Pédagogie Médicale **19**, 15-25 (2018)

https://doi.org/10.1051/pmed/2019002

Pédagogie Médicale

Disponible en ligne :
www.pedagogie-medicale.org

Recherche et perspectives

Jugement évaluatif : confrontation d'un modèle conceptuel à des données empiriques

Rater cognition: putting conceptual model to the test with empirical evidence

Geneviève GAUTHIER[1], Simonne COUTURE[2], et Christina ST-ONGE[1,*]

[1] Département de médecine, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Québec, Canada
[2] Département de psychologie, Université de Sherbrooke, Québec, Canada

Manuscrit soumis à la rédaction le 6 décembre 2017 ; commentaires éditoriaux formulés aux auteurs le 29 novembre 2018 et le 31 janvier 2019 ; accepté pour publication le 19 février 2019

Résumé – Contexte : Le recours au jugement des évaluateurs est de plus en plus présent en contexte d'utilisation d'une approche de formation par compétences ; toutefois sa subjectivité a souvent été critiquée. Plus récemment, les perspectives variées des évaluateurs ont commencé à être traitées comme source d'information importante et les recherches sur le jugement évaluatif (*rater cognition*) se sont multipliées. Lors d'une synthèse d'études empiriques sur le sujet, Gauthier *et al.* ont proposé un modèle conceptuel englobant une série de résultats concourants. **Objectif** : Dans le cadre de cette étude à devis mixte concomitant imbriqué (quan/QUAL), nous confrontons ce modèle théorique à des données empiriques issues d'entrevues semi-dirigées d'évaluateurs hors pair. Cette analyse vise à valider le modèle théorique et déterminer son utilité pour mieux comprendre le jugement évaluatif. **Méthodes :** Les verbatim d'entrevues audio-enregistrées de 11 participants observant et jugeant la vidéo d'une résidente lors d'une consultation avec un patient standardisé ont été codés en utilisant le modèle théorique comme arbre de codage. Les données quantitatives portant sur l'occurrence et la co-occurrence de chaque code, en général et par individu, ont été extraites et analysées. **Résultats :** Les données corroborent que l'ensemble des neuf mécanismes du modèle conceptuel sont bien représentés dans le discours des évaluateurs. Toutefois, les résultats suggèrent que le modèle avec ses neuf mécanismes indépendants ne rend pas justice à la complexité des interactions entre certains mécanismes et qu'un des mécanismes, le concept personnel de compétence, semble soutenir une grande partie des autres mécanismes.

Mots clés : jugement évaluatif , évaluation de la performance clinique , superviseurs cliniques , évaluateur

Abstract. Context: The use of rater-based assessment is increasingly present in health professions education but the subjectivity of their judgment has often been criticized. More recently, the multiple perspectives they provide have started to be treated as an important source of information. In a synthesis of empirical studies on rater cognition, Gauthier *et al.* have proposed a conceptual model encompassing a series of concurrent results. **Goal:** As part of this concomitant imbricated mixed design (quan/QUAL) study, we test the Gauthier *et al.* theoretical model on empirical data from semi-structured interviews of outstanding evaluators. This analysis aims at validating the theoretical model, refining it and determining its utility to better understand evaluative judgment. **Methods:** We coded the verbatim interviews of 11 participants observing and judging a resident's video during a consultation with a standardized patient using the theoretical model as a coding tree. Quantitative data on the occurrence and co-occurrence of each code, in general and per individual, were extracted and analyzed. **Results:** The data corroborate that all nine mechanisms of the conceptual model are well represented in the evaluator's discourse. However, the results suggest that the model with its nine independent mechanisms does not do justice to the complexity of the interactions between certain mechanisms and that one of the mechanisms, the personal concept of competence, seems to underlie many of the other mechanisms.

Keywords: rater-based assessment, performance-based assessment, rater cognition

*Correspondance et offprints : Christina ST-ONGE, 3001, 12e avenue Nord. J1H5N4 Sherbrooke, Québec, Canada.
Mailto : christina.st-onge@usherbrooke.ca.

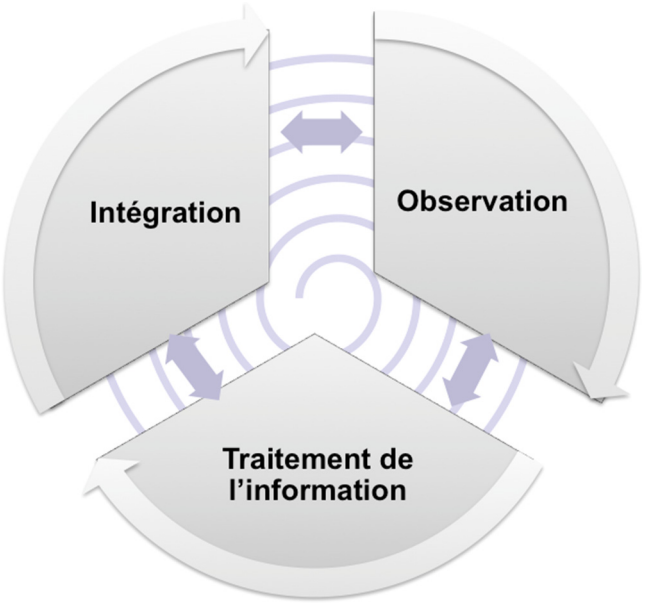

Figure 1. Phases du jugement évaluatif au cours de la formation clinique.

Introduction

Le jugement évaluatif des formateurs est de plus en plus sollicité dans le cadre de programmes qui s'inscrivent dans une approche de formation par compétences. La subjectivité inhérente aux évaluateurs est toutefois souvent critiquée. Dans le cadre de cette étude, nous tentons de comprendre davantage les processus cognitifs mobilisés lors du raisonnement évaluatif. Plus spécifiquement, nous confrontons un modèle théorique du jugement évaluatif à des données empiriques, soit des verbatim de cliniciens-enseignants qui évaluent la performance d'un résident.

Problématique et cadre conceptuel

L'approche de formation par compétences est de plus en plus présente en pédagogie des sciences de la santé [1]. Plusieurs organisations se sont même dotées de cadres de compétences ; c'est notamment le cas du Collège royal des médecins et chirurgiens du Canada, qui a développé le cadre de compétence CanMEDS pour baliser les compétences à acquérir durant la formation [2]. Dans une approche de formation par compétences, l'évaluation des apprenants est de plus en plus axée sur les attributs généraux attendus chez les futurs professionnels de la santé plutôt que centrée spécifiquement sur les connaissances [3]. Cette philosophie de l'évaluation introduit de nouveaux défis concernant l'opérationnalisation de l'évaluation et requiert le recours aux évaluateurs pour observer, interpréter et juger la performance d'apprenants [4]. Ce jugement implique un processus d'inférence de la compétence ou des compétences d'un individu à partir d'un ensemble d'actes situés ou de manifestations tangibles que constitue une performance en contexte d'interaction [5].

Le recours à des formateurs dans leur rôle d'évaluateurs pour porter un jugement évaluatif sur la performance des apprenants n'est pas nouveau en pédagogie des sciences de la santé. Toutefois, ce phénomène prend de plus en plus d'ampleur. L'évaluation de la performance clinique à partir d'observations est considérée comme une composante nécessaire de la formation des professionnels de la santé [6,7]. Dans la littérature, il a jusqu'à présent souvent été reproché à l'évaluateur d'être une source d'erreurs [8–11]. Désormais, ce changement de perspective sur l'évaluation conduit de plus en plus à le considérer comme une source importante d'informations [12–15]. Corollairement, ce changement de paradigme donne lieu à un nouveau champ de recherche en pédagogie médicale, dédié spécifiquement à la problématique du jugement évaluatif (*rater cognition*).

Afin de développer une conceptualisation et une terminologie communes des ressources cognitives des évaluateurs, sollicitées et mobilisées au cours du processus d'évaluation, Gauthier *et al.* [16,17] ont procédé à une revue intégrative de la littérature [18,19]. Le modèle proposé organise les résultats autour d'une séquence de phases, en l'occurrence une séquence d'événements au cours desquels différents mécanismes (à savoir, des processus sous-jacents spécifiques décrits par l'utilisation d'un langage similaire dans toutes les études) interviennent lors de l'évaluation de la performance du stagiaire. Le modèle résultant, qui identifie trois phases, est illustré sur la figure 1, et les neuf différents mécanismes répertoriés dans chacune de ces phases sont résumés dans le tableau I.

Le modèle proposé par Gauthier *et al.* [16] se base uniquement sur des écrits scientifiques. Les articles concernés, pour la plupart, mettent l'accent sur des mécanismes spécifiques, ou encore sur un sous-ensemble des mécanismes identifiés. Il est donc pertinent de vérifier si l'ensemble des mécanismes est mis à l'œuvre dans une situation donnée, et si l'intégration d'un plus grand ensemble de mécanismes peut révéler de nouvelles compréhensions par rapport au jugement évaluatif. Cette étude propose de vérifier empiriquement le modèle de Gauthier *et al.* [16] avec des données provenant d'entrevues semi-dirigées, réalisées auprès d'évaluateurs considérés hors pair par leurs collègues et leurs résidents [16]. Plus spécifiquement, nos objectifs étaient : (1) de vérifier si le modèle théorique est identifiable dans le corpus des données empiriques analysées et (2) d'examiner comment il peut aider à capturer la complexité du raisonnement évaluatif lors de l'observation et de l'évaluation de la performance d'une résidente.

Méthodes

Devis

Nous avons réalisé une étude à devis mixte concomitant imbriqué (quan/QUAL) [20,21] où des données qualitatives (QUAL, pour signifier leur dominance) ont été utilisées pour générer des données quantitatives (quan, pour signifier leur moindre importance) pour répondre à la question de recherche. Plus spécifiquement, nous avons

Tableau I. Cadre conceptuel du jugement évaluatif de Gauthier *et al.* [16], élaboré à partir d'une synthèse de la littérature.

Phase 1 : Observation	Phase 2 : Traitement de l'information	Phase 3 : Intégration de l'information
1. Génération automatique d'impressions des personnes	4. Le concept personnel de compétence	7. Stratégies de pondération et synthétisation de l'information
2. Formulation d'inférences de haut niveau	5. La comparaison avec des schémas d'exemples de provenance variée	8. Production de jugement en forme narrative
3. Mettre l'accent sur différentes dimensions des compétences	6. La spécificité de la tâche et du contexte	9. Traduction du jugement narratif en chiffre pour une grille d'évaluation

utilisé les verbatim de protocoles de pensée à voix haute [22] des 11 participants à l'étude de St-Onge *et al.* [15]. Ces verbatim ont été codés de manière déductive et inductive en utilisant le modèle théorique de Gauthier *et al.* [16]. Le codage a été quantifié par la suite et confronté au modèle de Gauthier *et al.* [16].

Données

Les données utilisées dans cette étude proviennent d'une étude précédente [15]. Brièvement, nous disposions de 11 entrevues audio-enregistrées et transcrites en mode verbatim, où les participants avaient observé une vidéo d'une résidente de deuxième année qui effectuait une rencontre avec un patient standardisé consultant au sujet d'une perte de poids. La vidéo a été divisée en quatre sections (entrevue, examen physique, suivi avec patient, conclusion avec médecin traitant). Entre chaque section, la vidéo était mise sur pause, et les participants devaient commenter la performance de l'étudiante ainsi que donner une note globale finale. Les 11 participants (sept hommes et quatre femmes) étaient des cliniciens-enseignants. Ils avaient une moyenne de 17,6 années d'expérience en pratique clinique de la médecine et de 15,0 ans d'expérience comme évaluateur.

Matériel et procédures

Dans le cadre de la présente étude, nous avons utilisé le modèle théorique proposé par Gauthier *et al.* [16] comme arbre de codage. Cet arbre a été traduit en français et ensuite révisé par les différents auteurs avant d'être testé sur une première entrevue. Ce codage a mené à certains ajustements de l'arbre, à l'ajout de nouveaux codes ainsi qu'à des éclaircissements dans la manière d'appliquer les codes. Toutes les entrevues ont été codées, en alternance avec des processus itératifs de validation intrajuge (cohérence interne du codeur principale) et interjuge (cohérence entre les membres de l'équipe quant à la codification des entrevues à l'aide de l'arbre de code). L'arbre de code final est présenté dans le tableau III (codage déductif utilisant le modèle conceptuel de Gauthier *et al.* [16]) et dans le tableau IV (codage inductif).

Analyses

Une fois le processus de codage complété, les données quantitatives portant sur l'occurrence et la co-occurrence de chaque code, en général et par individu, ont été extraites au moyen du logiciel Dedoose. Des analyses de fréquence ont été réalisées afin d'établir la prévalence des différents mécanismes proposés par le modèle.

Résultats

Reflet du modèle théorique dans les analyses empiriques

Les premières analyses effectuées à partir des arbres de codes montrent que les trois phases – observation, traitement et intégration de l'information – ainsi que les neuf mécanismes sont présents dans l'échantillon. En effet, comme cela est attesté par les données présentées dans le tableau II, l'ensemble des trois phases ont été observées chez tous les participants. Pour les mécanismes, neuf sont présents chez cinq des participants (1, 2, 5, 8 et 10), huit d'entre eux sont présents chez cinq autres participants (3, 6, 7, 9 et 11). Pour le participant 4, deux des mécanismes n'ont pas été observés, soit la formulation d'inférences de haut niveau et des références aux stratégies de pondération et synthétisation de l'information. Ce dernier mécanisme, qui fait référence aux différentes méthodes ou systèmes qu'emploient les évaluateurs pour assigner un poids relatif aux composantes de la performance, n'est pas toujours fait de manière consciente.

Pour chacun des neuf mécanismes, une description avec des extraits représentatifs est présentée dans le tableau III.

Éléments présents dans les données, mais non présents dans le modèle

Trois thèmes, non inclus dans le modèle ont été identifiés par l'équipe lors de l'analyse des verbatim : la pratique évaluative en rotation (lors des stages/sur les étages) (7 %), la rétroaction (5 %) et la conscience de ses propres biais (4 %). Leur description ainsi que certains extraits sont présentés dans le tableau IV.

Complexité du raisonnement évaluatif

Plusieurs extraits englobent plus d'un mécanisme. L'analyse des données empiriques à l'aide dudit modèle nous a permis d'identifier des patrons de co-occurrences entre différents mécanismes, qui suggèrent des interactions entre ces derniers, comme indiqué dans la figure 2. Une co-occurrence égale ou supérieure à 10 fois

Tableau II. Mécanismes du jugement évaluatif identifiés pour chacun des participants.

Participants	Observation			Traitement			Intégration			Nombre de mécanismes observés dans le verbatim
	1	2	3	4	5	6	7	8	9	
1	x	x	x	x	x	x	x	x	x	9
2	x	x	x	x	x	x	x	x	x	9
3	x	x	x	x	x	x		x	x	8
4	x		x	x	x	x		x	x	7
5	x	x	x	x	x	x	x	x	x	9
6		x	x	x	x	x	x	x	x	8
7	x	x	x	x	x	x	x	x		8
8	x	x	x	x	x	x	x	x	x	9
9	x	x	x	x	x	x		x	x	8
10	x	x	x	x	x	x	x	x	x	9
11	x	x	x	x	x	x		x	x	8

1 : génération automatique d'impression des personnes ; 2 : formulation d'inférences de haut niveau ; 3 : accent mis sur différentes dimensions des compétences ; 4 : concept personnel de compétence ; 5 : comparaison avec des schémas d'exemples de provenance variée ; 6 : prise en compte de la spécificité de la tâche et du contexte ; 7 : pondération et synthétisation de l'information ; 8 : production de jugement en forme narrative ; 9 : traduction du jugement narratif pour une grille d'évaluation.

entre deux mécanismes a été catégorisée comme fréquente et analysée de nouveau pour mieux comprendre les liens entre ces mécanismes. Le mécanisme qui est le plus fréquemment en co-occurrence (avec un autre mécanisme) est le cinquième, le concept personnel de compétence. Sa co-occurrence est relevée avec sept des autres mécanismes. La figure 3 est une représentation visuelle des interactions entre les différents mécanismes. Les interactions les plus fréquentes et des citations exemplaires sont présentées dans le tableau V.

Discussion

Signification des résultats

L'objectif principal de cette étude à devis mixte concomitant imbriqué (quan/QUAL) était de vérifier si le modèle conceptuel proposé par Gauthier *et al.* [16] était identifiable dans des discours d'évaluateurs commentant la performance d'une résidente junior, et si le modèle rendait adéquatement compte de la complexité du jugement évaluatif. Cette étude est une étape importante dans l'approfondissement de nos connaissances et la compréhension du jugement évaluatif de plus en plus sollicité en contexte de formation par compétences. Les données de cette étude corroborent que le processus de jugement évaluatif implique davantage que l'application de critères et de standards explicites et que le modèle conceptuel de Gauthier *et al.* [16] est bien documentable dans le discours d'évaluateurs qui commentent la performance d'une résidente. Toutefois, le modèle avec ses neuf mécanismes indépendants ne semble pas rendre justice à la complexité de la tâche du jugement évaluatif.

Les résultats de cette étude suggèrent que le processus du jugement évaluatif est encore plus complexe que ne le laissait croire le modèle proposé par Gauthier *et al.* [16]. En fait, il semblerait que le tout (jugement évaluatif dans son ensemble) est supérieur à la somme de ses mécanismes ! Les études qui avaient été retenues par Gauthier *et al.* [16] s'étaient concentrées sur un ou quelques mécanismes à la fois. Ainsi, il se peut qu'il ait été impossible d'obtenir une vision plus englobante ou holistique du jugement évaluatif, conformément à la nature des données utilisées pour élaborer ledit modèle.

Nonobstant la complexité des interactions entre les mécanismes, un mécanisme se démarque des autres – le concept personnel de compétence – si l'on en juge par sa co-occurrence avec sept autres mécanismes, c'est-à-dire presque tous les autres mécanismes du modèle. Le concept personnel de compétence, qui a été identifié de façon explicite dans de nombreuses recherches [12,23–25], semble soutenir l'articulation de nombreux mécanismes. Une meilleure compréhension de ce mécanisme clé est donc essentielle dans le contexte actuel où l'évaluation des apprenants, qui peut conditionner l'accès à la profession, repose de plus en plus sur le jugement des évaluateurs.

Selon le modèle de Gauthier *et al.* [16], il est attendu que nous retrouvions – au cœur de ce mécanisme – un rôle de facilitation dans l'évaluation de la performance d'apprenants. Dans la présente étude, nous avons observé une composante expérientielle, c'est-à-dire une opérationnalisation contextuelle de compétences qui prend en considération non seulement la performance, mais aussi l'enseignement et l'apprentissage de ces compétences dans un contexte précis. Bien que l'impact de la composante expérientielle ait été bien démontré par Kogan *et al.* [26],

Tableau III. Descriptions et manifestations des mécanismes documentés dans les verbatim.

Descriptions	Manifestations
1. Génération automatique d'impression des personnes	
Biais cognitif de généralisation (effet de halo) où des aspects sociaux ou traits personnels influencent inconsciemment le jugement	« Dans le fond, euh... t'sais, on essaie de voir si on peut lui faire confiance et tout ça, là, t'sais, quand on voit... pis comme médecin pis comme résident. Donc, clairement, c'est une fille consciencieuse, là, qui... qui fait un assez bon contact » (E5)
2. Formulation d'inférences de haut niveau	
Inférences à propos des caractéristiques ou des compétences qui sont basées sur d'autres faits qui ne sont pas articulés, justifiés ou observables et qui diffèrent selon les évaluateurs	« On voit qu'elle est nerveuse, elle se brasse sur sa chaise tout le temps, et puis on voit que c'est pas complètement fluide dans sa tête, ce qu'elle demande, parce qu'elle a... elle répète souvent la même chose, là. » (E9)
3. Mettre l'accent sur différentes dimensions des compétences	
Porter une attention à différents éléments d'une performance en fonction d'une compréhension différente des composantes de compétences	« J'ai porté mon attention sur tout. Donc, premièrement, là il y a plein de choses, là, mais une des choses, c'est l'attitude du médecin par rapport au patient, donc si elle a l'air à l'aise, si elle met le patient confortable, si elle est... t'sais, si elle est attentive au confort pis à la... au patient, pis si elle a l'air concernée, pis, bon, t'sais, toute le... ce qui englobe le... le bon rôle de... de... de médecin, pis j'ai trouvé ça correct. » (E9)
4. Le concept personnel de compétence	
Englobe une composante expérientielle avec des éléments de conception personnelle de bonne performance combinée avec des définitions opérationnelles des compétences ciblées et des façons de les enseigner	« ce que je veux qu'elle fasse, par exemple, devenue R2, c'est qu'elle soit capable de cibler, donc qu'elle attrape pas toute l'information un peu par la bande en allant chercher n'importe quoi, mais plutôt qu'elle aille le chercher en voulant aller le chercher. Fait qu'elle est pas obligée de le faire de la même façon, mais tant que ce que moi, je pense, je vois qu'elle aussi, elle y a pensé, maintenant comment est-ce qu'elle ira le chercher, ça, c'est ... ce sera propre à elle pis à comment est-ce qu'elle organise ses affaires, mais tant que je m'aperçois qu'on a voulu couvert... couvrir les mêmes sujets, c'est déjà bon. » (E1)
5. La comparaison avec des schémas d'exemples de provenance variée	
Utilisation d'exemples tel que les étudiants précédents, des collègues, d'anciens étudiants ou soi-même au même stade, pour établir des points de référence et des standards auquel on peut comparer des éléments de la performance en fonction du niveau de l'étudiant	« Ouais, mais t'sais, j'ai dans ma tête, moi, déjà où est-ce qu'un résident 1 devrait être rendu, un résident 2 devrait être rendu, un résident 3, un externe, même, quand on fait des plus jeunes. ... » (E1)
6. La spécificité de la tâche et du contexte	
La compréhension de la nature d'une tâche spécifique et de ses contraintes sur la performance se développe à travers l'expérience d'évaluation à quoi s'ajoutent les buts et le contexte de l'évaluation qui eux jouent un rôle d'éléments médiateurs sur l'évaluation	« Donc, il y a les exigences minimales auxquelles on s'attend. Il va avoir les degrés de difficulté des situations cliniques rencontrées, donc une situation clinique facile, on s'attend à avoir une bonne performance par rapport à une situation clinique plus difficile, on va être plus tolérants ou moins exigeants peut-être... » (E3)
7. Stratégies de pondération et synthétisation de l'information	
Ces stratégies sont diverses et variées ; elles incluent une moyenne de tous les éléments pertinents, une priorisation par rapport à l'importance relative des composantes d'une compétence donnée, une relativisation des aspects d'un comportement ou d'une procédure sous optimal, et une caractérisation d'un geste ou manquement qui équivalent à un échec ou perte de confiance face à la compétence de l'étudiant	« Pis à la fin de l'entrevue, ben là je regarde le nombre de lumières rouges, le nombre de lumières vertes... Q : Vous faites le décompte ? Pis là je me dis : « Bon... » Parce qu'il faut faire attention, des fois il y a des lumières rouges qui sont tellement rouges, c'est pas juste le nombre. » (E11)
8. Production de jugement en forme narrative	
Le jugement se développe en forme narrative qui prend plus souvent une forme négative et que parfois les participants relativisent en fonction de ce qui peut se	« C'est des éléments qui s'accumulent pour me faire dire que son approche clinique, elle est... elle est partielle. Pour ce problème-là, en tout cas, elle a été partielle. Son... son... sa formulation de

Tableau III. (suite).

Descriptions	Manifestations
corriger facilement, du niveau de l'étudiant ou du fait que ce n'est qu'une seule observation.	problème, elle est partielle. Euh... Elle a une hypothèse diagnostique en tête, elle a de la difficulté à... à élaborer sur d'autres hypothèses, elle a de la difficulté à aller plus loin, elle aurait besoin de beaucoup d'aide, là, pour... Je vous dirais qu'à ce niveau-là de résidence, résidence... une résidente 2, à mon point de vue, devrait être capable d'aller plus loin que ça, là. » (E10)
9. Traduction du jugement narratif pour une grille d'évaluation	
La traduction d'un jugement global en description narrative qui est subséquemment traduite en pointage ou en catégorie pour les différentes échelles d'une grille d'évaluation	« Tu vois, là, elle va être difficile... juste sur c'te cas-là... Tu vois, comme elle travaille bien, ce serait une résidente à qui je mettrais « conforme aux attentes », même si en réalité il y avait un ou deux points qui sont moins bons, pis je mettrais des commentaires ici, là. Je mettrais des commentaires du type « consciencieuse », euh... « appliquée, bon contact », pis là je... « enrichir sa réflexion diagnostique et son différentiel ». En fait, j'aurais probablement gardé « bon contact » à la fin, pour faire ma sandwich. Euh... euh... Ouais, c'est ça. Bon, les discussions et le temps de suivi. » (E5)

Tableau IV. Description et extraits pour les thèmes identifiés dans les données, mais absents dans le modèle théorique de Gauthier *et al.* [1].

Descriptions	Manifestations
La pratique évaluative en rotation (stages)	
Les évaluateurs font référence à ce qu'ils auraient fait en contexte d'évaluation en rotation (poser des questions, consulter l'équipe ou des collègues, etc.), un contexte qui est un peu différent d'une évaluation formelle ou de l'évaluation vidéo qu'ils ont à faire	T'sais, j'aurais été valider pour les questions... pour essayer de voir c'est-tu un problème de connaissances, donc essayer d'identifier un peu où est le problème. Donc, c'est des observations répétées, pis d'une fois à l'autre, ben nous, on fait la même démarche qu'on demande d'eux, là, on réfléchit sur où est le problème, pour essayer de voir c'est... qu'est-ce qui en est plus exactement. Pis c'est sûr que plus ils sont en difficulté, plus le temps de... si le temps d'observation est court, ça va être difficile de faire une rétroaction vraiment bien construite, là. » (E5)
La rétroaction	
L'impact de la rétroaction chez l'étudiant ainsi que la visée pédagogique constructive de l'évaluation joue un rôle important dans le choix des termes et des notes pour la plupart des évaluateurs	« Ben, c'est la rétroaction comme telle pis de doser, leur laisser la chance de... de s'améliorer pis de dire : « Ben, vous êtes en début de formation, pis tout ce que je vous dis là, ben ça va peut-être vous aider à vous améliorer ultérieurement », pis que c'est normal aussi. On peut atténuer un peu, là, sans noyer le poisson. » (E6)
La conscience de ses propres biais	
Les évaluateurs semblent conscients de l'existence de leur propre biais et mentionnent prendre certaines dispositions pour en contrer l'impact	« Évidemment, la personnalité va entrer en ligne de compte aussi. Il y a des gens avec qui ça va être beaucoup plus facile de... de... d'entrer en contact puis, euh... ça va être... on va être plus compatible, je devrais dire. Euh, je crois pas que ça influe beaucoup, mais... l'évaluation finale, parce qu'on va s'en rendre compte souvent pis on va essayer de faire abstraction de ça, puis avant de mettre des... si on est en désaccord ou si on pense que l'étudiant performe pas comme on devrait, je pense qu'on va se remettre beaucoup plus en question que lorsqu'on pense que l'étudiant performe bien. On va se poser la question. Est-ce que la personn... sa personnalité, est-ce que ça va entrer en ligne de compte ? Est-ce que... ? C'est-tu moi qui a... qui a un transfert négatif ou un contre-transfert négatif par rapport à cet étudiant-là, pis est-ce qu'il performe bien malgré tout ça ? Je pense que là on va se mettre beaucoup plus en réfle... en mode réflexion là-dessus. » (E3)

Mécanismes	1	2	3	4	5	6	7	8	9
1		4	5	1	24				2
2	4		11	2	19	1		1	1
3	5	11		8	34	4	2	1	1
4	1	2	8		53	24	4	7	5
5	24	19	34	53		23	16	5	15
6		1	4	24	23		3	1	1
7			2	4	16	3		6	
8		1	1	7	5	1	6		10
9	2	1	1	5	15	1		10	

Figure 2. Nombre de co-occurrences entre les différents mécanismes du jugement évaluatif. 1 : génération automatique d'impression des personnes ; 2 : formulation d'inférences de haut niveau ; 3 : accent mis sur différentes dimensions des compétences ; 4 : concept personnel de compétence ; 5 : comparaison avec des schémas d'exemples de provenance variée ; 6 : prise en compte de la spécificité de la tâche et du contexte ; 7 : pondération et synthétisation de l'information ; 8 : production de jugement en forme narrative ; 9 : traduction du jugement narratif pour une grille d'évaluation.

les autres sous-composantes de ce mécanisme demeurent à étudier de manière plus détaillée afin d'être en mesure de mieux comprendre et optimiser le processus du jugement évaluatif des superviseurs cliniques. Les résultats font écho aux discussions théoriques et pratiques reliées au concept de compétence [5,27,28], et plus particulièrement aux défis d'attitrer avec certitude une compétence à un étudiant qui produit une action dans un contexte donné.

Limites

La présente étude comporte certaines limites. L'étude repose sur une analyse secondaire de données. La collecte de données initiales n'avait pas été prévue afin de tester un modèle théorique ; il s'agissait plutôt d'une étude s'appuyant sur la philosophie de la théorisation ancrée [29]. Toutefois, le fait que nous ayons pu observer le modèle dans ces données est indicateur du fait que le modèle est représentatif de la réalité. L'utilisation d'un protocole de pensée à voix haute comporte certaines limites [30,31], mais ce protocole demeure un outil essentiel pour explorer la pensée humaine [32]. Enfin, comme tous les participants

Tableau V. Extraits représentatifs de l'interaction entre un mécanisme principal et un mécanisme secondaire lors du jugement évaluatif.

Mécanisme principal	Interactions avec mécanisme secondaire	Manifestations
Le concept personnel de compétence	Mettre l'accent sur différentes dimensions des compétences	« Mais à ce stade-ci, je dirais, je vais devoir faire très, très attention à : 1) comment elle va parler à son patient ; pis 2) comment elle va présenter. Parce que si jamais elle le présente d'une manière désorganisée, que son... sa recherche de diagnostic différentiel est... est faible, c'est-à-dire qu'elle pense à un ou deux trucs pis qu'elle laisse beaucoup de diagnostics... euh... euh... non explorés, à ce moment-là je vais dire : « Cette... cette résidente n'atteint pas les objectifs. » Donc, là je serais en... en... je serais très, très, très attentif à ce qui va arriver pour voir de quel côté elle va. » (E11)
	Formulation d'inférences de haut niveau	« On dirait qu'elle a une petite routine de base d'examen, pis t'sais, tout le monde a une espèce de séquence de base, pis là faut que tu rajoutes par-dessus ça les choses par rapport à la raison de consultation, pis là on dirait qu'elle a une routine de base pis qu'elle a pas vraiment élargi trop sa raison de consultation à elle. » (E9)
	Génération automatique d'impression des personnes	« En fait, j'essayais de voir qu'est-ce qui se passait dans sa tête pour deviner est-ce qu'elle a des hypothèses diagnostiques ou si elle fait juste un peu aller à la pêche, pis c'est en général ce qui nous permet de voir quelqu'un qui a déjà un... un processus d'organisation. » (E8)
	La comparaison avec des schémas d'exemples de provenance variée	« Fait qu'on compare à nous, pis on compare aussi avec la... T'sais, moi, je les vois, effectivement, en deuxième, troisième, quatrième année, pis toutes les cinq années de résidence, donc la progression attendue, on a déjà quand même une bonne idée. » (E8)

Tableau V. (suite).

Mécanisme principal	Interactions avec mécanisme secondaire	Manifestations
	La spécificité de la tâche et du contexte	« T'sais, si c'était quelqu'un qui venait parce qu'il avait... je sais pas... un abcès au pied, ça pourrait éventuellement passer, mais son problème était pulmonaire. Elle pouvait pas faire ça. T'sais, elle laisse de côté des indices importants pour faire un diagnostic différentiel. Peut-être qu'elle trouverait rien, mais peut-être qu'elle le trouverait, pis à ce moment-là ce serait un outil diagnostic de plus. C'est pour ça que je dis qu'elle est insuffisante. » (E7)
	Production de jugement en forme narrative	« Bien, compte tenu que pour moi, c'est pas ésotérique comme situation, la perte de poids, euh... bon, je me serais attendue qu'au début de la résidence 2, elle soit en mesure de... de me présenter ça de façon plus étoffée. » (E6)
	Stratégies de pondération et synthétisation de l'information	« je dirais que ça va faire plus une personne pense telle chose avec les bons points et les mauvais points. Quelqu'un qui met beaucoup, beaucoup de valeur sur les connaissances fondamentales, pour dire quelque chose, pis que le résident a pas nécessairement une bonne affaire là-dessus, ben il va certainement avoir une mauvai... moins bonne évaluation de ce patron-là, de cet évaluateur-là, alors que par exemple, un autre qui, lui, dit « ben moi, c'est beaucoup plus comment est-ce qu'elle fait son histoire, son examen », ben les deux auront pas la même vision, mais souvent ça va s'accorder parce qu'on voit tous la même affaire. Fait que je pense qu'une évaluation, c'est... un évaluateur donne déjà une bonne évaluation ; plusieurs évaluateurs, ça donne souvent une meilleure évaluation. » (E1)
Formulation d'inférences de haut niveau	Génération automatique d'impression des personnes	« En fait, j'essayais de voir qu'est-ce qui se passait dans sa tête pour deviner est-ce qu'elle a des hypothèses diagnostiques ou si elle fait juste un peu aller à la pêche, pis c'est en général ce qui nous permet de voir quelqu'un qui a déjà un... un processus d'organisation » (E8)
La comparaison avec des schémas d'exemples de provenance variée	La spécificité de la tâche et du contexte	« Ben, parce que là... euh... pour moi, perte de poids, c'est quelque chose qu'on pense que les... les externes, donc quand ils débutent leur résidence, ils devraient être capables de faire au moins un certain nombre de choses. Donc, s'il y a une année de fait... de faite, je m'attendrais à ce que ce soit plus... meilleur que ça. » (E6)
Production de jugement en forme narrative	Traduction du jugement narratif pour une grille d'évaluation	« Donc, élaboration du plan, identification, connaissances fondamentales, connaissances cliniques, anamnèse dirigée et pertinente, moi, je trouve qu'elle l'a pas, donc ça, elle est inconstante, inférieure aux attentes. Formulation du problème, moi, je mettrais « inférieure ». Élaboration d'hypothèses diagnostiques, c'est inférieur aux attentes. Formulation et justification des conduites, élaboration, tout est inférieur aux attentes au point de vue expertise. Au point de vue communication, habiletés de communication,

Tableau V. (suite).

Mécanisme principal	Interactions avec mécanisme secondaire	Manifestations
		ça, ça va. Professionnalisme, empathie, ça va. Hum... Donc, maintenant que je l'ai revu un peu point par point, moi, je mettrais une « inconstante, inférieure aux attentes » (E11)
Génération automatique d'impression des personnes	La comparaison avec des schémas d'exemples de provenance variée	« Ben, en fait, souvent... en fait, ça, j'ai pas réfléchi à ça avant aujourd'hui, mais c'est comme quand on voit des... des étudiants ou tout ça, on voit s'ils vont à la pêche ou s'ils sont vraiment organisés. Donc, elle, elle va pas juste à... elle a vraiment une structure de base vraiment organisée, pis quand elle est p'us certaine d'où aller, là elle... elle utilise des petites techniques de pêche, là, si on veut, là, mais qui sont par ailleurs pas n'importe où, là. » (E5)
La comparaison avec des schémas d'exemples de provenance variée	Traduction du jugement narratif pour une grille d'évaluation	« Donc, si c'était une R1 en début d'année, ça serait une bonne performance... euh... mais pour une R2, moi, qui vient... qui a quelques stages... euh... qui a fait quelques stages, moi, je dirais que... j'hésite, là, entre « conforme aux attentes » dans la globalité avec beaucoup de points à... à... à... comment dirais-je... à améliorer, ou « inférieure aux attentes. » (E11)

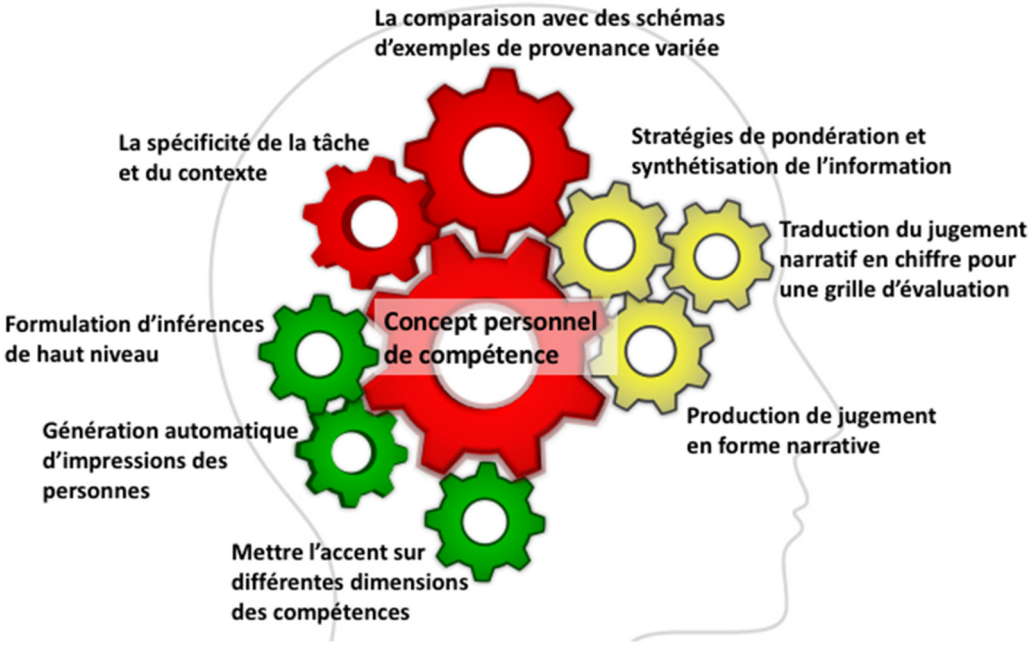

Figure 3. Représentation des interactions entre les mécanismes en jeu lors du jugement évaluatif.

sont issus de la même université, il est possible qu'ils partagent une même culture universitaire qui ait influencé indûment les résultats.

Forces

La fiabilité de nos résultats (reproductivité des résultats dans des conditions similaires [33] ; stabilité et constance des résultats [34]) a été assurée par les efforts de standardisation de la compréhension et de l'application de l'arbre de code. La crédibilité des résultats (justesse, pertinence et concordance entre les observations empiriques et leur interprétation [35]) est assurée par le travail de révision constante du codage par des membres de l'équipe. Nous avons tenté de produire une description riche de notre échantillon et de notre contexte pour que les lecteurs puissent juger de la transférabilité des résultats à leurs propres contextes. Finalement, nous avons tenté d'instaurer le plus possible une objectivité dans notre analyse des données en utilisant un arbre de code détaillé.

Conclusion

La conceptualisation du jugement évaluatif comme unité d'analyse, plutôt qu'outil, permet aux chercheurs de bien se situer les uns par rapport aux autres, et permet aussi une meilleure contextualisation de leurs découvertes les uns par rapport aux autres. Les résultats de cette étude nous permettent de suggérer que le modèle théorique du jugement évaluatif proposé par Gauthier *et al.* [16] est, somme toute, une bonne représentation du processus employé par les superviseurs cliniques lors de l'évaluation de la performance de leurs apprenants. Même s'il faut admettre que des études ultérieures seront nécessaires pour raffiner notre compréhension du jugement, il est important de mobiliser ces connaissances afin de les investir dans l'élaboration d'outils d'évaluation (utilisables par les superviseurs cliniques) ainsi que dans le cadre de la formation de superviseurs cliniques pour la tâche complexe qu'est l'évaluation de la performance des apprenants, dans un contexte où il faut souligner qu'une telle tâche devient particulièrement essentielle avec l'adoption d'approches de formation par compétences.

Contributions

Geneviève Gauthier et Christina St-Onge ont participé à la conception du protocole de recherche, à l'analyse et l'interprétation des résultats ainsi qu'à l'écriture du manuscrit. Simonne Couture a participé au recueil et à l'analyse des données, ainsi qu'à l'écriture du manuscrit.

Financement

Nous tenons à remercier la Chaire de recherche en pédagogie médicale Paul Grand'Maison de la Société des médecins de l'Université de Sherbrooke ainsi que le Conseil de recherches en sciences humaines du Canada pour le soutien financier de ce projet.

Avertissement éditorial

Les données qui font l'objet du présent manuscrit ont été utilisées dans un autre projet de recherche publié dans *Advances in Health Sciences Education*. (St-Onge C., Chamberland M., Lévesque A., Varpio L., Expectations, Observations, and the Cognitive Processes that Bind Them: Expert Assessment of Examinee Performance. Adv Health Sci Educ Theory Pract 2016;21:627-642.) Cependant, elles ont été analysées sous un nouvel angle dans le présent manuscrit.

Remerciements

Nous souhaitons remercier Lara Varpio et Martine Chamberland pour leur participation à la collecte de données initiales, Linda Bergeron pour son soutien dans la réalisation de l'étude et ses révisions critiques du manuscrit, ainsi que Marilyne Bolduc pour la mise en page du manuscrit.

Conflits d'intérêts

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

Approbation éthique

Ce projet a été approuvé par le Comité d'éthique de la recherche – Éducation et sciences sociales en date du 26 mars 2013 (n° de l'enregistrement : CER-ESS 2013-09)

Références

1. Irby DM, Cooke M, O'Brien BC. Calls for reform of medical education by the Carnegie Foundation for the Advancement of Teaching: 1910 and 2010. Acad Med 2010;85:220-7.
2. Royal College of Physicians and Surgeons of Canada. Competency-based medical education. 2011 [On Line]. Disponible sur http://www.royalcollege.ca/portal/page/portal/rc/common/documents/educational_initiatives/cbme.pdf.
3. ten Cate O, Scheele F. Competency-based postgraduate training: Can we bridge the gap between theory and clinical practice? Acad Med 2007;82:542-7.
4. ten Cate TJO, Snell L, Carraccio C. Medical competence: The interplay between individual ability and the health care environment. Med Teach 2010;32:669-75.
5. Kahn S, Rey B. La notion de compétence : une approche épistémologique. Éduc Francoph 2016; 44(2):4-18.
6. Harris P, Bhanji F, Topps M, Hart D, Snee S, Touchie C, *et al.* Evolving concepts of assessment in a competency-based world. Med Teach 2017;39:603-8.
7. Lockyer J, Carraccio C, Chan M-K, Ross S, Lieberman S, Franck J, *et al.* Core principles of assessment in competency-based medical education. Med Teach 2017;39:609-16.
8. Downing SM. Threats to the validity of clinical teaching assessments: What about rater error? Med Educ 2005;39:353-355.

9. Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. Constructing a validity argument for the mini-clinical evaluation exercise: A review of the research. Acad Med 2010;85:1453-61.
10. Norcini JJ. Current perspectives in assessment: The assessment of performance at work. Med Educ 2005;39:880-9.
11. Pelgrim EA, Kramer AWM, Mokkink HG, Van den Elsen L, Grol RPTM, Van der Vleuten CPM. In-training assessment using direct observation of single-patient encounters: A literature review. Adv Health Sci Educ Theory Pract 2011;16:131-42.
12. Berendonk C, Stalmeijer RE, Schuwirth LWT. Expertise in performance assessment: Assessors' perspectives. Adv Health Sci Educ Theory Pract 2013;18:559-71.
13. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the "black box" differently: Assessor cognition from three research perspectives. Med Educ 2014;48:1055-68.
14. Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G. Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. Acad Med 2010;85:780-6.
15. St-Onge C, Chamberland M, Lévesque A, Varpio L. Expectations, observations, and the cognitive processes that bind them: Expert assessment of examinee performance. Adv Health Sci Educ Theory Pract 2016;21: 627-42.
16. Gauthier G, St-Onge C, Tavares W. Rater cognition: Review and integration of research findings. Med Educ 2016;50:511-22.
17. Gauthier G, St-Onge C, Dory V. Synthèse et conceptualisation des processus cognitifs du jugement évaluatif de l'enseignant clinicien. Pédagogie Médicale 2016;17:261-7.
18. Cooper HM. Scientific guidelines for conducting integrative research reviews. Rev Educ Res 1982;52:291-302.
19. Whittemore R, Knafl K. The integrative review: Updated methodology. J Adv Nurs 2005;52:546-53.
20. Fortin M-F., Gagnon J. Fondements et étapes du processus de recherche : méthodes quantitatives et qualitatives. 3e édition. Montréal : Chenelière Éducation, 2015.
21. Creswell JW. Research design: Qualitative, quantitative, and mixed methods approaches. Thousand Oaks: SAGE Publications, 2014.
22. Ericsson KA, Simon HA. Verbal reports as data. Psychol Rev 1980;87(3):215-51.
23. Ginsburg S, Regehr G, Lingard L. Basing the evaluation of professionalism on observable behaviors: A cautionary tale. Acad Med 2004;79:S1-S4.
24. Govaerts MJB, Van de Wiel MWJ, Schuwirth LWT, Van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: Raters' performance theories and constructs. Adv Health Sci Educ Theory Pract 2013;18:375-96.
25. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E, Holmboe E. Opening the black box of clinical skills assessment *via* observation: A conceptual model. Med Educ 2011;45:1048-60.
26. Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. Acad Med 2010;85:S25-S28.
27. Rey B. « Compétence » et « compétence professionnelle ». Rech Form 2009;60:103-16.
28. Rey B. La notion de compétence en éducation et formation : enjeux et problèmes. Bruxelles : De Boeck, 2014.
29. Charmaz K. Grounded theory in the 21st century: Applications for advancing social justice studies, in The Sage handbook of qualitative research, Denzin NK, Lincoln YS, Editors. Thousand Oaks, California: Sage Publications, 2005, Vol. 3, p. 507-535.
30. van Someren MW, Barnard YF, Sandberg J. The think aloud method: A practical guide to modelling cognitive processes. London: Academic Press, 1994.
31. Eva K, Brooks L, Norman G. Forward reasoning as a hallmark of expertise in medicine: Logical, psychological, phenomenological inconsistencies, in Advances in psychology research, Shohov SP, Editor. New York, NY: Nova Science Publishers, Inc., 2002, Vol. 8, p. 41-69.
32. Hoppmann TK. Examining the "point of frustration". The think-aloud method applied to online search tasks. Qual Quant 2009;43:211-24.
33. Lincoln YS, Guba EG. Naturalistic Inquiry. Newbury Park (CA): Sage, 1985.
34. Fortin MF. Fondements du processus de recherche : méthodes quantitatives et qualitatives. Montréal : Chenelière Éducation, 2010.
35. Laperrière A. Les critères de scientificité des méthodes qualitatives, in La recherche qualitative : enjeux épistémologiques et méthodologiques, Poupart J, Groulx LH, Deslauriers JP, Laperrière A, Mayer R, Pires AP, Editors. Boucherville : Gaétan Morin, 1997, p. 365-389.

Citation de l'article : Gauthier G., Couture S., St-Onge C., Jugement évaluatif : confrontation d'un modèle conceptuel à des données empiriques. Pédagogie Médicale 2018:19;15-25