Pédagogie Médicale
REVUE INTERNATIONALE FRANCOPHONE D'ÉDUCATION MÉDICALE

# Adjonction de descripteurs qualitatifs à l'échelle ordinale des épreuves de format mini-CEX pour optimiser l'évaluation formative des compétences cliniques

## *Addition of Qualitative Descriptors to the Ordinal Scale of a Mini-CEX-Type Examination to Improve Formative Assessment of Clinical Competence*

*Diem-Quyen NGUYEN[1], Florence WEBER[1], Robert GAGNON[2]*

**Résumé** ***Contexte :*** *Les échelles de mesure qui servent actuellement à évaluer des compétences cliniques sont souvent de type Likert. Or, dans une perspective d'évaluation formative où un grand nombre de professeurs et d'étudiants sont impliqués, l'utilisation de ces échelles pour la rétroaction est très inégale. Il est alors suggéré que les standards de performance soient décrits explicitement sur ces échelles pour faciliter la rétroaction.* ***But :*** *Vérifier si l'adjonction à l'échelle de mesure globale des mini-CEX de descripteurs explicitant ces standards affecte la validité et la fidélité des mesures enregistrées lors des sessions d'observation directe des performances cliniques.* ***Méthode :*** *Pendant une période de 18 mois, 86 résidents des niveaux I à III en médecine interne ont eu des séances d'observation directe par 50 professeurs-cliniciens.* ***Résultats :*** *L'analyse de la consistance interne montre un coefficient alpha de Cronbach à 0,90. Cette échelle modifiée permet de plus la distinction des performances entre les résidents I et III et montre une progression de performance sur 18 mois.* ***Conclusion :*** *Lorsqu'elle est utilisée dans une perspective d'évaluation formative, l'ajout de descripteurs qualitatifs à une échelle de mesure globale ne emble affecter ni la fidélité des mesures, ni leur validité. Cette modalité pourrait notamment être explorée lorsque la formation professorale à des outils de mesure s'avère difficile par manque de temps.*

**Mots clés** Évaluation formative ; échelle de mesure.

**Abstract** ***Background:*** *Global rating scales, presently used to assess clinical competencies, are often of the Likert type. In a perspective of formative assessment where numerous clinical teachers and students are involved, the use of Likert scales for feedback is very unequal. It has been suggested that performance standards explicitly described on these scales would improve feedback.* ***Aim:*** *To verify if the addition of descriptors on scales of the mini-CEX-type examination affects the reliability and the validity of measurements recorded during sessions of direct observations of clinical performances.* ***Method:*** *Desirable behaviours are incorporated in the mini-CEX rating scale. Over a period of 18 months, 50 clinical teachers used that scale to observe directly 86 first to third year post-graduate students in internal medicine for formative assessment purposes.* ***Results:*** *Statistical analysis has shown a Cronbach's alpha coefficient of 0.90. Furthermore, this modified scale helps to distinguish training levels and shows a tendency of improvement in the residents' performance over the 18-months period.* ***Conclusion:*** *This study shows that the addition of descriptors to a global rating scale does not seem to affect its reliability and validity when it is used in a structured formative assessment.*

**Key words** Formative assessment; measurement scale.

*Pédagogie Médicale 2007;8:207-16*

*1- Département de Médecine - Faculté de médecine - Université de Montréal.*
*2- Centre de pédagogie appliquée aux sciences de la santé (CPASS)- Université de Montréal et Collège des médecins du Québec*
*Correspondance : Diem-Quyen Nguyen - Centre hospitalier universitaire de Montréal - Hôpital Saint-Luc - Service de Médecine interne. 1058 Saint-Denis - H2X 3J4 Montréal (Québec) Canada. Téléphone : (514) 890-8000, poste 32584. Télécopieur : (514) 412-7308. Mailto:diem.quyen.nguyen@umontreal.ca*

## Introduction

La formation clinique par des stages dans différents milieux cliniques a toujours été une méthode privilégiée pour préparer les étudiants à leur future pratique professionnelle. En médecine, comme dans d'autres disciplines, la base de cette formation consiste à exposer ces étudiants de différents niveaux de formation à des problèmes cliniques variés et complexes, afin qu'ils puissent développer et consolider leurs habiletés de recueil des données cliniques et de résolution efficace de ces problèmes.

Pendant plusieurs décennies, des procédures d'évaluation clinique exploitant notamment des situations cliniques à partir desquelles les étudiants étaient exhaustivement et longuement interrogés (« cas longs » – *long cases* –) ont été utilisées[1] ; elles répondaient à divers formats codifiés tels que, par exemple, *les Clinical EXercise (CEX)* ou les *Objective Structure Long Examination Record (OSLER)*. Pour satisfaire les critères de validité et de fidélité, et compte tenu des limites liées à la « spécificité de contexte ou de cas » *(case specificity)*, l'attention a été attirée sur la nécessité d'exposer les étudiants à un échantillonnage de situations d'évaluation suffisamment riche, ce qui posait des problèmes en termes de disponibilité des enseignants. Des recommandations ont alors été formulées invitant à recourir plutôt à des évaluations courtes et récurrentes qu'à des évaluations longues et isolées. C'est dans ce contexte que les « cas longs » ont pratiquement été abandonnés en Amérique du Nord et que se sont développées, entre autres, les procédures d'évaluation de type examen clinique objectif structuré *(ECOS ; Objective Structured Clinical Examination – OSCE –)* ou *mini-CEX* (ces dernières remplaçant les CEX). En 1998, à l'issue d'une revue détaillée de ces méthodes d'évaluation des compétences cliniques, Holmboe[2] a insisté sur la nécessité d'une observation effective des performances des étudiants, tout en reconnaissant que cette activité requiert beaucoup de ressources professorales. Epstein[3], récemment, a aussi examiné les problèmes soulevés par l'observation directe des performances cliniques comme méthode d'évaluation ; selon lui, la contrainte principale rencontrée est la disponibilité des professeurs.

Pour pallier certaines difficultés liées à ce problème, Norcini a proposé une version abrégée des CEX, les *mini-CEX*, pour évaluer les performances cliniques. Il s'agit de courtes sessions d'environ 15 à 30 minutes, au cours desquelles les professeurs observent les résidents en train soit d'interroger un patient, soit de l'examiner de façon pertinente, soit de lui communiquer une nouvelle.

Concernant les qualités psychométriques de ce format d'épreuve, Norcini rapporte un indice de reproductibilité (qui rend compte du critère de fidélité) de 0,80 lorsque 12 à 14 sessions de *mini-CEX* sont mises en œuvre pour un étudiant pendant un an[4, 5]. Pour guider les professeurs dans leur observation directe, Norcini propose une grille d'observation de type Likert à neuf niveaux[4]. Dans un contexte d'évaluation sommative, plusieurs études ont démontré la validité de cette méthode en la comparant à des examens de certification validés, qu'ils soient écrits[6] ou oraux[7]. Holmboe a également rapporté une étude destinée à vérifier la validité de construit des *mini-CEX*[8], qui confirme que les *mini-CEX*, qui utilisaient des scripts de situations d'évaluation prédéterminés, permettent de départager les résidents forts et les résidents faibles. Ces études ont contribué à conforter le grand engouement dont jouissent les *mini-CEX* auprès des directeurs de programmes, en raison de la facilité avec laquelle cette méthode peut être incorporée aux activités cliniques habituelles. Cependant, on peut aussi noter qu'il y a une grande variabilité des scores pour des items évalués. En effet, l'étude de Holmboe rapporte un étalement des scores de 1 (insatisfaisant) à 6 (satisfaisant) pour la même performance d'un étudiant simulant le rôle d'un « résident faible ». Dans une perspective d'évaluation formative, lors de la rétroaction, il est alors difficile d'expliquer à l'étudiant pourquoi il a obtenu, par exemple, un score de quatre au lieu de six pour une performance jugée satisfaisante. De la même manière, il peut être difficile de s'assurer de la même compréhension des professeurs vis-à-vis des exigences respectives de différents niveaux de performance (« insatisfaisant », « satisfaisant » et « supérieur ») et pour différents niveaux de formation des étudiants. Une formation des professeurs visant à optimiser et homogénéiser leur utilisation de l'outil de mesure est souhaitable mais, dans un contexte de programmes de formation impliquant de nombreux professeurs, une telle formation requiert beaucoup de temps professoral et n'est pas toujours réalisable[9]. Crossley[10] suggère alors que les outils de mesure soient sélectionnés selon les besoins en tenant compte de ces contraintes. Selon lui, l'évaluation formative devrait viser à fournir une rétroaction en se basant sur des profils de forces et de faiblesses avec une définition claire des critères de performance.

C'est dans ce contexte que nous avons modifié les *mini-CEX*, en ajoutant quelques descripteurs qualitatifs à l'actuelle échelle globale de mesure, pour que l'utilisation de cette échelle par les professeurs soit moins variable, même s'ils n'ont pas reçu de formation spéciale préalable. L'ajout des descripteurs qualitatifs à une échelle ordinale

n'est pas un concept nouveau. Même si elle demeure restreinte à ce stade de développement, notre contribution s'inscrit ainsi dans le courant qui souligne l'intérêt de développer des rubriques *(rubrics* ou *scoring rubrics)*, tel que l'ont explicité notamment Huba et Freed[11] en enseignement supérieur. Comme le rappelle Tardif[12], les rubriques sont des critères d'évaluation qui distinguent des degrés de maîtrise de ressources internes et externes en termes qualitatifs, en recourant à des dimensions jugées essentielles. En revanche, en médecine, peu de données existent sur la validité de ces types d'échelle. Il est ainsi licite de chercher à vérifier si ces descripteurs peuvent affecter la validité ou la fidélité des mesures.

Nous rapportons dans cette étude les résultats de l'utilisation, dans un cadre d'évaluation formative des performances cliniques, d'épreuves de type *mini-CEX* recourant à une échelle de mesure modifiée par l'adjonction de tels descripteurs qualitatifs.

# *Méthodes*

## Création de l'échelle descriptive globale

La grille d'évaluation des *mini-CEX* a d'abord été modifiée en y ajoutant des descripteurs qui reflètent les caractéristiques de standards de performance, tels que l'on peut les retrouver dans des recommandations académiques telles que l'*« ACGME Outcome Project »*[13], le *« Can MEDS 2000 »*[14] et le *« Good Medical Practice »*[15]. Cette première version de l'échelle a été ensuite soumise et discutée au Comité de programme de médecine interne de l'Université de Montréal. Ce projet s'inscrit dans le désir des responsables du programme de résidanat en médecine interne d'instituer formellement l'observation directe structurée comme méthode d'évaluation formative des compétences cliniques pour tous les résidents des trois premières années. A l'issue de ces deux étapes, une grille d'évaluation comportant une échelle avec sept items, reflétant respectivement chacun les tâches cliniques lors d'une entrevue médicale, et neuf niveaux de performance par item, a été créée. Les neuf niveaux sont les mêmes que ceux retrouvés dans l'échelle originale de *mini-CEX*.

Un processus visant à déterminer les niveaux de performance a ensuite été développé. Au cours de ce processus, douze professeurs ont visionné deux enregistrements audiovisuels de deux résidents qui questionnent et examinent un patient. Ils ont utilisé la grille fournie pour attribuer des scores de performance, comparé ensuite leurs scores respectifs et discuté afin d'atteindre un consensus. Cette discussion a amené une clarification des comportements désirables et indésirables et constitue la base des descripteurs. Seules les performances de niveau « satisfaisant » ou « exceptionnel » sont décrites, le but étant de transmettre aux résidents les comportements désirables pour les encourager à atteindre ces niveaux de performance. De plus, un consensus a aussi été atteint quant au niveau de performance attendu pour chaque niveau de résidence. Ces attentes sont notées sur l'échelle de mesure *(figure 1)*. Une mise à l'essai de cette échelle a eu lieu pendant 18 mois.

## Description des sessions d'observation directe

Le protocole de la mise à l'essai de la grille d'évaluation a été approuvé par le Comité de programme en 2004. Selon ce protocole, chaque résident serait soumis à deux sessions d'observation directe dans la même journée, tous les six mois. Pour souligner le caractère formatif de l'exercice, la méthode a été renommée Évaluation formative des performances cliniques (ÉFPC). Chaque session d'ÉFPC dure au total soixante minutes. Les premières trente minutes de l'épreuve se déroulent auprès du patient où, sous observation directe, le résident interroge, examine et répond à une question du patient en se limitant à son problème médical principal. Au cours des trente dernières minutes, en l'absence du patient, le résident résume et propose une résolution du problème clinique rapporté par le patient. Il bénéficie par la suite d'une rétroaction détaillée de la part du professeur sur sa performance concernant ce cas clinique.

Afin de diminuer les biais de la complaisance, les professeurs et les résidents ont été assignés de façon aléatoire afin de s'assurer qu'aucun des résidents ne soit évalué par un professeur qui l'a connu préalablement. Aucun professeur n'a eu une formation préalable ; une description de la méthode et de l'outil de mesure a été envoyée à chacun des participants deux semaines avant les sessions d'ÉFPC. Les patients sont sélectionnés par les professeurs et le niveau de difficulté des cas cliniques est estimé *a priori* par ces mêmes professeurs.

## Analyse statistique

Afin de s'assurer que cette grille, malgré l'ajout des descripteurs, continue à être un outil valide pour évaluer les compétences cliniques, l'analyse de la validité de construit a été effectuée sous deux angles : 1) déterminer si l'échelle peut détecter les différents niveaux de performance (par une analyse transversale) et 2) vérifier si l'échelle permet de détecter la progression du résident sur 18 mois.

**Figure 1 :**
**Grille d'appréciation des performances cliniques des étudiants aux épreuves *mini-CEX***

- **Evaluateur** : _____ **Date** : _____
- **Résident** : _____ **Niveau** : ☐ R-I ☐ R-II ☐ R-III
- **Patient** : Type de problème _____

Niveau de complexité : Faible ☐ Modéré ☐ Élevé ☐

***Niveau attendu de performance :*** *R-I = 4* *R-II = 5* *R-III = 6*

**A. Anamnèse** — **Commentaires**

- **Le questionnaire médical**

1 2 3 — Amélioration nécessaire
4 5 6 — Questions pertinentes, concises, incluant des données psychosociales (valeurs personnelles, niveau socio-économique…)
7 8 9 — Exceptionnel

- **Relation thérapeutique**

1 2 3 — Amélioration nécessaire
4 5 6 — A su établir une relation de confiance, de compréhension, empathique et respectueuse
7 8 9 — Exceptionnel

**B. Examen physique**

1 2 3 — Amélioration nécessaire
4 5 6 — Examen pertinent et efficace
7 8 9 — A su faire un examen exhaustif de façon efficace, afin de vérifier d'autres diagnostics alternatifs

**C. Synthèse de cas**

1 2 3 — Amélioration nécessaire
4 5 6 — A réussi à bien synthétiser le problème clinique en intégrant les données de l'histoire et de l'examen physique
7 8 9 — Exceptionnel

**Figure 1 (suite)**

**Commentaires**

**D. Diagnostic proposé**

| 1 2 3 | 4 5 6 | 7 8 9 |
|---|---|---|
| Doit s'améliorer | A su proposer une solution avec des arguments convaincants en se basant sur des données cliniques recueillies | A su trouver des diagnostics élaborés avec une argumentation claire et précise en se basant sur des données scientifiques intégrant les données cliniques |

**E. Plan d'investigation**

| 1 2 3 | 4 5 6 | 7 8 9 |
|---|---|---|
| Doit s'améliorer | Le plan d'investigation est approprié, utile et rentable | A su bien choisir en distinguant ce qui est idéal et ce qui est nécessaire, tenant compte de restrictions médicales et financières |

**F. Plan de traitement (pour R-III)**

| 1 2 3 | 4 5 6 | 7 8 9 |
|---|---|---|
| Doit s'améliorer | A su expliquer le traitement proposé au patient | A expliqué de façon claire le traitement et ses effets secondaires |

**G. Impression globale et commentaires**

| A améliorer | A maintenir |
|---|---|
| | |
| | |
| | |
| | |
| | |

**H. Durée de l'entrevue médicale :** _____

**Durée de la session :** _____

Le score global est un score combiné additif, représentant la somme des scores obtenus par chacun des résidents pour chacun des sept items de la grille. Ce procédé vise à souligner l'importance de chacune des composantes lors d'une performance clinique.

Le score moyen des deux sessions d'ÉFPC est calculé pour chaque résident de façon semestrielle lors de l'étude afin de documenter la progression des résidents et de vérifier la différence entre les niveaux de résidence.

La fidélité de l'outil, ou consistance interne, est mesurée par le cœfficient alpha de Cronbach. Une analyse de variance (ANOVA) des scores est effectuée en plus afin de détecter la possible présence d'un mode de cotation systématique de la part des évaluateurs *(« response-set »)* en calculant le ratio de variance inter- et intra- sujet. La corrélation des scores entre les deux sessions d'ÉFPC à chaque semestre est mesurée par le coefficient de corrélation intra-lasse.

L'étude ANOVA est aussi utilisée pour vérifier une différence éventuelle entre les scores obtenus par les trois groupes de résidents. Le test *a posteriori* de Scheffé sert par la suite à identifier les différences significatives (l'analyse transversale). La progression des résidents est mesurée par le test t pour données appariées (analyse longitudinale).

Tous les tests sont considérés comme statistiquement significatifs au seuil alpha $p < 0,05$.

## Résultats

Au total, 50 professeurs ont participé à ce projet. Quatre-vingt-six résidents ont eu des sessions d'ÉFPC pendant 18 mois. Il y a eu 48 résidents de niveau I (RI), 21 résidents de niveau II (RII) et 17 résidents de niveau III (RIII), avec un nombre égal de femmes et d'hommes. Le plus grand nombre de RI résulte de l'ajout d'un deuxième groupe de RI étant donné que le projet est étalé sur 18 mois et que l'année académique n'est que 12 mois.

Le niveau de complexité des cas cliniques a été estimé pour 16 % d'entre eux comme facile, pour 62 % comme étant de difficulté modérée et pour 22 % comme étant de difficulté élevée.

### Étude de la fidélité de mesure

Le cœfficient alpha de Cronbach pour les scores moyens de cette échelle dépasse 0,90 et est de 0,94 ; 0,93 ; et 0,91 pour les semestres 1, 2 et 3 respectivement. Quant aux cœfficients de corrélation entre les deux sessions pour chaque semestre, il est respectivement de 0,26 ; 0,36 et 0,06. Ce cœfficient de corrélation est positif et statistiquement significatif pour les deux premiers semestres, mais non significatif au troisième semestre. Une analyse additionnelle des sources de variance démontre une plus grande variance dans la performance entre les résidents qu'entre les différents items pour chaque résident. Ce dernier résultat suggère que la performance à un item semble être prédictive de la performance à un autre item.

### Étude de la validité de l'échelle

Le *tableau 1* montre la progression des scores globaux pour la cohorte de huit résidents de chaque niveau de formation, chez qui des données complètes sont disponibles pour les trois semestres. On observe une tendance à la progression du semestre 1 au semestre 3, même si ce n'est pas significatif, à la fois pour le groupe des RI (de 31,3 à 34,1) et des RII (35,1 à 37,1).

Le *tableau 2* montre les performances de chaque groupe de résidents selon leur niveau de formation à différents semestres. Leur performance varie selon leur niveau de formation et l'analyse montre une différence significative de performance entre le niveau des RI et des RIII. De plus, l'analyse détaillée de leur performance selon le niveau de difficulté des cas cliniques rencontrés a aussi été effectuée *(tableau 3)*, l'hypothèse étant que si l'échelle de mesure devait être utilisée pour évaluer les compétences cliniques, il devrait y avoir une différence significative selon les niveaux d'expérience des résidents pour chaque niveau de difficulté des problèmes cliniques. Comme on peut, de nouveau, le constater, la différence de performance entre les RI et les RIII est hautement significative ($p < 0,01$) pour les trois niveaux de difficulté.

## Discussion

Les résultats de cette étude tendent à démontrer que l'ajout de descripteurs qualitatifs à une échelle de mesure globale ne semble pas affecter la fidélité des mesures ni sa validité, quant à son utilisation dans un contexte d'évaluation formative des compétences cliniques.

Ainsi, même si les professeurs n'ont pas une formation préalable, la consistance interne de l'instrument dépasse 0,80. Ceci est comparable aux valeurs rapportées par

**Tableau 1 :**
**Évolution des scores globaux des performances cliniques des résidents aux épreuves *mini-CEX*, en fonction du nombre de semestres de fonction**

| Semestre | Résident I (n=8) | Résident II (n=8) |
|---|---|---|
| 1 | 31,3 (4,6) | 35,1 (4,4) |
| 2§ | 32,2 (4,6) | 37,6 (7,7) |
| 3§ | 34,1 (4,4) | 37,1 (5,2) |
| Valeur de p (entre le semestre 1 et 3) | 0,18 | 0,08 |

*RI : résidents en premier semestre ; RII : résidents en deuxième semestre ; RIII : résidents en troisième semestre.*
*§: puisque l'étude est étalée sur 18 mois, au semestre 2 et 3, les résidents I et résidents II sont devenus résidents II et résidents III respectivement.*

**Tableau 2 :**
**Performances des étudiants aux épreuves *mini-CEX* selon le niveau de résidence**

| | Semestre 1 Moyenne (ds) | Semestre 2 Moyenne (ds) | Semestre 3 Moyenne (ds) |
|---|---|---|---|
| RI | 31,1 (5,3) n = 17 | 29,7 (4,8) n = 25 | 30,1 (4,4) n = 23 |
| RII | 34,4 (3,9) n=13 | 32,9 (5,2) n = 17 | 34,6 (4,7)** n = 17 |
| RIII | 36,2 (5,0)* n = 13 | 37,7 (6,3)* n = 14 | 37,3 (4,3)* n = 13 |

*RI : résidents en premier semestre ; RII : résidents en deuxième semestre ; RIII : résidents en troisième semestre ; ds : déviation standard*
** p < 0,05 entre RI et RIII*
*** p < 0,05 entre RI et RII*

**Tableau 3 :**
**Performances des étudiants aux épreuves *mini-CEX* selon le niveau de difficultés des cas cliniques**

| | Faible<br>Moyenne (ds) | Modéré<br>Moyenne (ds) | Haut<br>Moyenne (ds) |
|---|---|---|---|
| RI(§) | 29,5 (5,6) | 30,5 (6,6) | 30,2 (6,0) |
| RII(§) | 31,6 (4,8) | 34,0 (5,0)* | 33,1 (6,5) |
| RIII(§) | 37,4 (5,9)** | 36,7 (4,3)** | 37,4 (6,8)** |

** p < 0.05 entre RI et RII*
*** p < 0.01 entre RI et RIII*
§ *RI : résident 1 ; RII : résident II ; RIII : résident III*

Norcini lorsqu'il a présenté des données sur les *mini-CEX*[4]. De plus, tel que rapporté dans une étude ultérieure par la même équipe[5], même si l'échelle permet de discriminer la performance entre les individus, les scores attribués à chaque résident semblent être en très grande corrélation entre les items. Nos données confirment aussi cette tendance. Les limites de notre recherche ne permettent pas de déterminer la cause de cette observation. Des études additionnelles seraient utiles afin de vérifier s'il s'agit de chevauchement entre différents domaines sous-tendant la compétence clinique, de redondance d'informations ou d'effet de halo. En effet, il serait alors intéressant de vérifier s'il y a un élément transversal à chacun de ces sept items (par exemple les connaissances cognitives) ou tout simplement si un évaluateur peut être influencé par les performances à une dimension lorsqu'il évalue une autre dimension. Quant à la redondance des items, comme ces sept items représentent des tâches cliniques courantes, à la fois retrouvées dans la littérature médicale et unanimement acceptées par les membres du comité de programme et par les 12 professeurs qui ont été sollicités pour valider la grille, il serait difficile de déterminer lequel de ces sept items serait superflu.

Par ailleurs, il existe une corrélation positive entre les professeurs, malgré l'absence de formation préalable. Cette dernière donnée implique que pour un même résident, deux professeurs différents évaluent sa performance de façon comparable. Même si ce résident n'a pas le même score, dans une situation d'évaluation formative, il reçoit quand même le même message : sa performance est jugée soit satisfaisante ou insatisfaisante par les deux professeurs, ce qui facilite sa compréhension de ses propres performances.

Comme l'esprit de l'évaluation formative est de faire progresser les étudiants, les résultats obtenus avec cette grille tendent aussi à montrer une progression dans les performances, comme cela a déjà été constaté dans d'autres études. En revanche, selon nos résultats, si ces sessions d'observation directe structurée et formelle se prolongent, avec une partie importante consacrée à la rétroaction, deux sessions avec deux problèmes cliniques différents par session, chaque semestre, pourraient suffire pour vérifier la progression des performances des résidents. Ce dernier constat découle des résultats qui semblent confirmer la progression des résí-

dents au cours de la période étudiée. Ceci pourrait alors être une solution partielle au problème de manque de temps professoral consacré à l'observation directe des résidents au cours de la formation clinique. Toutefois, et conformément aux mises en garde de Van der Vleuten et Schuwirth[16], il est nécessaire d'être prudent dans la prise de décision quant au nombre de cas nécessaire pour inférer de façon juste les compétences cliniques à partir d'un nombre restreint de cas. Comme ils l'ont suggéré, toute méthode d'évaluation doit être planifiée et mise en œuvre dans un contexte spécifique d'enseignement, en tenant compte de l'objectif pour lequel elle est conçue ; dans cette perspective, cette échelle descriptive utilisée dans le contexte d'évaluation formative avec des cas cliniques authentiques pourrait répondre aux besoins d'évaluer les résidents pour faciliter leurs apprentissages.

Finalement, ces résultats ouvrent un nouveau champ de recherche en identifiant plusieurs questions : l'adjonction de descripteurs qualitatifs aux échelles ordinales de mesure pourrait-elle faciliter l'évaluation formative et diminuer les difficultés inhérentes à l'utilisation des échelles globales traditionnelles lors de l'appréciation des performances cliniques[17] ? Quel est le rôle des rubriques d'évaluation en pédagogie des sciences de la santé ?

De nouveau, nos données doivent être interprétées et utilisées dans un contexte d'évaluation formative. L'échelle a été modifiée et le temps a été alloué aux professeurs pour discuter avec les résidents de façon approfondie afin de les aider dans le développement des compétences cliniques. L'échelle sert alors comme point d'ancrage pour initier le processus de rétroaction et les descripteurs pour faciliter la transmission aux résidents des standards de performance désirés.
Les limites de notre étude, à savoir qu'il s'agit d'un groupe d'étudiants d'une seule université, nous empêchent de généraliser nos résultats à d'autres programmes et d'autres universités. Il serait intéressant de vérifier si, dans le contexte d'un usage élargi, les données sur la fidélité des mesures s'amélioreraient.

***Remerciement***
*Nous aimerons remercier le Dr Khue Huu Nguyen pour sa révision de ce manuscrit.*

## *Contributions*

Diem-Quyen Nguyen a conçu le projet et effectué le recueil et l'analyse des données ; elle est la rédactrice principale de l'article. Florence Weber a apporté une contribution importante à la rédaction de l'article. Robert Gagnon a assuré la responsabilité de la saisie et de l'analyse des données.

## Références


1. *Wass V, Van der Vleuten CPM. The long case. Med Educ 2004;38:1176-80.*
2. *Holmboe ES, Hawkin RE. Methods for evaluating the clinical competence of residents in Internal medicine: a review. Ann Intern Med 1998;129:42-8.*
3. *Epstein RM. Assessment in medical education. N Engl J Med 2007;356:387-96.*
4. *Norcini JJ, Blank LL, Aznold JK, Kimball HR. The mini-CEX (Clinical Evaluation Exercise) : a preliminary investigation. Ann Intern Med 1995;123:795-9.*
5. *Norcini JJ, Blank LL, Duffy S, Fortna G. The mini-CEX : a method for assessing clinical skills. Ann Intern Med 2003;138:476-81.*
6. *Durning SJ, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the mini-clinical evaluation exercise for Internal Medicine residency training. Acad Med 2002;77:900-4.*
7. *Hatala JR, Slie M, Kassen BO, Mackie I, Roberts JM. Assessing the mini-clinical evaluation exercise in comparison to a national specialty examination. Med Educ 2006;40:950-6.*
8. *Holmboe ES, Huot S, Chung J, Norcini J, Hawkins RE. Construct validity of the mini-clinical evaluation exercise. Acad Med 2003;78:826-30.*
9. *Jolly BC. Faculty development for curricular implementation. In: Norman GR, van der Vleuten CPM & Newble DI (Eds). International handbook of research in medical education. Boston: Kluwer Academic Publishers, 2002;945-68.*
10. *Crossely J, Humphris G, Joly B. Assessing health professionals. Med Educ 2002;36:800-4.*
11. *Huba ME, Freed JE. Using Rubrics to provide feedback to students. In: Huba & JE Freed. Learner centered assessment on college campuses. Shifting the focus from teaching to learning. Needham Heights (MA) :Allyn and Bacon, 2000:150-200.*
12. *Tardif J. Ensuite des rubriques. In Tardif J. L'évaluation des compétences. Documenter le parcours de développement. Chenelière Éducation : Montréal, 2006;183-241.*
13. *ACGME (1999). ACGME Outcome Project, General competencies. [On-line]. Disponible sur: http//www.acgme.org/outcome/comp/compfull.asp.*
14. *Frank KJR, Jablour M, Tugwell P & al. Skills for the new millennium; report of the societal needs working group; CanMEDS 2000 project. Annals of Royal College of physicians and surgeons of Canada 1996;29:206-16.*
15. *General Medical Council. Good Medical Practice. 3e éd. [On-line]. Disponible sur : http//www.gmc-uk.org/guidance/good_medical_practice/index.asp#top .*
16. *Van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Med Educ 2005;9:309-17.*
17. *Williams R, Klamen DA, Mc Graghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med 2003; 15:270-92.*


*Manuscrit reçu le 26 mars 2007 ; commentaires éditoriaux formulés aux auteurs le 28 août 2007 ; accepté pour publication le 22 septembre 2007.*