/sites/default/files/shared/publications/docs/ROPS.GEISER._SAT_6.13.07.pdf

Goova, M. T., Hollett, L. A., Tesfay, S. T., Gala, R. B., Puzziferri, N., Kehdy, F. J., & Scott, D. J. (2008). Implementation, construct validity, and benefit of a proficiency-based knot-tying and suturing curriculum. *Journal of Surgical Education*, *65*(4), 309–315. doi: 10.1016/j.jsurg.2008.04.004

Griffin, P. (2007). The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation*, *33*, 87–99. doi: 10.1016/j.stueduc.2007.01.007

Guédé, V. (2009). Comparatif des applications informatiques de gestion des compétences. *Cahiers pédagogiques*, *476*. Repéré à http://www.cahiers-pedagogiques.com/Comparatif-des-applications

Harvey, L. (2012). Évaluation des compétences dans un programme de formation en enseignement : Validité de construit curriculaire. *Mesure et évaluation en éducation, 35*(2), 69–97.

Harvey, L. (2009). L'échafaudage lors de la supervision en milieu professionnel : études des modalités et un modèle. *Mesure et évaluation en éducation, 32*(1), 55–83.

Jewell, R. T., & McPherson, M. A. (2012). Instructor-specific grade inflation: Incentives, gender, and ethnicity. *Social Science Quarterly, 93*(1), 95–109. doi: 10.1111/j.1540–6237.2011.00827.x

Jones, E. B., & Jackson, J. D. (1990). College grades and labor market rewards. The *Journal of Human Resources, 25*(2), 253–266. doi: 10.2307/145756

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for predicting first-year college grade point average* (Research report no-2008-5). New York, NY: The College Board.

Komarraju, M., Ramsey, A., & Rinella, V. (2013). Cognitive and non-cognitive predictors of college readiness and performance: Role of academic discipline. *Learning and Individual Differences, 24*, 103–109. doi : dx.doi.org/10.1016/j.lindif.2012.12.007

Kuncel, N. R., Credé, M., & Thomas, L. L. (2007). A meta-analysis of the predictive validity of the graduate management admission test (GMAT) and undergraduate grade point average (UGPA) for graduate student academic performance. *Academy of Management Learning & Education, 6*(1), 51–68.

Kuncel, N. R., Wee S., Serafin, L., & Hezlett, S. A. (2010). The validity of the graduate record examination for master's and doctoral programs: A meta-analytic investigation. *Educational and Psychological Measurement, 70*(2), 340–352. doi: 10.1177/0013164409344508.

Lekholm, A. K., & Cliffordson, C. (2008). Discrepancies between school grades and test scores at individual and school level: Effects of gender and family background. *Educational Research and Evaluation, 14*(2), 181–199. doi: 10.1080/13803610801956663

Lewis, R. (2011). L'avenir de l'assurance-qualité au sein du système mondial d'enseignement supérieur. Dans OCDE (Éd.), *L'enseignement supérieur à*

l'horizon 2030 – Volume 2 : Mondialisation, La recherche et l'innovation dans l'enseignement. Éditions OCDE. doi: 10.1787/9789264075405-fr

Louis, R., Jutras, F., & Hensler, H. (1996). Des objectifs aux compétences : implications pour l'évaluation de la formation initiale des maîtres. *Revue canadienne de l'éducation, 21*(4), 414–432.

Luthy, T. L. (1996). *Validity and prediction bias of grade performance from Graduate Record Examination scores for students at Northern Illinois University: Age and gender considerations.* Northern Illinois University, ProQuest: UMI Dissertation Publishing.

Ma, X., Ma, L., & Bradley, K. D. (2008). Using multilevel modeling to investigate school effects. In A. A. O'Connell, & D. B. McCoach (Éds), *Multilevel modeling of educational data* (pp. 59–110). Charlotte, NC: Information Age Publishing.

Mason, G., & Dragovich, J. (2010). Program assessment and evaluation using student grades obtained on outcome-related course learning objectives. *Journal of Professional Issues in Engineering Education and Practice, 136*(4), 206–214. doi: 10.1061/(ASCE)EI.1943-5541.0000029

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. doi: 10.1037//0003-066X.50.9.741

Millman, J., Slovacek, S. P., Kulick, E., & Mitchell, K. J. (1983). Does grade inflation affect the reliability of grades? *Research in Higher Education, 19*(4), 423–429. doi: 10.1007/BF01418444

Ministère de l'Éducation de l'Ontario. (2000). *Dossier scolaire de l'Ontario : Guide, 2000.* Repéré à http://www.edu.gov.on.ca/fre/document/curricul/osr/osrf.html

Ministère de l'Éducation du Québec (MEQ). (2003). *Politique d'évaluation des apprentissages.* Repéré à http://www.mels.gouv.qc.ca/fileadmin/site_web/documents/publications/EPEPS/Formation_jeunes/Evaluation/13-4602.pdf

Morlaix, S. (2009). *Compétences des élèves et dynamique des apprentissages.* Rennes, France : Presses universitaires de Rennes.

Muis, K. R., & Winne P. H. (2012). Assessing the psychometric properties of the achievement goals questionnaire across task contexts. *Canadian Journal of Education, 35*(2), 232–248.

O'Connell, A. A., & McCoach, D. B. (Éds). (2008). *Multilevel modeling of educational data.* Charlotte, NC: Information Age Publishing.

Pattison, E., Grodsky, E., & Muller, C. (2013). Is the sky falling? Grade inflation and the signaling power of grades. *Educational Researcher, 42*(5), 259–265. doi: 10.3102/0013189X13481382

Pfister, C. (1975). *La validité de la note scolaire* (Thèse de doctorat inédite). Université de Neuchâtel, Berne, Suisse.

Pike, G. R., & Saupe, J. L. (2002). Does high school matter? An analysis of three methods of predicting first-year grades. *Research in Higher Education, 43*(2), 187–207. doi: 10.1023/A:1014419724092

Rexwinkel, T., Haenen, J., & Pilot, A. (2013). Quality assurance in higher education: Analysis of grades for reviewing course levels. *Quality and Quantity, 47*(1), 581–598. doi: 10.1007/s11135-011-9481-6

Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940–2009. *Teachers College Record, 114*(7), 1–23.

Ryan, K. E., & Cousins, J. B. (2009). Introduction. In K. E. Ryan, & J. B. Cousins (Éds), *The SAGE international handbook of educational evaluation* (pp. ix–xvii). Los Angeles, CA: SAGE.

Saupe, J. L., & Eimers, M. T. (2012). *Alternative estimates of the reliability of college grade point averages.* Annual Forum of the Association for Institutional Research, June 2–June 6, 2012. New Orleans: Louisiana.

Smith, D. L. (1992). Validity of faculty judgments of student performance: Relationship between grades and credits earned and external criterion measures. *The Journal of Higher Education, 63*(3), 329–340. doi: 10.2307/1982018

Thorsen, C., & Cliffordson, C. (2012). Teachers' grade assignment and the predictive validity of criterion-referenced grades. *Educational Research and Evaluation: An International Journal on Theory and Practice, 18*(2), 153–172. doi: 10.1080/13803611.2012.659929

Wingard, J. R., & Williamson, J. W. (1973). Grades as predictors of physicians' career performance: An evaluative literature review. *Journal of Medical Education, 48*(4), 311–322.