La nouvelle sanction des études sera-t-elle juste et équitable ?

Christian Rousseau*
Faculté des sciences de l'éducation
Université de Montréal

Pour peu qu'on se soucie de traiter les élèves de façon juste et équitable, il faudra porter une attention particulière aux instruments de mesure qui seront utilisés pour les épreuves synthèse de programme et assurer une certaine uniformité de ces épreuves dans l'ensemble du réseau collégial.

Comme le faisait remarquer Jacques Laliberté (1995a) dans un numéro récent de *Pédagogie collégiale*, « Aux États-Unis, les limites inhérentes à l'utilisation, fort répandue, des tests standardisés ont amené des spécialistes et des éducateurs à chercher d'autres façons d'évaluer les apprentissages. » Ces « autres façons d'évaluer » – nous parlerons ici d'une évaluation basée sur la performance – soulèvent présentement de nombreuses questions en ce qui concerne, notamment, les exigences de qualité auxquelles elles doivent répondre lorsqu'il s'agit de sanctionner les études.

Dans les collèges du Québec, ces questions ne trouvent pas beaucoup d'échos ; la sanction des études, cours par cours, à partir d'évaluations basées sur la performance, est de règle et elle semble aller de soi. Or, l'introduction d'une épreuve synthèse de programme pose le problème de la sanction des études dans une nouvelle perspective et nous oblige à faire nôtres les interrogations de nos voisins du Sud sur le sujet. Ainsi, il importe de préciser le bien-fondé des caractéristiques d'une évaluation basée sur la performance ; il importe aussi de se pencher sur la qualité des instruments de mesure : on doit en effet assurer la validité, la fidélité et une certaine uniformité des épreuves synthèse de programme pour pouvoir sanctionner, de façon juste et équitable, les apprentissages des élèves.

Le contexte des épreuves synthèse

On ne s'est pas beaucoup interrogé jusqu'ici sur la qualité technique des méthodes d'évaluation des apprentissages au collégial. Cela n'a rien d'étonnant. Lorsqu'il s'agit d'une dimension de la charge d'enseignement du professeur, l'évaluation demande à ce dernier de porter un jugement sur les accomplissements de ses élèves à partir des différentes données dont il dispose (non seulement les données recueillies par plusieurs instruments de mesure, mais aussi une connaissance de l'élève et du contexte particulier à l'intérieur duquel se sont déroulés l'apprentissage et l'évaluation). L'évaluation relève alors de la responsabilité professionnelle du professeur. Cela ne veut pas dire que les questions de validité et de fidélité des examens soient pour autant accessoires, mais plutôt qu'elles sont subordonnées à un concept plus fondamental, celui de la compétence du professeur. Les priorités concernent donc globalement le perfectionnement des éducateurs en matière d'évaluation des apprentissages, plutôt que strictement la qualité des instruments de mesure (Howe et Ménard, 1994).

On peut comprendre que les préoccupations techniques prennent une importance accrue depuis qu'on parle d'ÉSP, car la problématique des caractéristiques de l'instrument de mesure se pose tout autrement. En effet, cette épreuve est indépendante de la charge d'enseignement des professeurs et est placée sous la responsabilité de l'établissement. Qu'advient-il du processus *mesure-évaluation-décision* ? Comme dans la plupart des situations où des tests (au sens large du terme) débordent du cadre de la classe et sont administrés à plus grande échelle, la mesure n'est plus soumise au jugement professionnel d'un éducateur et n'est plus encadrée par une connaissance contextuelle approfondie. De fait, la mesure entraîne directement la décision, sans la médiation de ce qu'on nomme l'évaluation, c'est-à-dire l'interprétation des résultats à la lumière de toutes les informations pertinentes. Comme le lien entre la mesure et la décision est beaucoup plus direct, on doit se soucier au plus haut point des qualités métrologiques des instruments utilisés, par souci de justice et d'équité dans les décisions concernant les élèves. Ces décisions, faut-il le rappeler, sont lourdes de conséquences, car il est question de l'attribution ou non du diplôme d'études collégiales suite à la réussite ou à l'échec d'une épreuve unique.

Pour que ces décisions soient valables, elles doivent s'appuyer sur des données

* L'auteur est chargé de cours et étudiant au doctorat.

solides, c'est-à-dire être fondées sur des épreuves qui mesurent bel et bien ce qu'elles prétendent mesurer. La validité de l'utilisation de ces dernières constitue donc un rempart contre l'arbitraire, afin que le droit de chacun à une évaluation équitable soit respecté. La validité n'est donc pas un concept éthéré ayant pour unique fonction de divertir les spécialistes ; il s'agit d'une question bien réelle de justice.

Les valeurs de justice et d'équité liées aux décisions concernant les élèves ne sont toutefois pas les seules que l'on doive considérer à l'intérieur de la problématique de l'ÉSP. Il faut aussi comprendre les valeurs qui sous-tendent l'imposition de l'épreuve elle-même. Aux États-Unis, le vent de changement qui souffle sur les pratiques d'évaluation des apprentissages est alimenté par la volonté de développer des procédures qui amènent l'élève à appliquer sa compréhension des concepts et ses habiletés de résolution de problèmes. On met ainsi l'accent sur l'acquisition de *raisonnements d'ordre supérieur*, et l'on espère induire, par le biais de l'évaluation basée sur la performance, une modification du curriculum, de l'enseignement et des apprentissages (Resnick et Resnick, 1992 ; Wiggins, 1989). Voilà possiblement un des buts poursuivis par la nouvelle exigence d'une ÉSP dans l'ordre collégial au Québec (Forcier, 1994 ; Goulet, 1994). Si l'on vise effectivement un tel but, qui est de l'ordre du développement pédagogique, quelle place faut-il faire au souci d'équité qui devrait présider aux décisions découlant de ces épreuves ? Avant de tenter de répondre à cette question, il est nécessaire de s'entendre sur ce qui est en cause quand on parle d'évaluation basée sur la performance.

L'évaluation basée sur la performance

Quelle que soit la forme d'un test, qu'il s'agisse d'une expérience en laboratoire, de la rédaction d'un texte ou d'une suite d'items à réponse choisie, ce qu'on demande à l'individu demeure toujours de fournir une performance. En ce sens, on peut dire qu'il « n'y a aucune distinction absolue entre les tests basés sur la performance et les autres classes de tests » (Fitzpatrick et Morrison, 1971, p. 238). Qu'est-ce qui caractérise alors cette forme d'évaluation ? Tenter de définir l'évaluation basée sur la performance n'est pas commode, car la diversité terminologique et méthodologique rencontrée dans les textes qui traitent de ce sujet porte à confusion.

Pour bien cerner ce qu'est l'évaluation de la performance, il faut aussi parler d'évaluation authentique et d'évaluation des compétences.

Habituellement, on utilise l'expression *évaluation basée sur la performance* dans un sens général, qui englobe à la fois l'évaluation de l'accomplissement d'une tâche, c'est-à-dire du processus, et de la réalisation qui en résulte, c'est-à-dire du produit. Selon le cas, l'accent est mis sur l'un ou l'autre des deux éléments ou sur les deux à la fois. Outre le fait que cette méthodologie implique le jugement d'un correcteur, ce qui caractérise la performance, ici, serait son degré de réalisme ou de simulation (Fitzpatrick et Morrison, 1971) ; on parle alors d'authenticité. L'évaluation authentique des performances se réalise en plaçant l'élève dans une situation comparable à celles qu'il rencontrera dans « la vraie vie ».

Une mise en garde s'impose toutefois. L'emploi du terme *authentique* véhicule implicitement le message que les autres formes d'évaluation sont nécessairement *non authentiques*, ce qui est abusif. En effet, on doit se demander en regard de quel critère se définit l'authenticité : la réalité de la vie que l'élève aura à affronter plus tard ou la réalité présente des études qu'il poursuit ? Plusieurs apprentissages scolaires, surtout lorsqu'ils nécessitent un degré élevé d'abstraction, n'ont de véritable contexte d'application « dans la vraie vie » que par une longue chaîne d'inférences. L'évaluation de ces apprentissages, pour être authentique, se doit donc d'être décontextualisée (Messick, 1994). Dans le milieu collégial, on pourrait dire qu'il en va ainsi de certains apprentissages liés à la formation générale (en philosophie, par exemple) ou préuniversitaire (en mathématiques, par exemple).

Par *évaluation des compétences*, on réfère à l'objet sur lequel porte l'évaluation, c'est-à-dire aux états internes dont on suppose le développement à la suite d'une séquence d'apprentissage ou d'un programme d'études. Comme une compétence, par définition, n'est pas observable, on doit trouver des comportements observables (des performances) à l'intérieur

L'évaluation authentique des performances se réalise en plaçant l'élève dans une situation comparable à celles qu'il rencontrera dans « la vraie vie »

desquels on postule que la compétence intervient. Pour évaluer des compétences, on procède donc par inférence, à travers une évaluation qui est basée en premier lieu sur la performance.

L'évaluation des compétences est souvent reliée, comme c'est le cas au collégial, à des programmes d'études qui définissent ces dernières en termes d'habiletés complexes et englobantes. Cela constitue une sorte de contrepartie aux programmes par objectifs, dont le niveau de précision implique une décomposition des compétences en habiletés fragmentées, souvent enseignées et évaluées séparément. En posant que le tout est plus grand que la somme des parties, on vise l'intégration des habiletés dans une conception qui tient compte de leurs interactions. On postule ainsi que les compétences ne sont pas décomposables (Resnick et Resnick, 1992). C'est donc dire que, dans cette perspective, l'évaluation basée sur la performance repose sur des tâches complexes et souvent longues visant la synthèse des habiletés développées par l'élève : des rédactions, des exposés oraux, des expositions, des dossiers d'apprentissage (ou portfolios), des constructions de modèles… La liste des possibilités pourrait s'étirer longuement.

L'analyse des qualités métrologiques des méthodes utilisées pour évaluer les performances est encore, comme certains auteurs le soulignent, relativement peu avancée (Frechtling, 1991 ; Miller et Legg, 1993 ; Shavelson, Baxter et Gao, 1993). Néanmoins, lorsqu'on songe à utiliser une épreuve dans un contexte où les enjeux sont importants, on doit, par souci d'équité dans les décisions concernant les élèves, se pencher sur les considérations essentielles que sont la validité, la fidélité et la standardisation.

La validité

L'utilisation qu'on fait de l'évaluation basée sur la performance a généralement l'air valide aux yeux des éducateurs, car elle repose habituellement sur des tâches qui paraissent significatives. Selon Anastasi (1988), cette validité apparente (*face validity*)

> n'est pas la validité au sens technique; cela réfère, non pas à ce que le test mesure vraiment, mais à ce qu'il paraît superficiellement mesurer. Cela est lié au fait que le test semble valide ou non au yeux des sujets qui y répondent, du personnel administratif qui décide de son utilisation et des autres observateurs qui ne possèdent pas la formation technique pertinente. Fondamentalement, la question de la validité apparente concerne la sympathie et les relations publiques. *(p. 144)*

Il ne faut pas cependant sous-estimer cet aspect d'une procédure de mesure, car sa recevabilité par le milieu éducatif détermine souvent sa viabilité.

Néanmoins, lorsqu'on étudie la validité des évaluations basées sur la performance, on doit se poser la question classique : mesure-t-on vraiment ce qu'on prétend mesurer ? Pour y répondre, différentes avenues sont possibles, qu'on a traditionnellement regroupées sous trois grandes catégories d'analyses : certaines études peuvent s'intéresser à la représentativité du contenu d'une épreuve ; d'autres, au lien qui existe entre des critères externes et les résultats obtenus ; d'autres enfin, à la correspondance existant entre les construits cognitifs visés (les compétences telles que définies par un programme, par exemple) et ceux qui sont effectivement mesurés par l'épreuve (APA, AERA, NCME, 1985).

La représentativité du contenu

Pour l'évaluation basée sur la performance, on peut donc se demander en premier lieu si les scores obtenus à partir d'une tâche spécifique permettent d'inférer des conclusions portant sur un domaine plus général, c'est-à-dire si le contenu de cette tâche est représentatif du domaine visé. Deux sujets différents d'écriture, pour une épreuve de langue et littérature par exemple, permettent-ils de mesurer de la même manière la ou les compétences visées ? Ici, on peut demander à des experts du contenu de porter un jugement sur ce que semblent mesurer les tâches proposées. On demeure cependant au niveau des apparences. C'est donc surtout la corrélation entre des tâches différentes censées mesurer le même domaine (la même compétence) qui permet de porter un jugement sur ces dernières quant à leur représentativité réelle. Les auteurs ayant effectué une recension des recherches sur l'analyse de ce type de corrélation rapportent qu'on y retrouve des coefficients généralement faibles (Mehrens, 1992 ; Linn, 1993). Et plus les domaines sont largement définis, plus les tâches sont hétérogènes, plus les coefficients chutent. C'est dire qu'il existe une forte probabilité, pour une même compétence visée, que certains élèves réussissent bien la tâche X, mais échouent à la tâche Y, et inversement.

Concrètement, ces résultats de recherche laissent croire que, pour une épreuve portant même sur une seule matière, telle l'épreuve ministérielle en langue et littérature par exemple, les résultats d'un élève risquent d'être fort différents selon qu'on lui demande de rédiger une dissertation critique sur Rabelais ou sur Boris Vian. Comme les ÉSP annoncées porteraient sur les multiples disciplines qui contribuent à la formation dans un programme, le domaine de mesure risque de s'avérer encore plus vaste que dans l'exemple qu'on vient de mentionner. On devra donc porter un soin particulier à analyser la représentativité de l'épreuve en regard du domaine visé (les compétences propres à ce programme), car les données dont on dispose indiquent à tout le moins que rien ne devrait être pris pour acquis à ce chapitre.

Le lien avec un critère

En second lieu, on peut tenter d'établir la validité d'une évaluation en établissant un rapport avec un critère valable. Dans le cas des ÉSP, on pourrait bien sûr tenter de voir si l'information fournie par le test correspond à celle qui provient des dossiers scolaires des élèves (la moyenne des notes aux cours de formation spécifique, par exemple). Toutefois, si l'on croit utile de développer et d'administrer de telles épreuves, c'est qu'on présume que l'information fournie par une moyenne de notes à des cours doit être complétée par un autre type d'information, qui témoigne de l'intégration de l'ensemble des habiletés. Logiquement, l'une peut donc difficilement servir de critère de validité pour l'autre.

On a suggéré des indicateurs du succès ultérieur des élèves afin d'établir la validité prédictive : pour les programmes préuniversitaires, la réussite à l'université, par exemple, ou pour les programmes techniques, l'intégration au marché du travail (Howe, 1995). Ces approches sont sans doute les plus valables, mais à notre connaissance, aucun résultat de recherche de ce genre n'a encore été publié. On réalise de plus que les études longitudinales nécessaires à ce genre d'analyse posent d'importants problèmes de design, de temps et de coûts.

L'analyse des construits cognitifs

À l'intérieur de la troisième catégorie d'études de validation, on se penche sur l'analyse des construits cognitifs impliqués par la mesure. Autrement dit, on pose ainsi la question de départ : « Est-ce qu'on ne mesure pas aussi autre chose que ce qu'on prétend mesurer ? » On tente alors de voir si l'information fournie par une évaluation basée sur la performance dépend de variables non pertinentes. Les auteurs ayant effectué la recension de telles études montrent qu'on y rapporte l'influence importante de facteurs comme le sexe, l'ethnie, les pratiques éducatives ou les connaissances antérieures (Dunbar, Koretz et Hoover, 1991 ; Mehrens, 1992 ; Miller et Legg, 1993). Dans la validation des ÉSP, l'examen de ces biais potentiels mérite donc d'être effectué.

On doit aussi inclure dans cette troisième catégorie les analyses visant à examiner si le construit visé (une compétence) est entièrement reflété par le résultat d'une épreuve. À cet égard, Messick (1994) rappelle que ce qui est crucial au plan de la validité, ce n'est pas de chercher à démontrer qu'une compétence est réellement sollicitée par une tâche, mais bien de s'assurer que les critères utilisés pour apprécier la performance témoignent du fait que l'ensemble des éléments de cette compétence sont à l'œuvre. Selon cet auteur, cela risque de s'avérer problématique, parce que les compétences visées par ce type d'épreuve sont par définition complexes et englobantes ; il est donc difficile d'en préciser les caractéristiques.

Pour que les concepteurs d'une ÉSP soient en mesure de formuler des critères qui reflètent bel et bien la compétence visée, ils devront donc disposer préalablement d'une définition explicite des composantes de cette dernière. Si l'on tente de définir des critères uniquement en fonction d'une tâche particulière de mesure, il n'est pas du tout évident que l'ensemble de la compétence (rappelons qu'on parle ici d'habiletés complexes d'intégration et de transfert des apprentissages) soit captée par l'ÉSP. Or, il n'et pas certain que le Ministère sera en mesure de présenter de telles définitions avant l'échéance de 1998 (Goulet, 1994). Les concepteurs d'épreuves devront donc les produire localement…

Les conséquences liées à l'utilisation

Devant la difficulté d'établir la validité des instruments d'évaluation basée sur la performance, Moss (1992) rapporte que certains auteurs ont tenté de définir d'autres critères selon lesquels on devrait analyser les qualités techniques. L'argumentation en ce sens repose en grande partie sur l'importance, soulignée par Messick (1989), de considérer les conséquences de l'évaluation à l'intérieur du concept de validité. On a donc proposé des critères de validation qui maximiseraient les impacts positifs présumés liés à l'utilisation d'évaluations basées sur la performance (Frederiksen et Collins, 1989 ; Linn, Baker et Dunbar, 1991 ; Wiggins, 1993). Selon cette conception, on suppose (cela demeure en effet à être démontré) que cette forme d'évaluation induirait une amélioration du curriculum, de l'enseignement et des apprentissages, comme on l'a mentionné précédemment.

Si l'on souhaite, en misant sur l'intégration et le transfert des apprentissages à l'intérieur de l'ÉSP, provoquer des améliorations éducatives et pédagogiques, il faut alors que le cours collégial favorise le développement de ces habiletés dites supérieures. Certains, comme Jacques Laliberté (1995b) ou Jean-Pierre Goulet (dans le présent numéro de *Pédagogie collégiale*), expriment cependant leur inquiétude en constatant qu'en 1998, l'ÉSP, dans certains programmes du moins, risque d'être la seule occasion où l'élève aurait à manifester de telles habiletés, celles-ci n'ayant pas fait l'objet de formation antérieure. Si tel est le cas, on est en droit de se demander où sont les impacts positifs pour les élèves qui seront soumis à cette épreuve. De plus, si l'on veut vraiment tenir compte des conséquences dans une perspective de validation, on doit se souvenir, comme le dit Paul Forcier (1994), « que de telles épreuves peuvent porter en elles, de par leur présence même dans un curriculum, des effets pervers sur lesquels il faudrait au moins prendre le temps de réfléchir » (p. 22).

Quoi qu'il en soit, il semble que le fait de tenir compte de ces nouveaux critères ne diminue pas l'importance des éléments traditionnels de validation. On doit éviter de poser la question en termes de compromis byzantins entre les qualités métrologiques et les qualités pédagogiques ou entre l'équité des décisions qui dépendent d'une épreuve et la valeur éducative de cette dernière. En effet, selon Messick (1994), ces aspects sont indissociables et l'on ne saurait les subordonner les uns aux autres dans l'analyse de la validité.

« De telles épreuves peuvent porter en elles, de par leur présence même dans un curriculum, des effets pervers sur lesquels il faudrait au moins prendre le temps de réfléchir. »

La fidélité

Legendre (1993), dans son *Dictionnaire actuel de l'éducation*, définit le concept de fidélité comme la « qualité qu'a un instrument de mesurer avec la même exactitude chaque fois qu'il est administré » (p. 609). Il s'agit donc d'un concept distinct de celui de validité, mais nécessaire à ce dernier : une épreuve qui ne fournit pas des résultats fidèles ne peut être utilisée de façon valide.

Les recherches sur la fidélité de l'évaluation basée sur la performance abordent généralement la question de la validité de la mesure selon les deux sources de variabilité suivantes : la variabilité due au jugement dans la correction et la variabilité due à l'échantillonnage des tâches.

L'accord interjuges constitue un problème majeur, mais on sait désormais qu'il est possible d'obtenir des indices de fidélité relativement élevés à cet égard, en soignant les conditions de formation et d'encadrement des correcteurs et les échelles employées (Linn et Burton, 1994). On doit souligner cependant que le respect de ces conditions suppose une préparation minutieuse et entraîne des coûts parfois importants. Quant à la solution consistant à augmenter le nombre de correcteurs pour chaque performance évaluée, son effet dépend, selon Dunbar *et al.* (1991), du degré d'accord interjuges initial : s'il est élevé, cette stratégie donne peu de résultats ; s'il est faible par contre, des gains significatifs peuvent être réalisés.

Pour ce qui est de la variabilité due à l'échantillonnage des tâches, Linn (1993), dans sa revue des recherches portant sur le sujet, rapporte que la fidélité de l'évaluation basée sur la performance est généralement faible si l'épreuve est conçue à partir d'une seule tâche, complexe et englobante, visant la synthèse des habiletés acquises par l'élève (la rédaction d'un texte long, par exemple). Autrement dit, comme les performances sont étroitement liées à la tâche particulière soumise à l'élève, une seule de ces dernières ne permet pas d'obtenir un résultat auquel on peut vraiment faire confiance.

Il est à noter que l'augmentation du nombre de tâches semble avoir un effet positif considérable sur la situation qu'on vient de décrire. Ainsi, si l'on fait porter la mesure sur plusieurs tâches, l'information recueillie sera plus stable et on pourra lui faire davantage confiance que si l'épreuve comporte une seule tâche. Toutefois, dans le contexte de l'évaluation basée sur la performance, les tâches sont souvent longues, ce qui rend cette avenue difficilement praticable. On peut illustrer cet état de fait en reprenant l'exemple de langue et littérature. Dans l'épreuve qui sera expérimentée à l'hiver 1996, l'élève disposera de quatre heures et demie pour rédiger une dissertation critique d'au moins 900 mots. Il semble peu réaliste d'envisager que cette épreuve comporte plusieurs tâches de ce genre.

La fidélité de l'évaluation basée sur la performance pose donc certains problèmes importants qui risquent de compromettre la validité de cette dernière. Les responsables du développement d'ÉSP se doivent donc d'en tenir compte, à la fois dans l'élaboration de l'épreuve, en la faisant porter sur plusieurs tâches, dans la correction, en assurant une formation et un encadrement rigoureux aux correcteurs, et dans l'utilisation des résultats, en sachant que si la fidélité d'une épreuve est faible, l'erreur de mesure, elle, sera élevée (il sera alors difficile de fixer un seuil de réussite, par exemple).

La standardisation

L'analyse des mémoires présentés à la Commission parlementaire sur l'enseignement collégial porte à croire qu'en imposant des ÉSP, on poursuivrait le but de rendre comparable l'enseignement dispensé d'un collège à l'autre (Forcier, 1994). Si tel est le cas, quelqu'un doit les standardiser, c'est-à-dire mettre en place des conditions uniformes à l'intérieur desquelles le rendement des individus sera jugé : des épreuves équivalentes, administrées et corrigées selon des modalités similaires. Cette nécessité entre cependant en conflit avec la notion d'*authenticité*. La diversité des contextes rencontrés « dans la vraie vie » est peu compatible avec la décontextualisation des tâches exigée par le contrôle exercé sur elles lors de la standardisation (Resnick et Resnick, 1992). Face à ce dilemme, certains auteurs croient qu'on doive accepter la diversité des standards et remettre la responsabilité de l'évaluation entre les mains des éducateurs (par exemple Wiggins, 1991). Dans l'enseignement collégial, cela reviendrait à maintenir la situation qui prévaut actuellement.

Une autre question liée à la standardisation risque d'être soulevée plus directement par les ÉSP. En effet, chaque collège est responsable de leur imposition. Des collèges offrant un même programme peuvent donc décider d'administrer des épreuves très différentes. La comparabilité d'épreuves distinctes (en termes de performances mesurées, de degré de difficulté ou de capacité discriminante, par exemple) peut être très difficile à établir, et cela pourrait donner lieu à des contestations de la part d'élèves se jugeant lésés. Certains à qui l'on refusera de décerner un diplôme parce qu'ils auront échoué à une ÉSP

Mettre en place des conditions uniformes à l'intérieur desquelles le rendement des individus sera jugé

pourront, en toute légitimité, s'interroger sur la valeur de cette décision s'il s'avère qu'il existe ailleurs dans le réseau des épreuves qui semblent plus faciles et qui mènent à l'obtention du même diplôme.

Si la standardisation des ÉSP n'est assurée par aucun organisme intercollégial, quelle qu'en soit la nature ou la composition, on s'expose donc non seulement à une absence d'équivalence rendant impossible toute comparaison entre établissements, mais aussi à des aléas juridiques ; soulignons à ce propos que si, pour l'instant, les tribunaux sont très réticents à se mêler d'évaluation pédagogique (Proulx, 1994), il n'est pas dit que ce domaine échappera indéfiniment au processus de judiciarisation. On n'a qu'à regarder ce qui se passe ailleurs pour s'en convaincre (voir Mehrens et Popham, 1992).

Conclusion

L'ÉSP doit répondre à des exigences de qualité élevées afin que la sanction des acquis des élèves se fasse en toute équité. Or, il est loin d'être sûr que l'évaluation basée sur la performance, quoiqu'on la dise plus authentique, puisse répondre à ces exigences. Cela devrait inciter le milieu collégial à la plus grande prudence. Baker (1993) précise que « les évaluations basées sur la performance sont de bonnes stratégies pour améliorer la compréhension qu'ont les enseignants des élèves dans leurs classes », mais « qu'on n'en sait pas assez pour employer l'évaluation basée sur la performance comme procédure unique afin de certifier les accomplissements individuels d'un élève » (p. 14). Ce point de vue est partagé par plusieurs auteurs (voir entre autres Dunbar, Koretz et Hoover, 1991 ; Frechtling, 1991 ; Mehrens, 1992). En ce sens, certains voient principalement l'évaluation basée sur la performance comme un outil d'évaluation formative (Moss, Beck, Ebbs, Matson, Muchmore, Steele et Tayler, 1992).

Si l'échéance de 1998 est maintenue telle quelle, il faudra plus que de bonnes intuitions pour développer des épreuves justes et équitables. On devra en démontrer la validité, en établir la fidélité et assurer une certaine uniformité à ce processus à travers le réseau collégial. Il semble pourtant ne pas être encore beaucoup question de ces aspects dans les discussions actuelles. C'est un peu comme si l'on tentait de concevoir un moteur sans parler de mécanique.

Références

ANASTASI, A. (1988), *Psychological Testing* (6e ed.), New York, Macmillan.

AMERICAN PSYCHOLOGICAL ASSOCIATION, AMERICAN EDUCATIONAL RESEARCH ASSOCIATION ET NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (1985), S*tandards for Educational and Psychological Testing*, Washington, American Psychological Association.

BAKER, E. L. (1993), « Questionning the Technical Quality of Performance Assessment » dans *The School Administrator*, vol. 50, n° 11, p. 12-16.

DUNBAR, S. B., D. M. KORETZ et H.D. HOOVER (1991), « Quality Control in the Development and Use of Performance Assessment » dans *Applied Measurement in Education*, vol. 4, n° 4, p. 289-303.

FITZPATRICK, R. et E. J. MORRISON (1971), « Performance and Product Evaluation » dans R. L. Thorndike (Ed.*), Educational Measurement* (2e ed.), Washington (DC), American Council on Education.

FORCIER, P. (1994), « À la recherche de la baguette magique » dans *Pédagogie collégiale*, vol. 8, n° 2, p. 18-23.

FRECHTLING, J. A. (1991), « Performance Assessment: Moonstruck or the Real Thing » dans *Educational Measurement: Issues and Practices,* vol. 11, n° 1, p. 23-25.

FREDERIKSEN, J. R. et A. COLLINS (1989), « A Systems Approach to Educational Testing » dans *Educational Researcher*, vol. 18, n° 9, p. 27-32.

GOULET, J.-P. (1994), « L'épreuve synthèse de programme : pour éviter l'épreuve » dans *Pédagogie collégiale*, vol. 7, n° 4, p. 19-22.

HOWE, R. (1995), *Guide méthodologique pour la mise en œuvre de l'épreuve synthèse de programme*, Collège Montmorency, mai 1995.

HOWE, R. et L. MÉNARD (1994), « Croyances et pratiques en évaluation des apprentissages » dans *Pédagogie collégiale*, vol. 7, n° 3, p. 21-27.

LALIBERTÉ, J. (1995a) « D'autres façons de concevoir et de faire l'évaluation des apprentissages ? dans *Pédagogie collégiale*, vol. 8, n° 3, p. 9-13.

LALIBERTÉ, J. (1995b) « L'épreuve synthèse de programme : gage ou facteur de l'intégration des apprentissages ? dans *Pédagogie collégiale*, vol. 8, n° 3, p. 18-23.

LEGENDRE, R. (1993), *Dictionnaire actuel de l'éducation*, (2e éd.), Montréal, Guérin.

LINN, R. L. (1993), « Educational Assessment: Expanded Expectations and Challenges » dans *Educational Evaluation and Policy Analysis,* vol. 15, p. 1-16.

LINN, R. L. et E. BURTON (1994), « Performance-Based Assessment: Implications of Task Specificity » dans *Educational Measurement: Issues and Practices*, vol. 13, n° 1, p. 5-8, p. 15.

LINN, R. L., E. L. BAKER et S. B. DUNBAR (1991), « Complex, Performance-Based Assessment: Expectations and Validation Criteria » dans *Educational Researcher*, vol. 20, n° 8, p. 15-21.

MEHRENS, W. A. (1992), « Using Performance Assessment for Accountability Purposes » dans *Educational Measurement: Issues and Practices*, vol. 11, n° 1, p. 3-9, p. 20.

MEHRENS, W. A. et W. J. POPHAM (1992), « How to Evaluate the Legal Defensibility of High-Stakes Tests » dans *Applied Measurement in Education*, vol. 5, n° 3, p. 265-283.

MESSICK, S. (1994), « The Interplay of Evidence and Consequences in the Validation of Performance Assessments » dans *Educational Researcher*, vol. 23, n° 2, p. 13-23.

MESSICK, S. (1989), « Validity » dans R. L. Linn (Ed.), *Educational Measurement* (3e ed.), New York, Macmillan.

MILLER, M. D. et S. M. LEGG (1993), « Alternative Assessment in a High-Stakes Environment » dans *Educational Measurement: Issues and Practices*, vol. 12, n° 2, p. 9-15.

MOSS, P. A. (1992), « Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment » dans *Review of Educational Research*, vol. 62, n° 3, p. 229-258.

MOSS, P. A., J. S. BECK, C. EBBS, B. MATSON, J. MUCHMORE, D. STEELE et C. TAYLER (1992), « Portfolios, Accountability, and an Interpretive Approach to Validity » dans *Educational Measurement: Issues and Practices*, vol. 11, n° 3, p. 12-21.

PROULX, J.-P. (1994), « Perspectives juridiques et judiciaires de l'évaluation pédagogique » dans *Mesure et évaluation en éducation*, vol. 17, n° 1, p. 5-43.

RESNICK, L.B. et D. P. RESNICK (1992), « Assessing the Thinking Curriculum: New Tools for Educational Reform, dans B.R. Gifford et M.C. O'Connor (Eds.), *Changing Assessments: Alternative Views of Aptitude, Achievement, and Instruction*, Boston, Kluwer.

SHAVELSON, R. J., G. P. BAXTER et X. GAO (1993), « Sampling Variability of Performance Assessment » dans *Journal of Educational Measurement*, vol. 30, n° 3, p. 215-232.

WIGGINS, G. (1993), « Assessment: Authenticity, Context, and Validity » dans *Phi Delta Kappan*, vol. 74, n° 3, p. 200-214.

WIGGINS, G. (1991), « Standards, Not Standardization: Evoking Quality Student Work » dans *Educational Leadership*, vol. 48, n° 5, p. 18-25.

WIGGINS, G.(1989), « Teaching to the (Authentic) Test » dans *Educational Leadership*, vol. 46, n° 7, p. 41-47.

L'auteur remercie le Conseil de recherches en sciences humaines du Canada (CRSH) pour son aide financière dans la période pendant laquelle cet article a été rédigé (bourse de doctorat nº 752-93-1174).