Mesure et évaluation en éducation

Élaboration d'une grille d'évaluation formative des performances dans une perspective d'approche par compétences : un exemple du programme de formation postdoctorale en médecine interne

Diem-Quyen Nguyen et Jean-Guy Blais

Volume 30, numéro 2, 2007

URI : https://id.erudit.org/iderudit/1085886ar
DOI : https://doi.org/10.7202/1085886ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Nguyen, D.-Q. & Blais, J.-G. (2007). Élaboration d'une grille d'évaluation formative des performances dans une perspective d'approche par compétences : un exemple du programme de formation postdoctorale en médecine interne. *Mesure et évaluation en éducation*, *30*(2), 73–97. https://doi.org/10.7202/1085886ar

Résumé de l'article

L'approche de formation axée sur le développement des compétences est actuellement répandue au stade de la formation universitaire. L'évaluation formative des compétences devient dès lors une composante importante du curriculum. Nous rapportons ici nos réflexions et les démarches de validation suivies lors du développement d'une grille d'évaluation des compétences cliniques. Cette grille est l'outil de recueil de données qui sera incorporé à une nouvelle méthode d'évaluation des compétences appelée « Évaluations formatives des performances cliniques », créée à l'intention des étudiants du programme de formation postdoctorale en médecine interne.

Élaboration d'une grille d'évaluation formative des performances dans une perspective d'approche par compétences : un exemple du programme de formation postdoctorale en médecine interne

Diem-Quyen Nguyen

Jean-Guy Blais

Université de Montréal

Mots clés : Évaluation des compétences, évaluation formative, évaluation authentique, échelle de mesure

L'approche de formation axée sur le développement des compétences est actuellement répandue au stade de la formation universitaire. L'évaluation formative des compétences devient dès lors une composante importante du curriculum. Nous rapportons ici nos réflexions et les démarches de validation suivies lors du développement d'une grille d'évaluation des compétences cliniques. Cette grille est l'outil de recueil de données qui sera incorporé à une nouvelle méthode d'évaluation des compétences appelée « Évaluations formatives des performances cliniques », créée à l'intention des étudiants du programme de formation postdoctorale en médecine interne.

Key words : Assessment of competencies, formative assessment, authentic assessment, rating scale

The competency-based approach in higher education is being adopted on a large scale by most of professional training programs. Assessment of competencies is thereby becoming an area of great importance in such reform. We report here steps leading to the creation of a global descriptive scale to be used in an authentic method of formative assessment of clinical competencies.

Note des auteurs –Nous remercions madame Huguette Bernard pour sa contribution lors de la conception initiale de la grille d'évaluation. Toute correspondance peut être adressée comme suit : Diem-Quyen Nguyen, Service de médecine interne, CHUM–Hôpital Saint-Luc, 1058, rue Saint-Denis, Montréal (Québec) H2X 3J4, ou par courriel à l'adresse suivante : [diem.quyen.nguyen@umontreal.ca].

Palavras-chave: Avaliação de competências, avaliação formativa, avaliação autêntica, escala de medida

A abordagem da formação baseada no desenvolvimento de competências está, actualmente, alargada à formação universitária. A avaliação formativa das competências torna-se, assim, uma importante componente do currículo. Relatamos, aqui, as nossas reflexões e procedimentos de validação seguidos no desenvolvimento de uma grelha de avaliação das competências clínicas. Esta grelha é o utensílio de recolha de dados que será incorporada a um novo método de avaliação das competências chamado "Avaliações formativas das performances clínicas", construída para os estudantes do programa de formação pós-doutoral em medicina interna.

Introduction

La principale méthode de formation utilisées au stade des études médicales postdoctorales, tant en Amérique du Nord que dans plusieurs pays européens, consiste en des stages cliniques en hôpitaux universitaires ou communautaires. L'étudiant est ainsi confronté au contexte dans lequel il exercera et il apprend à résoudre des problèmes réels à complexité variable. La résolution de ces problèmes, sous la supervision d'un médecin compétent du même domaine, devrait permettre à l'étudiant de bâtir et de consolider ses bases de connaissances ainsi que d'atteindre éventuellement le niveau de compétence nécessaire pour lui permettre de pratiquer la médecine de façon indépendante. Cette formation requiert des méthodes valides d'évaluation formative mais les études actuelles portent principalement sur l'évaluation sommative. En effet, l'évaluation formative au cours de la formation postdoctorale est un domaine qui a reçu peu d'attention en recherche. La plupart des recherches porte plutôt sur des méthodes d'évaluation sommative (Harlen & James, 1997).

C'est dans ce contexte que nous cherchons à créer une grille d'évaluation formative des performances. L'élaboration d'une telle grille dans une approche par compétences est une démarche complexe. Elle nécessite plusieurs étapes tant sur le plan de la réflexion que sur celui de la réalisation. La présente démarche a ainsi pour but de rapporter les résultats découlant du processus d'élaboration et de validation du contenu d'une telle grille.

En première partie de ce texte, nous situons donc le contexte de formation et ses contraintes, en deuxième partie, nous décrivons nos réflexions sur l'évaluation formative dans une approche de développement de compétences et, finalement, en troisième partie, nous décrivons le processus de validation du contenu d'une grille d'évaluation des performances cliniques.

Les stages et le contexte de l'évaluation formative

Le stage clinique consiste à exposer des étudiants à la pratique dans différents milieux cliniques tels que les salles d'urgence, les unités de soins intensifs ou de soins hospitaliers et les cliniques ambulatoires. Au cours de ces stages, chaque étudiant apprend à prendre soin de véritables patients sous la supervision de différents cliniciens-enseignants. Le type de stages obligatoires et optionnels dépend de la future spécialité choisie par les étudiants. À titre d'exemple, pour des étudiants qui se destinent à la spécialité de médecine interne, les stages en cardiologie sont obligatoires, mais ceux en dermatologie sont considérés comme optionnels. Si les avantages de la formation par stages sont indéniables, il n'en demeure pas moins qu'ils présentent des contraintes importantes pour des activités d'apprentissage et d'évaluation, à la fois sommative et formative (Toohey, Ryan & Hughes, 1996). Par exemple, la capacité d'accueil limitée de chaque service hospitalier ne permet pas à tous de faire les mêmes stages en même temps dans le même milieu. L'ordre des stages au cours de la formation est donc différent d'un étudiant à un autre. Chaque étudiant profite alors de chaque stage différemment selon ses acquis précédents. L'aspect aléatoire des problèmes cliniques rencontrés au cours des stages rend aussi difficile la gestion de la planification exacte du contenu du stage. Finalement, la variabilité des compétences pédagogiques des maîtres de stages, souvent choisis pour leur compétence clinique plutôt que pédagogique, rend aussi inégal l'apprentissage et l'évaluation de chaque étudiant. Malgré ces problèmes, tout étudiant doit, à la fin de sa formation, atteindre un niveau de compétences prédéterminé (Holm, 2002).

Dans ce contexte, la difficulté de planifier des activités d'évaluation formative a toujours été un défi en formation clinique postdoctorale. Les méthodes d'évaluation actuelles ne semblent pas satisfaire les besoins de formation des étudiants (Usatine, Edelstein, Yajima, Slavin & Wilkes, 1997). Dans toutes les approches d'évaluation des compétences cliniques, l'instrument de recueil des données favorisé est la fiche d'évaluation de stage. Il s'agit de l'instrument le plus utilisé, le plus étudié et le plus critiqué (Turnbull & Van Barneveld, 2002). Si cet instrument est celui qui est retenu dans le contexte actuel, c'est

parce que la fiche d'évaluation est souple et facile à utiliser. La recherche sur sa fiabilité montre en effet qu'un niveau de 0,6 à 0,8 de corrélation inter-juge peut être atteint si elle est utilisée par différents professeurs au moins sept fois l'an dans une perspective sommative (Carline, Paun, Thiede & Ramsey, 1992). Il n'en demeure pas moins que la fiche d'évaluation de stage n'est pas sans soulever certains problèmes. Sur le plan conceptuel, il demeure que les performances cliniques dans des situations complexes et variables requièrent une intégration des connaissances, des habiletés psychomotrices et affectives pour interroger les patients, pour les examiner, pour établir une relation thérapeutique et pour résoudre le problème clinique. Cette démarche de résolution de problème comprend un diagnostic différentiel logique, ainsi qu'un plan d'investigation et de traitement approprié. Quant à son application, les utilisateurs (notamment les professeurs) ignorent souvent les objectifs de stage, ne connaissent pas suffisamment les étudiants et finissent par la remplir inadéquatement (Turnbull & Van Barneveld, 2002). De plus, il y a une certaine confusion dans l'utilisation de cette grille. En effet, lorsque les professeurs remplissent cette grille à la fin du stage, les résultats consignés ont alors une valeur sommative, mais par contre si elle est remplie au cours du stage, cette grille est alors considérée un instrument de mesure à des fins formatives. Lorsqu'une telle situation prévaut, la compréhension que l'étudiant a de ce que l'on attend de lui peut être affectée. La planification des activités de régulation deviennent aussi difficile.

Élaboration des « Évaluations formatives des performances cliniques » (EFPC)

Nous avons donc élaboré une grille pour les « Évaluations formatives des performances cliniques » (EFPC), en tentant de tenir compte de la variabilité du bagage de connaissances de chacun des étudiants, du temps que les professeurs peuvent consacrer à l'enseignement (toujours dans le contexte de formation en stages) et de la complexité des problèmes cliniques à résoudre.

L'EFPC est une méthode structurée et formelle d'évaluation formative selon l'approche de l'évaluation authentique des performances cliniques. Chaque session d'EFPC dure environ une heure et se déroule soit à l'hôpital, soit dans une clinique ambulatoire. Pendant une demi-heure, sous observation directe du professeur, l'étudiant interroge et examine le patient, et il répond ensuite à une ou plusieurs questions du même patient. Au cours de la deuxième demi-heure, en l'absence du patient, l'étudiant présente de façon synthétique ce qu'il a retenu du problème et propose un diagnostic différentiel,

un plan d'investigation ainsi qu'un traitement lorsque cela s'avère pertinent. Cette étape devrait durer dix minutes. Par la suite commence la rétroaction qui combine l'autoévaluation et l'évaluation par le professeur. L'étudiant et le professeur remplissent ainsi indépendamment une grille d'évaluation de la performance de l'étudiant. La discussion qui suit entre l'étudiant et le professeur porte principalement sur les points consignés sur les grilles respectives. L'accent est mis sur les raisons qui pourraient avoir amené l'étudiant à performer tel qu'il l'a fait.

L'observation directe au cours de la performance est le moyen privilégié pour recueillir des données au cours de l'activité puisque cette stratégie permet de constater la performance réelle de l'étudiant. Une grille d'observation doit par contre être créée pour servir de support à l'opération de recueil des données et à la rétroaction. La grille d'observation doit tenir compte des multiples contraintes inhérentes à la formation par stages : l'évolution variable des résidents au cours de leur formation ; le nombre élevé de professeurs engagés dans la formation ; la difficulté d'inférer les compétences à partir des performances avec un nombre limité de cas cliniques. La grille devrait en outre faciliter l'évaluation formative de l'étudiant en lui servant de repère pour ses lectures, sa façon d'agir et même sa façon d'être.

La démarche d'élaboration de la grille d'évaluation formative des compétences cliniques doit donc permettre de préciser les tâches qui y seront intégrées, d'expliciter les critères et standards de performances, de déterminer le nombre de niveaux de performance à intégrer, et de favoriser l'autoévaluation et la rétroaction.

Cadre théorique

L'élaboration de cette grille se réfère à des concepts tels que l'évaluation des compétences dans un contexte formatif ainsi que les caractéristiques des grilles d'évaluation et des échelles de mesure.

Évaluation des compétences

Le concept de compétence a été depuis 15 ans le sujet de plusieurs débats mais un consensus semble actuellement sur le point d'émerger. La compétence serait une manifestation des capacités stabilisées d'un individu à mobiliser ses connaissances cognitives, affectives et psychomotrices de façon intégrée, en faisant des choix pour résoudre efficacement des problèmes complexes dans des familles de situations (Le Boterf, 2000 ; Perrenoud, 1997 ; Voorhees, 2001).

Mais si le consensus sur le concept de compétence semble se profiler, peu d'études portent par contre sur ce que devrait être l'évaluation formative des compétences dans un programme de formation professionnelle universitaire. Quel devrait être l'aspect d'un instrument élaboré pour observer des compétences dans une perspective d'évaluation formative ? Comment peut-on s'assurer de la validité et de la fidélité des données dans l'évaluation formative des compétences basée sur des performances complexes ?

Les compétences sont inférées à partir de l'observation des performances complexes pour résoudre des problèmes qui s'apparentent à ceux que l'étudiant rencontrera dans sa vie professionnelle (Gagné, Wager, Grolas & Leller, 2005). L'évaluation authentique est une approche émergente qui prend actuellement une grande importance dans l'évaluation des compétences. Il s'agit un concept proposé par Wiggins en 1989, et bonifié en 1998. Pour que l'évaluation se situe dans un contexte authentique, elle devrait inclure des tâches réalistes et demander à l'étudiant de faire la preuve de son jugement et de sa capacité à innover. Il faudrait aussi que la performance amenant à la résolution du problème dépasse une simple démonstration. De plus, le contexte des performances devrait être proche du milieu de travail éventuel, le problème à résoudre devrait faire appel à l'intégration des habiletés et des connaissances. Finalement, un mécanisme de rétroaction devrait être intégré à l'évaluation pour améliorer la performance. Selon cette approche, la précision des attentes repose sur des échelles descriptives globales (terme utilisé par Scallon, 2004, pour traduire le terme «*rubrics*» utilisé en anglais) où sont décrits clairement des standards de performances et des standards de contenu (termes utilisés par Messick, 1994). Ce type d'échelle semble favoriser le développement de l'autoévaluation (Wiggins, 1998).

Évaluation des compétences dans un contexte formatif

Bloom, Hasting et Madaus (1971) définissaient l'évaluation comme une collection systématique d'évidences pour déterminer la quantité et le degré des changements survenus chez l'étudiant après qu'il ait suivi des activités d'enseignement. Cette définition influença longtemps l'évaluation dans l'approche basée sur les objectifs, où seuls des résultats observables devraient être évalués.

La portée de l'évaluation a été élargie avec l'avènement de la pédagogie basée sur les compétences, au cours des années 1980. Ainsi, le cheminement cognitif de l'étudiant lors de l'accomplissement des tâches complexes serait tout aussi important à considérer lors de l'évaluation que les résultats

obtenus. Glasner (1999) soutient de plus qu'évaluer est une activité formelle dont le but principal devrait être de fournir à l'étudiant une rétroaction significative. Selon cet auteur, seule une démarche d'évaluation qui tient aussi compte des processus cognitifs pour arriver à ces résultats serait susceptible de donner une information plus complète sur les capacités de l'étudiant. Huba et Freed (2000) considèrent d'ailleurs que l'évaluation fait partie des tâches du professeur d'université.

Généralement, il y a trois fonctions de l'évaluation des apprentissages qui sont reconnues dans les écrits sur le sujet : les fonctions diagnostique, sommative et formative (Harlen & James, 1997). L'évaluation formative est fondamentalement un processus d'évaluation pour aider l'étudiant à progresser. Plusieurs auteurs anglo-saxons et francophones estiment que différents aspects caractérisent l'évaluation formative quant à ses enjeux, aux rôles de l'étudiant et du professeur et aux implications pédagogiques (Allal 1991 ; Black & William, 1998 ; Gipps & Stobart, 2003 ; Saddler, 1998 ; Scallon, 2000). Tous croient que l'évaluation formative devrait avoir lieu au cours de la formation, être accompagnée d'un processus de rétroaction et être sous la responsabilité du professeur et de l'étudiant. Certains de ces auteurs ajoutent qu'un processus de régulation devrait aussi faire partie de la définition (Allal, 1991 ; Saddler, 1998).

Dans une perspective de développement des compétences, l'évaluation formative est un processus continu intégré aux activités d'apprentissage. L'observation des performances de l'étudiant au cours de ces activités est un des moyens privilégié pour recueillir l'information pertinente nécessaire à l'évaluation et à la rétroaction (Scallon, 2004). La très grande majorité des auteurs ayant étudié l'évaluation formative considère d'ailleurs que la rétroaction en est une composante essentielle.

Les études faites dans différents domaines comme en enseignement secondaire (Butler, 1988), en formation clinique (Hodder, Rivington, Calcutt & Hart, 1989) et en formation des professeurs (Brinko, 1993) démontrent l'importance de la rétroaction et son impact positif sur l'apprentissage de l'étudiant. Dans l'approche par compétences, lors de cette rétroaction, l'étudiant est encouragé à discuter les raisons qui sous-tendent ses performances. Cette discussion a pour but de mieux comprendre le processus cognitif de l'étudiant pour réajuster les activités d'apprentissage (Scallon, 2004).

Le souci de développer la métacognition chez l'étudiant devrait être une caractéristique additionnelle de l'évaluation formative selon l'approche par compétences. Cette dimension semble être fondamentale dans le processus du développement de l'autonomie de l'étudiant. L'autoévaluation constitue alors une étape importante du développement de la métacognition. L'autoévaluation est définie comme l'évaluation par l'étudiant de sa propre performance en se servant d'un référentiel externe, tandis que la métacognition fait référence à la capacité de l'étudiant de connaître et de juger son propre processus d'apprentissage pour s'améliorer (Gipps & Stobart, 2003 ; Laurier, Toussignant & Morissette, 2005).

En conclusion, soulignons que l'évaluation formative des compétences devrait être un processus dont l'élément central est l'évaluation des compétences au cours de la formation, en mettant l'accent sur l'individualisation de l'enseignement et de l'évaluation. Le développement de l'autoévaluation et de la métacognition devrait de plus occuper une place importante dans l'évaluation formative pour développer des compétences.

Grille d'évaluation pour des performances et échelle de mesure

Plusieurs méthodes d'évaluation des performances complexes ont été rapportées. Glasner (1999) en a décrit quelques exemples pour le secteur de l'enseignement supérieur : le portfolio, le rapport des projets, l'évaluation des performances dans des milieux de travail, les études de cas, les journaux de bord de recueils des incidents critiques et les évaluations cliniques objectives et structurées (ECOS). Un des principaux instruments de collecte des données utilisés par ces méthodes d'évaluation est constitué de grilles d'évaluation comprenant différents types d'échelles pour consigner la performance, qu'il s'agisse de listes de vérification, d'échelles graduées ou d'échelles descriptives (Laurier, Toussignant & Morissette, 2005).

Pour des performances complexes, une échelle avec des choix obligatoires ayant des catégories dichotomiques permet difficilement d'évaluer les compétences. Hodges, Regher, McNaughton, Tiberius et Hanson (1999) ainsi que Regher , MacRae, Reznick et Szalay(1998) ont ainsi illustré qu'il a été difficile de prédire des niveaux de compétence avec une liste de vérification comme on a pu le faire avec une échelle de mesure ordinale au cours des ECOS en médecine. Dans le contexte de l'évaluation des compétences par une approche authentique, le modèle d'échelle d'appréciation recommandé est de type « échelle descriptive globale ». Wiggins (1998) présente cette échelle

comme un outil qui décrit la tâche demandée et qui donne les dimensions de cette tâche. Notons que l'échelle descriptive globale diffère de l'échelle descriptive («*descriptive scale*») comme l'entendent Morgenstern et Keeves (1997). Les échelles classiquement appelées descriptives contiennent des items constitués de critères unidimensionnels et d'une échelle d'appréciation où chaque catégorie de réponse est un comportement descriptif. Selon Scallon (2004), ce qui distingue l'échelle descriptive de l'échelle descriptive globale est que les critères descriptifs séparés en items dans la première sont regroupés pour former une suite d'échelons dans la deuxième. Plus d'une vingtaine d'ouvrages et de nombreux sites Internet donnent des exemples variés de l'utilisation de l'échelle descriptive globale en éducation. Le point de convergence de ces écrits semble être que cette grille permettrait un meilleur recueil des données pour s'assurer d'une rétroaction de qualité, mais peu d'études ont été rapportées pour confirmer cette hypothèse.

Il n'y a pas non plus véritablement de consensus sur le nombre de catégories qu'une échelle de mesure devrait avoir. Selon Andrich (1997), le nombre de catégories devrait être assez grand pour profiter de la capacité de jugement et de discrimination d'un évaluateur. Ainsi, la majorité des grilles contiennent un nombre qui varie entre quatre et neuf niveaux. L'état d'imprécision se reflète aussi dans les études sur l'observation directe et l'évaluation de performances cliniques. Dans une perspective d'évaluation sommative, les listes de vérification de type dichotomique « fait ou non fait » ont été utilisées au tout début de la création des ECOS. Progressivement, ces listes de vérification ont fait de la place aux échelles d'appréciation allant de trois à neuf niveaux, comprenant des niveaux de fiabilité atteignant 0,68 avec dix sessions d'observation directe (Regher, MacRae, Reznick & Szalay, 1998 ; Norcini, Blank, Duffy & Fortna, 2003).

L'évaluation formative des compétences implique à la fois une appréciation de la réalisation d'une tâche et une appréciation du processus menant à la réalisation de cette tâche, puisque la notion de savoir-faire efficace est au centre du concept de compétence. Dans cette perspective, une grille d'évaluation formative des compétences basée sur des performances complexes devrait également permettre d'apprécier la démarche de résolution de problème au cours de la performance.

Validité et fidélité des mesures pour des performances complexes

L'évaluation des performances complexes présente un défi pour au moins deux raisons. Les performances complexes sont généralement plus longues à évaluer et, par conséquent, le nombre de situations d'évaluation des performances est restreint par la contrainte du temps disponible (Kane, 1999). L'évaluation des performances fait, de plus, appel aux appréciations de qualité et la subjectivité dans les jugements peut en devenir importante et entraîner de potentielles erreurs de mesure (Messick, 1994 ; Swanson, Norman & Linn, 1995).

Afin de s'assurer de la validité du contenu lors de l'évaluation des performances, Messick (1994) suggère d'utiliser l'analyse des tâches pour établir des standards de contenu et de performance. Cette étape fait appel au jugement des personnes reconnues compétentes dans le domaine.

En ce qui concerne l'aspect de la structure interne de la validité, Messik propose de se servir des modèles de mesure basés sur l'utilisation des échelles descriptives globales contenant des critères d'appréciation bien définis. Au cours de cette étape, la comparabilité de performances devrait être établie non pas en fonction des individus mais plutôt selon les standards de performance.

La généralisation de la performance observée à d'autres tâches demeure encore difficile à établir, pour ce qui est de l'évaluation des compétences basées sur des performances complexes puisqu'elle n'est effectuée qu'avec un nombre limité de tâches. Messick (1994) ainsi que Friedman et Mennin (1991) estiment cependant que l'utilisation des critères génériques pour décrire des tâches complexes pourrait permettre de généraliser la performance à une tâche effectuée dans un contexte à une tâche réalisée dans d'autres contextes. La répétition des tâches à accomplir dans le temps permet de plus de dégager un profil de performance de l'étudiant et ainsi d'inférer sur ses capacités dans d'autres situations non évaluées (Messick, 1994 ; Wiggins, 1998). Comme les données de recherche manquent actuellement pour soutenir cette hypothèse, l'échantillonnage des tâches à accomplir reste encore une stratégie recommandée pour des examens à enjeu élevé (Messick, 1994).

La fidélité des appréciations lors de l'évaluation des performances complexes reste quant à elle une question non résolue, mais certaines études suggèrent d'utiliser deux personnes pour établir la fiabilité « inter-juge » ou d'utiliser la consistance interne pour s'assurer de l'homogénéité des items (Petrusa, 2002).

Concernant le contrôle des biais, Friedman et Mennin (1991) ainsi que Fleming (1999) ont identifié plusieurs types d'erreurs dont deux importantes : les erreurs aléatoires et les erreurs systématiques. Ces auteurs proposent différentes stratégies pour minimiser ces deux types d'erreurs. Les erreurs aléatoires pourraient être contrôlées par le recours à l'échantillonnage des situations d'évaluation et les erreurs systématiques pourraient être contrôlées en tenant compte du jugement d'experts du domaine lors de l'analyse des résultats et de la prise de décision.

Découlant de ces concepts théoriques, nous cherchons à créer une grille d'évaluation qui permettrait d'apprécier la qualité des performances afin de pouvoir inférer sur les compétences cliniques. Dans une perspective d'évaluation formative, nous cherchons principalement à s'assurer de la validité du contenu de la grille, celle-ci représentant les standards de performance qui devraient être transmis aux étudiants. La conceptualisation et l'élaboration de cette grille feront aussi appel aux jugements des experts de différents domaines cliniques pour diminuer des erreurs systématiques.

Méthode

Étapes de développement d'une grille d'évaluation des performances cliniques

Nous inspirant de travaux de Streiner et Norman (2003) et de Van der Maren (1999), nous avons créé une démarche de validation qui se résume en trois étapes : la décision quant au contenu de la grille, la validation du contenu et la mise à l'essai de la grille. Ces opérations sont discutées dans les paragraphes suivants.

Première étape : décision quant au contenu et au format de la grille

Sur le plan du contenu, une recension des écrits permet de déterminer les besoins pédagogiques, de vérifier ce qui existe déjà et d'en apprécier les qualités afin d'en tirer des bénéfices.

Quant au format de la grille, les réflexions à propos de l'évaluation des compétences dans un contexte formatif et les caractéristiques inhérentes aux échelles de mesure nous ont amenés à favoriser une échelle d'évaluation descriptive globale pour apprécier la qualité de la performance tout en tentant de nous assurer de la fidélité des données. À la suggestion de Messick (1994), nous avons utilisé des critères et des standards de performances génériques

pour favoriser une éventuelle généralisation des résultats. Ces critères reflètent les tâches professionnelles et les standards de performances tiennent compte des compétences devant être développées à la fin de la formation.

Deuxième étape : établir la validité du contenu de la grille

Plus spécifiquement, la seconde phase vise à établir des standards de performance. Le processus principal visé était d'identifier le degré de performance jugé adéquat pour être reconnu compétent. Plusieurs processus pour établir des standards et des scores de césure ont été décrits dans le passé. Des méthodes les plus connues ayant besoin des jugements critiques, notons celles de Angoff, de Nedelsky et de Ebel (Erwin & Wise, 2001). Ces méthodes ont été principalement utilisées pour établir les seuils de passage d'examens contenant des items à réponses choisie. D'autres méthodes ont aussi été proposées pour évaluer des performances. Il s'agit des méthodes combinées où un jugement critique sur des données et un regroupement empirique des groupes d'étudiants aux niveaux opposés de performances ont été utilisées. Ce sont les méthodes soit avec un « groupe limite », soit avec des « groupes contrastes ». Ces deux méthodes reposent de façon importante sur le jugement des évaluateurs pour classifier les étudiants (Erwin, 2001).

Une méthode émergente proposée par Lewis et ses collaborateurs (1998) est la méthode du signet (« *bookmark standard setting* »). Elle a été proposée pour déterminer les standards de réussite ou d'échec en précisant le rang assigné à chaque item (du niveau le plus facile au niveau le plus difficile) en fonction des résultats de performance antérieurs. Selon ce modèle, les « juges » discutent en premier lieu, en petits groupes, de chaque item et de ce qui fait que l'item en question est plus difficile qu'un autre. Après cette première discussion, les juges cotent ces items de façon indépendante et il peut en découler un réarrangement des items. Au deuxième tour, les juges prennent connaissance de la cotation des autres et discutent de nouveau des items où il y a divergence d'opinion. Après cette discussion, une deuxième cotation est effectuée pour arriver à un consensus. Au troisième tour, en grand groupe, chaque juge prend connaissance de la médiane des cotations. De plus, si cette médiane est adoptée comme niveau de note de passage, l'impact sur le nombre d'étudiants qui échoueraient sera estimé. Faisant suite à cette information, une troisième cotation aura lieu et la médiane des cotes attribuées sera utilisée comme le seuil de passage recommandé pour chaque item. Toutes ces méthodes requièrent des données de performances antérieures qui proviennent de mises à l'essai sous forme de projet pilote pour établir des standards.

Comme nous désirons élaborer une nouvelle grille d'évaluation, les professeurs n'ont pas de données antérieures. La méthode d'EFPC est de plus créée dans un but d'évaluation formative, de sorte que nous cherchons à établir les standards considérés comme des niveaux de performance adéquats et inadéquats plutôt que d'établir un score de césure. Les trois étapes de la méthode de Lewis ont donc servi de source d'inspiration pour établir la validité du contenu des standards de performance.

Déroulement du processus d'établissement de standards de performances

Au cours de cette phase, vingt professeurs cliniciens ont été sollicités pour agir comme juges lors du processus d'établissement des standards. Ils furent recommandés par des membres du comité de programme de médecine interne à l'Université de Montréal. Les critères de choix des juges étaient basés sur les recommandations d'Erwin (2001) selon lesquelles les professeurs agissant comme juges devraient avoir des connaissances pertinentes du domaine évalué, de la population étudiante visée, et des méthodes d'évaluation à visée formative. De plus, engager les professeurs directement responsables de l'enseignement des étudiants faciliterait une meilleure collaboration pour l'utilisation future de cette grille d'évaluation. Douze de ces vingt professeurs ont accepté l'invitation. Ces médecins pratiquent des spécialités médicales diversifiées (neurologie, cardiologie, médecine interne, etc.) et ont au moins cinq ans d'expérience d'enseignement clinique. Ils sont aussi familiarisés avec le programme de formation postdoctorale en médecine interne.

Deux enregistrements audiovisuels d'entrevues entre des résidents[1] et des patients facilitent le processus. Les professeurs sont invités à utiliser la grille, dont une partie est illustrée à titre d'exemple à l'annexe 1 pour coter les performances du résident enregistrées sur vidéocassettes. Le but de la session est d'établir un consensus sur les critères et les niveaux de performance attendus d'un résident. Le visionnement des enregistrements a eu lieu pendant quatre heures et le processus pour établir des standards s'inspire des trois étapes du modèle de Lewis (1998).

En premier lieu, ces douze professeurs ont été divisés en trois groupes de quatre. Ils ont visionné un premier enregistrement d'entrevue, et ont coté de façon indépendante, à l'aide de la grille fournie, la performance du résident. Ils ont par la suite comparé leur cotation pour chaque critère et la discussion a eu lieu sur les divergences d'opinion. Un premier consensus du petit groupe a eu lieu sur la cotation de performance du résident. En deuxième lieu, chaque

membre du même petit groupe a visionné une deuxième entrevue et a coté de nouveau la performance du résident de façon indépendante et une deuxième discussion a eu lieu car il y avait encore des divergences d'opinion. Par la suite, un deuxième consensus a été atteint sur la cotation donnée pour chaque critère. Cette deuxième étape sert principalement à expliciter et à renforcer les attentes pour chaque niveau de performance. Au cours de la troisième étape, au moment où les douze professeurs sont réunis en grand groupe, un porte-parole de chaque sous-groupe rapporte par la suite la cotation donnée pour chaque critère et une discussion complète à nouveau le processus.

Troisième étape : la mise à l'essai de la grille

Cette étape vise surtout à s'assurer de la clarté des items lors de son utilisation. C'est alors que des expressions ambiguës, des questions à double sens, des jargons, des mots incompréhensibles et équivoques seront clarifiés ou éliminés. Crocker et Algina (1986) ajoutent que même si cette étape est souvent informelle, elle permet l'amélioration par une révision qui peut être parfois extensive.

Dix professeurs, recommandés par le comité de programme, ont été invités à participer au projet pilote pour valider cette dernière grille. Les critères sont leur dévouement et leur expérience dans l'enseignement clinique. Cinq ont accepté l'invitation. Les étudiants ont été recrutés parmi les soixante résidents du programme de médecine interne à l'Université de Montréal. Le processus de recrutement s'est effectué par lettre ouverte. Treize ont accepté de participer.

Nous avons envoyé la grille résultant de la deuxième phase aux participants pour vérifier auprès d'eux la clarté des énoncés. Par la suite, pendant un mois au début de l'année 2005, nous avons mené une étude pilote où les mêmes utilisateurs (cinq professeurs et treize résidents) ont expérimenté la méthode. Les professeurs et les étudiants sont aléatoirement assignés l'un à l'autre et, une fois jumelés, le professeur et l'étudiant fixent à leur convenance la date et l'heure de la séance d'EFPC. Le professeur prévoit un patient qu'il connaît et qui accepte volontairement de participer au projet d'étude.

Avant la séance, chaque professeur et chaque étudiant reçoit une copie de la description détaillée de la méthode (son but et son déroulement) et une copie de la grille d'évaluation. Les professeurs ont aussi reçu un texte se basant sur Brinko (1993) expliquant comment faire une rétroaction constructive. Les patients sont choisis par les professeurs et un seul critère est retenu : que le patient ait un problème médical véritable courant et stabilisé (afin d'éviter une détérioration clinique au cours de l'observation directe).

Résultats

Première étape : décision quant au contenu et au format de la grille

En ce qui concerne l'analyse des tâches professionnelles d'un médecin lors de sa première consultation médicale avec un patient, la recension des écrits nous a permis de relever un corpus de publications aussi abondant que pertinent. L'article de Frank, Jabbour, Tugwell et al. (1996), qui relate une consultation élargie menée auprès de multiples associations de spécialités médicales et chirurgicales, illustre le propos en explicitant des critères d'évaluation. Jourriles, Burdick et Hobgood (2002) ainsi que Petrusa (2002) identifient les huit composantes d'une performance clinique. Il s'agit du questionnaire du patient, des techniques d'entrevue, de l'examen physique, de la communication de l'information médicale, de la présentation de cas aux collègues, de la proposition d'un diagnostic différentiel et d'un plan thérapeutique d'investigation pour résoudre le problème du patient.

En ce qui concerne le format, la recension des écrits favorise l'utilisation d'une échelle globale. Cette échelle d'appréciation, si nous demeurons dans la logique de l'évaluation des compétences, est une échelle ordinale avec des standards descriptifs qui devraient permettre d'apprécier la qualité de la performance, du niveau le plus faible (novice) jusqu'au niveau exceptionnel (compétence expert). Pour déterminer le nombre de catégories de niveaux de performance, une révision de la documentation sur les méthodes d'évaluation de performances cliniques et sur les grilles descriptives nous a permis de constater qu'il n'y a aucun consensus. Dans des publications récentes par contre, Norcini et al. (2003) rapportent l'utilisation d'une échelle à neuf niveaux de type Likert dans un contexte de mini-exercice clinique (mini-CEX) avec observation directe au cours des sessions de performance clinique de courte durée à des fins sommatives. Ces auteurs ont obtenu des indices de fidélité jusqu'à 0,81 et de consistance interne jusqu'à 0,86 après quatorze sessions.

L'intégration des contenus ainsi que du format a permis de mettre en place une grille d'appréciation composée de huit items et de neuf niveaux de performance (*cf.* l'annexe 1 qui fournit une illustration de la grille) où le caractère « général » de l'échelle laisse beaucoup de place au jugement individuel des professeurs.

Deuxième étape : établir la validité du contenu de la grille auprès des experts

On se souvient que la seconde phase consiste à établir les standards de performances en faisant appel à douze professeurs-cliniciens. Ils ont eu à discuter et à déterminer, en sous-groupes puis en grand groupe, après le visionnement de deux enregistrements d'audio-visuel d'entrevue clinique, les niveaux de performance acceptable et inacceptable.

De ces discussions, les éléments suivants ont été dégagés :

1. Tous s'entendent que les sessions d'observation directe ne doivent pas dépasser une heure et que l'accent devrait être mis sur le problème clinique principal du patient (ex. : si un patient se plaint surtout de mal de tête, une douleur abdominale vague ne sera pas abordée) à moins que le résident ne juge que c'est pertinent. Il lui reviendra par la suite d'expliquer cette pertinence. Le problème en question doit être vague et mal défini, mais réaliste et représenter des problèmes courants en pratique médicale. Il est jugé préférable d'utiliser un vrai patient.
2. La tâche du résident doit englober toutes les démarches cliniques pour arriver à une résolution plausible du problème du patient. Les huit items sont donc retenus. De plus, trois professeurs ont tenu à expliciter le questionnaire pour tenir compte des impacts de la maladie sur la vie des patients. Cet item a été ajouté comme neuvième item.
3. Même si tous les professeurs sont d'accord pour utiliser une échelle d'appréciation globale de type Likert, tous se sont rendu compte, lors de la discussion, qu'avec une grille d'appréciation globale complètement ouverte, sans description des critères, les attentes de chaque professeur sont très variables. Ils se sont mis d'accord pour une meilleure explicitation des attentes à chaque niveau de performance.
4. Ils ont aussi reconnu qu'il leur est difficile de distinguer les neuf niveaux de performance. Ils suggèrent de retenir trois niveaux : excellent, adéquat et non acceptable. Comme il s'agit d'une activité d'évaluation formative, ils ne voient pas l'utilité d'une échelle à plusieurs niveaux.

Cette deuxième phase a donc donné lieu à l'élaboration d'une grille, où nous retrouvons les neuf items et les trois niveaux de performance avec la description de chaque niveau. Les commentaires généraux et les suggestions pédagogiques d'amélioration sont aussi inclus.

Troisième étape : la mise à l'essai

Quatre des cinq professeurs ont terminé les séances qui leur ont été assignées et neuf résidents ont eu des séances d'EFPC.

Après un mois d'utilisation de la grille modifiée, les professeurs et les étudiants ont été rencontrés dans le cadre d'une entrevue individuelle et semi-structurée afin de recueillir leur perception de la grille d'EFPC ainsi que de ses modalités d'utilisation. Les questions portaient sur la compréhension des niveaux de performance, les critères et la perception de l'utilité de la grille pour soutenir une rétroaction. Les données recueillies ont été consignées et analysées. Cette analyse a permis de regrouper les commentaires en deux grandes familles : ceux relatifs aux aspects positifs de la grille et de la méthode d'EFPC et ceux concernant les aspects plus négatifs.

Les aspects positifs sont regroupés en trois catégories :

- Sur les critères : tous les professeurs se reconnaissent dans cette démarche et sont d'accord qu'elle comprend les neuf étapes de base d'une première entrevue médicale avec un patient. Ainsi, aucune modification n'a été suggérée.
- Sur le format de l'EFPC et de la grille : les professeurs sont unanimes à dire que cette grille permet de centrer l'observation et facilite la rétroaction.
- Sur les standards de performance : les professeurs trouvent que l'échelle descriptive semble évoluer dans la bonne direction et qu'elle les aide en clarifiant mieux les attentes.

Les aspects négatifs suivants concernent surtout l'échelle de standards de performance et l'autoévaluation :

- De l'échelle des standards de performance : Les professeurs et les étudiants ont été aussi unanimes à juger négativement le nombre de niveaux de performance. Ils estiment qu'une grille à trois niveaux de performance est trop restrictive. Ainsi, les professeurs ont rapporté que « […] certains étudiants affichent une performance adéquate, mais il me semble qu'ils pourraient faire mieux en faisant une certaine manœuvre. »

  Les professeurs estiment que la distinction est parfois floue avec des chevauchements entre les tâches et les niveaux de performance. Ainsi, certains trouvent que « […] toutes les données sont rapportées, mais c'est trop long, comment faire pour souligner qu'à leur niveau, il faut être plus synthétique ? [...] »

Les quatre professeurs ont alors proposé un niveau additionnel intitulé : « adéquat, mais des améliorations sont nécessaires ».

De l'autoévaluation : Les étudiants appelés à s'évaluer eux-mêmes avec cette grille auraient aimé avoir des exemples pour mieux comprendre les niveaux de performance. De plus, remettre la grille sans explications préalables sème beaucoup de confusion.

Les étudiants insistent pour s'assurer du caractère formatif de l'activité. Ils trouvent que les critères aident à mieux comprendre les attentes mais ignorent si leur capacité à s'autoévaluer s'améliore véritablement.

Synthèse

Ces étapes de conception d'un prototype d'échelle de mesure, de validation du contenu auprès des experts puis de mise à l'essai auprès d'un groupe restreint d'usagers ont permis de créer une grille d'évaluation des performances complexes dans un contexte d'évaluation formative (annexe 2). Cette échelle, qui est constituée au départ de huit items et de neuf niveaux de performance, contient dans son format final neuf items et quatre niveaux de performance. De plus, ces niveaux de performance sont aussi devenus plus descriptifs qu'au départ. Le niveau excellent décrit le médecin compétent, le niveau adéquat est celui d'une performance minimale acceptable. Les deux autres niveaux représentent ceux de novices.

Forces et limites de l'étude

Cette étude cherche à créer une grille d'évaluation formative des performances complexes dans un domaine où très peu d'études existent. En suivant rigoureusement un processus de validation, nous avons pu constater que l'intégration des données rapportées et des opinions des professeurs ont amené une évolution dans le contenu et le format de la grille. Il est par contre important de souligner que notre étude se base sur l'opinion d'un groupe de professeurs cliniciens d'une même université, ce qui peut amener un biais de sélection. De plus, au cours de la triangulation du contenu, le nombre restreints de professeurs et d'étudiants peut rendre les résultats moins crédibles, même si les commentaires vont tous dans la même direction.

Discussion et conclusion

Les étapes de la conception, de l'établissement des standards de performance et de la validation du contenu ont permis de créer un modèle d'échelle descriptive globale pour évaluer des compétences cliniques. Les étapes de l'analyse des tâches professionnelles et la validation du contenu auprès des experts ont réussi à déterminer les neuf items qui représentent leur travail clinique habituel. La mise à l'essai par la suite confirme que les usagers s'identifient avec ses items et ne suggèrent plus de changements.

Le processus d'établissement des standards de performances par un groupe d'experts avec l'utilisation des enregistrements audiovisuels a permis une discussion approfondie et la prise de décision de réduire les niveaux de performance de neuf à trois, même si aucune donnée publiée ne soutienne cette décision. Ce même groupe a aussi insisté pour qu'une grille d'évaluation formative soit distincte de celle sommative. Par contre, c'est à la suite de la mise à l'essai que le nombre de niveaux désirés et perçus comme nécessaires pour une grille d'évaluation formative est de quatre quand elle a comme type d'échelle une échelle ordinale avec des descriptions pour chaque niveau de performance. De plus, l'utilisation d'une échelle descriptive semble clarifier les attentes et selon des usagers, aiderait à éviter des biais potentiels tels que l'effet de halo[2] et l'effet de la complaisance[3].

Par contre, lors de la mise à l'essai, nous n'avons pas suffisamment insisté sur l'autoévaluation ; ce qui a été source de confusion à la fois chez les professeurs, qui devraient s'assurer du processus, et chez les étudiants, qui devraient apprendre à s'autoévaluer. Ceci nous a permis de constater qu'une grille d'évaluation descriptive globale n'est en soi qu'un outil de mesure, et c'est l'usage qu'on en fait qui est déterminant.

En conclusion, nous retenons que dans le processus d'établissement des standards, une grille d'évaluation formative des performances complexes, adaptée de la méthode « onglet » de Lewis, semble avoir été nécessaire pour départager les niveaux de performances « adéquat » et « inadéquat ». De plus, cette grille d'évaluation des compétences cliniques à des fins formatives ne semble pas requérir un nombre élevé de niveaux de performance. Selon nos résultats, une grille avec quatre niveaux répondrait aux besoins de formation. Par ailleurs, ces catégories bien explicitées semblent aussi faciliter la tâche des professeurs pour observer directement et rétroagir. Il est tout aussi important de noter que les utilisateurs acceptent bien cette grille utilisée dans un contexte

précis d'évaluation formative. Par contre, comme les commentaires des étudiants le soulignent, cette grille n'a pas été conçue à des fins d'évaluation sommative et il faut mener d'autres études si l'on désire la transformer en un outil servant à cette fin.

Finalement, le recueil des données nécessaires pour évaluer des compétences cliniques semble être facilité par l'utilisation d'une grille d'évaluation descriptive globale intégrée à une méthode d'évaluation authentique. Les items sur cette grille reflètent des tâches professionnelles auxquelles des utilisateurs s'identifient facilement, tel que le préconise Messick (2004). De plus, des niveaux de performance qui traduisent les compétences à développer à des fins de formation semblent aussi être acceptables par les utilisateurs. Ces deux éléments contribuent à augmenter l'acceptabilité de cette grille d'évaluation. D'ailleurs, cette grille, en se basant sur des critères génériques pour faciliter éventuellement l'établissement de la comparabilité de performance du résident avec différents cas cliniques, semble aussi être utile à tous les professeurs, quel que soit leur domaine d'expérience, et aux étudiants à différents niveaux de formation. De plus, même avec différents professeurs, le résident peut toujours mieux percevoir ses forces et ses faiblesses pour cibler l'apprentissage afin de s'améliorer, puisque les mêmes critères ont été appliqués.

Notre recherche menée auprès d'un groupe restreint d'utilisateurs ne permet pas d'établir la fidélité des données avec un nombre limité de performances. Comme cette étude vise surtout à établir la validité de contenu de la grille, il est donc difficile de tirer une conclusion sur les autres aspects de validité. Par ailleurs, lors d'une utilisation à plus grande échelle, il serait intéressant de vérifier si une analyse des items apportait d'autres modifications. De plus, l'utilité de cette grille intégrée à une méthode d'évaluation de performance lors d'une utilisation répétée à une population estudiantine élargie reste aussi à vérifier dans d'autres études.

NOTES

1. Résident est l'appellation pour désigner l'étudiant de médecine au stade de la formation postdoctorale.
2. L'effet de halo fait ici référence au fait où un évaluateur est influencé par des facteurs tels que le rendement antérieur de l'étudiant ou son comportement général lors d'une situation d'évaluation indépendante.
3. L'effet de complaisance s'observe lorsqu'un évaluateur accorde une meilleure note par gentillesse.

RÉFÉRENCES

Allal, L. (1991). *Vers une pratique de l'évaluation formative ; Matériel de formation continue des professeurs*. Belgique : De Boeck & Larcier.

Andrich, D. (1997). Rating scale analysis. In J.P. Keeves (dir.), *Educational research, methodology and measurement: an international handbook* (2e éd., pp. 874-880). Oxford : Pergamon.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-14.

Bloom, B.S., Hasting, J.T., & Madaus, G.F. (1971). *Handbook on formative and summative evaluation of student learning*. New York : McGraw Hill.

Brinko, K.T. (1993). The practice of giving feedback to improve teaching : what is effective ? *Journal of Higher Education, 64*(5), 574-593.

Butler, R. (1988). Enhancing and undermining intrinsic motivation : the effects of task-involving and ego- involving evaluation on interest and performance. *British Journal of Educational Psychology, 58*, 1-14.

Carline, J.D., Paun, D.S., Thiede, K.W., & Ramsey, P.G. (1992). Factors affecting the reliability of ratings of students' clinical skills in medicine clerkship. *Journal of General Internal Medicine, 7*, 506-510.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. California : Wadsworth.

Erwin, T.D., & Wise, S. (2001). Standard setting. In R.A.Voorhees (éd.), Measuring what matters : competency-based learning models in Higher Education. *New directions for institutional research, 110*, 55-64.

Fleming, N.D. (1999). Biaises in marking of student's written work quality. In S. Brown & A. Glassner (dir.), *Assessment matters in Higher Education : choosing and using diverse approaches*. Londres : SRHE & Open University Press.

Frank, K. J.R., Jabbour, M., Tugwell, P., et al. (1996). Skills for the new millenium : report of the societal needs working group, CanMEDS 2000 Project. *Annals Royal College of physicians and surgeons of Canada, 29*, 206-216.

Friedman, M., & Mennin, S.P. (1991). Rethinking critical issues in performance assessment. *Academic Medicine, 66*, 390-395.

Gagné, R.M., Wager, W.W., Grolas, K.C., & Leller, J.M. (2005). *Principles of instructional design* (5e éd.). New York : Thomson Wadsworth.

Gipps, C., & Stobart G. (2003). Alternative assessment. In T. Kellaghan & D.L. Stufflebeam (dir.), *International handbook of educational evaluation*. Boston : Kluwer Academic Publishers.

Glasner, A. (1999). Innovations in student assessment : a system wide perspective. In S. Brown & A. Glasner (dir.), *Assessment matters in Higher Education : choosing and using diverse approache*. USA : SRHE & Open University Press.

Harlen, W., & James, M. (1997). Assessment and learning : differences and relationships between formative and summative assessment. *Assessment in Education, 4*(3), 365-379.

Hodder, R.V., Rivington, R.N., Calcutt, L.E., & Hart, I.R. (1989). The effectiveness of immediate feedback during the Objective Structured Clinical Examination. *Medical Education, 23*, 184-188.

Hodges, B., Regehr, G., McNaughton, N., Tiberius, R., & Hanson, M. (1999). OSCE checklists do not capture increasing levels of expertise. *Academic Medicine, 74*, 1129-1134.

Holm, H.A. (2002). Post-graduate Education. In G.R. Norman, C.P.M. Van der Vleuten & D. I. Newble (dir.), *International handbook of research in medical education* (pp. 381-413). Londres : Kluwer Academic Publisher.

Huba, M.E., & Freed, J.E. (2000). *Learner–centered assessment on college campuses, shifting the focus from teaching to learning*. Boston : Allyn & Bacon.

Jourriles, N., Burdick, N., & Hobgood, C. (2002). Clinical assessment in emergency medicine. *Academic Emergency Medicine, 9*(11), 1289-1294.

Kane, M. (1999). Validating measures of performance. *Educational measurement, issues & practice, 18*(2), 5-17.

Laurier, M.D., Toussignant, R., & Morissette, D. (2005). *Les principes de la mesure et de l'évaluation des apprentissages* (3e éd.). Québec : Gaétan Morin.

Le Boterf, G. (2000). *Construire les compétences individuelles et collectives*. Paris : Les Éditions d'Organisation.

Lewis, D.M., et coll. (1998). The Bookmark standards setting procedure : methodology and recent implementation. Cité par T.D. Erwin & S.L. Wise (2001), Standard setting. In R.A. Voorhees (dir.), Measuring what matters : competency-based learning models in higher education. *New direction for institutional research, 110*, 55-64.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Morgenstern, C., & Keeves, J.P. (1997). Descriptive scales. In J. . Keeves (dir.), *Educational research, methodology and measurement : an international handbook* (2e éd., pp. 900-908). United Kingdom : Elsevier Science Ltd.

Norcini, J.J., Blank, L.L., Duffy, S.,& Fortna, G. (2003). The mini-CEX : A method for assessing clinical skills. *Annals of Internal Medicine, 138*, 476-481.

Perrenoud, P. (1997). *Construire des compétences dès l'école*. Paris : ESF.

Petrusa, E.R. (2002). Clinical performance assessments. In G.R. Norman, C.P.M. Vander Vleuten & D.I. Newble (dir.), *International handbook of research in Medical Education* (pp. 673-710). United Kingdom : Kluwer Academic Publishers.

Regehr, G., MacRae, H., Reznick, R.K., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine, 73*, 993-999.

Roegiers, X. (2001). *Une pédagogie de l'intégration : compétences et intégration des acquis dans l'enseignement* (2e éd.). Bruxelles : De Boeck.

Saddler, R. (1998). Formative assessment : revisiting the territory, *Assessment in education, 5*(1), 77-84.

Scallon, G. (2000). *L'évaluation formative. Canada* : Éditions du Renouveau pédagogique.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*. Québec : Éditions du Renouveau pédagogique.

Streiner, D.L., & Norman, G.R. (2003). *Health measurement scales, a practical guide to their development and use* (3e éd.) Great Britain : Oxford Medical Publication.

Swanson, D.B., Norman ,G.R., & Linn, R.L. (1995). Performance-based assessment : lessons from the Health Professions. *Educational Researcher, 24*(5), 5-11.

Toohey, S., Ryan, G., & Hughes, C. (1996). Assessing the practicum. Assessment & Evaluation. *Higher Education, 21*(3), 215-227.

Turnbull, J., & Van Barneveld, C. (2002). Assessment of clinical performance, in-training evaluation. In G.R. Norman, C.P.M. Van der Vleuten & D.I. Newble (dir.*), International Handbook of Research in Medical Education* (pp. 793-810). Boston : Kluwer Academic Publishers.

Usatine, R.P., Edelstein, R.A., Yajima, A., Slavin, S.J., & Wilkes, M.S. (1997). Medical student perceptions of the accuracy of various new clinical evaluation methods. In A.J.J. Scherpbier, C.P.M. Van der Vleuten, J.J. Rethans & A.F. Van der Steeg (dir.), *Advances in medical education* (pp. 200-202). Boston : Kluwer Academic Publishers.

Van der Maren, J.M. (1999). *La recherche appliquée en pédagogie, des modèles pour l'enseignement*. Bruxelles : De Boeck Université.

Voorhees, R.A. (2001). Competency based learning models : a necessary future. In R.A.Voorhees (dir.), Measuring what matters : competency-based learning models in higher education. *New directions for institutional research, 110*, 5-13.

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership, 46*(7), 41-47.

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

ANNEXE 1

Illustration de la grille d'AFPC – version initiale

ÉVALUATEUR : _____ 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

• **Le questionnaire médical**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	Amélioration nécessaire			Questions pertinentes, concises, incluant des données psychosociales (valeurs personnelles, niveau socioéconomique ...)			Exceptionnel		
• **Relation thérapeutique**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	Amélioration nécessaire			A su établir une relation de confiance, de compréhension, empathique et respectueuse			Exceptionnel		

Commentaires

G. Impression globale et commentaires :

À améliorer :	**À maintenir :**

H. Durée de l'entrevue médicale : _____

ANNEXE 2

Illustration de la grille d'EFPC (4 niveaux)

Date : _____ Grille remplie par : _____

Résident : _____ Niveau : ☐ R-I ☐ R-II ☐ R-III

I. Recueillir des données cliniques avec le patient

	Inadéquat	Amélioration nécessaire	Adéquat	Excellent	Exemple, commentaires et suggestions
1. Questionner le patient sur le problème médical principal	☐ Les détails cliniques sont obtenus de façon disparate, sans fil conducteur apparent	☐ Les informations sont obtenues de façon concises, faisant preuve de connaissances physiopathologiques, mais, ☐ la chronologie des interventions antérieures n'a pas été obtenue	☐ Les informations sont obtenues de façon concise, et ☐ la chronologie des interventions antérieures est obtenue de façon complète	☐ Toutes les informations ont été obtenues de façon concise avec un fil conducteur logique faisant preuve de connaissance approfondie de physiopathologie logiques sous-jacentes	
2. Explorer des aspects psycho-sociaux de ce même problème mécical	☐ Aucune vérification des impacts de la maladie sur la vie du patient	☐ Des habitudes de vie pertinentes ont été obtenues (ex. : tabagisme en relation avec la maladie cardiaque) ☐ Peu a été demandé sur les impacts de la maladie sur la vie du patient	☐ Les habitudes de vie ainsi que les impacts de la maladie sur la vie du patient ont été adroitement obtenus	☐ toutes les informations sur l'impact de la maladie ainsi que la perception du patient vis-à-vis sa maladie ont été obtenues de façon adroite et respectueuse	

Commentaires généraux : Points à améliorer Points à maintenir

Suggestions pédagogiques :

Suivi pédagogique :

Ce qui a été fait depuis un mois :

Ce qui pourrait être fait au cours du prochain mois :