Pédagogie Médicale **18**, 17-24 (2017)

https://doi.org/10.1051/pmed/2017023

Pédagogie Médicale

Disponible en ligne :
www.pedagogie-medicale.org

Recherche et perspectives

Validation d'un outil critérié d'évaluation des compétences des résidents en médecine familiale : étude qualitative du processus de réponse

Validation of a criterion-referenced competency assessment tool for family medicine residents: a qualitative study of the response process

Marie-Lee SIMARD, Miriam LACASSE*, Caroline SIMARD, Jean-Sébastien RENAUD, Christian RHEAULT, Isabelle TREMBLAY, et Luc CÔTÉ

Département de médecine familiale et de médecine d'urgence, Université Laval, Québec, Canada

Manuscrit soumis à la rédaction le 24 novembre 2016, commentaires éditoriaux formulés aux auteurs le 7 mai et le 31 octobre 2017, accepté pour publication le 1er novembre 2017

Résumé – Contexte : L'implantation de l'approche par compétences prescrit aux programmes de formation médicale postdoctorale un ajustement de leurs stratégies d'évaluation. Un outil critérié d'évaluation des compétences (OCÉC) a été élaboré à partir de jalons ayant fait l'objet d'une validation de contenu, mais le processus de réponse restait à valider avant son implantation. **Objectifs :** Évaluer le processus de réponse à l'OCÉC et identifier les difficultés rencontrées par ses usagers pouvant introduire des biais dans l'évaluation. **Méthodes :** Étude qualitative auprès de dix cliniciens-enseignants volontaires. La collecte de données s'est effectuée au cours d'entrevues individuelles utilisant la méthode de la pensée à voix haute. Une analyse de contenu des verbatims a été réalisée par trois chercheurs, ce qui a permis la triangulation des données. La structure de codification se compose de quatre thèmes : compréhension, récupération de l'information, jugement et sélection de la réponse. **Résultats :** La compréhension et la récupération de l'information posaient généralement peu problème. Des difficultés aux étapes du jugement (indicateurs hétérogènes et évaluation normative) et de sélection de la réponse (échelle de réponse inadéquate) ont été relevées. **Discussion et conclusion :** Cette étude a permis de vérifier le processus de réponse à l'OCÉC, qui s'est avéré adéquat pour la compréhension et la récupération de l'information, mais à améliorer en regard du jugement et de la sélection de la réponse. L'OCÉC a été révisé en fonction des résultats obtenus avant son implantation. L'évaluation en cours de la structure interne, des relations avec d'autres variables et des conséquences du test complétera le processus de validation de l'OCÉC.

Mots clés : éducation médicale par compétences, évaluation, résidence en médecine familiale, protocole de pensée à voix haute

Abstract - Background: The implementation of competency-based medical education requires postgraduate medical training programs to adjust their assessment strategies. A criterion-referenced competency assessment tool (C-CAT) was developed based on content-validated milestones, but the response process had to be validated before implementation. **Objectives:** To evaluate the C-CAT response process and to identify difficulties encountered by users that could bias the assessment. **Methods:** Qualitative study involving 10 volunteer clinical teachers. Data was collected during individual interviews using the think-aloud protocol. Three researchers independently carried out a content analysis of the verbatim responses, hence data were triangulated. The coding structure consists of four themes: *comprehension, information recall, judgment* and *response selection*. **Results:** Comprehension and information retrieval generally did not cause any problem. Difficulties at the judgment (heterogeneous indicators and normative evaluation) and response selection

*Correspondance et offprints : Miriam LACASSE, Département de médecine familiale et de médecine d'urgence, Faculté de médecine, Pavillon Ferdinand-Vandry, bureau 4486, Université Laval, 1050 avenue de la Médecine, G1V 0A6 Québec (QC).
Mailto : miriam.lacasse@mfa.ulaval.ca

(inadequate response scale) stages were identified. **Discussion and conclusion:** This study validated the C-CAT response process, which proved adequate for comprehension and information retrieval, but required improvements for judgment and response selection. Based on these results, the C-CAT was revised before implementation. The ongoing evaluation of the internal structure, relations with other variables and consequences of testing will complete the validation process.

Keywords: competency-based medical education, assessment, family medicine residency, think-aloud protocol

Introduction

Les programmes de formation médicale postdoctorale au Canada et aux États-Unis font actuellement l'objet d'un processus d'ajustement aux principes de la formation médicale axée sur les compétences et à ses exigences relatives à l'évaluation de l'interne ou du résident. Holmboe *et al.* [1] recommandent notamment que les processus d'évaluation soient critériés, qu'ils s'inscrivent dans une perspective de développement des compétences et qu'ils comportent des méthodes et des outils d'évaluation qui répondent aux exigences minimales de qualité. Les programmes de résidence en médecine utilisent une variété de méthodes pour évaluer les résidents, mais ils continuent de s'appuyer sur des échelles d'évaluation globale en fin de stage [2] en raison de leur flexibilité, de leur capacité d'exploiter l'ensemble des compétences, de leur faible coût et de leur potentiel de rétroaction [3].

Considérant l'ancienne échelle normative de la fiche d'évaluation de fin de stage et la nécessité de l'adapter aux principes d'évaluation axée sur les compétences, le programme de médecine familiale de l'Université Laval (Québec, Canada) a développé un nouvel outil critérié d'évaluation des compétences (OCÉC) de fin de stage, en s'ajustant aux objectifs de chaque stage et au cheminement du résident à l'intérieur de sa formation [4] (http://www.fmed.ulaval.ca/fileadmin/documents/programmes-etudes/etudes-medecine/post-md-residence/medecine-famille/indicateurs-developpement-medecine-familiale.pdf). Les 34 compétences évaluées dans l'OCÉC sont issues du cadre de référence CanMEDs-médecine familiale (CanMEDS-MF), qui comptabilise plus d'une centaine de compétences et à partir duquel sont élaborés les programmes de médecine familiale canadiens. Ce cadre décrit sept rôles et les compétences qui leur sont associées, attendues à la fin du programme [5]. Ces rôles sont : professionnel, communicateur, érudit, expert en médecine familiale, promoteur de la santé, collaborateur et gestionnaire. Au sein de chaque rôle se trouvent à la fois des compétences principales (capacités), lesquelles sont indispensables au succès du stage, ainsi que des compétences habilitantes (manifestations) complémentaires.

Dans le cadre de l'OCÉC, afin de faciliter la compréhension des évaluateurs, chaque niveau de développement des compétences (peu autonome, partiellement autonome et autonome) est détaillé par des rubriques apparaissant automatiquement au-dessus du choix de réponse. Ces niveaux de maîtrise des compétences ont été situés sur des jalons (*milestones*) échelonnés dans le temps, afin de situer les résidents par rapport aux attentes de leur programme en cours de formation, lesquelles ont fait l'objet d'une étude antérieure [6]. Ces « jalons » servent de critères et doivent être connus des superviseurs afin qu'ils puissent juger de la performance du résident par rapport à ce qui est attendu, plutôt qu'en comparaison avec celle des autres résidents. Ceci constitue d'ailleurs une distinction entre l'évaluation critériée et l'évaluation normative [7].

Dans une perspective d'évaluation par compétences, l'OCÉC évalue aussi des activités professionnelles confiables (APC – *entrustable professional activities*). Telle que conceptualisée par Ten Cate [8], une APC est une tâche/responsabilité pour laquelle la direction de programme peut avoir confiance qu'elle soit exécutée par un résident ayant développé les compétences requises. Ces APC correspondent à des activités et des situations professionnelles à partir desquelles est documentée la maîtrise des compétences (clés et habilitantes) ; les APC font le pont entre les compétences et la pratique [8,9].

L'interface qui permet à l'utilisateur d'exploiter l'OCÉC prend la forme d'une fiche d'évaluation de fin de stage (ou d'un rapport trimestriel de progrès), par laquelle on demande d'abord au clinicien-enseignant d'établir le niveau d'autonomie du résident en fonction de trois rubriques décrivant les comportements observés pour chaque compétence à évaluer [10]. Par la suite, le système informatique relie ce niveau d'autonomie aux attentes du programme [11] afin d'établir un résultat (précoce, attendu, limite ou retard). Pour suggérer à l'évaluateur l'issue du stage (échec, limite, réussite), le système informatisé calcule une proportion, soit le nombre de compétences non maîtrisées (par exemple : limite ou retard) par rapport au total des compétences évaluées. L'importance relative de la compétence, dont rend compte le fait qu'elle est ou non considérée comme à réussite obligatoire (et donc identifiée comme telle en tant que compétence-clé), est également considérée dans le calcul. Toutefois, la décision finale quant à l'issue du stage reste la prérogative de l'évaluateur, qui peut ou non accepter la proposition du système.

L'*American Educational Research, l'American Psychological Association* et le *National Council on Measurement in Education* ont conjointement établi des standards dans la conception d'outils d'évaluation. Selon ces derniers, cinq sources de preuves de validité servent à

déterminer la qualité d'un outil d'évaluation ; elles concernent respectivement le contenu du test, le processus de réponse, la structure interne, les relations avec d'autres variables et les conséquences du test [12]. La présente étude vise à valider le processus de réponse à l'OCÉC, les autres sources de validité ayant déjà été documentées ou faisant l'objet d'études en cours. Plus précisément, l'évaluation du processus de réponse a consisté à analyser les étapes cognitives menant l'évaluateur à choisir une réponse à l'OCÉC, soit : peu autonome, partiellement autonome et autonome. La validation du processus de réponse permet aux concepteurs de l'outil de mesure de s'assurer que les étapes cognitives entreprises par les répondants correspondent effectivement aux caractéristiques visées par l'outil d'évaluation [12].

Méthodes

Cette étude a suivi un devis descriptif, de nature qualitative. Dix cliniciens-enseignants (CE) volontaires provenant de cinq unités de médecine familiale (UMFs) ont évalué chacun un résident de leur UMF respective à l'aide de l'OCÉC pour recueillir des données sur le processus de réponse. Les données ont été obtenues lors d'entrevues individuelles cognitives semi-structurées utilisant la méthode de la pensée à voix haute (*think aloud*) [13]. Cette méthode éprouvée [14] permet une compréhension du processus sous-jacent aux réponses données à une question, d'où sa pertinence lorsqu'elle est jointe aux méthodes usuelles de validation de questionnaires [15]. Pour cette raison, les méthodes cognitives, dérivées de la psychologie sociale et cognitive, sont de plus en plus utilisées dans la validation pilote de questionnaires. Lorsqu'elles sont mises en relation avec les modèles théoriques du processus de réponse à une question, ces méthodes contribuent grandement à la compréhension des sources d'erreurs de mesure [16]. Plusieurs auteurs proposent un modèle décrivant quatre étapes cognitives impliquées dans le processus de réponse à une question : compréhension, récupération de l'information, jugement et sélection/formulation de la réponse [15–21].

Dans les instructions données à chaque clinicien-enseignant, il était indiqué de « tout verbaliser, incluant tout ce qui se passe dans sa tête en accomplissant la tâche ». Pendant l'entrevue, la tâche du clinicien-enseignant consistait à évaluer un résident en utilisant le prototype informatisé de l'OCÉC, basé sur les rubriques discutées précédemment (http://www.fmed.ulaval.ca/les-programmes-detudes/etudes-en-medecine/residences-etudes-medicales-postdoctorales/residence-en-medecine-familiale/evaluations/), tout en exprimant à voix haute sa compréhension de chaque élément de la fiche et les raisons sous-tendant son choix du niveau d'autonomie (peu autonome, partiellement autonome, autonome) pour chaque compétence. Les entrevues conduites par l'auxiliaire de recherche suivaient un guide d'entrevue contenant des questions complémentaires. Le guide d'entrevue a été élaboré par les auteurs de manière à recueillir les informations jugées les plus importantes pour répondre à l'objectif de recherche, soit l'explication du rationnel de leurs réponses, la clarté des énoncés, les difficultés ainsi que les questions émergeant lors de la réalisation de la tâche. Les considérations pratiques décrites dans l'ouvrage de van Someren « *The think aloud method: a practical guide to modeling cognitive process* » [22] ont été appliquées lors des entrevues. Tel que suggéré par cet auteur [22], un court exercice d'entraînement semblable à la tâche à effectuer a été élaboré afin de s'assurer d'un niveau de verbalisation adéquat et de permettre un ajustement préalable si nécessaire.

Les entrevues ont été transcrites intégralement. Les verbatims ont été codifiés de manière indépendante et analysés par trois personnes, parmi lesquelles l'auxiliaire ayant réalisé les entrevues. Une classification des problèmes en quatre rubriques correspondant aux étapes cognitives impliquées dans le processus de réponse a été utilisée : compréhension, récupération de l'information, jugement et sélection/formulation de la réponse. Les codes assignés aux catégories de problèmes potentiels ont été déterminés par une approche mixte déductive/inductive. Au départ, les dix protocoles verbaux ont fait l'objet d'une première analyse de contenu avec la structure de codification initiale basée sur les écrits scientifiques [15–21]. Les catégories de problèmes ne pouvant être encodées grâce aux codes existants ont mené à la création de nouveaux codes. Une fois tous les codes créés, au moins un exemple de problème a été assigné à chaque catégorie d'encodage pour faciliter l'attribution ultérieure de chaque code aux extraits de rapports verbaux. L'élaboration de la structure de codification et l'attribution des exemples s'y rapportant a fait l'objet d'un processus itératif à plusieurs cycles impliquant les auteurs et la consultation continue des deux analystes, jusqu'à l'obtention d'une structure de classification suffisamment complète et détaillée.

Résultats

Les entrevues ont été réalisées entre juillet et août 2015 dans les cinq UMF participantes. La durée moyenne des entrevues fut de 85 minutes.

Les principales forces de la fiche ayant été évoquées sont sa convivialité, ainsi que la présence des onglets descriptifs (détaillant chaque niveau de d'autonomie), alimentant la réflexion et soutenant l'objectivité de la démarche d'évaluation. Les principales limites soulevées sont le nombre important d'éléments à évaluer, l'ambiguïté des termes « habituellement » et « systématiquement » utilisés pour l'échelle de réponse, ainsi que l'improbabilité qu'un résident de première année soit en échec, malgré la présence de difficultés, selon l'algorithme de calcul des résultats.

Le processus d'analyse étant de nature itérative, la catégorisation des problèmes cognitifs rencontrés dans le processus de réponse à l'OCÉC a évolué au fil des analyses des verbatims. L'annexe 1 présente sous forme de tableau la structure de la classification finale selon une forme

hiérarchique, avec les grandes catégories et les problèmes qui s'y rapportent. Les quatre grandes catégories de problèmes proposés par le cadre théorique ont été conservées (numéros 1 à 4) et deux catégories ont été ajoutées pour relever des éléments observables mais dont on ne pouvait tirer d'inférence (résultats -5- et autres -6). Plusieurs catégories de problèmes ont été ajoutées (voir énoncés italiques en annexe 1 ; par exemple : 3.7. effet de halo), alors que certaines ont été regroupées (par exemple : « formulation de réponse ambiguë » et « dissonance entre réponse désirée et échelle de réponse », regroupées sous 4.1. « échelle de réponse inadéquate ») alors que d'autres ont été retirées lorsque le problème n'avait pas été rencontré ou ne s'appliquait pas (par exemple : « références temporelles ambiguës »). Les problèmes identifiés par les cliniciens-enseignants en fonction de l'étape cognitive impliquée se détaillent comme suit : compréhension (124 occurrences pour 7 catégories de problèmes identifiés), récupération de l'information (93 occurrences pour 3 catégories), jugement (133 occurrences pour 7 catégories) et sélection/formulation de la réponse (48 occurrences pour 2 catégories). Le nombre de participants ayant rapporté un problème en lien avec l'OCÉC a été compilé par item (compétence) et par code de la structure de codification. Les problèmes les plus souvent évoqués (catégories de problèmes rapportées dans quatre rapports verbaux ou plus pour une compétence) sont présentés dans l'annexe 1, ainsi que des extraits d'entrevue s'y rapportant (annexe 2). L'échantillon n'est toutefois pas d'ampleur suffisante pour tirer des conclusions de ces indicateurs quantitatifs.

Afin de déterminer la fidélité des données recueillies (ce qui correspond au critère de crédibilité en recherche qualitative) [23,24], le degré d'accord inter-juge a été calculé sous forme de pourcentage d'accord. Cet indice a été choisi compte tenu du petit nombre d'observations disponibles. Il tend toutefois à être surestimé puisqu'il ne tient pas compte de l'accord dû au hasard [25]. Le pourcentage d'accord entre les codeurs varie entre 60,34 % et 98,28 %, à l'exception de « réponse/lecture intuitive » (24,14 %) et « réfère à une compétence précédente » (44,83 %).

Discussion

Cette étude a permis de valider le processus de réponse en lien avec l'OCÉC par la méthode de la pensée à voix haute. Ceci a permis d'identifier les catégories de problèmes rencontrés lors de quatre étapes cognitives. La compréhension et la récupération de l'information étaient généralement appropriées. Toutefois certaines difficultés ont été identifiées en lien avec les étapes de jugement et de sélection de la réponse.

Cette étude est novatrice par sa méthodologie, laquelle propose une structure de codification conçue spécifiquement pour évaluer le processus de réponse à un outil d'évaluation. En effet, plusieurs structures de codification sont disponibles, mais aucune d'entre elles ne permet l'évaluation spécifique du processus de réponse à un outil d'évaluation dans un contexte d'évaluation des compétences.

L'OCÉC a été apprécié par les usagers de l'étude pour sa facilité d'utilisation et la richesse de l'information qu'il contient, soutenant la décision des évaluateurs. Compte tenu des limites évoquées, il est recommandé de revoir l'utilisation du terme « systématiquement » dans le libellé des rubriques de développement. De plus, il faudrait offrir une formation professorale soutenue pour faciliter l'appropriation de l'OCÉC par ses usagers, notamment pour prévenir la propension des utilisateurs à évaluer les résidents selon une impression générale et à attribuer une côte similaire à toutes les compétences (effet de halo) [26].

Les résultats obtenus démontrent que le problème principal identifié se trouve à l'étape de sélection/formulation de la réponse. De fait, 72 extraits ont été encodés à travers l'ensemble des rapports verbaux dans la catégorie de problème 4.1. « échelle de réponse inadéquate ». La lacune principale de l'échelle de réponse utilisée est l'emploi de qualificatifs fréquentiels mal définis et ambigus (par exemple : systématiquement). En effet, les qualificatifs fréquentiels utilisés pour décrire les différents niveaux de développement ne reflètent pas la fréquence absolue d'un comportement, mais plutôt la fréquence relative de ce dernier en fonction des attentes du répondant [27,28]. Ainsi, le clinicien-enseignant qui ne s'attend pas à ce qu'un résident adopte un comportement relevant de la perfection a tendance à interpréter le qualificatif « systématique » moins sévèrement et fait preuve de plus de flexibilité dans son évaluation du résident. Par ailleurs, l'interprétation de ces qualificatifs fréquentiels est modulée spontanément par les comportements des répondants eux-mêmes, ainsi que par les comportements de ses pairs [28]. Ainsi, un clinicien-enseignant qui n'adopte pas un comportement de façon systématique dans sa propre pratique interprète le qualificatif « systématique » de façon moins stricte dans son évaluation du résident. Pour cette raison, une meilleure définition des qualificatifs fréquentiels utilisés, voire une révision de ces derniers, s'impose pour délimiter clairement les niveaux de développement et permettre une évaluation plus objective et uniforme des résidents. En effet, plutôt que de laisser le répondant calibrer lui-même l'échelle de réponse en fonction de son interprétation personnelle du sens des qualificatifs fréquentiels, il est plus utile et rigoureux de fournir des définitions claires qui ne laissent pas place à l'interprétation [21–28].

Les problèmes identifiés en lien avec la difficulté de rappel, lesquels se situent à la deuxième étape cognitive, celle de la récupération de l'information, sont cohérents avec ce que l'on retrouve dans les écrits scientifiques. En effet, la difficulté éprouvée par les participants à se rappeler des exemples de comportements précis dans un contexte commun est un phénomène courant lors de la réponse à des questionnaires portant sur la fréquence comportementale [29]. Les capacités de rappel sont toutefois améliorées lorsque les répondants ont à leur

disposition un temps suffisant pour réfléchir [30]. Des tentatives répétitives de rappel peuvent être nécessaires pour récupérer de nouvelles informations et ce, malgré un nombre possiblement élevé de tentatives de rappel antérieures infructueuses [31]. Compte tenu de l'étendue de la tâche qui était demandée et des contraintes temporelles de l'entrevue, les cliniciens-enseignants avaient peu de temps pour réfléchir à leurs réponses. Dans un contexte pratique courant, les cliniciens-enseignants auront un outil plus bref, auront plus de temps pour réfléchir à leurs réponses et n'auront pas pour tâche de penser à voix haute, ce qui alourdit la réflexion. Il est donc possible que cette contrainte soulevée en entrevue n'apparaisse pas comme un obstacle important lors de l'utilisation réelle de l'outil.

Le problème identifié en lien avec l'hétérogénéité des indicateurs, lequel se situe à la troisième étape cognitive, celle du jugement, découle du fait qu'une série de comportements spécifiques sont décrits pour chaque compétence et pour chaque niveau de développement. Le résident répondant à certains comportements, mais pas à d'autres, pour un niveau de développement donné se trouve difficile à positionner dans l'échelle de réponse. Le poids de chaque élément d'évaluation nécessite alors le jugement du clinicien-enseignant, lequel déterminera le niveau d'autonomie attribué. Cet aspect subjectif de l'évaluation pourrait expliquer des divergences entre deux évaluations faites par différents cliniciens-enseignants pour un même résident. Certaines directives accompagnant l'OCÉC ou fournies lors de la formation offerte seront essentielles afin de guider les cliniciens-enseignants confrontés à ce problème.

L'autre problème rencontré dans la troisième étape cognitive (jugement) se rapporte au changement du format de l'échelle de réponse, désormais critériée et non plus normative. Évaluer les résidents en comparant leurs aptitudes avec l'ensemble des étudiants du même niveau continue d'être le réflexe initial de certains cliniciens-enseignants. Il serait important de rappeler clairement à ces derniers qu'ils doivent utiliser les indicateurs de développement comme base de leur jugement pour l'évaluation, et non pas les comportements des autres résidents du même niveau. Une mise en garde préalable spécifiant que le résultat de stage suggéré ne fait office que d'indicateur et doit être utilisé avec prudence dans la décision finale du clinicien-enseignant quant à la passation ou non du stage aurait pour effet d'éviter toute hésitation des cliniciens-enseignants à respecter le format d'évaluation critérié de l'OCÉC.

Les résultats qui découlent de la présente étude ont d'importantes implications pratiques et théoriques. D'abord, l'objectif principal de l'étude était de relever les défis cognitifs rencontrés par les cliniciens-enseignants lors de l'utilisation de l'OCÉC. Les résultats ont permis d'en relever plusieurs, ce qui permettra d'améliorer la qualité de l'outil en limitant les biais de réponse et en allégeant le processus d'utilisation de l'outil. En effet, un outil difficile et laborieux à compléter peut épuiser les répondants ou les décourager, augmentant ainsi les risques de réponses erronées [14]. Par ailleurs, certaines implications théoriques peuvent être envisagées, puisque la méthodologie employée se veut novatrice. Elle répond directement aux standards en psychométrie pour explorer les biais pouvant s'insérer dans les processus de réponse en utilisant une méthode rarement employée, soit celle des entrevues cognitives. Critiquée pour son caractère subjectif et artificiel, Drennan [15] soutient que cette méthode peut fournir des informations complémentaires aux autres approches quant à la validité de l'outil, lesquelles ne pourraient être obtenues autrement. Par ailleurs, une structure de codification a été développée à partir de l'adaptation et de l'intégration de plusieurs modèles théoriques sur les problèmes rencontrés lors de l'accomplissement d'une tâche [15–21]. Adaptée au contexte de la validation d'outil d'évaluation, cette structure peut constituer un repère pertinent pour les recherches futures portant sur d'autres outils.

Certaines limites sont à noter puisqu'elles réduisent la portée des résultats obtenus. Premièrement, quoique certains auteurs affirment qu'un échantillon aussi petit que dix participants puisse être suffisant pour ce type d'étude [32,33], la saturation des données n'a pas été atteinte avec dix entrevues. Le recrutement d'un plus grand nombre de participants aurait été souhaitable. Deuxièmement, cet échantillon restreint de cliniciens-enseignants ne permet pas le calcul d'un indice d'accord inter-juges. Il est impossible d'isoler la variabilité attribuable aux caractéristiques de l'évaluateur de celle attribuable aux caractéristiques du résident. En effet, il aurait fallu que chaque paire de clinicien-enseignant évalue plusieurs résidents. En raison de ces limites et à défaut d'avoir pu réunir les conditions pour utiliser un indice d'accord plus précis, le pourcentage d'accord inter-juge est présenté à des fins descriptives, mais ne permet pas de tirer de conclusions quant à la fiabilité de l'analyse. Il ne permet pas non plus de se prononcer quant à la validité de la structure de classification. Il faut donc lui attribuer une portée strictement indicative et non concluante. L'échantillon actuel est très petit pour que les pourcentages d'accord et les fréquences d'occurrence des problématiques rencontrées puissent être généralisables. Il est à noter que même si la crédibilité des résultats est un critère de scientificité en recherche qualitative, les exigences et les critères de rigueur scientifique en recherche qualitative diffèrent considérablement de la recherche quantitative. La vérification de l'accord entre les juges est toutefois une méthode d'usage (voire incontournable [34]) de la fiabilité d'une analyse de contenu [23], quoique les exigences méthodologiques d'une analyse quantitative diffèrent de celles d'une analyse qualitative. Troisièmement, le contexte expérimental de l'étude n'a pas reproduit fidèlement le contexte réel d'utilisation de la fiche, ce qui aurait été préférable pour une recherche qualitative. Habituellement, dans les UMF, les cliniciens-enseignants se réunissent périodiquement pour évaluer les résidents de manière formative et sanctionnelle lors de réunions prévues à cet effet. Ensemble, les cliniciens-enseignants peuvent colliger leurs expériences d'évaluation respectives avec le résident évalué et produire une évaluation plus juste et complète

pour chaque compétence visée par l'OCÉC. Le contexte d'évaluation en groupe n'a pas été reproduit dans cette étude pour des raisons pratiques.

En conclusion, cette étude a permis la validation préimplantation du processus de réponse de l'OCÉC à l'aide d'une approche par entrevue cognitive de pensée à voix haute rétrospective, ce qui est peu évoqué dans les écrits scientifiques. Cette vérification préalable à l'implantation de l'outil est toutefois nécessaire pour déceler tout problème en lien avec l'utilisation d'un questionnaire [16]. Les résultats de l'étude ont permis d'améliorer l'OCÉC selon les problèmes identifiés, favorisant une intégration réussie dans les milieux d'enseignement. Plus spécifiquement, l'échelle de réponse utilisée a été revue et améliorée. Toutefois, de plus amples analyses de validation de l'OCÉC seront à effectuer une fois implantée sur le terrain, soit par l'évaluation de la structure interne, des relations avec d'autres variables et les conséquences du test.

Contributions

Marie-Lee Simard et Caroline Simard ont participé au recueil et à l'analyse des données et rédigé le manuscrit, Miriam Lacasse a élaboré l'idée originale, conçu le protocole de recherche, participé au recueil et à l'analyse des données en vue du développement et de la validation de l'outil, conçu et développé le système informatisé et participé à la rédaction du manuscrit. Jean-Sébastien Renaud a participé au travail en tant qu'expert-conseil en évaluation (processus de validation) et participé à la révision du manuscrit. Christian Rheault et Isabelle Tremblay ont participé à l'analyse des résultats et à la révision du manuscrit. Luc Côté a participé à l'étude en tant qu'expert-conseil en recherche qualitative, participé à l'analyse des données qualitatives et à la révision du manuscrit.

Approbation éthique

Non sollicitée.

Conflits d'intérêts

Les auteurs déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

Annexe 1: Catégories de codification élaborée selon une approche inductive/déductive basée sur le modèle théorique du processus de réponse proposé par plusieurs auteurs [14–20].

<table>
<tr><th>1 Compréhension</th><th>2 Récupération de l'information</th><th>3 Jugement</th><th>4 Sélection/formulation de la réponse</th><th>5 Résultantes</th></tr>
<tr><td>1.1. Compétence ambiguë
1.2. Indicateur de développement ambigu
1.3. Terme ambigu
1.3.1. Problème d'inclusion/exclusion
1.3.3. Vocabulaire technique/non adapté
1.4. Instructions imprécises
1.4.1. Tâche imprécise
1.4.2. Instructions incomplètes
1.5. Énoncé lourd/complexe
1.6. Erreur de lecture
1.7. Non reliée à la compétence</td><td>2.1. Indicateur incorrect/présomptueux
2.2. Difficulté de rappel
2.3. Manque d'information pour répondre
2.3.1. Non applicable
2.3.2. Improbable de connaître la réponse
2.3.3. Non applicable dans le contexte expérimental</td><td>3.1. Indicateurs hétérogènes
3.2. Incertitude
3.3. Problème computationnel
3.3.1. Surcharge cognitive
3.4. Potentiellement sensible/compromettant
3.4.1. Réponse socialement acceptable
3.5. Évaluation normative
3.5.1. Référence au niveau de l'étudiant
3.5.2. Référence à soi-même
3.5.2.1. Référence aux traits pers.
3.5.2.2. Références aux valeurs pers.
3.6. Réponse inconsistante avec réponse précédente
3.7. Effet halo</td><td>4.1. Échelle de réponse inadéquate
4.1.1. Réponse « entre deux »
4.2. Incapacité de répondre</td><td rowspan="3">5.1. Silence
5.2. Sons de remplissage
5.3. Phrases inachevées
5.4. Phrases incompréhensibles
5.5. Auto questionnement
5.6. Relecture
5.7. Réponse/lecture intuitive
5.7.1. Lecture intuitive
5.7.2. Lecture incorrecte
5.7.3. Réponse intuitive
5.7.4. Réponse spontanée
5.8. Niveau d'aisance
5.8.1. Facile
5.8.2. Difficile
5.9. Réfère à une compétence précédente
5.9.1. Modification d'une réponse précédente
5.9.2. Retour sur une compétence précédente</td></tr>
<tr><td colspan="3">6 Autre</td><td>7 Non catégorisé</td></tr>
<tr><td>5.9.2.1. Redondance entre les compétences
5.10. Dissonance réponse-explication</td><td>6.3. Distraction
6.4. Question dirigée à l'auxiliaire de recherche
6.4.1. Question sur la tâche</td><td>6.5. Réaction émotionnelle
6.6. Compréhension personnelle</td><td>8 Commentaires et impressions additionnels</td></tr>
</table>

Note : les énoncés en italique sont le résultat d'un ajout, un déplacement ou une fusion de deux catégories de façon inductive, soit au fil de l'analyse de contenu.

Annexe 2: Catégorisation des problèmes rapportés dans le processus de réponse à l'outil critérié d'évaluation des compétences (OCÉC) avec extraits de verbatims d'entrevues.

Catégorie de problèmes	Extraits de verbatim d'entrevues
Première étape cognitive : *compréhension*	
1.2. Indicateur de dévelop-pement ambigü	***Gérer son apprentissage de façon autonome (4 répondants)*** • «J'ai de la difficulté à comprendre pourquoi UpToDate est considéré comme une référence générique. Je suis confuse puisque je ne l'aurais pas classé dans cette catégorie de référence» (P3[†]) • «Je ne comprends pas en quoi les lignes directrices sont supérieures à UpToDate. J'ai l'impression que les résidents ne peuvent pas consulter l'un sans l'autre. Je ne comprends pas la nuance qu'on a voulu apporter avec cette précision.» ***Assumer des rôles administratifs et agir comme leader (4 répondants)*** • «Je ne suis pas certaine de comprendre tout… la lecture de l'énoncé ne m'aide pas nécessairement à répondre» (P1) • «Les items décrits ne sont pas très clairs. Concrètement, je ne suis pas certaine de savoir ce que ça représente.» (P6)
1.3. Terme ambigü	***Respecter les dimensions éthiques de la prise de décision clinique (4 répondants)*** • «Je ne comprends pas le sens du mot "générique". Veut-on dire "général" plutôt que "générique"? » (P2) • «… de manière générique, c'est un peu bizarre. Je dirais peut-être "de façon générale"? » (P7)
1.5. Énoncé lourd/ complexe	***Communiquer verbalement (4 répondants)*** • «Il y a beaucoup trop d'items.» (P3) • «Je trouve que le texte est long. S'il était possible de le mettre plus court ça serait mieux.» (P2)
1.7. Réponse non reliée à la compétence	***Respecter les dimensions éthiques de la prise de décision clinique (4 répondants)*** • «Les termes médicaux utilisés ne sont pas nécessairement adaptés à ceux qui ont un faible niveau d'éducation.» (P10) • «Pour cette compétence, je crois que le résident est partiellement autonome puisqu'il présente des barrières de langue.» (P3)
Deuxième étape cognitive : *récupération de l'information*	
2.3. Information insuffisante pour répondre	***Assurer son développement professionnel continu (4 répondants)*** • «Je trouve que cet aspect est difficile à évaluer si on n'a pas nécessairement toutes les informations en main» (P1) • «En tant qu'enseignant, je ne sais pas ce que le résident a fait comme activité Apprentis [Centre de simulation] ou ce qu'il a fait comme modules… Probablement que certains médecins ont l'information, mais encore là, je ne sais pas à quel point ces données sont colligées.» (P8) ***Réaliser une évaluation de la qualité de l'acte/projet d'érudition/de recherche (7 répondants)*** • «Il y a une personne qui est dédiée pour accompagner les résidents dans la réalisation de leur projet. Cette personne serait en mesure d'évaluer le résident, mais moi n'ai pas l'information.» (P8) ***Enseigner aux étudiants et à ses collègues (7 répondants)*** • «Le résident n'a pas d'occasion de faire de l'enseignement comme tel» (P1) • «Pour pouvoir dire que quelqu'un est systématique, il faut qu'il l'ait fait plus que trois fois. Les résidents n'ont pas suffisamment d'exposition pour que je puisse dire que c'est systématique.» (P5) • «Je n'ai jamais pu voir un résident enseigner à d'autres étudiants ou à ses collègues.» (P7) ***Assumer les rôles administratifs et agir comme leader (7 répondants)*** • «Je ne suis pas capable de corréler avec des exemples concrets qui me permettraient d'évaluer le résident sur cet aspect.» (P1) • «En tant que patron je ne sais pas comment le résident agit avec ses collègues. Je ne sais pas quel nibveau de *leadership* il adopte à l'intérieur de son équipe.» (P8) • «Je n'ai pas eu l'occasion de voir le résident dans ces rôles là.» (P9)
Troisième étape cognitive : *jugement*	
3.1. Indicateurs hétérogènes	***Assurer son développement professionnel continu (4 répondants)*** • «Le problème c'est qu'il y a plusieurs points d'évaluation pour cette compétence. Un résident qui ne répond pas à un seul point, mais à tous les autres, serait automatiquement considéré comme partiellement autonome. Nous ne sommes pas capables de nuancer» (P2) • «Je suis embêté car le résident répond à 3 critères sur 4. Je ne sais pas si je dois le mettre autonome puisqu'il ne répond pas aux 4 critères. (P5)
3.5. Évaluation normative	***Réfléchir sur sa pratique (4 répondants)*** • «Comme tous les résidents, je pense que ce résident n'a pas intégré de façon systématique dans sa pratique la tenue de journaux réflexifs.» (P8) • «Je trouve que le niveau "autonome" est difficile à atteindre à la fin de la première année de résidence. […] Je n'ai même pas l'impression qu'après 5 ans de pratique je suis moi-même rendu là.» (P10) «Même en pratique il nous arrive de ne pas faire ça» (P3) ***Assurer son développement professionnel continu (4 répondants)*** • «On ne s'attend pas qu'ils aient complété l'ensemble des modules rendus à la fin de la première année de résidence.» (P8) • «Le libellé est problématique parce qu'il est impossible que résident de première année ait complété l'ensemble des modules. L'exigence est qu'ils les aient complétés à la toute fin de la résidence.» (P3)
Quatrième étape cognitive : *sélection/formulation de la réponse*	
4.1. Échelle de réponse inadéquate	***Pour l'ensemble de la fiche (10 répondants)*** • «"Systématiquement" c'est trop lourd.» (P2), «Le mot systématique est très fort.» (P4) • «Dans ma tête "systématiquement" c'est toujours, mais ç'est rare qu'on fait les choses 100% du temps. (P8) • «À chaque fois où c'est écrit systématiquement, ça me fait toujours rire, parce qu'on s'entend là, en médecine familiale, des choses où c'est tout blanc tout noir, c'est très rare.» (P6) • «Quand il est écrit "systématiquement" est-ce qu'on veut dire tout le? Pour ma part, je l'interprète comme la majorité du temps. Mais ensuite est-ce que majoritairement du temps correspondrait plutôt à "habituellement" dans ce cas là?» (P8) • «"Habituellement", "systématiquement"… c'est pas coupé au couteau entre les deux. Je trouve que la nuance n'est pas évidente.» (P5) • «Il n'y a pas une grosse différence entre "habituellement" et "systématiquement"… Je ne peux pas dire que pour ce résident c'est 100% du temps, bien que c'est quand même assez bien. Je le mettrait comme dans le milieu.» (P2)

[†] Correspond au numéro du participant ayant rapporté cet extrait.

Références

1. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach 2010;328:676-682.
2. Chou S, Lockyer J, Cole G, McLaughlin K. Assessing postgraduate trainees in Canada: are we achieving diversity in methods? Med Teach 2009;322:58-63.
3. Streiner DL. Assessing Clinical Competence. New York: Springer Publishing Company, 1985.
4. Lacasse M, Rheault C, Tremblay I, Renaud JS, Coché F, St Pierre A, *et al.* Développement d'un outil novateur critérié d'évaluation des compétences à partir d'indicateurs de développement pour les résidents en médecine familiale. Pédagogie Médicale 2016; soumis pour publication.
5. Tannenbaum D, Konkin J, Organek A, Parsons E, Saucier D *et al.* 2011. Cursus Triple C axé sur le développement des compétences. Rapport du Groupe de travail sur la révision du cursus postdoctoral – Partie 1. Mississauga (ON) : Collège des médecins de famille du Canada.
6. Ten Cate O. Medical education: trust, competence, and the supervisor's role in postgraduate training. BMJ 2006;333:748.
7. Ten Cate O. AM last page: what entrustable professional activities add to a competency-based curriculum. Acad Med 2014;89:691.
8. Lacasse M, Théorêt J, Tessier S, Arsenault L. Expectations of clinical teachers and faculty regarding development of the CanMEDS-Family Medicine competencies: Laval developmental benchmarks scale for family medicine residency training. Teach Learn Med 2014;26:244-251.
9. Oandasan I, Saucier D. (Eds). 2013. Rapport sur le Cursus Triple C axé sur le développement des compétences – Partie 2 : Faire progresser la mise en œuvre. Mississauga, ON : Collège des médecins de famille du Canada, vol. 10.
10. Tardif J. L'évaluation des compétences : documenter le parcours de développement. Montréal : Chenelière éducation, 2006.
11. Lacasse M, Rheault C, Tremblay I, Renaud J-S, Coché F, St-Pierre A, Théorêt J, Tessier S, Arsenault L, Simard M-L, Simard C, Savard I, Castel J, Côté L. Indicateurs de développement en médecine familiale : Intervalles attendus pour développer un niveau de compétence autonome durant le programme de résidence [En ligne]. Québec : Département de médecine familiale et de médecine d'urgence, Université Laval; 2016. [cité le 1er décembre 2017] Disponible sur http://www.fmed.ulaval.ca/fileadmin/documents/programmes-etudes/etudes-medecine/post-md-residence/medecine-famille/indicateurs-developpement-medecine-familiale.pdf.
12. Joint Committee of the American Educational Research. Standards for Educational and Psychological Testing. Washington (DC): American Psychological Associatio, 2014.
13. Forget MH. Le développement des méthodes de verbalisation de l'action : un apport certain à la recherche qualitative. Recher Qual 2013;32(1):57-80.
14. Durning SJ, Artino AR, Beckman TJ, Graner J, VanDerVleuten E *et al.* Does the think-aloud protocol reflect thinking? Exploring functional neuroimaging differences with thinking (answering multiple choice questions) *versus* thinking aloud. Med Teach 2013;35:720-726.
15. Drennan, J. Cognitive interviewing: verbal data in the design and pretesting of questionnaires. J Adv Nurs 2003;42:57-63.
16. Collins D. Pretesting survey instruments: an overview of cognitive methods. Qual Life Res 2003;12:229-238.
17. Willis GB, Royston P, Bercini D. The use of verbal report methods in the development and testing of survey questionnaires. Appl Cogn Psychol 1991;5:251-267.
18. Forsyth B, Levin K, Fisher S. Test of an appraisal method for establishment survey questionnaires, in Proceedings of the ASA Section on Survey Research Methods. Alexandria (VA): American Statistical Association, 1999, p. 145-149.
19. Bolton RN. Pretesting questionnaires: content analyses of respondents' concurrent verbal protocols. Mark Sci 1993;12:280-303.
20. Graesser AC, Bommaredy S, Swamer SS, Golding JM. Integrating questionnaire design with a cognitive computational model of human question answering, in Answering questions: methodology for determining cognitive and communicative processes in survey research, Schwarz N, Sudman S, Editors. San Francisco: Jossey-Bass, 1996, p. 143-175.
21. Tourangeau R, Rasinski KA. Cognitive processes underlying context effects in attitude measurement. Psychol Bull 1988;103:299.
22. Van Someren MW, Barnard YF, Sandberg JA. The think aloud method: a practical guide to modeling cognitive processes. London: Academic Press, 1994.
23. Fortin M. Fondements du processus de recherche : méthodes quantitatives et qualitatives. Montréal : Chenelière Éducation, 2010.
24. Golafshani N. Understanding reliability and validity in qualitative research. Qual Rep 2003;8(4):597-606.
25. Hayes JR, Hatch JA. Issues in measuring reliability correlation *versus* percentage of agreement. Writ Commun 1999;16:354-367.
26. Thorndike EL. A constant error in psychological ratings. J appl psychol 1920; 4: 25-29.
27. Schwarz N, Oyserman D. Asking questions about behavior: cognition, communication, and questionnaire construction. Am J Eval 2001;22:127-160.
28. Wänke M. Conversational norms and the interpretation of vague quantifiers. Appl Cogn Psychol 2002;16:301-307.
29. Schwarz N. Self-reports: how the questions shape the answers. Am Psychol 1999;54:93-105.
30. Reiser BJ, Black JB, Abelson RP. Knowledge structures in the organization and retrieval of autobiographical memories. Cogn Psychol 1985;17:89-137.
31. Williams MD, Hollan JD. The process of retrieval from very long-term memory. Cogn Sci 1981;5:87-119.
32. Hoppmann TK. Examining the 'point of frustration'. The think-aloud method applied to online search tasks. Qual Quant 2007;43:211-224.
33. Virzi RA. Refining the test phase of usability evaluation: how many subjects is enough? Hum Factors J Hum Factors Ergon Soc 1992;34:457-468.
34. Lombard M, Snyder-Duch J, Bracken CC. Content analysis in mass communication: assessment and reporting of intercoder reliability. Hum Commun Res 2002;28(4): 587-604.

Citation de l'article : Simard M.-L., Lacasse M., Simard C., Renaud J.-S., Rheault C., Tremblay I., Côté L., Validation d'un outil critérié d'évaluation des compétences des résidents en médecine familiale : étude qualitative du processus de réponse. Pédagogie Médicale 2017:18;17-24