Mesure et évaluation en éducation

Utilisation du Mini Entrevues Multiples en contexte francophone
Étude de généralisabilité

Christina St-Onge, Daniel J. Côté et Carlos Brailovsky

Volume 32, numéro 2, 2009

URI : https://id.erudit.org/iderudit/1024954ar
DOI : https://doi.org/10.7202/1024954ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada - Université Laval

ISSN

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

Découvrir la revue

Citer cet article

St-Onge, C., Côté, D. J. & Brailovsky, C. (2009). Utilisation du Mini Entrevues Multiples en contexte francophone : étude de généralisabilité. *Mesure et évaluation en éducation*, *32*(2), 49–69. https://doi.org/10.7202/1024954ar

Résumé de l'article

Le succès académique antérieur s'avère le meilleur prédicteur du succès académique en médecine, mais cet indicateur n'informe en rien quant aux habiletés non cognitives des candidats. Le Mini Entrevues Multiples (MEM), une mesure d'habiletés non cognitives, démontre une bonne fidélité et validité prédictive. L'objectif de la présente étude était de mesurer la fidélité du MEM UdeS élaboré et administré dans le cadre du processus de sélection au doctorat en médecine de l'Université de Sherbrooke. Les résultats observés démontrent la fidélité de l'outil, et ce, dans un contexte d'administration différent des études précédentes.

Utilisation du Mini Entrevues Multiples en contexte francophone : étude de généralisabilité

Christina St-Onge
Daniel J. Côté
Université de Sherbrooke

Carlos Brailovsky
Université Laval

MOTS CLÉS : Admission, évaluation, habiletés non cognitives, médecine

Le succès académique antérieur s'avère le meilleur prédicteur du succès académique en médecine, mais cet indicateur n'informe en rien quant aux habiletés non cognitives des candidats. Le Mini Entrevues Multiples (MEM), une mesure d'habiletés non cognitives, démontre une bonne fidélité et validité prédictive. L'objectif de la présente étude était de mesurer la fidélité du MEM UdeS élaboré et administré dans le cadre du processus de sélection au doctorat en médecine de l'Université de Sherbrooke. Les résultats observés démontrent la fidélité de l'outil, et ce, dans un contexte d'administration différent des études précédentes.

KEY WORDS: Admission, evaluation, non-cognitive abilities, medicine

Previous academic success is the best predictor for success in medical schools; however that indicator cannot be used to assess candidates' non cognitive abilities. The Multiple Mini Interview, a tool to assess non cognitive abilities, has demonstrated good reliability and predictive validity. The purpose of this study was to study the MMI's reliability when used to select students for the MD program at Université de Sherbrooke. The results suggest that the tool is reliable, even when used in a different context.

Palavras-chave: Admissão, avaliação, habilidades não-cognitivas, medicina

O sucesso académico anterior é o melhor preditor do sucesso académico em medicina, mas este indicador nada informa quanto às habilidades não-cognitivas dos candidatos. O Mini Entrevistas Múltiplas (MEM), uma medida das habilidades não-cognitivas, demonstra uma boa fidelidade e validade preditiva. O objectivo do presente estudo era medir a fidelidade do MEM no quadro de um processo de selecção de estudanters para o doutoramento em medicina da Université de Sherbrooke. Os resultados observados demonstram a fidelidade do instrumento mesmo quando usado num contexto de administração diferente dos estudos anteriores.

Note des auteurs – Toute correspondance peut être adressée comme suit : Christina St-Onge, Ph. D., Titulaire de la Chaire de recherche en pédagogie médicale de la Société des médecins de l'Université de Sherbrooke, Centre de pédagogie des sciences de la santé, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, 3001, 12e Avenue Nord, Sherbrooke, QC, J1R 5N4, téléphone : 819.821.8000, poste 75047, ou par courriel à l'adresse suivante : [Christina.St-Onge@USherbrooke.ca].

Introduction

Le taux d'attrition des programmes de médecine est très faible (Eva, Reiter, Rosenfeld & Norman, 2004a) et le taux de réussite aux examens de certification est très élevé ; ainsi, la probabilité de devenir un médecin est extrêmement élevée une fois admis dans un programme de médecine (Albanese, Snow, Skochelak, Huggett & Farrell, 2003). Les directeurs des processus de sélection des programmes de doctorat en médecine ont une responsabilité importante (Roberts et al., 2008) ; il est donc essentiel que les outils de sélection utilisés soient fidèles et valides.

Les facultés de médecine canadiennes font une première sélection des candidats en s'appuyant sur leurs résultats scolaires antérieurs. Cette prise en compte des habiletés cognitives se justifie par le fait qu'à plusieurs occasions le succès académique antérieur a été démontré comme le meilleur prédicteur du succès académique dans un programme de formation prédoctorale en médecine (Salvaroti, 2001). Les mesures des habiletés cognitives couramment utilisées au Canada sont le Grade Point Average (GPA), la cote R et le Medical College Admission Test (MCAT).

Norman (2004) suggère qu'il faut recruter des candidats qui ont non seulement de bonnes habiletés cognitives, mais qui possèdent aussi les habiletés et qualités non cognitives essentielles aux futurs professionnels de la santé. Par exemple, selon le Collège royal des médecins et chirurgiens du Canada (CRMCC), être un bon médecin exige d'être un expert dans le domaine des connaissances médicales, mais aussi d'être un communicateur, un collaborateur, un gestionnaire, un professionnel, et un promoteur de la santé. Cette vision du CRMCC a été mise de l'avant dans le cadre de référence CanMEDS (Frank, 2005). Une vision semblable du médecin est partagée par d'autres organismes, tels que le Collège des médecins de famille du Canada, le Collège des médecins du Québec ainsi que l'*American Association of Medical Colleges*.

Kulatunga-Moruzi et Norman (2002) rapportent que le GPA, une mesure des habiletés cognitives, est un pauvre prédicteur de l'habileté non cognitive, la communication, si cruciale dans l'exercice de la médecine. Qui plus est, plusieurs mesures des habiletés et qualités non cognitives utilisées jusqu'à maintenant se révèlent peu fidèles ou ayant une faible validité prédictive (Salvatori, 2001). Par exemple, Reiter, Eva, Rosenfeld et Norman (2007) ont observé de très faibles corrélations (– 0,14 à 0,26) entre des mesures

d'habiletés non cognitives administrées durant le processus d'admission (notes autobiographiques, entrevues traditionnelles et simulation de tutoriaux) et celles administrées durant le programme de formation (Examen clinique objectif structuré [ECOS], évaluation de stages et certaines parties de l'examen d'aptitude du Conseil médical du Canada [CMC] partie 1 : CLEO et PHELO).

Les résultats quant à la fidélité des notes autobiographiques, des lettres d'intérêts, des essais, des lettres de recommandation, ainsi que de certains tests d'aptitudes (comme le California Psychological Inventory, le Watson Glazer Critical Thinking Appraisal, le Otis Quick-Scoring Mental Abilities Test) sont aussi faibles (Salvatori, 2001). De plus, une fidélité faible ou modérée a été observée pour les entrevues de sélection (Kreiter, Yin, Solo & Brennan, 2004; Mann, 1979).

Au début des années 2000, une nouvelle mesure des habiletés et qualités non cognitives a été développée à la *Micheal deGroote Faculty of Health Sciences, McMaster University*. Il s'agit du *Multiple Mini Interviews* (MMI). Le Mini Entrevues Multiples (MEM) est une adaptation francophone du MMI. Cet outil a progressivement été adopté par 12 des 17 facultés de médecine du Canada et par près de 30 institutions universitaires du monde entier. Plusieurs recherches publiées ont démontré que le coefficient de généralisabilité relatif du MMI était de l'ordre de 0,75 (Reiter et al., 2007). De plus, les résultats présentés par *McMaster University* suggèrent que cet outil a une bonne validité prédictive (Eva, Reiter, Rosenfeld & Norman, 2008; Eva, Rosenfeld, Reiter & Norman, 2004b; Reiter et al., 2007). Par exemple, Eva et al. (2008) ont observé que la performance au MMI 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, quand on considère que le score de communication au MCCQE PII est un bon prédicteur des habiletés relationnelles en pratique (Tamblyn et al., 2007).

Convaincue de la valeur du MMI, la Faculté de médecine et des sciences de la santé de l'Université de Sherbrooke (UdeS) a introduit le MEM en avril 2008 dans le processus de sélection du programme de formation prédoctorale en médecine. Cette administration du MEM est, à la connaissance des auteurs, la première administration francophone du MMI. Les standards d'évaluation mis de l'avant par l'*American Psychological Association* (APA), l'*American Educational Research Association* (AERA) et le *National Council on Measurement and Evaluation* prescrivent que toute nouvelle évaluation ou

toute évaluation utilisée dans un nouveau contexte fasse l'objet d'études afin d'en assurer la fidélité et la validité (AERA, APA & NCME, 1999). Il y a donc lieu de vérifier la fidélité du MEM UdeS 2008 avant d'utiliser le MEM à plus grande échelle.

Le Mini Entrevues Multiples

Le MEM est une adaptation francophone du MMI et a pour but d'évaluer des habiletés et qualités non cognitives telles que la collaboration, la communication, l'empathie, l'intégrité et la maîtrise de soi. Le MEM s'apparente aux Examens cliniques objectifs et structurés (ECOS) (voir, *e.g.*, Grand'Maison, Brailovsky & Lescop, 1996; Grand'Maison, Lescop, Rainsberry & Brailovsky, 1992; Smee et al., 2003) fréquemment utilisés comme moyen d'évaluation des étudiants en cours et à la fin du cursus médical. Lors d'évaluations de ce type, qui sont de la catégorie *performance assessment,* les candidats doivent démontrer leurs habiletés dans le cadre de tâches complexes effectuées en situations dites authentiques. Ces tâches sont présentées dans des stations, incluses dans un parcours. Typiquement, l'évaluation de la performance a lieu dans le cadre d'une mise en situation (médicale pour les ECOS en médecine, non médicale pour le MEM) où le candidat interagit avec un évaluateur ou encore un acteur. Avant chaque station, le candidat consulte une vignette qui décrit la mise en situation. Par exemple, lors du MEM, les candidats peuvent être appelés à discuter avec un collègue (rôle joué par l'acteur) qui annonce qu'il a peur de prendre l'avion au moment du départ vers l'aéroport. Les candidats peuvent également être appelés à interagir avec un évaluateur qui leur demande de s'exprimer sur une question donnée. Voir l'annexe A pour un exemple de chaque type de station.

L'ECOS et le MEM s'appuient sur un principe fondamental de la mesure : une évaluation sera plus fidèle si elle repose sur plusieurs observations. Ainsi, le MEM comprend généralement de neuf à douze stations. Qui plus est, ce genre d'examen permet aux candidats d'avoir un nouveau « départ » à chaque station, leur permettant ainsi de se reprendre si leur performance était inférieure dans une autre station.

Dans les examens de type ECOS, l'évaluation de la performance des candidats se fait à l'aide de listes de vérifications ou d'échelles descriptives. Dans le cas particulier du MEM, ces échelles sont typiquement de type Likert

et visent à évaluer une ou quelques habiletés telles que l'habileté non cognitive spécifiquement mesurée dans la station, les habiletés communicationnelles et la performance globale lors de la station.

Eva, Rosenfeld, Reiter et Norman (2004) ont présenté les premiers résultats quant à la fidélité du MMI. Dans un premier temps, leur MMI a été administré à des étudiants gradués de la *Micheal deGroote Faculty of Health Sciences, McMaster University*. Ce MMI était composé de six stations. Deux évaluateurs devaient juger la performance des candidats sur quatre items évalués à partir d'échelles de type Likert à 7 points. Les items mesuraient les habiletés de communication (*communication skills*), la force de l'argumentation (*strength of arguments raised*), la pertinence pour les sciences de la santé (*suitability for health sciences*), et la performance globale (*overall performance*). Le coefficient de généralisabilité relatif obtenu pour cette première étude pilote était de 0,81. Eva et ses collaborateurs ont observé des corrélations élevées entre les items d'une même station, ils ont donc décidé de retenir un seul item, soit celui de la performance globale. De plus, ils ont fait une étude d'optimisation dont les résultats ont permis de déterminer qu'un devis d'évaluation de 12 stations, avec un seul évaluateur par station, permettrait d'obtenir un coefficient de généralisabilité relatif de 0,85. Dans le même article, ces auteurs présentent les résultats d'une deuxième étude pilote, dans laquelle 117 des 396 candidats ayant pris part au processus d'admission 2002 à leur programme de formation prédoctorale en médecine ont accepté de participer à leur étude. Ce MMI était composé de 10 stations, avec un évaluateur par station qui notait la performance globale des candidats à partir d'une échelle de type Likert en 7 points. Le coefficient de généralisabilité relatif était de 0,65. Dans le cadre de ces deux études pilotes, les candidats ainsi que les évaluateurs ont démontré une attitude favorable envers l'utilisation de cet instrument.

Eva et al. (2004) en ont conclu que le MMI était un outil prometteur, car son utilisation permettait :

1. d'obtenir de multiples échantillons des habiletés des candidats,
2. de diminuer la probabilité de chance et de biais liés aux évaluateurs,
3. d'évaluer les candidats de manière standardisée quant aux vignettes et aux items,
4. de concevoir des stations et des items de manière à évaluer des habiletés non cognitives en accord avec les valeurs prônées par l'institution,

5. d'obtenir des évaluations indépendantes des candidats,
6. d'avoir recours à moins de ressources qu'une entrevue traditionnelle.

Depuis la publication de ce premier article, le nombre d'articles au sujet du MMI a explosé. On peut recenser, depuis 2004, 22 articles portant sur cet instrument de mesure. Toutefois, dans le cadre de la présente recension de la documentation, seuls les résultats des articles portant sur l'étude de la fidélité du MMI sont présentés.

Eva et al. (2004a) ont étudié la relation entre les caractéristiques des évaluateurs et leur évaluation lors d'un MMI. Ils avaient recruté 54 des 198 candidats invités à passer une entrevue dans le cadre du processus d'admission 2003 à leur programme de formation prédoctorale en médecine. Les évaluateurs de ce MMI, composé de neuf stations, étaient des professeurs en sciences de la santé, des résidents et des membres de la communauté. Ils devaient, à partir d'échelles de type Likert en 7 points, évaluer la performance des candidats sur quatre items. Eva et ses collaborateurs ont observé un coefficient de généralisabilité relatif de 0,78. Ils ont aussi observé qu'il y avait davantage de variance inter-évaluateurs chez les membres de la communauté que chez les professeurs et les résidents. Finalement, comme dans le cadre de l'étude précédente (Eva et al., 2004), ils ont observé des corrélations très élevées entre les items d'une station (la corrélation moyenne entre des paires d'items était de 0,96). Sur la base de ces résultats et des résultats d'une étude d'optimisation, les auteurs suggèrent d'utiliser un seul item par station (celui de la performance globale lors de la station), d'augmenter le nombre de stations et d'avoir un seul évaluateur par station.

Roberts et ses collaborateurs (2008) ont étudié l'effet que pouvaient avoir certains facteurs sur la fidélité du MMI. Pour ce faire, ils ont élaboré et administré un MMI de huit stations aux 485 candidats retenus par le consortium *Australian graduate-entry schools,* dans le cadre du processus de sélection à leurs programmes de formation prédoctorale en médecine. Il y avait un évaluateur par station ; toutefois, les évaluateurs étaient appelés à changer de station entre les itérations afin d'éviter un biais de fatigue. Roberts et ses collaborateurs ont observé un coefficient de généralisabilité relatif de 0,70. De plus, ils ont observé que la plus grande source de variance était associée aux évaluateurs. En conclusion, ils ont considéré que le MMI avait une fidélité modérée.

Le contexte d'administration du MEM UdeS 2008

Le contexte d'administration du MEM UdeS 2008 semble différer des contextes recensés, sur quatre éléments importants :

1. les candidats,
2. les évaluateurs,
3. la culture et les besoins du programme,
4. le devis même du MEM comparé à celui du MMI.

Dans les études recensées, les candidats soumis au MMI avaient généralement complété un baccalauréat ou étaient inscrits aux études graduées. Par contre, les candidats invités au MEM 2008 de l'UdeS avaient soit un baccalauréat universitaire, soit un minimum de 45 crédits universitaires en sus d'un diplôme du collégial. Ils n'avaient donc pas à avoir terminé un baccalauréat. On peut donc supposer une plus grande hétérogénéité entre les candidats.

La formation des évaluateurs MEM offerte par les organisateurs de l'UdeS était similaire à celle offerte aux évaluateurs de *McMaster University*. Toutefois, ces évaluateurs avaient l'habitude des entrevues traditionnelles pour sélectionner leurs étudiants en médecine, alors que la Faculté de médecine et des sciences de la santé de l'UdeS abandonnait ces dernières dans les années 1970. On peut donc postuler une plus grande hétérogénéité parmi les évaluateurs sherbrookois.

Le MEM UdeS 2008 est administré en français. La culture des facultés de médecine québécoises francophones, et plus particulièrement celle de l'UdeS, peut différer de celles des institutions anglophones ayant utilisé le MMI : par exemple, la valeur accordée à des expériences de bénévolat ou encore de recherche est nettement moindre à la Faculté de médecine et des sciences de la santé de l'UdeS. Ceci peut affecter le nombre de dimensions dans le tableau de spécification de l'instrument de mesure, le contenu des vignettes et des grilles de correction, et peut donc possiblement augmenter la variance de l'interaction candidats-stations.

Le MEM UdeS 2008 évaluait huit habiletés ou des qualités non cognitives (empathie, respect, intégrité, jugement, conscience sociale, tolérance à l'incertitude, motivation, collaboration) tandis qu'il n'y en avait que quatre dans celui de *McMaster University* (raisonnement critique [*critical thinking*], prise de décisions éthiques [*ethical decision making*], habiletés communicationnelles [*communication skills*], ainsi que la connaissance du système de la

santé [*knowledge of the health care system*]). Ces habiletés ou qualités non cognitives étaient évaluées par 10 stations (voir le tableau 1). Chaque station était composée de trois items de type Likert à 7 points mesurant :

1. l'habileté ou la qualité non cognitive spécifique,
2. la communication,
3. la performance globale,

comparativement à *McMaster University* qui n'utilise plus qu'un seul item par station (performance globale). Ces différences pourraient elles aussi influencer la variance inhérente à l'interaction candidats-stations, bref à ce qui est autrement appelé le phénomène de « spécificité de contexte ». Eva et al. (2004a) indiquent que la spécificité de contexte se réfère au caractère contextuel de la performance des candidats, indiquant que l'habileté des candidats peut varier d'un contexte à un autre. Une différence importance relative de la variance due à l'interaction entre la facette Candidat et la facette Station pourrait influencer la fidélité de l'instrument de mesure.

L'étude avait pour objectifs spécifiques de vérifier, en comparant les résultats observés dans le cadre de cette étude à ceux des études recensées, l'influence que le MEM UdeS 2008 a pu avoir sur :

1. la variance due aux candidats,
2. la variance due aux stations,
3. la variance due à la spécificité de contexte,
4. la fidélité de l'instrument de mesure.

Il a été impossible d'étudier la variance due aux évaluateurs étant donné le devis d'évaluation choisi par les organisateurs du MEM UdeS 2008.

Méthode

Participants

Les données des 86 étudiants ayant passé le MEM UdeS 2008 ont été utilisées dans le cadre de cette étude. L'attribution des candidats aux circuits et aux itérations a été déterminée de façon quasi aléatoire, soit en suivant l'ordre alphabétique. Les 86 candidats invités au MEM 2008 de l'UdeS avaient au moins un baccalauréat universitaire ou un minimum de 45 crédits

universitaires en sus d'un diplôme du collégial. Parmi les candidats invités, 57 % étaient des femmes. Les candidats invités avaient entre 19 et 29 ans, avec une moyenne de 22,5 ans et un écart-type de 2,5 ans.

Instrument de mesure

Le titre MEM est une traduction du titre MMI, toutefois, les stations MEM ne sont pas des traductions des stations MMI : elles sont des créations originales adaptées au contexte québécois. Le MEM UdeS 2008 était composé de 10 stations, de 10 minutes chacune. La table de spécification utilisée dans le cadre du MEM UdeS 2008 est présentée au tableau 1. Chaque station était composée de trois items de type Likert à 7 niveaux. Ces items mesuraient :

1. l'habileté ou la qualité non cognitive spécifiquement évaluée dans la station,
2. la communication,
3. la performance globale (*i.e.*, 10 mesures d'habiletés, 10 mesures de communication et 10 mesures de performance globale).

Tableau 1
Table de spécification

Habiletés	Stations
Collaboration	Station 1
Conscience sociale	Station 2
Empathie	Station 3
Jugement	Station 4, Station 9
Motivation	Station 5
Apprentissage	Station 6
Respect	Station 7
Intégrité	Station 8, Station 10

Devis d'évaluation

Pour administrer le MEM à 90 personnes dans une journée, trois circuits parallèles avaient été mis en place et il y avait trois itérations par circuit. Ce devis permettait l'évaluation de 90 étudiants en une journée ; toutefois, seulement 86 candidats se sont présentés au MEM UdeS 2008.

Un circuit est composé de 10 stations évaluées chacune par un juge ; le MEM UdeS 2008 ayant trois circuits, au total 30 juges ont participé à l'évaluation des candidats. Une itération représente une plage horaire où un groupe d'étudiants était évalué à l'intérieur d'un circuit. Les évaluateurs restaient dans la même station pour les trois itérations. La figure 1 illustre la logistique de cette évaluation. Pour éviter que les candidats ne se divulguent le contenu des stations, un confinement stratégique avait été mis en place. Le diagramme de Venn correspondant au devis d'évaluation est présenté à la figure 2.

	Circuit Jaune	**Circuit Rose**	**Circuit Bleu**
Itération 1 (10h00)	Station 1 – 10 Juges 1-10 Étudiants 1-10	Station 1 – 10 Juges 11-20 Étudiants 11-20	Station 1 – 10 Juges 21-30 Étudiants 21-30
Itération 2 (13h00)	Station 1 – 10 Juges 1-10 Étudiants 31-40	Station 1 – 10 Juges 11-20 Étudiants 41-50	Station 1 – 10 Juges 21-30 Étudiants 51-60
Itération 3 (15h00)	Station 1 – 10 Juges 1-10 Étudiants 61-70	Station 1 – 10 Juges 11-20 Étudiants 71-80	Station 1 – 10 Juges 21-30 Étudiants 81-90

Figure 1. ***Organisation du MEM UdeS 2008 pour accommoder 90 candidats***

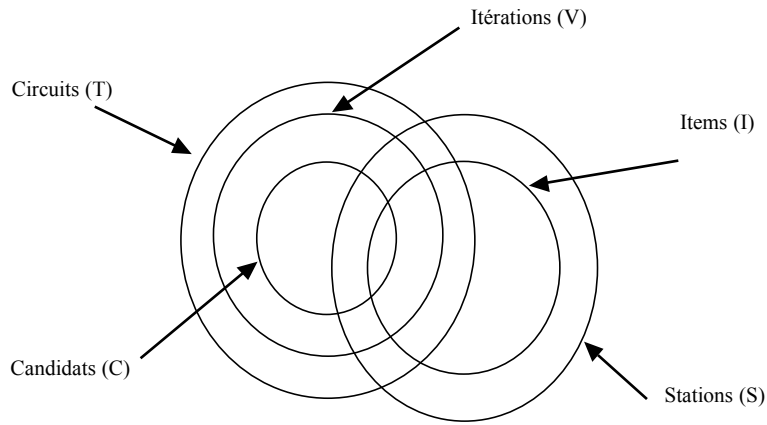

Figure 2. ***Diagramme de Venn illustrant le devis d'évaluation***

Analyses

Des analyses de généralisabilité ont été effectuées afin d'estimer la fidélité du MEM UdeS 2008, ainsi que pour déterminer l'importance relative de certaines facettes du devis d'évaluation. Pour ce faire, deux modèles de généralisabilité ont été analysés : le modèle Candidat × Station et le modèle Candidat × (Item : Station). Le logiciel urGenova (voir Brennan, 2001a, 2001b) a été utilisé pour estimer les composantes de variance. Ce logiciel a été utilisé car il permet l'analyse de devis non balancé. Les facettes d'instrumentations suivantes ont été retenues dans le cadre des analyses : les items et les stations.

Résultats

L'analyse descriptive du MEM UdeS 2008 est présentée au tableau 2. Les scores aux 10 stations variaient de 4 à 21, 21 étant le score maximum possible. Les moyennes des scores pour les 10 stations variaient de 11,52 à 15,52.

Tableau 2
Moyenne et écart-type des scores par station

Stations	*Moyenne*	*Écart-type*	*Minimum*	*Maximum*
Station 1	14,00	4,14	5	21
Station 2	12,60	4,19	5	21
Station 3	13,50	4,17	4	21
Station 4	13,79	4,56	4	21
Station 5	13,79	3,48	5	20
Station 6	11,52	4,42	4	21
Station 7	13,36	3,90	4	20
Station 8	13,38	3,38	7	21
Station 9	14,01	4,00	6	21
Station 10	15,52	2,92	7	21

Les résultats des analyses de généralisabilité sont présentés selon l'ordre croissant de complexité des modèles analysés. Des coefficients de généralisabilité relatifs ont été calculés, étant donné que les résultats du MEM étaient utilisés pour ordonner les candidats.

Un modèle à deux facettes a été analysé en premier lieu, soit le modèle Candidat × Station où Candidat est la facette de différentiation et Station est la facette d'instrumentation et représente la somme des trois items utilisés pour l'évaluer. Les résultats de cette analyse sont présentés au tableau 3. Il peut être observé que la facette de différentiation explique 21 % de la variance des scores MEM. La facette Station explique 6 % de la variance. L'interaction entre la facette Candidat et la facette Station est combinée à l'erreur résiduelle dans ce modèle ; ensemble elles expliquent presque les trois-quarts de la variance du score MEM, soit 73 %. Le coefficient de généralisabilité relatif pour ce modèle est de 0,745. Ainsi, il est possible de se fier à 75 % au classement des étudiants par les stations.

Tableau 3
Analyse de variance pour le modèle C × S

Facteurs	*d.l.*	*Carré moyen*	*Estimation de la composante de variance*[a]	*%*
Candidats (C)	85	47,23	3,52	21 %
Stations (S)	9	90,71	0,92	6 %
C × S *(+ l'erreur résiduelle)*	765	12,05	12,05	73 %

Coefficient de généralisabilité relative :

$$\sigma^2(candidats)/[\sigma^2(candidats) + (\sigma^2(candidats \times stations)/10)] = 0{,}745$$

Note a – Les erreurs standards ne sont pas présentées car il existe peu de documentation au sujet de l'estimation des erreurs standards de mesure des composantes de variance de devis non balancés (Brennan, 2001a, p. 388). De plus, les erreurs standards de mesure calculées par urGENOVA sont seulement appropriées pour les devis balancés (p. 4, version 24-08-2001 du logiciel urGENOVA – Brennan, 2001b).

Une facette d'instrumentation, soit les Items nichés dans les Stations, a été ajoutée au modèle 1 pour créer le modèle 2 : Candidat × (Item : Station). Les résultats, obtenus lors de l'analyse du deuxième modèle de généralisabilité, sont présentés au tableau 4. Le pourcentage de variance des scores MEM associé à la facette de différenciation est de 18 %. L'interaction entre la facette Candidat et la facette Station explique 52 % de la variance des scores MEM. La facette Station et l'erreur résiduelle expliquent, respectivement, 5 % et 26 % de la variance des scores MEM. Les items nichés dans les stations ne

contribuent pas à la variance des scores MEM. Le coefficient de généralisabilité relatif pour ce modèle est de 0,744. Ainsi, il est possible de se fier à 74% au classement des étudiants lorsque leurs scores représentent la somme des items.

Tableau 4
Analyse de variance pour le modèle C × (I : S)

Facteurs	*d.l.*	*Carré moyen*	*Estimation de la composante de variance*[1]	*%*
Candidats (C)	85	15,74	0,39	18%
Stations (S)	9	30,24	0,10	5%
Items : Stations (I:S)	20	1,77	0,01	0%
C × S	765	4,01	1,15	52%
C × I:S *(+ l'erreur résiduelle)*	1700	0.57	0.57	26%

Coefficient de généralisabilité relative :

$\sigma^2(candidats)/[\sigma^2(candidats) + (\sigma^2(candidats \times stations)/10) + (\sigma^2(candidats \times items:stations)/30)] = 0,744$

Discussion

L'objectif général de cette étude était d'évaluer la fidélité du MEM utilisé dans un nouveau contexte : le processus d'admission d'un programme québécois francophone de formation prédoctorale en médecine. Des analyses de généralisabilité ont été effectuées afin de répondre à cette question et déterminer l'importance relative qu'ont certaines facettes du devis d'évaluation.

La fidélité du MEM UdeS 2008

Les coefficients de généralisabilité relatifs obtenus, 0,744 et 0,745, se comparent à ceux observés dans la documentation (0,65 à 0,81 ; Eva et al., 2004a, 2004b ; Roberts et al., 2008). Nous en concluons que le MEM UdeS 2008 est un instrument de mesure fidèle. Néanmoins, il serait avantageux d'obtenir un coefficient de généralisabilité relatif plus élevé, étant donné les enjeux élevés associés au processus d'admission.

Variance due aux candidats

Il a été observé que la facette de différentiation Candidat expliquait de 18% à 21% de la variance selon le modèle analysé. Eva et al. (2004, 2004a), ainsi que Roberts et al. (2008), avaient observé que la facette de différentiation

expliquait, respectivement, 25%, 15% ou 22% de la variance des scores. Les résultats de la présente étude sont comparables à ceux recensés dans la documentation. Il semble donc que l'hypothétique hétérogénéité de notre échantillon de candidats a eu peu d'influence sur la variance due aux candidats.

Variance due aux stations

Dans les études recensées, la facette Station expliquait de 5% à 10% de la variance. Dans le cadre de la présente étude, cette même facette expliquait de 5% à 6% de la variance. Les résultats de cette étude sont donc comparables à ceux observés dans la documentation. Il est possible de supposer que les stations étaient de niveaux de difficulté similaires.

Variance due à la spécificité de contexte

La variance due à la spécificité de contexte pourrait être plus grande lors de l'administration d'un MMI ayant une plus grande multidimensionnalité, plus d'items dans chaque station ainsi qu'un échantillon de candidats hétérogènes. Il a été observé que l'interaction entre les facettes Candidat et Station expliquait 52% de la variance des scores MEM, tandis que Eva et al. (2004, 2004a) avaient observé que l'interaction Candidat et Station expliquait de 16 à 17% de la variance des scores MMI. Les résultats de la présente étude semblent indiquer que le phénomène de spécificité du contexte est prépondérant dans la performance des candidats au MEM UdeS 2008. Enfin, les résultats observés semblent supporter l'hypothèse que la variance due à la spécificité de contexte est plus grande lors de l'administration d'un MMI francophone.

Limites de l'étude

Les évaluateurs francophones peuvent différer des évaluateurs anglophones quant à leurs valeurs, attentes et expériences, et ces différences peuvent influencer l'évaluation qu'ils font des candidats et ainsi la variance due aux évaluateurs. Il aurait donc été important d'évaluer la part de variance que les évaluateurs contribuent aux scores MEM. Toutefois, le devis d'évaluation utilisé dans le cadre de l'administration du MEM UdeS 2008 n'a pas permis de quantifier la part de variance associée aux évaluateurs. En effet, les évaluateurs étant nichés dans les stations, il était impossible de dissocier la part de variance qu'ils pouvaient contribuer aux scores MEM. Les études sur ce sujet sont très rares : en fait, à notre connaissance, seulement deux équipes de chercheurs ont étudié, dans le cadre de MMI, la part de variance due aux évaluateurs : Eva et al. (2004a) ainsi que Roberts et al. (2008). Ces chercheurs

avaient observé que la variance due aux évaluateurs expliquait, respectivement, 25 % et 14 % des scores MMI. Ils avaient tout de même conclu, sur la base du coefficient de généralisabilité relatif, que leurs instruments étaient fidèles.

Cette étude a été effectuée uniquement auprès de candidats ayant un diplôme de l'ordre universitaire ou encore un diplôme de l'ordre collégial et au moins 45 crédits universitaires. Les résultats de notre étude ne peuvent donc être étendus d'emblée aux autres candidats postulant à la Faculté de médecine et des sciences de la santé de l'UdeS, particulièrement à ceux n'ayant qu'un diplôme de l'ordre collégial.

Pistes de recherches futures

Des 86 candidats à qui le MEM UdeS 2008 a été administré, 33 ont été admis au programme de formation prédoctorale en médecine de l'Université de Sherbrooke. D'autres analyses, notamment des analyses de validité prédictive, seront effectuées afin de mieux évaluer l'efficacité du MEM. Le comité d'admission utilisera ces résultats, et ceux qui suivront, pour continuellement évaluer l'efficacité du MEM.

Une étude de fidélité devra être reproduite lorsque le MEM sera administré aux candidats du contingent collégial afin d'évaluer si cet instrument est pertinent pour cette population. Qui plus est, il faudra évaluer la part de variance expliquée par l'interaction entre les facettes Station et Candidats afin d'évaluer si la spécificité de contenu a un impact particulier chez les candidats du contingent collégial.

Finalement, étant donné le peu de connaissances quant à l'influence que peuvent avoir les évaluateurs sur la variabilité des scores MEM, il serait important d'étudier ce phénomène davantage.

Conclusion

Considérant l'ensemble des résultats obtenus dans le cadre de cette première administration du MEM dans un contexte québécois francophone, cet instrument de mesure s'avère fidèle ; ainsi, son utilisation peut être favorisée à plus grande échelle. La présente étude montre aussi l'importance d'analyser rigoureusement tout nouvel instrument de mesure, ou tout instrument de mesure utilisé dans un nouveau contexte, que ce soit à l'aide de la théorie de la généralisabilité ou selon un autre modèle de mesure. En effet, malgré un coefficient de généralisabilité relatif conforme à ceux recensés dans la documentation, il a été observé que la variance due à la spécificité de contexte (interaction entre la facette Candidat et la facette Station) du MEM diffère de celles observées dans le MMI. Enfin, la variance due aux évaluateurs mérite d'être étudiée davantage.

RÉFÉRENCES

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 (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Albanese, M.A., Snow, M.H., Skochelak, S.E., Huggett, K.M., & Farrell, P.M. (2003). Assessing personal qualities in medical school admissions. *Academic Medicine, 78*, 313-321.

Brennan, R.L. (2001a). *Manual for urGenova*. Iowa City, Iowa: Testing Programs, University of Iowa. Disponible sur le site: [http://www.education.uiowa.edu/casma/GenovaPrograms.htm], consulté le 12 avril 2009.

Brennan, R.L. (2001b). *Generalizability theory*. New York: Springer-Verlag.

Eva, K.W., Reiter, H.I., Rosenfeld, J., & Norman, G.R. (2004a). The relationship between interviewers' characteristics and ratings assigned during a Multiple Mini-Interview. *Academic Medicine, 79*, 602-609.

Eva, K.W., Reiter, H.I., Rosenfeld, J., & Norman, G.R. (2004b). The ability of the Multiple Mini-Interview to predict preclerckship performance in medical school. *Academic Medicine, 79*, S40-S42.

Eva, K.W., Reiter, H.I., Rosenfeld, J., & Norman, G.R. (2008). *An update on the validity evidence pertaining to the Multiple Mini-Interview as a candidate selection strategy*. Paper presented at the Canadian Medical Education Conference, Montréal, Québec.

Eva, K.W., Rosenfeld, J., Reiter, H.I., & Norman, G.R. (2004). An admissions OSCE: The Multiple Mini-Interview. *Medical Education, 38*, 314-326.

Frank, J.R. (2005). *Le Cadre de compétences CanMEDS 2005 pour les médecins. L'excellence des normes, des médecins et des soins*. Ottawa: Le Collège royal des médecins et chirurgiens du Canada. Disponible sur le site [http://crmcc.medical.org/canmeds/CanMEDS2005/CanMEDS2005_f.pdf], consulté le 12 avril 2009.

Grand'Maison, P., Brailovsky, C.A., & Lescop, J. (1996). Content validity of the Quebec licensing examination (OSCE) assessed by practising physicians. *Canadian Family Physician*, *42*, 254-259.

Grand'Maison, P., Lescop, J., Rainsberry, P., & Brailovsky, C.A. (1992). Large-scale use of an objective, structured clinical examination for licensing family physicians. *Canadian Medical Association Journal*, *146*, 1735-1740.

Kreiter, C.D., Yin, P., Solow, C., & Brennan, R.L. (2004). Investigating the reliability of the medical school admissions interview. *Advances in Health Sciences Education, 9*, 147-159.

Kulatunga-Moruzi, C., & Norman, G.R. (2002). Validity of admissions measures in predicting performance outcomes: The contribution of cognitive and non cognitive dimensions. *Teaching and Learning in Medicine, 14*, 34-42.

Mann, W.C. (1979). Interviewer scoring differences in student selection interviews. *American Journal of Occupation Therapy, 33*, 235-239.

Moreau, K., Reiter, H.I., & Eva, K.W. (2006). Comparison of aboriginal and nonaboriginal applicants for admissions on the Multiple Mini-Interviews using aboriginal and nonaboriginal interviewers. *Teaching and Learning in Medicine, 18*, 58-61.

Norman, G. (2004). Editorial – The morality of medical school admission. *Advances in Health Sciences Education, 9*, 79-82.

Reiter, H.I., Eva, K.W., Rosenfeld, J., & Norman, G.R. (2007). Multiple Mini-Interviews predict clerkship and licensing examination performance. *Medical Education, 41*, 378-384.

Roberts, C., Walton, M., Rothnie, I., Crossley, J., Lyon, P., Kumar, K., & Tiller, D. (2008). Factors affecting the utility of the Multiple Mini-Interview in selecting candidates for graduate-entry medical school. *Medical Education, 42*, 396-404.

Salvatori, P. (2001). Reliability and validity of admissions tools used to select students for the health professions. *Advances in Health Sciences Education, 6,* 159-175.

Smee, S.M., Dauphinee, D.W., Blackmore, D.E., Rothman, A.I., Reznick, R.K., & Des Marchais, J. (2003). A sequenced OSCE for licensure: Administrative issues, results and myths. *Advances in Health Sciences Education, 8*, 223-236.

Tamblyn, R., Abrahamowicz, M., Dauphinee, D., Wenghofer, E., Jacques, A., Klass, D., et al. (2007). Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *Journal of the American Medical Association, 208*, 993-1001.

ANNEXE A

Les deux types de stations MEM

Les exemples présentés dans cet article ont été traduits et adaptés de la banque de stations de *McMaster University* (Eva et al., 2004a).

Premier exemple : **Station du type « Discussion »**

Vignette

Nos universités ont de la difficulté à équilibrer leur budget. C'est encore pire depuis que de plus en plus d'étudiants accèdent aux études universitaires. Parmi les diverses mesures envisagées pour optimiser les coûts de formation universitaire, certains proposent d'augmenter le nombre d'étudiants par classe. Tout de suite, des opposants à cette idée ont poussé les hauts cris…

Veuillez réfléchir à ce débat.

(Au signal de la cloche, vous entrez et vous aurez huit minutes pour discuter de votre point de vue avec l'intervieweur.)

Entrevue

Dans ce type de stations, la candidate ou le candidat discute directement avec l'évaluateur, qui fait aussi office d'intervieweur.

Deuxième exemple : **Station du type « Jeu de rôle »**

Vignette

Votre compagnie vous mandate, ainsi qu'une collègue (Sara) d'une autre division, pour assister à une rencontre d'affaires d'une importance capitale. Cette rencontre a lieu à San Diego. Sara et vous habitez Montréal. Pour simplifier les choses, vous avez proposé à Sara de passer la prendre en auto chez elle, puis de vous rendre ensemble à l'aéroport Pierre-Elliott-Trudeau (de Dorval).
À l'heure convenue, vous sonnez à la porte de Sara.
Elle vous invite à entrer.

(Au signal de la cloche, vous entrez et vous aurez huit minutes pour effectuer cette rencontre.)

Entrevue

Dans ce type de stations, la candidate ou le candidat interagit avec un (ou des) acteur(s) jouant le(s) personnage(s) mentionné(s) dans la vignette (Sara dans le cas présent). Ainsi, lorsque le candidat entre dans la salle, l'actrice lui dit : « Je suis désolée de te dire cela maintenant, mais depuis 2001 je n'embarque plus dans aucun avion : J'ai trop peur ! Je n'irai donc pas à San Diego ».

L'évaluateur est ici en retrait et observe la performance de la candidate ou du candidat.