ÉCHELLE DESCRIPTIVE D'APPRÉCIATION DU RENDEMENT DE L'ÉLÈVE : ACQUIS ET PROBLÉMATIQUES

RENÉE FORGETTE-GIROUX & MARIELLE SIMON
Faculté d'éducation, Université d'Ottawa

RÉSUMÉ. Sous l'influence du constructivisme et du cognitivisme, l'évaluation en salle de classe se dégage progressivement de la notation quantitative et tente de mettre à profit des approches plutôt qualitatives. Ce virage a suscité récemment de nombreux écrits sur l'évaluation de la performance à partir d'une échelle descriptive d'appréciation. Le présent article fait état des acquis en matière de notation qualitative et analyse les principaux aspects problématiques.

DESCRIPTIVE RATING SCALE OF STUDENT PERFORMANCE: STATE OF THE ART

ABSTRACT. Under the influence of constructivism and cognitivism, classroom assessment is gradually moving from quantitative to qualitative grading methods. This change of direction has given rise to many papers on performance assessment using a descriptive rating scale such as the rubric. This article reports on the current trends in qualitative grading and analyzes the main problematic aspects.

Introduction

Sous l'influence du constructivisme et du cognitivisme, les nouveaux programmes scolaires proposent des modifications importantes dans la façon de concevoir l'apprentissage et dans ce qui doit être enseigné dans les salles de classe. À l'intérieur de ces théories, l'approche par compétences porte sur des apprentissages complexes, multidimensionnels, intégrés et transférables et constitue le principal cadre de référence de plusieurs programmes d'études. Cela entraîne nécessairement des changements de perspective majeurs en matière d'évaluation des apprentissages, plus précisément en ce qui concerne la nature des objets évalués, le rapport entre l'évaluation et l'apprentissage, les dispositifs d'évaluation et la façon d'interpréter les résultats obtenus. L'évaluation se dégage alors progressivement de la notation quantitative et tente de mettre à profit des approches plutôt qualitatives. Puisqu'elle provient initialement d'un besoin ressenti en salle de classe, la notation qualitative se conforme davantage que les mesures psychométriques au paradigme constructiviste de l'apprentissage. Ce virage a suscité récemment de nombreux écrits sur la mise

en œuvre, tant en salle de classe que dans les programmes d'envergure, de l'évaluation de la performance et de la notation qualitative à partir d'une échelle descriptive d'appréciation (Arter & McTighe, 2001; Brahier, 2001; Brualdi, 1998; Custer, 1998; Marzano, 2000; Moskal, 2000).

La réforme en éducation, désormais basée sur les normes de rendement, réoriente graduellement le curriculum, les pratiques d'enseignement et les modalités d'évaluation de plusieurs pays occidentaux, d'états américains et de provinces canadiennes (Stroble, 1993; Firestone, 2000). Dans cette foulée, la pratique de la notation qualitative à partir d'échelles descriptives d'appréciation prend également son envol dans les classes et les systèmes scolaires (Arter & McTighe, 2001), malgré le manque de modèles théoriques et cohérents guidant une mise en application uniformisée et fondée.

Même s'il existe une panoplie de ressources mises à la disposition des praticiens en matière d'évaluation du rendement à partir de tâches complexes de performance, l'application des échelles de notation qualitative continuent de susciter des problèmes en salle de classe (Brahier, 2001; Conseil de l'Europe, 2001; Depka, 2000; Harper, O'Connor & Simpson, 1999; Marzano, 2000; MEO, 2000; O'Connor, 1999; Phillips, 1998; Popham, 1999; Rogers & Graham, 2000; Simon & Forgette-Giroux, 2003; Stiggins, 2001; Wiggins, 1998). Selon les écrits, les praticiens et les dirigeants des programmes d'évaluation d'envergure éprouvent notamment certaines difficultés à :

- utiliser l'échelle descriptive de notation principalement aux fins d'évaluation des compétences,
- privilégier l'échelle générique,
- rendre explicites, à l'intérieur de l'échelle, les dimensions et les critères pertinents à la compétence cible,
- rapporter les résultats de la notation qualitative au bulletin,
- établir un équilibre entre la validité et la fidélité des résultats de notation.

Le présent article a pour objectif de faire état des acquis en matière de notation qualitative en salle de classe et de traiter des principaux aspects problématiques. Il met en évidence l'application souvent incohérente et inappropriée de l'échelle de notation qualitative. La première section traite de l'aspect transversal de l'échelle, c'est-à-dire de son utilisation aux fins d'évaluation de compétences appliquées par le biais de n'importe quelle discipline. La deuxième fait ressortir l'importance de présenter une échelle complète et claire, articulée à partir des critères de performance. La troisième décrit la tendance à vouloir quantifier les résultats qualitatifs obtenus en appliquant l'échelle. Enfin, la quatrième section discute des qualités métrologiques de l'échelle descriptive d'appréciation. En conclusion, les auteurs questionnent la viabilité de l'échelle comme véritable outil de communication de résultats qualitatifs de l'évaluation du rendement des élèves.

Échelle générique

Selon le paradigme constructiviste, l'évaluation en salle de classe s'intègre davantage à l'activité pédagogique et contribue ainsi au processus actif, constructif et métacognitif de l'apprentissage (Tardif, 1993). L'évaluation en vue d'un apprentissage remplace peu à peu l'évaluation de l'apprentissage (Biggs, 1995; Stiggins, 2001; Popham, 1999). En effet, l'évaluation encourage la maîtrise de compétences manifestées à partir de situations réelles qui font appel à la construction du savoir. Elle adopte la fonction formative, offre une rétroaction signifiante à l'élève par rapport à l'état de ses apprentissages et l'informe de ses résultats. Le tout se fait à partir d'une approche diversifiée et complémentaire combinant l'observation du processus et du produit et le questionnement traditionnel (Azwell & Shamr, 1995; Biggs, 1995; Tardif, 1993). De ce fait même, la notation des résultats d'évaluation sort éventuellement du carcan de la production de notes chiffrées complètement décontextualisées et prend plutôt une orientation qualitative et descriptive du rendement. Plusieurs écrits font ressortir bon nombre de caractéristiques requises pour l'élaboration d'une échelle descriptive efficace comme outil de notation dans un contexte d'évaluation de la performance (Arter, 1999; Arter & McTighe, 2001; Biggs, 1992; Brookhart, 1999; Frisbie & Waltman, 1992; Goodrich-Andrade, 2000; ministère de l'Éducation de l'Ontario, 2000; Popham, 1997; Stiggins, 2001; Wiggins, 1998) :

- prédominance des aspects qualitatifs et descriptifs,
- simplicité et transparence de l'échelle,
- pertinence et utilité des critères de rendement,
- accessibilité à des copies types de la performance attendue,
- adéquation de l'échelle à la compétence visée,
- ouverture à la participation des évalués à l'évaluation,
- sensibilité au progrès, à la rétroaction et au réinvestissement.

Outre ces caractéristiques, les auteurs s'entendent généralement pour dire que la notation qualitative à partir d'échelles descriptives d'appréciation s'applique tout particulièrement à l'évaluation des compétences disciplinaires et transversales telles que le processus d'écriture, la résolution de problèmes, la création artistique, l'investigation scientifique, la recherche expérimentale et le processus d'évaluation. Selon eux, elle s'utilise également pour noter des habiletés plus ou moins complexes à l'intérieur d'une discipline, dont la compréhension de concepts, le rapprochement et la mise en application en contextes variés.

Effectivement, plusieurs écrits situent l'utilisation de la notation qualitative dans une discipline quelconque. En langue, l'échelle descriptive d'appréciation s'applique par exemple à l'évaluation de l'écriture (Arter, 1993; Arter & McTighe, 2001; DeRemer, 1998; Mabry, 1999; Potter, 1995; Saunders, 1999; Stuhlman, Daniel, Dellinger, Denny & Powers, 1999; Swartz et al., 1999).

Le *Six trait analytical model* d'Arter sert davantage de modèle de notation de l'écriture. L'échelle, à double entrée, offre des niveaux de réussite, avec énoncés décrivant six dimensions de l'écriture : les idées et le développement du contenu, l'organisation, le style, le vocabulaire, la phrase et les conventions. Par ailleurs, en histoire, Baker et ses collaborateurs; (1992; 1994) ont développé une échelle à cinq niveaux basés sur les fondements théoriques relatifs au développement de compétences cognitives observées chez les historiens experts. Dans ce contexte, l'échelle sert principalement d'outil de notation des productions écrites des élèves. Elle comprend une dimension globale portant sur l'impression générale de la qualité du contenu et cinq échelles analytiques, chacune traitant d'une habileté secondaire : les connaissances préalables, le nombre et la qualité des principes et des concepts fournis, l'argumentation, la référence au texte et les idées préconçues. Quelques écrits présentent également des échelles descriptives d'appréciation en études technologiques (Custer, 1998), en arts visuels (Davis, 1993; Defibaugh, 2000) et en communication orale (Chaloub-Deville, 1995). En mathématiques, les échelles les plus populaires font appel à la compréhension d'un problème, aux connaissances procédurales, à la résolution de problèmes et à la communication (Arter, 1993; Brahier, 2001; Depka, 2000; Krulic & Rudnick, 1999; Suzuki, 1998; Taylor, 1998; Taylor & Bidlingmayer, 1998).

Les études en mathématiques et en sciences cependant présentent des opinions partagées en ce qui concerne la notation globale ou analytique, générale ou spécifique. Ceci s'explique en partie par la nature de ces disciplines où l'accent est mis traditionnellement sur les contenus, sur des habiletés de base ou sur des compétences disciplinaires plutôt que sur des compétences génériques. Plusieurs chercheurs continuent d'explorer diverses approches de notation à partir d'échelles descriptives d'appréciation dans ces matières (Busching, 1998; Francis & Hord, 1995; Solano-Flores & Shavelson, 1997; Solano-Flores, Shavelson, Ruiz-Primo, Schults, Wiley & Brown, 1997). La méthode de notation présentée par Solano-Flores et Shavelson, par exemple, met en évidence certaines habiletés typiques des scientifiques chevronnés, notamment la comparaison, l'identification à partir de preuves, l'observation et la classification. Mais l'application de l'échelle demeure encore fonction de la discipline. Les contraintes imposées par l'organisation des milieux scolaires actuels font en sorte que l'échelle descriptive d'appréciation s'utilise principalement à l'intérieur d'une discipline en vue de noter les habiletés de base et non de façon transversale pour évaluer des compétences génériques. Les recherches ultérieures devront se pencher davantage sur les raisons de la faible utilisation de l'échelle aux fins d'évaluation transdisciplinaire ainsi que sur les moyens d'y arriver.

Échelle explicite

Les études font ressortir des composantes essentielles de l'échelle descriptive d'appréciation à double entrée conçue dans le but d'évaluer des compétences

transversales : énoncé de la compétence cible, niveaux de rendement, critères de performance, descripteurs, norme de réussite et copies types (Arter, 1993; Arter & McTighe, 2001; Custer, 1998; Francis & Hord, 1995; Kahl, 1995; Moskal, 2000; Popham, 1999; Simon & Forgette-Giroux, 2001; Wiggins, 1998). Simon & Forgette-Giroux (2003) décrivent en détails les composantes principales dans un ouvrage sur l'utilisation de l'échelle dans le dossier d'apprentissage.

Le présent article met davantage l'accent sur la distinction entre les termes critères et attributs. En principe, l'échelle se réfère à plusieurs critères et attributs. Dans plusieurs écrits la notion de critères de performance porte parfois à confusion faute d'une définition claire et acceptée universellement par la communauté scientifique. Un critère se veut une « caractéristique de la dimension d'une performance, d'un produit ou d'une réponse élaborée. » (Simon & Forgette-Giroux, 2001, p. 100). À titre d'exemples de critères, il suffit de penser au respect de l'auditoire lors d'une présentation orale en tant que compétence ciblée ou au choix de stratégies lorsque la compétence porte sur la résolution de problèmes en mathématiques. Un attribut se veut une qualité propre ou particulière du critère de performance, comme la pertinence (attribut) du choix de stratégie (critère). La rigueur, l'utilité, l'originalité, l'exactitude, l'exhaustivité, la clarté en sont d'autres exemples. Certains attributs, tels que la pertinence, la clarté et la profondeur, s'appliquent à plusieurs performances peu importe la compétence retenue. En général, toutefois, la sélection des critères et des attributs se fait en fonction de la nature de l'objet d'évaluation, des contenus et des exigences spécifiques à la tâche d'évaluation. Dans l'échelle, les attributs progressent d'un échelon à l'autre en prenant des valeurs croissantes ou décroissantes. Stiggins (2001) ajoute que les critères et les attributs doivent être pertinents, clairs, exhaustifs, précis et conviviaux.

Les écrits dans le domaine suggèrent l'articulation explicite des critères et des attributs à l'intérieur de l'échelle (Simon & Forgette-Giroux, 2003; Tierney & Simon, 2004). Par exemple, l'énoncé du critère « démontre une compréhension approfondie d'une variété de concepts », comprend la profondeur et l'envergure comme attributs. Dès lors, le jugement porte sur le niveau d'atteinte de ces deux qualités du critère de compréhension. Une telle articulation permet de valider l'objet d'évaluation et de préciser les dimensions qui l'entourent. L'accompagnement d'une définition écrite, explicite et claire de chacun des critères et des attributs assure une interprétation davantage commune et constante de la part de toutes les personnes concernées. À ces explications peuvent s'ajouter des copies types de performance d'élèves représentatives des divers niveaux de l'échelle, ou du moins illustratives des niveaux supérieurs de performance. Les définitions se présentent sous forme de consignes rattachées à la tâche d'évaluation plutôt que par des explications détaillées fournies dans l'échelle même, afin de conserver son aspect générique et pratique (Tierney & Simon, 2004).

Dans une certaine mesure, le choix des attributs dicte la nature des niveaux de progression de l'échelle descriptive d'appréciation. (Rohrmann, 2002; Tierney & Simon, 2004). En effet, les attributs varient selon leur fréquence, leur quantité ou leur intensité. Certains termes viennent rapidement à l'esprit pour décrire la progression sur l'échelle. Par exemple, les termes « jamais », « rarement », « parfois », « quelquefois », « souvent », « presque toujours » et « toujours » s'associent automatiquement à l'attribut fréquence. Pour sa part, l'intensité prend des valeurs telles que « peu », « faible », « limité », « plus ou moins », « très », « forte » ou « beaucoup » dépendant du critère et de l'attribut de performance auxquels il se rapporte. Par contre, une certaine ambiguïté existe en ce qui concerne les valeurs que peut prendre l'échelle de progrès pour décrire la quantité d'un attribut retenu. Lorsqu'elle réfère à l'envergure par exemple, l'échelle adopte des termes comme « peu » et « beaucoup » tandis qu'elle utilise les termes « quelques » et « nombreux » pour décrire l'exactitude en termes de nombre d'erreurs.

Malgré l'utilisation fréquente et répandue de l'échelle descriptive d'appréciation en éducation, très peu d'études se sont attardées à cet aspect particulier et à leurs répercussions sur son efficacité. Étant donné la popularité grandissante de cet outil et le besoin particulier de préciser les attributs en termes de qualités prenant une valeur différente progressant d'un niveau à l'autre, les recherches devraient abonder davantage dans ce sens. À cet égard, l'étude empirique de Rohrmann (2002) explore, à partir d'analyses statistiques exhaustives, les données provenant d'un échantillon d'enseignants en ce qui concerne leur perception de l'adéquation des termes utilisés à chacun des niveaux de progression de l'échelle descriptive d'appréciation. L'étude s'intéresse précisément à la fréquence, à la quantité et à l'intensité de nombreux attributs. Ses résultats l'amènent à classer, à l'intérieur des niveaux appropriés de l'échelle, les valeurs les plus populaires reliées à chacun des attributs. Par exemple, la valeur *peu* pour décrire la fréquence de la clarté se situe le plus souvent au premier niveau de l'échelle selon la perception des personnes interrogées à cet effet. Cette classification sert de point de départ dans l'établissement de points d'ancrage aux fins d'une interprétation commune et universelle issue de l'utilisation de l'échelle descriptive d'appréciation. D'autres études semblables sont nécessaires afin de valider et solidifier les observations de Rohrmann.

Les quelques principes énoncés par Rohrmann (2002) et Tierney et Simon, (2004) en ce qui concerne l'échelle ne garantissent pas nécessairement son utilisation efficace en contextes d'apprentissage. En effet, il faut également se pencher sur la pertinence pratique des attributs et des critères à retenir lors de l'élaboration de l'échelle. Par exemple, l'attribut « clarté » peut s'observer à l'aide d'une échelle de fréquence ou d'intensité. Toutefois, avant de pouvoir affirmer que la performance soit claire sur la base de la fréquence d'une performance observée, il faut avoir à sa disposition un nombre suffisant

d'occasions de l'observer. Bien que les praticiens aient tendance à opter pour cette échelle en raison de sa nature plutôt objective, ces derniers n'ont pas toujours accès à de multiples occasions d'observer la performance ciblée. Dans de tels cas, l'échelle de fréquence perd de sa pertinence et doit faire place à une échelle d'intensité. Bref, lors de la détermination de l'échelle descriptive d'appréciation, il y a lieu d'expliciter ses critères de performance, ses attributs et ses échelles sous-jacentes afin d'assurer sa pertinence et son efficacité au plan pratique.

Échelle qualitative quantifiée

La responsabilisation des systèmes éducatifs à partir de programmes d'évaluation à grande échelle laisse présager un danger imminent, soit celui du besoin de comparer l'efficacité des systèmes aux dépens de celui du développement réel des compétences chez l'élève. Les politiques relatives à ces programmes privilégient des résultats quantitatifs même lorsqu'ils font appel à une échelle descriptive d'appréciation. Souvent, cela signifie une transformation des résultats qualitatifs en lettres nuancées ou carrément en notes chiffrées.

La présente section traite des variantes de la notation qualitative à partir d'échelles descriptives d'appréciation. Plusieurs facteurs comme l'horaire scolaire, le milieu de travail, le nombre d'élèves dans la classe, la forme du bulletin scolaire et l'expérience de l'enseignant ainsi que les valeurs, l'approche pédagogique et la formation de ce dernier influencent la conversion de la notation qualitative en résultats quantitatifs (Defibaugh, 2000; Mabry, 1999). Les contraintes administratives des programmes d'évaluation à grande échelle encouragent également l'application de pratiques de notation qui vont à l'encontre des intentions de la communication explicite des résultats d'évaluation en salle de classe. Plutôt que de renseigner qualitativement sur les acquis et le progrès de l'élève dans son apprentissage, ces mesures incitent les enseignants à convertir les résultats qualitatifs en notes globales et chiffrées sans pour autant fournir des méthodes de conversion fondées et adéquates. Cette pression amènent les praticiens en salle de classe à adopter des procédures de conversion plus ou moins complexes et douteuses, d'un descriptif qualitatif à une note numérique. Il existe donc actuellement toute une gamme de procédures de conversion de la note qualitative à la note chiffrée, en partant d'une quantification interne à l'échelle descriptive jusqu'à la quantification du résultat global qualitatif en pourcentage. Les prochaines sections présentent trois catégories de modalités de conversion et traitent des difficultés inhérentes à chacune d'elles.

Premièrement, la transformation quantitative se fait généralement à partir d'algorithmes spécifiques et plus ou moins sophistiqués. Selon Cizek (2000), les systèmes de notation qualitatifs dégénèrent souvent en une échelle quantitative. Par exemple, dans le cadre d'une évaluation du respect des conventions

grammaticales en écriture, certaines échelles présentent comme descripteurs de progression, des énoncés portant sur la quantité d'erreurs, tels que plus que 10 erreurs = 0 point, entre 6 et 10 = 1 point, entre 3 et 5 erreurs = 2 points et 2 erreurs ou moins = 3 points pour un maximum de trois points. Chacune des échelles est ainsi quantifiée de manière à produire un total quantitatif pour l'ensemble. Les épreuves pan-canadiennes du Programme d'indicateurs du rendement scolaire (PIRS) en lecture et écriture, en mathématiques et en sciences, régies par le Conseil des Ministres de l'Éducation, Canada (CMEC), font toutes appel à une échelle descriptive d'appréciation. Bien que les épreuves du PIRS en sciences et en mathématiques mesurent des connaissances et des habiletés, elles tendent à privilégier l'application manuelle ou informatisée d'un algorithme plus ou moins complexe plutôt que l'attribution globale d'un des cinq niveaux de l'échelle pour les performances. Voici l'algorithme tiré du rapport technique du CMEC en mathématiques :

> Pour être classé à un niveau de performance donné, un élève devait satisfaire à deux critères : avoir répondu correctement à 15 items de ce niveau sur 25 et avoir répondu à un nombre minimal de questions dans chaque domaine. (1998, p. 46)

Dans cette même catégorie, d'autres écrits font ressortir des pratiques où l'évaluateur attribue une note alphanumérique, établie de A à *n* ou de 1 à *n*, selon le nombre de niveaux de l'échelle et accorde une note par rapport à chacun des critères de l'échelle, ce qui résulte en une série de notes, par exemple, 4, 3, 3, 2 pour quatre critères de performance retenus (Arter & McTighe, 2001; Cizek, 2000; Northwest Regional Educational Laboratory, 1998; Trumbull & Farr, 2000). La note finale se calcule alors par la fréquence des notes définie à partir du mode, donc la note 3; à partir de la moyenne, donc 3; ou de la médiane qui donne encore 3. Cette note, sur un maximum de 4, en supposant un total de quatre niveaux, est ensuite convertie en une note chiffrée, soit 75%. Certaines pratiques considèrent le nombre total de points à additionner, à pondérer au besoin et à les transformer directement en pourcentage. En supposant qu'il y ait quatre niveaux, il s'agirait d'additionner 4+3+3+2=12/16 (note maximale) X 100%, ce qui donnerait 75%. Lorsqu'un critère, disons le premier dans ce cas-ci, se voit accordé une pondération double, il y a donc lieu de multiplier par deux ce qui donne 8+ 3+3+2=16/20, ou 80%. De plus, des pratiques transforment directement, à partir d'une règle de conversion plutôt arbitraire la note accordée en note numérique. Par exemple, un rendement qui reçoit deux 3 et un 4 se voit attribué un A, jugé équivalent à 90% (Arter & McTighe, 2001; Trumbull & Farr, 2000). Des auteurs suggèrent également la pratique d'étendre le potentiel de notation de l'échelle afin de pouvoir mieux discriminer la performance des élèves en ajoutant un « + » et un « – » autour de la valeur attribuée à chaque niveau de l'échelle (Academic Senate for California Community Colleges, 1996; Penny, Johnson & Gor-

don, 2000). Voici un exemple d'une échelle de conversion : A+=4.3; A=4.0; A–=3.7; B+=3.3; B=3.0; B–=2.7; C+=2.3; C= 2; C–=1.7; D+=1.3; D=1; D–=0.7; F=0.

Dans le même ordre d'idées, Biggs (1992; 1995) quantifie une échelle descriptive d'appréciation basée sur une matrice taxonomique qui se compose de cinq types de performance transversale et cognitive, chacun ayant trois niveaux de maîtrise. Dans le but d'atteindre des fins administratives, il attribue une note chiffrée à chacune des cellules de cette matrice de 3X5 cellules selon le niveau de complexité et le niveau de la performance. En supposant que toutes les cellules soient pondérées également, la cellule A1 reçoit une note chiffrée de 15, la cellule A2=14, la cellule A3=13, la cellule A4=12, la cellule A5=11, la cellule B1=10, la cellule B2=9 et ainsi de suite. Biggs suppose que ces scores s'additionnent à l'intérieur et à travers des matières pour donner un résultat final chiffré. À première vue, l'échelle numérique utilisée dans l'exemple ci-dessus se veut ordinale avec des exigences mathématiques particulières plus ou moins respectées. La manipulation mathématique du résultat doit donc se limiter à la nature de l'échelle numérique sous-jacente.

Deuxièmement, certains auteurs proposent de conserver le rôle qualitatif et descriptif de l'échelle jusqu'à la fin du processus d'évaluation et de ne transformer le résultat en note chiffrée que pour répondre à des besoins administratifs, organisationnels ou politiques (Marzano, 2000; O'Connor, 1999; Suzuki, 1998; Waltman, Kahn & Koency, 1999). Marzano présente une technique dite « score émergent » ou *power law* qui permet de mieux estimer le score réel de l'élève. Cette approche requiert l'examen de l'ensemble des notes qualitatives attribuées au cours d'une période d'études afin de détecter une tendance dans le progrès des apprentissages. Cette tendance peut, par exemple, faire ressortir des notes plus élevées vers la fin de la période d'évaluation. Une fois la note attribuée, elle peut ensuite subir une transformation quantitative aux fins administratives selon une règle logique établie plus ou moins arbitrairement, par exemple, 4 représente un score entre 90 % et 100%. Cette méthode tient compte également de toute pondération adoptée. Dans sa recherche, Suzuki rapporte une note qualitative par dimension cognitive d'une échelle en mathématiques mais transforme les résultats en notes chiffrées à des fins psychométriques, c'est-à-dire afin de déterminer les qualités techniques de l'échelle lorsque celle-ci est appliquée à divers contextes disciplinaires.

Dans la troisième catégorie, des chercheurs et praticiens insistent pour rapporter des résultats plutôt qualitatifs que quantitatifs (Kulm, 1994; Simon & Forgette-Giroux, 2001; Kahl, 1995; Stroble, 1993). Le *process score* de Kulm fournit à l'élève un profil de notes associées à chacune des dimensions d'une échelle à quatre niveaux, soit, par exemple, compréhension – 4, solution – 4, communication – 3 et transfert – 2, ce qui renseigne ainsi davantage l'élève

que le fait une simple note de 10/12, ou 83% ou B. Encore plutôt abstrait, le profil de notes alphanumériques permet aux élèves d'identifier leurs forces et leurs faiblesses relatives aux dimensions de la compétence cible. Par ailleurs, Simon et Forgette-Giroux (2001), Kahl (1995) et Stroble (1993) présentent des échelles descriptives d'appréciation qui donnent des résultats purement qualitatifs. Les écrits font parfois référence à cette approche comme étant descriptive et portant sur les concepts de développement, de croissance ou de progrès (Simon & Forgette-Giroux, 2001; 2003; Trumbull & Farr, 2001; Wiggins, 1998). Les niveaux de progression indiquent la différence perçue dans la maîtrise ou dans l'atteinte de la compétence et prennent des valeurs comme «novice » à « expert » ou de « bien » à « exceptionnel », selon une échelle de fréquence ou d'intensité. Ces qualificatifs sont ensuite consignés tels quels au relevé de notes.

La notation descriptive des résultats d'évaluation a l'avantage d'articuler en termes narratifs la performance de l'élève, ce qui facilite la communication aux parties concernées. Elle assure un plus grand lien entre les attentes du curriculum et l'apprentissage de l'élève. Inversement, comme l'exprime Sadler (2004), dès que l'on codifie la performance de l'élève, la connexion disparaît entre celle-ci et les attentes visées par les programmes d'études (p. 2). Malgré l'apparence d'une certaine subjectivité, la notation descriptive, comparativement à la note chiffrée, offre plus de clarté, de précision et de rétroaction directe à l'élève. Elle s'inscrit davantage dans la l'esprit de l'interprétation critériée puisque l'élève s'évalue uniquement sur la base d'une performance attendue plutôt que par rapport à ses pairs. Étant donné la situation actuelle en ce qui concerne la notation à partir de l'échelle descriptive d'appréciation, la consignation qualitative des résultats d'évaluation n'est qu'à ses débuts et des recherches de son impact éventuel sur la qualité des apprentissages s'avèrent essentielles.

Échelle métrologiquement équilibrée

Les échelles de notation qualitative, conçues d'abord comme substituts aux échelles quantitatives devenues inadéquates et instables dans leur forme traditionnelle, particulièrement lors de l'évaluation de tâches complexes de performance, doivent nécessairement conduire à des résultats significatifs et stables, peu importe leur contexte d'application. La pertinence et la constance de ces échelles renvoient aux qualités métrologiques de validité et de fidélité. Il existe de nombreuses façons de vérifier les diverses facettes de la validité de l'échelle descriptive d'appréciation. La validité se préoccupe de la congruence entre le contenu de la tâche d'évaluation et la performance attendue (Moskal & Leydens, 2000). Il importe que les niveaux, les critères, les attributs et les descripteurs de l'échelle tiennent compte des éléments de la tâche, de la matière enseignée ainsi que des composantes de la compétence et de ses différents stades de maîtrise (Della-Piana, 1993; Hunter,

Jones & Randhawa, 1996; Tardif, 1993). Selon Moskal et Leydens, cette validité s'établit à partir des liens clairement établis entre ces paramètres. Il n'est cependant pas simple pour les enseignants d'établir cette concordance. Dans leur étude, Pomplun, Capps et Sundbye (1998) mettent en évidence la difficulté qu'éprouvent les enseignants comme juges à respecter les dimensions sous-jacentes à l'échelle descriptive d'appréciation lors de la notation globale. Ces derniers ont tendance à miser sur le détail ou à ne s'en tenir qu'à des critères implicites, indépendants de ceux retenus dans l'échelle. Leur tâche se complique lorsque les dimensions et les stades de développement de la compétence visée ne s'appuient pas nécessairement sur des fondements théoriques. En effet, Potter (1995) démontre que les correcteurs à une épreuve d'écriture, ont réussi à ne reconnaître que deux dimensions sur cinq de la compétence. Dans ce cas, l'identification initiale des niveaux de l'échelle, lors de son élaboration, peut se faire arbitrairement mais à l'aide d'échantillons de travaux d'élèves représentatifs des divers niveaux de performance. Turner (2000) confirme, à partir de l'étude empirique d'une méthode d'élaboration d'une échelle de notation en français langue seconde, l'effet significatif des échantillons de travaux d'élèves dans l'élaboration d'échelles conformes aux intentions d'évaluation et aux stades de maîtrise de la compétence.

Parallèlement, lors de l'examen de la validité des trois niveaux de performance de l'échelle du *National Assessment of Educational Progress* en mathématiques 4e, 8e et 12e années aux États-Unis, Burstein, Koretz, Linn, Sugrue, Novak, Baker et Harris (1995/1996) soulignent l'importance d'interpréter les résultats obtenus illustratifs de ce que les élèves « peuvent » faire en fonction de ce que les experts jugent qu'ils « devraient » faire. Dans le même sens, Elser (1997) et Baker (1994) établissent la validité des dimensions et des stades de développement de la compétence à l'intérieur de leurs échelles respectives en écriture et en histoire en s'appuyant sur la théorie dans ces domaines. Dans un contexte de notation globale, Elser fait appel à l'opinion de 45 enseignants experts et consulte les écrits théoriques. Non seulement cette approche cherche-t-elle à établir le lien entre l'échelle et le construit, mais elle démontre également à quel point la note attribuée réussit à capter les dimensions de la compétence et leur relation entre elles. Selon Abedi et Baker (1995), la validation des niveaux de l'échelle descriptive d'appréciation en histoire s'améliore avec l'utilisation d'une approche de modélisation des variables. Simon et Forgette-Giroux (2001), pour leur part, assurent un degré de correspondance entre les dimensions d'une échelle d'appréciation utilisée pour évaluer les travaux d'étudiants en contexte universitaire et les critères d'évaluation d'écrits académiques adoptés par la communauté scientifique à des fins de publication d'articles. Ces études illustrent le besoin d'examiner davantage l'étape d'identification des stades de développement et de les distancer plus ou moins également dans le but d'élaborer des échelles descriptives d'appréciation davantage valides.

Lorsqu'il traite de la validité de ces échelles, Marzano (2000) se montre très positif en ce qui concerne le jugement que l'enseignant porte à l'aide de l'échelle. Pour lui, les évaluations réalisées par l'enseignant à l'aide de l'échelle descriptive d'appréciation sont presque toujours plus exactes que celles obtenues avec des notes chiffrées. À ce chapitre, il rejoint l'opinion de Wiggins (1998) et de Guskey et Bailey (2001) qui disent que, puisque les enseignants connaissent leurs élèves, comprennent comment ils travaillent et ont une notion claire des progrès réalisés, leurs perceptions les amènent à offrir une description exacte de ce que les élèves ont appris. Toutefois, lorsqu'ils communiquent les résultats de notation aux parents, les enseignants doivent faire en sorte de fournir toutes les informations relatives à la signification de la note, notamment des copies types, une description détaillée des critères et une explication des niveaux afin que le message véhiculé puisse améliorer l'apprentissage de l'élève ou entraîner un ajustement positif de sa part (Wiggins, 1998).

Pour ce qui est de la fidélité, celle-ci s'attarde à la stabilité et à la constance des résultats. Peu importe à quel moment l'épreuve est administrée, quand elle est notée et qui la note, les résultats obtenus devraient, en principe, être semblables. La principale forme de fidélité à considérer lors de l'élaboration d'une échelle descriptive d'appréciation concerne la concordance interjuges et intrajuge (Moskal & Leydens, 2000). Une échelle élaborée selon certaines règles de l'art réduit la possibilité d'obtenir des scores différents avec plusieurs correcteurs. D'abord, une définition claire de chacun des critères dans l'échelle et des consignes précises incitent le correcteur à se référer à l'échelle régulièrement lors de son application (Burstein et al., 1995-1996). Les questions ci-dessous réfèrent à d'autres règles clés (Della-Piana, 1993; Moskal & Leydens., 2000; Turner, 2000) et contribuent ainsi à la clarté et à la précision de l'échelle descriptive d'appréciation :

- la description des niveaux s'interprète-t-elle de façon constante?
- la distinction des niveaux de notation contribue-t-elle à une interprétation stable?
- des échantillons de travaux d'élèves ou des copies-types illustrent-ils les niveaux de notation?

La participation des praticiens à l'élaboration de l'échelle, le partage des critères de l'échelle avec les élèves et les exercices de médiation, ou d'établissement de consensus entre juges au moment de la notation, augmentent également la fidélité interjuges. Aussi, un nombre suffisant de tâches d'évaluation ainsi que leur pertinence à l'objet d'évaluation contribuent-ils à atteindre un niveau de fidélité davantage acceptable. De plus, l'interprétation commune des diverses composantes de l'échelle descriptive d'appréciation s'améliore sensiblement lorsque l'élève consulte des exemples de travaux modèles, reçoit une rétroaction régulière et participe aux activités de médiation. Finale-

ment, les correcteurs formés produisent des scores plus stables et le fait de connaître au préalable les échelles descriptives d'appréciation augmente la performance des élèves aux tâches d'évaluation parce qu'ils communiquent mieux aux élèves ce que réussir au niveau supérieur signifie (Johnson, Penny & Gordon, 2000; Marzano, 2000; Schafer, Swanson, Bene & Newberry, 2001; Zuzovsky, 1999).

Certains chercheurs continuent à étudier en profondeur les difficultés associées à la fidélité de l'échelle qualitative de notation appliquée en salle de classe. Penny, Johnson et Gordon (2000), par exemple, estiment que le fait de noter les élèves à partir d'une échelle à quatre niveaux avec l'ajout de + ou – à chacun des niveaux tend à améliorer le coefficient interjuges. Toutefois cette conclusion ne fait pas l'unanimité dans les écrits. L'utilisation de ces sous-catégories peut faciliter la répartition des performances mais l'absence d'articulations descriptives de chacune d'elles donne souvent lieu à des surévaluations. Aussi, la fidélité des échelles descriptives d'appréciation a-t-elle fait l'objet d'études empiriques afin de comparer la stabilité de la notation globale et de la notation analytique (Goulden, 1994; Hunter et al., 1996; Klein et al., 1998; Pomplun et al., 1998; Swartz et al., 1999; Waltman, Kahn & Koency, 1999). Les résultats de ces travaux tendent à confirmer que les deux approches produisent des niveaux de fidélité semblables mais que l'approche globale convient mieux aux évaluations à grande échelle tandis que l'approche analytique sert mieux les besoins des enseignants en salle de classe. L'absence de base conceptuelle constitue la plus grande faiblesse reprochée à la notation globale. Hunter et ses collaborateurs suggèrent qu'elle prenne ses assises théoriques dans la psychologie gestaltiste, cette dernière fournissant des éléments qui aident à comprendre la dynamique des méthodes de notation, méthodes qui peuvent être définies comme des exercices de perception contrôlés. Cependant, le peu d'études relatives à cet aspect de la fidélité incite à la plus grande prudence.

Conclusion

Tel que mentionné précédemment, l'échelle descriptive d'appréciation fut initialement conçue comme substitut à l'échelle numérique devenue désuète et instable, particulièrement lors de l'évaluation d'une compétence en salle de classe. Par ailleurs, son utilité dépend largement de sa capacité à fournir des renseignements valables, constants et fiables. Les preuves de validité et de fidélité de cette échelle suscitent encore beaucoup de préoccupations. La fonction de l'évaluation, l'arrimage de ses composantes à l'objet de l'évaluation et son articulation en contexte pratique se retrouvent au coeur de cette problématique. Si le milieu de la salle de classe possède des exigences propres et des besoins particuliers en ce qui concerne l'évaluation, les programmes à grande échelle imposent des paramètres spécifiques et souvent incompatibles avec les réalités de la salle de classe.

Effectivement, dans sa classe, l'enseignant utilise à des fins principalement formatives l'échelle descriptive d'appréciation afin de rendre compte du cheminement de l'apprentissage de l'élève et en vue d'améliorer ou d'ajuster son enseignement. Ces échelles sont habituellement axées sur la tâche et les contenus d'apprentissages. Souvent, elles s'éloignent des échelles génériques et ne s'appliquent que dans un contexte unique, ce qui contribue à une notation spécifique et pointue. Dans ce cas, la fonction analytique de l'échelle prévaut. Lorsque l'échelle sert à évaluer des performances complexes à des fins sommatives, elle prend une forme plus générique et fait appel à une notation plutôt globale. Peu importe sa fonction formative ou sommative, l'efficacité de l'échelle générique dépend de la prédominance des aspects qualitatifs et descriptifs, de la simplicité et de la transparence de l'échelle, de la pertinence et de l'utilité des critères de rendement, de l'accessibilité à des copies types de la performance attendue, de l'adéquation de l'échelle à la compétence visée, de l'ouverture à la participation des évalués à l'évaluation, de l'articulation des critères et des attributs à l'intérieur de l'échelle et de la sensibilité de l'échelle au progrès, à la rétroaction et au réinvestissement.

Pour les administrateurs préoccupés par la question de responsabilisation du système et de ses couts, l'échelle descriptive d'appréciation s'inscrit dans des conditions fort différentes, telles que l'établissement de bilans des acquis, la standardisation et la comparaison. Dans ce contexte, les dirigeants privilégient surtout la notation sommative et globale et mettent souvent l'accent sur la fidélité aux dépens de la validité. Curieusement, toutefois, les deux contextes d'application de l'échelle se rejoignent dans leurs habitudes quasi-automatiques des praticiens de quantifier les résultats qualitatifs, faute probablement du manque de règles établies « de A à Z » guidant l'utilisation de la notation et la présence d'un vocabulaire souvent flou et non défini relatif à l'évaluation de la performance.

Le besoin de vouloir quantifier les résultats descriptifs de la notation qualitative demeure un problème majeur. Les pressions administratives actuelles des diverses institutions scolaires et les traditions en évaluation encouragent le recours à une lettre ou un chiffre pour quantifier un jugement qualitatif fait à partir de l'échelle. Selon les dirigeants scolaires, la note chiffrée apparaît comme étant une mesure plus objective, plus valable, plus fidèle et plus facilement communicable. Pourtant, Weiss (1996) montre à quel point la notation qualitative ne gagne rien à être réduite à une quantification :

> La notation chiffrée est commode, apparemment facile à lire et à comprendre. Chiffrée, jusqu'aux décimales quelquefois, elle est considérée généralement comme précise, juste, indiscutable. C'est ce qui fait sa force, sa résistance au temps. Elle rassure en effet parents, élèves, jusqu'aux enseignants eux-mêmes, qui voient en elle une précision mathématique bienvenue dans une école tellement conflictuelle, tellement complexe. Ces notes sont toutefois bien commodes pour prendre des décisions administratives, pour stimuler le travail des élèves ou pour maîtriser la gestion de la classe.

> L'analyse la plus superficielle montre pourtant combien elle est inadéquate et trompeuse. L'évidence est telle qu'en faire la démonstration serait faire injure au lecteur. Chacun sait combien l'usage de chiffres est abusif et combien la précision des évaluations est illusoire. Depuis de nombreuses années, les insuffisances de la notation chiffrée sont dénoncées et critiquées. (p. 23)

Toutes ces nouvelles perspectives en notation qualitative trouvent écho dans ce qui est communément appelé « évaluation de la performance. » C'est pourtant l'objet de l'évaluation qui détermine le choix du paradigme de notation, car vouloir utiliser l'échelle descriptive d'appréciation à toutes les intentions d'évaluation ne peut que nuire à sa validité. D'ailleurs, il existe d'autres dispositifs de notation qui servent à noter des connaissances ou des habiletés de base. Ceci cependant ne se fait pas toujours sans confusion et, pour arriver à des changements pertinents, cohérents et durables, il faut procéder à des clarifications, à des explications et à des ajustements susceptibles de rallier la communauté scolaire, tant administrative qu'enseignante. Pour donner à la notation qualitative sa juste et véritable place, ne faut-il pas s'accorder avec Marzano (2000) selon qui l'abandon graduel et à plus long terme des notes chiffrées est susceptible de mener éventuellement à une description riche de sens et d'informations pertinentes ?

RÉFÉRENCES

Abedi, J., & Baker, E. L. (1995). A latent-variable modeling approach to assessing interrater reliability, topic generalizability, and validity of a content assessment scoring rubric. *Educational and Psychological Measurement, 55*(5), 701-715.

Academic Senate for California Community College. (1996). *Plus and minus grading options: Toward accurate student performance evaluations*. ERIC Document Reproduction Service No ED395631)

Arter, J. (1993). *Designing scoring rubrics for performance assessments: The heart of the matter*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta. ERIC Document Reproduction Service No ED358143.

Arter, J. (1999). Teaching about performance assessment. *Educational Measurement: Issues and Practice, 18*(2), 30-44.

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom*. Thousand Oaks, CA: Corwin Press, Inc.

Azwell, T., & Schamr, E. (Eds). (1995). *Report card on report cards*. Portsmouth. NH: Heinemann.

Baker, E. L. (1994). Leaning-based assessments of history understanding. *Educational Psychologist, 29*(2), 97-106.

Baker, E. L., Aschbacher, P. R, Niemi, D., & Sato, E. (1992). *CRESST Performance assessment models*. Available online at http://cresst96.cse.ucla.edu/CRESST/Sample/CMODELS.PDF

Biggs, J. B. (1992). A qualitative approach to grading students. *HERSA New, 14*(3), 3-6.

Biggs, J. B. (1995). Assessing for learning: Some dimensions underlying new approaches to educational assessment. *The Alberta Journal of Educational Research, 41*(1), 1-17.

Brahier, D. (2001). *Assessment in middle & high school mathematics: A teacher's guide*. Larchmont, NY: Eye on Education.

Brookhart, S. M. (1999). Teaching about communicating assessment results and grading. *Educational Measurement: Issues and Practice, 18*(1), 5-13.

Brualdi, A. (1998). Implementing performance assessment in the classroom. *Practical Assessment, Research, & Evaluation*, 6(2). Available online at http:/pareonline.net/getvn.asp?v=7&n=10

Burstein, L., Koretz, D., Linn, R., Sugrue, B., Novak, J., Baker, E. L., & Harris, L. E. (1995/1996). Describing performance standards: Validity of the 1992 National Assessment of Educational Progress achievement level descriptors as characterizations of mathematics performance. *Educational Assessment, 3*(1), 9-51.

Busching, B. (1998). Grading inquiry projects. In Anderson, R. S., & Speck, B. W. (Eds.), *Changing the way we grade student performance: Classroom assessment and the new learning paradigm (p. 74)*. San Francisco: Jossey-Bass Publishers.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1),16-33.

Cizek, G. J. (2000). Pockets of resistance in the assessment revolution. *Educational Measurement: Issues and Practice, 19*(2), 16-23.

Conseil de l'Europe. (2001). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Conseil de la Coopération culturelle. Comité de l'éducation. Division des langues vivantes. Strasbourg : Didier.

Conseil des ministres de l'éducation, Canada (CMEC). (1998). *Rapport du programme d'indicateurs du rendement scolaire en mathématiques*. Toronto.

Custer, R. L. (1998). Rubrics: An authentic assessment tool for technology education. *The Technology Teacher*, 27-37.

Davis, T. M. (1993). *Development and use of a visual arts portfolio scoring procedure implemented by the California Art Education Association Student Assessment Project*. Unpublished doctoral dissertation, University of Oregon.

Defibaugh, K. L. (2000). *An examination of the evaluation practices of elementary visual arts teachers*. Unpublished doctoral dissertation, State University of New Jersey.

Della-Piana, G. (1993). Teacher, what does my writing test score mean? *Language Arts, 70*(7). 583-590.

Depka, E. (2000). *Designing rubrics for mathematics*. Arlington Heights, Illinois: Skylight Professional Development.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5*(1), 39-70.

Elser, T. L. (1997). *A descriptive correlation study of the Holistic Developmental Writing Scales designed for children in grades K-6*. Unpublished doctoral dissertation, University of Montana.

Firestone, W. A. (2000). *Format, Focus, and Frustation. The Policy and Politics of State Testing*. Paper presented at the annual conference of the British Educational Research Association, Cardiff, Wales. Also available online at http://www.cepa.gse.rutgers.edu/page2.htm

Francis, R., & Hord, S. (1995). *Designing scoring tools for authentic & alternative assessments: A common sense method*. (ERIC Document Reproduction Service No ED392804)

Frisbie, D. A., & Waltman, K. K. (1992). Developing a personal grading plan. *Educational Measurement: Issues and Practice, 11*(13), 35-42.

Goodrich-Andrade, H. (2000). Using rubrics to prompt thinking and learning. *Educational Leadership, 57*(5), 13-18.

Goulden, N. P. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *Journal of Research and Development in Education, 27*(2), 73-82.

Guskey, T. R., & Bailey, J. M. (2001). *Developing Grading and Reporting Systems for Student Learning*. Thousand Oaks, CA: Corwin Press – Sage Publications.

Harper, M., O'Connor, K., & Simpson, M. (1999). *Quality assessment: Fitting the pieces together.* Toronto: OSSTF.

Hunter, D. M., Jones, R. M., & Randhawa, B. S. (1996) The use of holistic versus analytic scoring for large-scale assessment of writing. *The Canadian Journal of Program Evaluation, 11*(2), 61-85.

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education, 13*(2), 121-138.

Kahl, S. R. (1995). *Scoring issues in selected statewide assessment programs using non-multiple-choice formats.* Paper presented at the annual conference of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Services No ED392846)

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., Comfort, K., & Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11(2), 121-137.*

Krulik, S., & Rudnick, J. A. (1999). *Assessing reasoning and problem solving: A sourcebook for the elementary school teachers.* Toronto: Allyn & Bacon-Prentice-Hall.

Kulm, G. (1994). *Mathematics assessment : What works in the classroom.* San Francisco: Jossey-Bass.

Mabry, L. (1999). Writing to the rubric - Lingering effects of traditional standardized testing on direct writing assessment. *Phi Delta Kappan, 80*(9), 673-679.

Marzano, R. J. (2000). *Transforming classroom grading.* Alexandria: ASCD.

Ministère de l'Éducation de l'Ontario. (2000). Série de cinq vidéocassettes intitulée *L'évaluation: une vision nouvelle* portant sur les thèmes suivants: *L'évaluation du rendement, La notation et la collecte de données, La communication du rendement, L'évaluation et les élèves ayant des besoins particuliers, L'amélioration du rendement.*

Moskal, B. (2000). Scoring rubrics: what, when and how? *Practical Assessment, Research and Evaluation, 7*(3). Available online: http://pareonline.net/getvn.asp?v=7&n=3

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research, & Evaluation, 7*(10). Available online: http://pareonline.net/getvn.asp?v=7&n=10

Northwest Regional Educational Laboratory. (1998). *Improving Classroom Assessment: A Toolkit for Staff Developers.* Regional Educational Laboratory, sponsored by U. S. Department of education, Office of Educational Research and Improvement (OERI). Available from NWREL. Portland, or http://www.nwrel.org

O'Connor, K. (1999). *How to grade for learning.* Arlington Heights, IL: Skylight.

Penny, J., Johnson, R. L., & Gordon, B. (2000). Using rating augmentation to expand the scale of an analytical rubric. *Journal of Experimental Education,* 68(3), 269-287.

Phillips, L. (1998). *Assessment handbook 3.* Scarborough, ON: Prentice Hall Ginn Canada.

Pomplun, M., Capps, L., & Sundbye, N. (1998). Criteria teachers use to score performance items. *Educational Assessment,* 5(2), 95-110.

Popham, W. J. (1997). What's wrong – and what's right – with rubrics. *Educational Leadership,* 55(2), 72-75.

Popham, W. J. (1999). *Classroom assessment: What teachers need to know* (2nd ed.). Toronto: Allyn & Bacon.

Potter, G. P. (1995). *Factor analysis of the Domain Scoring Rubric for the Arkansas Writing Assessment.* Unpublished doctoral dissertation, University of Arkansas.

Rogers, S. & Graham, S. (2000). *The high performance toolbox: Succeeding with performance tasks, projects, and assessments.* Evergreen, CO: Peak Learning Systems.

Rohrmann, B. (2002). *Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data*. Retrieved October 7, 2003, from University of Melbourne: http://www.psych.unimelb.edu.au/staff/br/vqs-report.pdf

Sadler, D.R. (2004). *How criteria-based grading misses the point.* Presentation to the Effective Teaching and Learning Conference, Griffith University, Southbank campus.

Saunders, P. I. (1999). *Primary trait scoring: A direct assessment option for educators*. Paper presented at the National council of Teachers of English Annual Conference. (ERIC Document Reproduction Services No ED444624)

Schafer, W. D., Swanson, G., Bene, N., & Newberry, G. (2001). Effects of teacher knowledge of rubrics on student achievement in four content areas. *Applied Measurement in Education, 14*(2), 151-170

Simon, M., & Forgette-Giroux, R. (2001). A rubric for scoring postsecondary academic skills. *Practical Assessment, Research, & Evaluation, 7*(18). Also available online at http://pareonline.net/getvn.asp?v=7&n=18

Simon, M., & Forgette-Giroux, R. (2003). Évaluer pour informer: l'utilisation du dossier d'apprentissage. Dans M. Laurier (Éd.), *Évaluation et communication*. Montréal: Éditions Logiques.

Solano-Flores, G., & Shavelson, R. (1997). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice, 16*(3), 16-25.

Solano-Flores, G., Shavelson, R. J., Ruiz-Primo, M. A., Schults, S. E., Wiley, E. W., & Brown, J. H. (1997). *On the development and scoring of classification and observation science performance assessments*. Paper presented at the annual conference of the American Educational Research Association. (ERIC Document Reproduction Services No ED411314)

Stiggins, R. J. (2001). *Student-involved classroom assessment* (3rd Ed). Upper Saddle River, NJ: Merrill/Prentice-Hall.

Stroble, E. J. (1993). Kentucky student portfolios: Expectations of success. *Equity & Excellence in Education, 26*(3), 54-60.

Stuhlmann, J., Daniel, C., Dellinger, A., Denny, K., Kenton, R., & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Reading Psychology, 20*(2), 107-127.

Suzuki, K. (1998). *Measuring "To think mathematically": Cognitive characterization of achievement levels in performance-based assessment.* Unpublished doctoral dissertation, Urbana: Illinois.

Swartz, C. W, Hooper, S. R., Montgomery, J. W., Wakely, M. B., DeKruif, R. E. L., Reed, M., Brown, T. T., Levine, M. D., & White, K. P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement, 59*(3), 492-506.

Tardif, J. (1993). L'évaluation dans le paradigme constructivisme. Dans R. Hivon (Éd.), *L'évaluation des apprentissages*. Sherbrooke: Éditions du CRP.

Taylor, C. S. (1998). An investigation of scoring methods for mathematics performance-based assessments. *Educational Assessment, 5*(3), 195-224.

Taylor, C. S., & Bidlingmaier, B. (1998). Using scoring criteria to communicate about the discipline of mathematics. *Mathematics Teacher, 91*(5), 416-426.

Tierney, R. & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment and Research in Education*, 9(2). Available at: http://pareonline.net/getvn&asp?v=9&n=2

Trumbull, E., & Farr, B. (2000). *Grading and reporting in an age of standards*. Norwook, MA: Christopher-Gordon Publishers, Inc.

Turner, C. (2000).Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *The Canadian Modern Language Review*, 56(4), 555-584.

Waltman, K., Kahn, A., & Koency, G. (1999). *Alternative approaches to scoring: The effects of using different scoring methods on the validity of scores from a performance assessment.* CSE Technical Report 488. (ERIC Document Reproduction Service No. ED427080)

Weiss, J. (1996). Évaluer plutôt que noter. *Revue internationale d'Éducation de Sèvres*, *11*, 3-34.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* San Francisco: Jossey-Bass Publishers.

Zuzovsky, R. (1999). Problematic aspects of the scoring system of the TIMSS practical performance assessment: Some examples. *Studies in Educational Evaluation*, *25*(3), 315-323.

RENÉE FORGETTE-GIROUX est professeure titulaire et vice-doyenne aux programmes à la Faculté d'éducation de l'Université d'Ottawa. Ses intérêts de recherche sont l'évaluation des apprentissages en salle de classe et à grande échelle. Elle s'intéresse particulièrement aux politiques, aux pratiques, aux stratégies et aux qualités des instruments d'évaluation.

MARIELLE SIMON est professeure en mesure et évaluation à la Faculté d'éducation de l'Université d'Ottawa depuis 1994. Son programme de recherche actuel comprend l'étude des pratiques de notation en salle de classe, l'évaluation formative et l'analyse des liens entre le contexte pédagogique et le rendement des élèves francophones minoritaires aux enquêtes nationales et internationales en mathématiques, lecture et écriture.

RENÉE FORGETTE-GIROUX is a full professor and Vice-Dean (Programs) in the Faculty of Education at the University of Ottawa. Her research interests are classroom and large-scale assessments. She focusses particularly on the study of policies, practices, and the technical qualities of assessment instruments and approaches.

MARIELLE SIMON has been a professor in measurement and evaluation in the Faculty of Education at the University of Ottawa since 1994. Her current research program includes the study of grading methods in the classroom, formative assessment and the links between teaching practices and the achievement of minority Francophone students on national and international large scale studies in mathematics, reading and writing.