Pédagogie Médicale 2011; 12 (1): 17–27
DOI: 10.1051/pmed/2011019

Évaluation édumétrique d'un dispositif d'entrevues structurées multiples pour la sélection de candidats dans un programme postgradué de dermatologie

Edumetric assessment of multiple structured interviews for admission in postgraduate dermatology programs

Linda BERGERON[1], Christina ST-ONGE[2], Sandra MARTEL[3] et Dominique HANNA[4]

1 Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Canada
2 Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Canada
3 Centre hospitalier de l'Université de Sherbrooke, Sherbrooke, Canada
4 Centre hospitalier de l'Université de Sherbrooke et Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Canada

Manuscrit reçu le 31 août 2010 ; commentaires éditoriaux formulés aux auteurs le 13 février 2011 ; accepté pour publication le 2 mai 2011

Mots clés :
Habiletés non cognitives ;
admission ;
études postdoctorales ;
qualité édumétrique ;
MEM

Résumé – Contexte : De plus en plus d'importance est accordée aux habiletés non cognitives dans le rôle du médecin. Cette réalité se reflète dans les programmes de formation ainsi que lors de la sélection des candidats dans les programmes de médecine. Les programmes de dermatologie de trois universités québécoises ont conjointement élaboré un instrument de sélection pour évaluer les habiletés non cognitives de leurs candidats. Afin de limiter les effets dus aux juges, un devis d'évaluation avec plusieurs juges par station a été réalisé. **But** : Le but de l'étude est d'évaluer l'impact de ce devis d'évaluation sur les qualités édumétriques de l'instrument. **Sujets/matériel/méthodes** : Dix-huit participants ont été rencontrés lors du processus de sélection des candidats de dermatologie. L'instrument utilisé, soit des mini-entrevues structurées, était organisé en trois stations comportant chacune quatre questions évaluées par quatre juges. **Résultats** : L'instrument offre un bon accord interjuge (M = 0,79, ET = 0,11) et une fidélité acceptable (alpha de Cronbach = 0,70). La performance des candidats est corrélée plus fortement entre les questions d'une même station qu'entre les questions de stations différentes. **Conclusion** : Ces derniers résultats suggèrent la présence d'un effet de halo chez les juges lorsqu'ils évaluent un même candidat sur différentes questions. Ainsi, des observations indépendantes ou un entraînement des juges devraient être privilégiées pour pallier cette influence. Les efforts mis dans le cadre de l'élaboration de l'instrument ont permis néanmoins de contribuer à un processus de sélection fidèle.

Keywords:
Non-cognitive abilities;
admission;
post-graduate education;
psychometrics;
MMI

Abstract – Context: The role of physicians has expanded over the last decade and there is increased recognition of the importance of non-cognitive abilities. Programs and admission processes are changing to reflect that new reality. Three French-speaking Quebec universities joined forces to develop a tool used in the selection process for their dermatology programs. A multiple-rater evaluation design was proposed to reduce rater effect. **Purpose**: The goal of this study was to investigate the impact of the evaluation design on the psychometric qualities of the tool. **Subjects/Materials/Methods**: Eighteen candidates were met during the selection process. The mini-structured interview tool was made up of three stations, each consisting of four questions. **Results**: The inter-rater reliability was good (mean *ICC* = 0.79, *SD* = 0.11), and the reliability was acceptable (Cronbach's Alpha = 0.70). Candidates' performance between questions of a given station correlated significantly more than their performance between questions of different stations. **Conclusion**: The results suggest that raters may have been subject to a halo effect when assessing students on multiple questions. Therefore, independent observations or rater training should be favoured. The tool development efforts are believed to have contributed to the reliability of the selection process.

Introduction

Problématique

La multi-dimensionnalité du rôle des médecins est au centre des préoccupations et des développements de divers organismes régissant l'enseignement de la médecine. L'Association des facultés de médecine du Canada (AFMC)[1], par exemple, soutient l'importance, pour les médecins, de développer à la fois des habiletés cognitives (comme le raisonnement clinique, les connaissances biomédicales, le jugement clinique, etc.) et des habiletés non cognitives (telles que l'empathie, la communication, la compassion, etc.) tout au long de leurs carrières. Certains cadres de référence ont été élaborés pour refléter et promouvoir cette réalité, notamment le projet CanMEDS du Collège royal des médecins et chirurgiens du Canada (CRMCC)[2]. Le cadre de référence CanMEDS propose sept rôles du médecin (expert dans le domaine des connaissances médicales, communicateur, collaborateur, gestionnaire, professionnel, érudit et promoteur de la santé). Certains rôles reflètent davantage des habiletés non cognitives, qui sont en lien avec l'expertise médicale au sens commun du terme mais qui vont nettement au-delà.

Par ailleurs, on retrouve un écho à cette réalité dans les processus de sélection des candidats pour les différents programmes de médecine. De surcroît, dans un projet sur l'avenir de l'éducation médicale au Canada, l'Association des facultés de médecine du Canada[1] a émis diverses recommandations quant à la formation en médecine, l'une d'entre elles concernant spécifiquement la sélection des candidats en médecine : « Compte tenu du large éventail d'attitudes, de valeurs et d'habiletés requises des médecins, les facultés de médecine doivent améliorer les processus d'admission pour inclure l'évaluation des valeurs-clés et des caractéristiques personnelles des futurs médecins – telles que les habiletés en communication, en relations interpersonnelles et en collaboration et une gamme d'intérêts professionnels – ainsi que des aptitudes cognitives. » (p. 5)[1].

Dans ce contexte où les attitudes, valeurs et habiletés relationnelles sont recherchées chez les candidats aux différents programmes d'études médicales, un outil de sélection a été élaboré pour évaluer de façon fidèle et valide les habiletés et qualités non cognitives des candidats : les Mini-Entrevues Multiples (MEM – *Multiple Mini Interviews*). Un processus rigoureux et standardisé a été jugé nécessaire pour favoriser une sélection des candidats faite de façon juste et équitable. Le MEM permet en effet de recueillir un échantillon des comportements

des candidats de façon systématique pour chacun, tout en permettant une économie maximale des ressources (plusieurs candidats rencontrés en même temps). Cette façon de faire permet d'obtenir une mesure fidèle de ces habiletés et qualités non cognitives, tout en s'approchant d'un contexte authentique. Élaboré au début des années 2000, cet instrument a grandement contribué à l'amélioration des processus de sélection des programmes pré-gradués en médecine[3].

La création de ce nouvel outil pouvant offrir une bonne fidélité et validité a été encouragée par la difficulté d'obtenir un portrait fidèle et valide des habiletés et qualités non cognitives à l'aide des autres outils déjà existants. Bien que les instruments utilisés afin de mesurer les habiletés ou qualités non cognitives et recensés dans les écrits scientifiques soient nombreux et variés (lettres de recommandation, essais, notes autobiographiques, lettres d'intérêts, entrevues, etc.), il reste admis que la fidélité ou la validité de ces instruments sont habituellement faibles[4] quand il s'agit de mesurer les connaissances, habiletés et attitudes requises dans le cadre des performances tant sur les plans académiques que cliniques[4]. Malgré une validité apparente[3] et une acceptabilité[5] importante, les entrevues présentent une fidélité relativement faible, surtout lorsqu'elles sont peu structurées.

Le problème de fidélité lié aux entrevues rend difficile l'évaluation de sa validité. La fidélité est une condition à la validité[6]. En conséquence, il est impossible d'assurer qu'un instrument mesure les bonnes habiletés (validité) s'il ne génère pas avec constance le même résultat (fidélité). Dès lors que les entrevues présentent une bonne validité apparente et une forte acceptabilité, l'objectif était d'améliorer sa fidélité. Eva *et al.*[3] ont émis l'hypothèse que la faible fidélité d'une entrevue pouvait être non seulement liée aux juges, mais aussi à l'influence du contexte dans lequel l'individu est interviewé. Effectivement, l'entrevue est un contexte particulier auquel est liée la performance du candidat et n'offre ainsi qu'une seule fenêtre d'observation sur ses habiletés et qualités. Eva *et al.*[3] ont proposé de répondre à ce problème en donnant à leur instrument une forme apparentée à celle d'un examen clinique objectif structuré (ECOS). Cette forme d'évaluation permet de multiples mesures de la performance de l'étudiant dans plusieurs situations cliniques différentes. Eva *et al.* ont modelé cette façon de faire en multipliant le nombre d'entretiens avec les candidats afin d'obtenir plusieurs échantillons de leur performance et ce, dans différents contextes, pour ainsi augmenter la fidélité de la mesure des habiletés et qualités non cognitives.

Le dispositif de MEM élaboré par Eva *et al.*[3] a fait l'objet de maintes études qui ont démontré ses qualités édumétriques pour la sélection des candidats dans les programmes prégradués. Dans une première étude, Eva *et al.*[3] rapportent un coefficient de généralisabilité moyen de 0,65. Reiter et ses collaborateurs rapportent le résultat d'analyses de la fidélité de différentes versions du MEM utilisées dans d'autres écoles. Ils ont observé des coefficients de généralisabilité relative moyens de 0,75[7]. Des études documentent la validité prédictive de l'instrument[7,8]. Par exemple, Eva *et al.*[8] ont observé que la performance au MEM et le score de communication à l'examen d'aptitude – partie 2 – du Conseil médical du Canada (MCCQE PII) étaient corrélés à 0,65. Ce résultat est très prometteur si l'on considère que le score de communication au MCCQE PII prédit les habiletés relationnelles en pratique[9].

Les études portant sur le MEM ont principalement été effectuées dans le cadre de processus de sélection pour les programmes de formation prégraduée en médecine. L'admission dans des programmes postgradués en médecine suscite des enjeux différents. Lors de la sélection aux programmes prégradués, l'objectif est de recruter des candidats qui pourront développer certaines qualités spécifiques à la profession. Dans le cadre de la sélection aux programmes postgradués, l'objectif est d'identifier les candidats qui possèdent certaines qualités propres au programme donné. Néanmoins, des études ont montré que le MEM est un outil qui peut s'adapter à ce contexte différent[5,10,11]. Notamment, Hofmeister *et al.*[10], Bandiera et Regehr[5] et

Onyon *et al.*[11] ont utilisé le MEM ou une méthode similaire lors de leur processus de sélection. Bandiera et Regehr[5] ont utilisé un instrument composé de quatre entretiens permettant d'obtenir une mesure d'habiletés dans quatre domaines différents (caractéristiques personnelles, potentiel pour la médecine d'urgence, potentiel pour le programme de formation, potentiel à être formé). Onyon *et al.*[11] ont élaboré un instrument composé de trois stations dont chacune sollicitait une habileté différente dans un contexte différent, soit : station 1- se présenter – ; station 2- avoir une entrevue structurée – et station 3- être évalué sur sa capacité à communiquer dans un jeu de rôle. Hofmeister *et al.*[10] ont opté, quant à eux, pour un MEM traditionnel composé de 12 stations lors de la sélection des candidats internationaux postulant au programme de médecine de famille.

Les protocoles d'évaluation mis en œuvre dans les études susmentionnées reflètent certaines différences quant au protocole original et aux suggestions proposés par Eva *et al.*[3], particulièrement en ce qui concerne le nombre de juges évaluant chaque station. Bandiera et Regehr[5] ont réalisé un protocole dans lequel chaque entretien était évalué par deux juges. Dans le même sens, Onyon *et al.*[11] ont eu recours à deux juges par station. Hofmeister *et al.*[10] ont opté pour une formule plus traditionnelle, au cours de laquelle la majorité des stations étaient évaluées par un juge. Toutefois, trois stations étaient évaluées par deux juges, dont un qui était présent pour vérifier l'accord interjuge.

Quoique les approches de développement et de validation de ces instruments aient été variées, les résultats observés sont, dans l'ensemble, positifs. Bandiera et Regehr[5] ont observé une cohérence interne de 0,86 (coefficient alpha de Cronbach pour les huit scores : quatre entretiens à raison de deux scores par entretien [1 score par juge]) et de 0,83 (pour les quatre moyennes des deux scores de chaque entretien). Les analyses réalisées dans l'étude de Onyon *et al.*[11] portaient sur quatre ensembles différents (en fonction du niveau des candidats) de trois stations et ont montré que trois de ces quatre ensembles devraient comprendre davantage de stations pour satisfaire une fidélité correcte. Finalement, les résultats de l'étude de Hofmeister *et al.*[10] ont révélé un coefficient de généralisabilité de 0,70, ce qui est similaire à celui trouvé dans Reiter *et al.*[7].

Seuls Hofmeister *et al.*[10] ont évalué la validité de leur instrument en étudiant sa corrélation avec d'autres scores (par exemple ceux obtenus lors de l'administration du MCCQE PII ou lors de la réalisation d'un ECOS à l'entrevue) et en évaluant les différences entre les groupes étudiés (genre, langue, temps passé depuis la validation des études médicales prégraduées). Leurs résultats suggèrent que l'instrument ne comprend pas de biais quant aux différences entre les personnes et qu'il possède une certaine validité critériée.

À la lumière de certains de ces résultats[5], force est de constater que le recours à plusieurs juges peut donner des résultats plus fidèles que ceux observés généralement (coefficient de généralisabilité relatif moyen = 0,75[7]). Roberts *et al.*[12], en utilisant un MEM dans le processus d'admission au programme prégradué, avaient obtenu un coefficient de généralisabilité similaire à celui des autres études (0,70). Toutefois, des analyses secondaires faites à l'aide d'études d'optimisation ont suggéré que le nombre de stations requis pour un coefficient approprié (0,80) pourrait être moindre si deux juges étaient présents par station plutôt qu'un seul.

Contexte

Les universités de Sherbrooke, de Montréal et l'Université Laval se sont jointes pour élaborer un nouvel instrument de sélection, qui, s'inspirant du MEM ainsi que des principes de l'entrevue structurée de Pettersen et Durivage[13], pourrait améliorer le processus de sélection pour leurs programmes postgradués de dermatologie.

Tout comme Bandiera et Regher[5] ainsi que Onyon *et al.*[11] le soulignent, l'effet lié à l'évaluateur est un facteur pouvant influencer la fidélité dans un contexte d'évaluation subjective[14]. La plupart des outils utilisés dans la sélection de candidats

(les entrevues, les notes autobiographiques, les essais, les lettres de recommandation et d'intérêts) requièrent un jugement pour donner une valeur à ce qui est présenté. Or, le score d'une évaluation se compose de l'habileté de la personne et de l'erreur de mesure, comprenant, notamment, la subjectivité qui appartient au juge. Plusieurs biais peuvent altérer la perception des juges, tels que l'effet de halo, l'effet de récence ou de primauté et le biais de contraste. Ces biais contribuent à l'erreur de mesure associée aux juges[15] et réduisent ainsi le niveau de fidélité des instruments nécessitant une évaluation subjective.

Afin d'augmenter la précision de la mesure, en diminuant les biais dus aux évaluateurs, un devis d'évaluation avec plusieurs juges par station a été mis en place.

Objectifs de l'étude

Reconnaissant la nécessité de vérifier les qualités édumétriques de tout nouvel instrument utilisé dans un processus d'évaluation ou de sélection lorsqu'il a été significativement modifié[6], l'objectif principal de l'étude est d'évaluer la fidélité des entrevues structurées multiples utilisées pour l'admission aux programmes postgradués de dermatologie. L'objectif secondaire est de vérifier spécifiquement l'impact du changement apporté au protocoles originaux des MEM (à savoir, le recours à de multiples juges et à de multiples questions par station) sur la fidélité de l'instrument.

Matériels et méthodes

Instrument de mesure

L'élaboration de l'instrument a été réalisée suivant une démarche en cinq étapes, soit respectivement : l'établissement du profil de compétences, l'opérationnalisation des compétences relevées en habiletés, l'élaboration de l'échelle de notation et la détermination de l'importance relative des habiletés. Ces cinq étapes ont été développées dans le cadre d'un processus itératif.

Dans un premier temps, un profil de compétences a été établi. Pour ce faire, les membres du comité de sélection ont, par le biais d'un sondage et d'une activité de remue-méninges, distingué les comportements typiques des résidents respectivement les plus et les moins performants. Deux résidents et dix cliniciens enseignants en dermatologie composaient ce comité de sélection. Leur expérience de pratique en dermatologie variait entre un et 30 ans. Une analyse de contenu de l'information recueillie a permis de regrouper les comportements en thèmes principaux et de créer ainsi un profil de compétences[16]. Ce dernier se composait de 12 compétences : le jugement, l'autonomie d'apprentissage, l'autonomie clinique, la motivation, l'empathie, la tolérance au stress, le sens des responsabilités, l'attitude positive, la capacité d'adaptation, la capacité à travailler en équipe, l'autocritique et l'habileté à décrire.

Chacune de ces compétences a ensuite été opérationnalisée en ayant recours aux définitions tirées d'un dictionnaire de compétences[17]. Ces définitions ont été révisées par le comité de sélection et des modifications y ont été apportées afin que chaque membre en ait une même compréhension.

Par la suite, l'élaboration de questions comportementales et situationnelles associées à chacune de ces habiletés[13] a été réalisée grâce à un processus itératif de révision et d'amélioration par les membres du comité de sélection.

L'étape suivante a permis d'établir une échelle de notation. En se référant à chacune des 12 définitions du profil de compétences, le comité de sélection a rédigé des exemples de comportements souhaités et non souhaités. Le résultat de ce travail a été utilisé pour élaborer les échelles descriptives, chacune ayant été divisée en six niveaux afin d'éviter des évaluations de tendance centrale et de renforcer la discrimination des réponses des évaluateurs. Pour chacune des échelles, des exemples de comportements souhaités ont été ajoutés en tant qu'indicateurs pour le niveau 6 tandis que des exemples

1- Jugement	Questions	Indicateurs
Prendre des décisions pertinentes et adaptées en se basant sur l'analyse et l'expérience	**Comportement passé** : Décrivez une décision que vous avez dû prendre et qui s'est révélée complexe. Nous sommes intéressés à connaître votre démarche et ce qui vous a permis de prendre votre décision. Décrivez nous le contexte, la démarche, la décision prise et le résultat.	***1/6***
		– prend des décisions dangereuses *– fait des liens erronés* *– ne reconnaît pas ses limites* *– est inefficace et manque de réflexion* *– passe trop vite à l'action ou délibère trop longuement* *– n'analyse pas en profondeur la situation* *– ne cerne pas l'essentiel*
	Mise en situation : Vous en êtes à votre première journée de stage de dermatologie, en clinique externe avec votre patron. Il s'agit d'une journée chargée. Vous êtes appelé à l'unité néonatale pour évaluer une éruption sévère chez un nouveau-né. Le pédiatre demande l'opinion de la dermatologie et éventuellement une biopsie. Il veut une réponse le plus rapidement possible. Le patron de garde aux consultations est absent de l'hôpital. Que faites-vous ?	**6/6**
		– reconnaît ses limites *– possède une vue d'ensemble* *– sait dégager l'essentiel et priorise* *– tient compte des particularités du contexte dans lequel s'inscrit la décision* *– analyse en profondeur tout en s'en tenant à l'essentiel*

Fig. 1. Exemple d'une compétence opérationnalisée, des questions posées en conséquence et des indicateurs qui la composent.

de comportements non souhaités ont été ajoutés en tant qu'indicateurs pour le niveau 1 (un exemple est présenté à la figure 1).

Enfin, l'importance relative de chacune des compétences du profil a été déterminée par les membres du comité de sélection. Ceux-ci ont exprimé par vote l'importance qu'ils accordaient à chacune des compétences et, ensuite, une pondération a été calculée en fonction des résultats obtenus à ce vote.

Déroulement de l'évaluation

Une différence majeure entre l'outil décrit ici et le protocole original du MEM était la multiplication du nombre de juges par station. Ce choix reposait premièrement sur une volonté de diminuer l'influence due aux juges, telle qu'elle est décrite dans les écrits scientifiques [14,15]. De plus, ce devis d'évaluation a été élaboré pour respecter les ressources disponibles au sein des trois universités engagées dans le projet. Douze juges (cinq de l'Université Laval, quatre de l'Université de Montréal et trois de l'Université de Sherbrooke) pouvaient faire partie du processus de sélection et par convenance il a été établi qu'au minimum un représentant de chaque université soit présent à chaque station. Les entrevues structurées multiples de dermatologie évaluaient 12 habiletés à l'aide de 12 questions. L'instrument était composé de trois stations comprenant chacune quatre questions auxquelles il fallait répondre dans

Tableau I. Nature des habiletés non cognitives évaluées dans chaque station.

Stations	Habiletés
Station 1	Jugement
	Autonomie de l'apprentissage
	Autonomie clinique
	Motivation
Station 2	Empathie
	Tolérance au stress
	Sens des responsabilités
	Attitude positive
Station 3	Capacité d'adaptation
	Travail d'équipe
	Autocritique
	Habileté à décrire

un délai de 20 minutes. À chaque station, quatre juges étaient présents pour évaluer la performance du candidat. La composition des stations est présentée dans le tableau I.

Vingt-quatre candidats ont postulé au programme de résidence en dermatologie pour l'année 2009–2010. Après un processus de présélection basé sur les notes, les évaluations de stage, les lettres de recommandation, les lettres de motivation et les activités extracurriculaires, 18 étudiants ont été conviés aux entrevues structurées. De façon à pouvoir rencontrer les 18 candidats en une journée, un circuit de six itérations a été mis en place, c'est-à-dire que trois candidats étaient rencontrés à la fois, chacun dans une station et ce, pendant 20 minutes. À la fin de cette période de questions, les candidats changeaient de station. Un circuit (trois stations) était effectué en une heure. Au total, six circuits eurent lieu au cours de la journée.

Analyse des données

Il importe de préciser tout d'abord que la quatrième question de la station 3 n'a pas été prise en compte dans les analyses, car elle faisait l'objet d'une évaluation pilote et donc aucun score ne lui a été attribué. À chacune des 11 questions retenues dans le cadre des analyses, un score moyen entre les quatre juges a été calculé pour chaque candidat. Un maximum de 6 points était possible par question pour un maximum de 66 points pour l'ensemble des questions. Les données brutes ont été utilisées pour faire les analyses.

Deux types d'analyses ont été effectués pour vérifier la fidélité de l'instrument, soit des analyses d'accord inter-juge et une analyse de la cohérence interne. L'accord inter-juge a été vérifié à l'aide de corrélations intra-classes entre les quatre juges d'une même station. Le résultat représente la moyenne des paires de corrélations intra-classes. La cohérence interne de l'instrument a été établie à l'aide du coefficient alpha de Cronbach. Ces analyses ont été effectuées sur la moyenne des scores des juges pour chaque question. Finalement, une analyse de moyenne a été effectuée à l'aide d'un test t de Student pour comparer les coefficients de corrélations moyens intra-stations et inter-stations, afin de vérifier l'influence liée à l'inclusion, respectivement, de plusieurs juges et de plusieurs questions dans une même station.

Résultats

Les scores moyens par question sont présentés dans le tableau II. La moyenne des scores à une question est calculée à partir de tous les scores obtenus à cette question. Les scores moyens aux questions varient entre 3,47 et 4,89 pour un maximum de 6. Dans l'ensemble, l'accord inter-juge est assez élevé ($M = 0{,}79$, $ET = 0{,}11$) à l'exception des questions 10 (0,565) et 11 (0,654). Les analyses de fidélité par la méthode de cohérence interne ont permis d'observer un coefficient de fidélité (alpha de Cronbach) de 0,70. Les corrélations des scores moyens entre les questions d'une même station ($M = 0{,}54$; $ET = 0{,}19$) sont significativement plus élevées que les corrélations des scores moyens des questions entre les différentes stations ($M = 0{,}13$; $ET = 0{,}29$) ($t(53) = 4{,}96$, $p = 0{,}000$).

Discussion

Les entrevues multiples structurées ont été élaborées dans le cadre d'une collaboration interuniversitaire

Tableau II. Moyenne (*M*) et écart type (*ET*) des scores obtenus à chaque question et moyenne des corrélations inter-juges observée pour chaque question.

Questions	Statistiques descriptives *M*(*ET*)	Corrélation inter-juges
Station 1 – question 1	3,47 (1,29)	0,854
Station 1 – question 2	3,96 (1,27)	0,836
Station 1 – question 3	3,68 (1,07)	0,787
Station 1 – question 4	3,72 (1,03)	0,770
Station 2 – question 5	4,24 (0,99)	0,936
Station 2 – question 6	4,60 (0,71)	0,896
Station 2 – question 7	4,89 (0,71)	0,714
Station 2 – question 8	4,74 (0,73)	0,863
Station 3 – question 9	4,32 (0,85)	0,820
Station 3 – question 10	4,53 (0,52)	0,565
Station 3 – question 11	4,51 (0,72)	0,654

afin d'améliorer le processus de sélection des candidats aux programmes postgradués de dermatologie respectifs de chacune des trois universités. Compte tenu de l'implantation de ce nouvel outil de sélection, il s'imposait d'en vérifier les qualités édumétriques. En outre, un intérêt particulier a été accordé à l'appréciation des effets liés à l'augmentation du nombre de juges par station sur la qualité de l'instrument.

Le processus mis en place pour l'élaboration des entrevues structurées multiples en dermatologie apporte des garanties quant à la validité de contenu. En l'occurrence, des experts du domaine de la dermatologie ont défini les thèmes à évaluer et ils ont élaboré les questions de l'outil. Ce travail rigoureux, adossé sur l'expertise des personnes ressources du domaine, permet d'assurer de façon crédible que le contenu de l'outil mesure les compétences et habiletés recherchées.

Une modification majeure du devis d'évaluation consistait à inclure plusieurs juges par station afin de réduire l'erreur de mesure due aux juges. Toutefois, ce changement a occasionné une augmentation du nombre de questions par station. Pour vérifier l'impact de cette modification, une comparaison de moyennes, réalisée à l'aide d'un test-t de Student, a permis de calculer la différence entre les corrélations intra-stations et les corrélations inter-stations. Le résultat obtenu reflète l'impact de l'inclusion de plusieurs juges et plusieurs questions dans une station. Il s'est avéré que la performance des candidats était davantage corrélée d'une question à l'autre dans le cadre d'une même station qu'entre les questions de stations différentes, indiquant ainsi qu'un même juge avait tendance à évaluer, de façon très similaire, la performance d'un candidat aux différentes questions d'une station. Cette tendance à se faire une opinion globale du candidat se réfère au biais de l'évaluateur connu sous le nom de l'effet de halo[18]. Bandiera et Regehr[5] ont observé des résultats similaires, les juges ayant tendance à évaluer de la même façon un candidat sur les cinq échelles d'un même entretien. Bien qu'on ait voulu minimiser les effets dus à l'évaluateur en augmentant le nombre d'évaluateurs par station, le protocole d'évaluation utilisé dans la présente étude n'a pas permis d'éviter l'effet de halo. Il est possible que cet effet d'évaluateur motive d'autres auteurs, tels que Eva *et al.*[3], à augmenter le nombre de stations, plutôt que le nombre de juges par station, afin d'améliorer la fidélité d'un instrument. En effet, les évaluateurs tendent à ne pas discriminer de façon appropriée des éléments qui devraient être indépendants lors de l'évaluation de la performance d'un candidat[19].

Deux types d'analyses ont été effectués pour vérifier la fidélité de l'instrument, soit les analyses

d'accord inter-juges et l'analyse de la cohérence interne. Les analyses de corrélations inter-juges ont révélé un bon accord inter-juges pour la majorité des questions. Ces résultats s'expliquent, en partie, par le travail itératif d'opérationnalisation des concepts, qui a rendu explicite les habiletés et les qualités non cognitives mesurées par l'instrument. Cette étape importante du processus d'élaboration de l'instrument a aussi eu comme conséquence de permettre aux juges participant à ce processus de s'approprier les grilles d'évaluation, les questions et les mises en situation. Cette rigueur démontrée dans l'élaboration de l'instrument a certainement contribué au bon accord inter-juges.

Par ailleurs, il faut souligner que l'accord inter-juges est plus faible pour les questions portant sur le travail d'équipe (question # 10) et sur l'autocritique (question # 11), toutes deux à la même station. Ces résultats peuvent être le reflet de la difficulté à évaluer des qualités de savoir-être auquel est associée une grande subjectivité. En effet, bien que des dispositions aient été prises pour limiter l'interprétation de l'échelle de notation (par l'attribution de comportements concrets à l'échelle de notation, par exemple), il se pourrait que l'appréciation des réponses à ces deux questions n'ait pas été consolidée au même degré que celles aux autres questions. Par exemple, les problèmes d'évaluation de l'autocritique peuvent être liés à la difficulté connue qu'ont les gens à s'auto-évaluer ou s'auto-critiquer[20]. Il peut être ainsi plus difficile d'évaluer cette qualité chez une autre personne. Une recommandation serait donc d'approfondir la définition de ces deux qualités et d'objectiver de façon plus concrète la grille de notation.

Le deuxième type d'analyse pour vérifier la fidélité de l'instrument porte sur la cohérence interne, mesurée par le coefficient alpha de Cronbach. Le résultat obtenu (0,70) est acceptable quoique légèrement inférieur au seuil de fidélité fréquemment retenu de 0,80 dans les contextes d'évaluation à enjeux élevés[21], ainsi qu'au coefficient moyen de généralisabilité relative observé pour les MEM, soit de 0,75[7]. Le processus itératif de l'élaboration de l'instrument a dû contribuer à obtenir une certaine cohérence de la mesure. Toutefois, l'accord inter-juges plus faible observé pour deux questions, tel que mentionné plus haut, a pu contribuer à diminuer la cohérence interne de l'instrument.

L'instrument faisant l'objet de la présente étude comporte certaines forces qu'il convient de mentionner. Le protocole d'évaluation comprenant plusieurs juges dans une même station a permis à chacune des trois universités de considérer chacun des candidats et ce, sur une seule journée et dans un même lieu. Étant donné que la majorité des juges a contribué à l'élaboration de l'instrument, ce dernier a permis de répondre à leurs besoins et leurs perceptions quant aux qualités à évaluer chez les candidats. De plus, ce type d'outil est reconnu et accepté par la communauté[3,22]. L'obtention d'une cohérence interne acceptable et d'un bon accord inter-juges, combinée au respect des ressources et des besoins des programmes de dermatologie et à l'acceptabilité de l'instrument sont autant d'éléments concourant à une évaluation globalement positive de l'instrument[21]. De futures recherches pourront vérifier certaines preuves de validité, en l'occurrence la validité prédictive de l'instrument, c'est-à-dire la capacité du score obtenu aux entrevues structurées multiples à prédire la performance en stage ainsi que la performance aux examens de certification.

L'étude comporte aussi certaines limites. Il est possible que certaines qualités évaluées soient moins corrélées, par exemple l'empathie et la motivation vis-à-vis du jugement et de l'autocritique, ce qui peut réduire la cohérence interne de l'instrument. Le recours à d'autres modèles de mesure, tels que ceux que propose la théorie de la généralisabilité, qui auraient permis de relever les différentes sources d'erreurs de mesure ainsi que leur importance relative, a été grandement limité en raison de la petite taille de l'échantillon.

De plus, cette faible taille de l'échantillon ainsi que son homogénéité peuvent avoir aussi contribué à diminuer le coefficient de cohérence interne observé dans cette étude. Le contexte entourant la sélection des candidats au programme de formation

postgraduée en dermatologie est différent de celui des études effectuées sur le MEM, dont les échantillons étaient plus grands et provenaient de programmes variés. Le nombre possible de candidats est beaucoup moins élevé (se limitant au nombre d'élèves qui valident les programmes prégradués et aux candidats provenant d'autres pays) et ces candidats ont tous obtenu une formation similaire dans le cadre de la présente étude.

Conclusion

L'utilisation de l'instrument permettant de mesurer les habiletés et qualités non cognitives des candidats au programme spécialisé post-gradué de dermatologie répond à la réalité de la multi-dimensionnalité du rôle des médecins. Qui plus est, le recours à de multiples observations des caractéristiques non cognitives contribue à une mesure fidèle. Néanmoins, des observations indépendantes (c.-à-d., un juge évaluant une seule question) sont à privilégier étant donné le risque potentiel que les juges soient influencés par la performance observée à une question lorsqu'ils évaluent la performance à une seconde question. Une autre façon de réduire l'influence des évaluateurs pourrait être de former les évaluateurs à réfléchir sur ce phénomène et d'insister sur l'importance d'appuyer leur évaluation sur les descripteurs leur étant fournis. Les résultats de cette étude suggèrent que les efforts accomplis dans le cadre de cette collaboration interuniversitaire ont porté fruit : la mise en commun des ressources a rendu possible la multiplication des observations et la réalisation d'un processus de sélection plus standardisé et rigoureux.

Contributions

Linda Bergeron et Christina St-Onge ont contribué à l'interprétation des résultats, à l'analyse statistique et à l'écriture du manuscrit. Sandra Martel a contribué à la conception du protocole de recherche, à l'interprétation des résultats et à la révision du manuscrit. Dominique Hanna a participé à la conception du protocole de recherche, au recueil des données, à l'interprétation des résultats et à la révision du manuscrit.

Remerciements. Nous tenons à remercier les Drs. Isabelle Auger, Marie-Michèle Blouin, Benoît Côté, Thérèse El-Hefou, Dominique Friedmann, Martin Gilbert, Louise Loranger, Bruno Maynard, Khue Nguyen, Élizabeth O'Brien, Mélissa Saber et Marie-Marthe Thibeault pour l'élaboration de l'instrument de sélection ainsi que le Dr Daniel J. Côté pour son rôle de consultant dans le processus d'élaboration de l'instrument.

Aides financières ou subventions

Les analyses effectuées dans cette étude ainsi que la rédaction de cet article ont été réalisées grâce à un financement reçu de la Chaire de recherche en pédagogie médicale de la Société des médecins de l'Université de Sherbrooke.

Cette recherche a été approuvée par le Comité d'éthique de la recherche en éducation et sciences sociales de l'Université de Sherbrooke.

Les résultats de cette recherche ont été présentés à la Conférence canadienne sur l'éducation médicale à Toronto en mai 2011.

Références

1. AFMC. L'avenir de l'éducation médicale au Canada (AEMC) : Une vision collective pour les études médicales prédoctorales. Ottawa : AFMC ; 2010.

2. Frank JR. Le Cadre de compétences CanMEDS 2005 pour les médecins. L'excellence des normes, des médecins et des soins. 2005 [On-line]. Disponible sur: http://crmcc.medical.org/canmeds/CanMEDS2005/CanMEDS2005_f.pdf

3. Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the multiple mini-interview. Med Educ 2004;38:314-26.

4. Salvatori P. Reliability and validity of admissions tools used to select for the health professions. Adv Health Sci Educ 2001;6:159-75.

5. Bandiera G, Regehr G. Reliability of a structured interview scoring instrument for a Canadian postgraduate emergency medicine training program. Acad Emerg Med 2004;11:27-32.

6. AERA, APA, NCME (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education) Joint Committee on Standards for Educational and Psychological Testing. Standards for educational and psychological testing. Washington (DC): AERA; 1999.

7. Reiter HI, Eva KW, Rosenfeld J, Norman G. Multiple mini-interviews predict clerkship and licensing examination performance. Med Educ 2007;41:378-84.

8. Eva KW, Reiter HI, Rosenfeld J, & Norman GR. An update on the validity evidence pertaining to the Multiple Mini-Interview as a candidate selection strategy. Canadian Medical Education Conference; 3 au 7 mai 2008; Montréal, Canada.

9. Tamblyn R, Abrahamowicz M, Dauphinee D, Wenghofer E, Jacques A, Klass D *et al.* Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. JAMA 2007;298:993-1001.

10. Hofmeister M, Lockyer J, Crutcher R. The multiple mini-interview for selection of international medical graduates into family medicine residency education. Med Educ 2009;43:573-9.

11. Onyon C, Wall D, Goodyear HM. Reliability of multi-station interviews in selection of junior doctors for specialty training. Med Teach 2009;31:665-7.

12. Roberts C, Walton M, Rothnie I, Crossley J, Lyon P, Kumar K *et al.* Factors affecting the utility of multiple mini-interviews in selecting candidates for graduate-entry medical school. Med Educ 2008;42:396-404.

13. Pettersen N, Durivage A. L'entrevue structurée. Pour améliorer la sélection du personnel. Québec : Presses Universitaires du Québec; 2006.

14. Scullen SE, Mount MK, Goff M. Understanding the latent structure of job performance ratings. J Appl Psychol 2000;85:956-70.

15. Harasym PH, Woloschuk W, Mandin H, Brundin-Mather R. Reliability and validity of interviewers' judgments of medical school candidates. Acad Med 1996;71:S40-2.

16. Green CP. Get Talents! Interview for Action, Select for Results. Boston: SkilFast Publishing; 2007.

17. Lombardo MM, Eichinger RW. For Your Improvement: Guide d'encadrement et de développement destinés aux débutants, superviseurs, dirigeants, conseillers et observateurs. Minneapolis : Lominger Limited; 2002.

18. Nisbett RE, DeCamp Wilson T. The halo effect: Evidence for unconscious alteration of judgments. J Pers Soc Psychol 1977;35:250-6.

19. Feeley TH. Evidence of halo effects in student evaluations of communication instruction. Communication Education 2002;51:225-36.

20. Eva KW, Regehr G. Knowing when to look ip up: a new conception of self-assessment ability. Acad Med 2007;82:S81-4.

21. van der Vleuten, CPM. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ 1996;1:41-67.

22. Brownell K, Lockyer J, Lemay J. Introduction of the multiple mini interview into the admissions process at the University of Calgary: acceptability and feasibility. Med Teach 2007;29:394-6.

Correspondance et offprints : Christina St-Onge, 3001, 12e Avenue Nord, Sherbrooke (Québec), Canada, J1H 5N4.
Mailto : christina.st-onge@usherbrooke.ca.