Les facteurs influençant la variabilité du jugement des enseignants-cliniciens, dans le contexte de l'évaluation de la performance des apprenants en médecine : une recension systématique de la littérature avec synthèse descriptive.

David Adjo, md.

Thèse soumise à l'Université d'Ottawa dans le cadre des exigences du programme de Maitrise en Art de l'Éducation, concentration Enseignement aux professionnels de la santé

Études supérieures
Faculté d'éducation
Université d'Ottawa

Table des matières

1. Liste des tableaux

2. Liste des Figures

3. Légende

AFMC : Association des facultés de médecine du Canada.

CanMEDS: Canadian Medical Education Direction for Specialists.

CRMCC : Collège Royal des Médecins et Chirurgien du Canada.

DPT : Théorie du double processus de la prise de décision.

USA : United States of America (États-Unis d'Amérique).

UK: United Kingdom (Royaume-Uni).

4. Résumé

L'intégration de l'« approche par compétences » dans l'éducation médicale a rencontré plusieurs défis, notamment en ce qui concerne l'utilisation de l'évaluation basée sur la performance pour mesurer les compétences. En milieu clinique, cette évaluation repose sur le jugement des évaluateurs, qui est souvent influencé par divers facteurs, entraînant des variations indésirables des scores. Peu d'études systématiques existent sur les facteurs influençant la variabilité du jugement des enseignants-cliniciens et leurs biais cognitifs dans la formation médicale. Cette recension systématique de la littérature avec synthèse descriptive a donc examiné ces facteurs et leur impact sur le processus décisionnel des enseignants-cliniciens. Elle a identifié plusieurs éléments clés influençant cette variabilité : manque d'uniformité des normes, impressions automatiques, approche déductive pour formuler des inférences, mise en avant de différentes dimensions des compétences, traduction du jugement narratif en chiffres, et influence de l'émotion de l'évaluateur. La variabilité du jugement résulte de diverses sources d'erreurs, de biais et de la combinaison de plusieurs critères d'évaluation.

5. Abstract

The integration of the "competency-based approach" into medical education has encountered several challenges, not least the use of performance-based assessment to measure competency. In clinical settings, this assessment relies on the judgment of assessors, which is often influenced by a variety of factors, leading to undesirable variations in scores. Few systematic studies exist on the factors influencing the variability of teacher-clinician judgment and their cognitive biases in medical training. This systematic literature review with descriptive synthesis therefore examined these factors and their impact on teacher-clinicians' decision-making processes. It identified several key elements influencing this variability: lack of uniformity of standards, automatic impressions, deductive approach to formulating inferences, emphasis on different dimensions of competence, translation of narrative judgment into numbers, and influence of the evaluator's emotion. The variability of judgment results from various sources of error, bias and the combination of several evaluation criteria.

6. Remerciements

Je souhaite exprimer ma gratitude à mon premier directeur de thèse, le Professeur Éric Dionne, pour son aide précieuse et ses conseils avisés tout au long de l'initiation de ce travail. Merci Éric d'avoir été à mon écoute et de m'avoir prodigué des enseignements sur la recherche, ce qui m'a permis de mener à bien cette œuvre.

Je tiens à exprimer ma profonde gratitude à ma directrice de thèse Professeur Isabelle Bourgeois pour son aide précieuse, son soutien indéfectible et ses conseils avisés durant la préparation de ce travail. Chère Isabelle, je te remercie très sincèrement pour l'enseignement de qualité que tu m'as dispensé dans le domaine de la recherche, pour ton accompagnement tout au long de ce projet, pour ton écoute attentive lorsque ma vie sociale a été perturbée par des situations difficiles et tes encouragements tout au long de ce travail. Ça été un privilège de t'avoir pour directrice de thèse. J'ai été profondément impressionné par tes qualités de respect et d'empathie. Tu as été pour moi un mentor et en tant que futur enseignant, tu continueras d'être ma source d'inspiration. Merci beaucoup Isabelle.

Je remercie le comité de thèse pour ses remarques pertinentes, ses critiques constructives et son accompagnement attentif, qui ont été des éléments clés dans l'aboutissement de cette recherche.

Chère Marie Hélène Chomienne, permet moi de te remercier pour ta disponibilité, tes conseils précieux, ton expertise et ta bonne humeur motivante et communicante qui ont été pour moi un atout indéniable.

Chère Mariette Théberge, merci beaucoup pour tes conseils, tes orientations éclairées, ta bienveillance et tes suggestions précieuses.

J'exprime toute ma reconnaissance à vous mes collègues pour votre soutien et vos conseils dans la réalisation de cette œuvre.

Je tiens à remercier ma famille, en particulier mes deux filles, Éliya et Yona, mon épouse, ma chère mère ainsi que mon père, pour leur soutien indéfectible tout au long de la préparation et de la réalisation de cette œuvre.

Chère Laurence, un grand merci pour ton soutien sans faille durant les situations difficiles que nous avons traversées et qui auraient pu entraver la réalisation de ce travail. J'espère que ce travail te rendra fière.

Un merci tout spécial à toi, mon père, qui m'as toujours encouragé et cru en mes qualités d'enseignant. Merci, chère maman, pour la personne extraordinaire que tu as été pour moi.

Chapitre 1– La problématique.

Depuis plus de vingt ans, une réforme mondiale est en marche dans les systèmes éducatifs en médecine pour intégrer « *l'approche par compétences* »[1](AFMC, 2012; Irby et al., 2010). Cette réforme est née de la volonté des éducateurs médicaux d'exploiter la notion de compétences à des fins pédagogiques. L'intégration de cette approche par compétences dans les programmes d'étude a amené des changements dans les pratiques pédagogiques et évaluatives, de telle sorte que sa mise en œuvre présente des défis (Lacasse et al., 2016).

L'un de ces défis majeurs vient du fait que la mise en œuvre de l'évaluation basée sur la compétence met en évidence le rôle de l'évaluation basée sur la performance[2] (Harris et al., 2017). En effet, l'évaluation des performances permet d'inférer le niveau de compétences, puisque les compétences ne peuvent pas être évaluées directement. Pour évaluer cette performance dans le milieu clinique, l'évaluation en milieu de travail est communément utilisée. Celle-ci « incorpore l'évaluation de tâches cliniques complexes dans la pratique quotidienne par l'observation directe des stagiaires lorsqu'ils interagissent de manière authentique avec les patients dans des contextes cliniques réels » (Gingerich et al., 2014, p. 1056). Autrement dit, en médecine, les évaluations des compétences cliniques qui sont essentiellement des évaluations globales, sont fondées sur les observations d'échantillons de preuves de performances cliniques (Williams et al., 2003). Ces observations fournissent à l'évaluateur les informations nécessaires pour juger de la progression des apprenants[3]. Par conséquent, l'évaluation de la performance

[1] L'approche par compétences est un concept pédagogique qui met l'accent sur l'acquisition de compétences plutôt que celui du contenu, qui était autrefois la norme dans l'approche par objectif.

[2] L'évaluation des performances se rapporte au Référentiel CanMEDS. Ce référentiel définit et décrit les normes en vigueur en matière de performance dans les différentes compétences. Les compétences dont il est question s'articulent autour des sept rôles du médecin compétent : « Expert médical (le rôle intégrateur), communicateur, collaborateur, leader santé, promoteur de la santé, érudit, professionnel. » (Frank et al., 2015, p. 8)

Le référentiel CanMEDS : était un sigle à l'origine, qui signifiait « *Canadian Medical Education Direction for Specialists* ». On a délaissé officiellement l'acronyme en 2004. Désormais on parle simplement de « CanMEDS ». (Frank et al., 2015)

[3] Le terme « les apprenants » fera référence aux étudiants de médecine et aux résidents en spécialité.

érigée sur le modèle des évaluations en milieu de travail est considérée comme un élément crucial de la stratégie d'évaluation en éducation médicale (Gingerich et al., 2014) et elle est basée essentiellement sur le jugement d'évaluateurs (Gomez-Garibello & Young, 2018 ; Lacasse et al., 2016). Or, de nombreux auteurs s'accordent pour dire que l'évaluation basée sur le jugement de l'évaluateur (qui est aussi appelé enseignant-clinicien)[4] est un processus complexe. Ce processus peut être influencé par une multitude de facteurs tels que les facteurs cognitifs, sociaux, émotionnels et contextuels (Frank et al., 2010; Gingerich et al., 2014; Gomez-Garibello & Young, 2018), qui constituent des sources indésirables de variations de scores dans les évaluations des performances cliniques. Williams et al. (2003) expliquent à ce sujet que les processus de notation contribuent de manière plus importante à la variance des cotes de performances que les problèmes de mesure qui sont liés à la conception des instruments (Williams et al., 2003).

La variabilité des jugements des enseignants-cliniciens dans le domaine de l'éducation médicale a suscité un intérêt croissant et a donné lieu à un nouveau domaine d'étude axé sur les processus cognitifs des évaluateurs. Cette variabilité est examinée à travers différentes perspectives conceptuelles telles que les trois perspectives de la variabilité du jugement des évaluateurs proposées par Gingerich et al. (2014), ainsi que les modélisations du raisonnement évaluatif de Gauthier et al. (2016) et St-Onge et al. (2016). Ces perspectives conceptuelles ont permis d'attribuer cette variabilité à des facteurs sociaux, environnementaux et cognitifs. Ces facteurs se manifestent dans les choix faits par l'enseignant quant aux situations d'évaluation, aux critères de correction ou à l'interprétation de ces critères (Fontaine & Loye, 2017; Romainville, 2011). Le concept de biais cognitif en tant que facteur de cette variabilité peut être défini comme « la manière dont l'être humain traite l'information et développe une hypothèse, en particulier lorsqu'il évalue des tâches complexes » (Feinsilber et al., 2019, p. 1). Autrement dit, un écart significatif par rapport à la réalité perçue par l'évaluateur lors du traitement de cette information pourrait créer des problèmes importants (Feinsilber et al., 2019), tels que l'effet de «

[4] Médecin chargé de superviser, de favoriser l'apprentissage, de donner une rétroaction et d'agir comme modèle en matière de valeurs, de rôles CanMEDS et d'enthousiasme pour la médecine dans le contexte de la pratique clinique. (CRMCC, 2015)

contraste[5] », de « **représentativité** », de « **halo** », et enfin de « **halo inversé** ou **effet du diable** », que nous aborderons plus en détail dans le cadre conceptuel.

Pour Gauthier et al. (2016), le processus d'évaluation d'une performance est complexe et multidimensionnel. Faherty et al. (2020) expriment des préoccupations plus marquées au sujet du processus cognitif. En effet selon ces auteurs, le processus cognitif conduisant à la variabilité du jugement des évaluateurs, en favorisant l'expression des biais cognitifs demeure flou. Pourtant, en comprenant mieux les processus cognitifs qui conduisent à l'expression de ces biais, on pourrait apporter des modifications aux pratiques d'évaluation afin d'améliorer la justesse des décisions d'évaluation et la valeur des rétroactions formatives échangées avec les apprenants, et *in fine,* augmenter la sécurité des patients (Gingerich et al., 2014). Une exploration de la littérature dans cette perspective nous a indiqué que bien que de telles recherches en vue de comprendre ce processus soient fréquentes en psychologie cognitive et sociale (Kocovski, 2009) et dans divers domaines professionnels, elles le sont cependant moins en éducation médicale. En effet, il semble y avoir peu d'études systématiques qui récapitulent et mettent à jour les connaissances actuelles des facteurs influençant la variabilité du jugement des enseignants-cliniciens et surtout comment ces facteurs favorisent leurs biais cognitifs dans le contexte particulier de la formation médicale.

C'est donc pour répondre à ce problème que cette étude vise à faire un état des lieux des facteurs qui influencent la variabilité du jugement des enseignants-cliniciens et la manière dont ces facteurs s'expriment au cours du processus décisionnel d'enseignants-cliniciens chargés de l'évaluation de la performance d'apprenants en médecine. L'étude porte sur l'ensemble des facteurs (c'est-à-dire, cognitifs, sociaux, émotionnels, contextuels et environnementaux) qui peuvent influencer la variabilité du jugement des évaluateurs. Dans cette perspective, nous avons abordé les questions de recherche suivantes :

[5] L'effet de contraste est un biais de jugement qui fait que la représentation d'une information est affectée par la représentation d'une information de nature opposée produite antérieurement ou en même temps. Cet effet ne s'applique pas uniquement à la représentation, il intervient également dans la cognition, car le jugement que l'on porte sur une personne ou sur soi-même ne l'est jamais de manière absolue, mais relative.

Par exemple : un évaluateur aura tendance à mieux noter la même personne si elle passe après un mauvais candidat qu'après un très bon candidat. Cet effet joue surtout si la personne est un candidat moyen.

Quels sont les facteurs et les biais qui influencent la variabilité du jugement des enseignants-cliniciens dans le contexte de l'évaluation de la performance clinique des apprenants en médecine ?

- De quelles manières ces facteurs influencent-ils la variabilité du jugement des enseignants-cliniciens?
- De quelles manières ces facteurs influencent le double processus de décision d'enseignants-cliniciens dans le contexte de l'évaluation de la performance d'apprenants en médecine?

Chapitre 2 – Le cadre théorique.

1. L'éducation médicale dans le contexte de l'approche par compétences

L'éducation médicale a pour objectif premier de former des médecins compétents[6] (Gingerich et al., 2014) par la mise en œuvre d'approches, de modalités et de stratégies pédagogiques pour la formation en médecine (CRMCC, 2015). Dans cette perspective, l'éducation médicale a subi des réformes qui ont abouti à l'adoption de l'approche par compétences. Celle-ci préconise une éducation basée sur les aptitudes, l'évaluation de la performance et la culture d'apprentissage continu (Naik et al., 2012). Selon Lacasse et al. (2016), la mise en œuvre de cette nouvelle approche dans les programmes d'étude en médecine représente des défis rencontrés à trois niveaux, tels que décrits dans les paragraphes qui suivent.

Le premier défi rencontré est la définition du terme de compétences qui varie selon les écrits recensés. Cependant, Lacasse et al. (2016) expliquent que la définition apportée par Tardif (2006) selon laquelle, « la compétence est un savoir-agir complexe prenant appui sur la mobilisation et la combinaison efficaces d'une variété de ressources internes et externes à l'intérieur d'une famille de situations » (p. 22), semble susciter une large adhésion dans la communauté des éducateurs médicaux dans la francophonie.

Le second défi est le bouleversement pédagogique lié à l'implantation de l'approche par compétences. En effet, Lacasse et al. (2016) expliquent que l'approche par compétences a permis la création du référentiel de compétences nationales *CanMEDS* et une réorganisation de l'enseignement pour amener l'apprenant à jouer un rôle actif [7] dans son apprentissage. Pour ce faire, on assiste à une réorganisation des stages cliniques et à des changements dans la posture de l'enseignant, qui au lieu d'être un simple dispensateur de cours, accompagne désormais

[6] Médecins compétents : Il s'agit de médecins qui répondent aux besoins des patients et qui sont capables d'améliorer les soins de santé de leurs communautés.

[7] Rôle actif dans son apprentissage en mobilisant des ressources internes (ses connaissances intellectuelles) et externes (la littérature, le matériel médical, le personnel soignant) en vue d'atteindre les compétences attendues.

l'apprenant dans sa démarche réflexive en participant à un processus d'évaluation continue, avec des rétroactions régulières.

Enfin, le troisième défi est la nécessité d'une progression du système d'évaluation (Lacasse et al., 2016). Ainsi, bien que le référenciel CanMEDS présente un recueil des compétences attendues en fin de programme de formation et des stratégies d'évaluation permettant un suivi de l'apprentissage des apprenants en contexte réel (Laurin et al., 2013), il n'en demeure pas moins que ce système d'évaluation est généralement considéré inefficace pour évaluer adéquatement la performance clinique (Gingerich et al., 2014). En outre, dans cette approche par compétences, l'évaluation prend un sens nouveau car elle implique de recourir à des critères multiples compte tenu du fait que le caractère intégrateur et combinatoire d'une compétence suppose que les différents domaines de connaissances soient évalués séparément (Nguyen & Blais, 2007)[8] par la pratique de l'observation et de rétroactions permanentes.

2. L'évaluation de la performance des apprenants.

2.1. La définition de l'évaluation de la performance.

Abernot (1993) définit l'acte d'évaluer comme une activité qui consiste à juger de la compétence d'un apprenant à travers sa performance, à extrapoler sa compétence à partir de comportements observables et/ou d'un projet réalisé. La définition proposée par Abernot (1993) intègre plusieurs dimensions de l'approche par compétences, dont celle de la compétence, qui est érigée au rang d'objet d'évaluation. La notion d'évaluation est intrinsèquement liée au concept de compétences. De ce fait, la tâche d'évaluer une compétence est ardue étant donné qu'il s'agit d'un concept polysémique faisant l'objet de nombreuses controverses dans le monde de l'éducation (Rey, 2014).

Enfin, il ressort de ces observations qu'évaluer une performance découlant d'une compétence représente un véritable défi, autant en raison de la complexité de l'objet d'évaluation

[8] Évaluation formative : il s'agit, à des moments variables ou de manière continue, de donner à l'apprenant et à l'enseignant des informations objectives sur la nature et la valeur des apprentissages réalisés, afin que l'un et l'autre ajustent et optimisent, respectivement, leurs stratégies d'apprentissage ou leurs interventions pédagogiques. Elle est généralement assimilée à la fonction pédagogique de l'évaluation (Jouquan & Bail, 2003, p. 40).

que pour les exigences de la mise en œuvre, qui selon certains experts devraient se traduire par une démarche bien planifiée avec des critères explicites. Legendre (2005) traduit bien ces exigences en affirmant que : « L'évaluation est une activité humaine basée sur une démarche permettant de porter un jugement, à partir de normes ou de critères explicitement établis, sur la valeur de la performance d'un apprenant en vue de décisions pédagogiques ou administratives » (p.631). Cette définition met en exergue la notion de démarche évaluative qu'on peut modéliser sous forme de continuum linéaire ou cyclique et d'étapes intégrées.

2.2. La démarche évaluative.

Les figures comparatives 2 et 3 (voir figure 2 et 3 en annexe) présentent quelques exemples de démarches évaluatives en vigueur en éducation médicale. Ceux-ci permettent de mieux cerner les similarités et les différences entre les approches évaluatives au cœur de ces démarches. Compte tenu du fort ancrage du processus d'évaluation de Prégent et al. (2009) dans le concept d'approche par compétences et qui de surcroit est en vigueur en éducation médicale, nous nous servirons du processus d'évaluation de Prégent comme type de description de la démarche évaluative en éducation médicale.

2.3. Le jugement évaluatif.

La démarche évaluative est caractérisée par sa complexité et sa multi dimensionnalité comme le démontrent les travaux de St-Onge et al. (2016). Au cœur de cette démarche se trouve un jugement évaluatif tout aussi complexe. Pour mieux comprendre ce jugement, de nombreux chercheurs ont proposé une modélisation du raisonnement évaluatif issue pour les uns des écrits scientifiques et pour les autres des données empiriques (Gauthier et al., 2016; St-Onge et al., 2016). Ainsi on distingue deux approches du raisonnement évaluatif qui tentent d'articuler les processus cognitifs sous-jacents:

- Premièrement, l'approche empirique de St-Onge et al. (2016), présente un modèle conceptuel des processus cognitifs impliqués dans une évaluation (voir figure 2). Cette approche démontre l'interaction de nombreux processus cognitifs au cours du raisonnement évaluatif pour gérer adéquatement les fortes tensions qui existent entre la représentation de subjectivité et d'objectivité de l'évaluation.

- L'approche théorique de Gauthier et al. (2016) présente les mécanismes employés lors de l'observation des performances dans le cadre d'une démarche évaluative (voir tableaux 9, 10 et 11, phase 1, 2, 3 en annexe).

3. La définition des biais cognitifs.

Le jugement est considéré comme une entité construite modulable et modifiable[9] , se traduisant par une combinaison pondérée d'informations (Chasseigne, 2007; Hammond et al., 1964; Morewedge & Kahneman, 2010). Dans cette perspective, Morewedge and Kahneman (2010), alignés sur les travaux de Kahneman and Frederick (2002) expliquent que « les biais de jugement peuvent toujours être décrits comme une surpondération de certains aspects de l'information et une sous-pondération ou une négligence d'autres aspects, par rapport à un critère d'exactitude ou de cohérence logique » (p. 435). Ainsi, selon un postulat relatif au traitement cognitif de l'information, il semble que les informations recevant le plus d'activation seraient plus pondérées qu'elles ne le méritent et les informations pertinentes qui sont moins activées par le contexte associatif seraient sous-pondérées ou négligées (Morewedge & Kahneman, 2010; Mussweiler, 2007; Weber & Johnson, 2006). Regardons maintenant comment surviennent et persistent ces erreurs de pondération de l'information traitée à l'origine des biais chez les enseignants-cliniciens, à travers la théorie du double processus de la prise de décision.

4. La théorie du double processus de la prise de décision (DPT).

La théorie du double processus de la prise de décision est largement utilisée par les chercheurs dans divers domaines pour expliquer les bases conceptuelles des biais du raisonnement humain (Croskerry et al., 2013; Morewedge & Kahneman, 2010). En effet, de nombreux modèles proposés pour rendre compte des erreurs de jugement invoquent une perspective à double processus de la prise de décision (DPT) ou à système double, dans lequel les processus automatiques ou intuitifs « *système un* » génèrent des représentations et des jugements provisoires qui peuvent être acceptés, bloqués ou corrigés par des processus contrôlés ou analytiques « *système deux* » (Croskerry et al., 2013; Morewedge & Kahneman, 2010).

[9] Modulable et modifiable en fonction des processus de traitement cognitifs appliqués aux informations.

Ainsi, les biais proviennent du fait que le mode de prise de décision intuitif est caractérisé par des méthodes heuristiques[10] et que les intuitions défaillantes[11] (issues de certaines caractéristiques principales de la mémoire associative) dans le « système un » ne parviennent pas à être détectées et corrigées par le « système deux »[12] (Croskerry et al., 2013; Morewedge & Kahneman, 2010). En résumé, nous sommes prédisposés à utiliser des heuristiques qui sont les plus souvent utiles, mais qui sont cependant vulnérables aux erreurs[13] (Croskerry et al., 2013).

5. Les facteurs environnementaux et sociaux.

Dans le cadre de l'évaluation en milieu de travail, l'enseignant et l'apprenant interagissent dans un environnement dynamique. Cet environnement n'est pas toujours neutre; bien au contraire, il peut être la source d'un certain nombre de facteurs de distraction susceptibles d'interférer dans le processus d'évaluation. Ces divers facteurs de distraction peuvent être de nature environnementale et sociale comme nous l'explique Williams et al. (2003). Selon ces auteurs, il s'agit notamment de la pression et du manque de temps lié aux exigences des activités cliniques; par exemple, l'environnement de travail surchargé de l'urgence qui contraste avec la clinique au bureau (Wood, 2014). Il émane de ces facteurs de pression et de manque de temps, une rareté de l'observation des performances et une négligence du temps accordé à la réflexion de l'évaluateur. Ceci a pour corollaire, l'accumulation de preuves anecdotiques et par conséquent, une réduction de la qualité des informations sous-jacentes intégrées à la réflexion de l'évaluateur sur les progrès réalisés par l'apprenant. Les auteurs concèdent cependant que, bien que l'influence néfaste des effets de la pression et du manque de temps n'ont pas encore été directement prouvés dans les notations en éducation médicale, ceux-ci ont été démontrés en psychologie.

---

[10] Les heuristiques sont des modes de pensée abrégés et représentent un mécanisme adaptatif qui nous permet d'économiser du temps et des efforts lors de la prise de décisions au quotidien (Croskerry et al., 2013).

[11] Les « systèmes Un », sont plus couramment utilisés. Ils sont rapides, généralement efficaces, mais aussi plus susceptibles d'échouer. Comme elles sont en grande partie inconscientes, les erreurs - quand elles se produisent, passe le plus souvent inaperçues et par conséquent, parviennent rarement à être corrigées(Croskerry et al., 2013; Evans & Frankish, 2009; Gilovich, Griffin, & Kahneman, 2002; Kahneman, 2011).

[12] Les « systèmes Deux », sont relativement fiables, sûrs et efficaces, mais lents et gourmands en ressources (Croskerry et al., 2013)

[13] Les erreurs systématiques sont appelées des biais et plus d'une centaine d'entre elles sont des biais cognitifs (Jenicek, 2010).

Par ailleurs, hormis la pression et le manque de temps en lien avec les exigences des activités cliniques, on distingue parmi les facteurs environnementaux et sociaux les contextes de travail suivants :

- Un environnement éducatif ne suscitant pas le perfectionnement du corps professoral peut faire en sorte que celui-ci ne s'approprie pas les outils d'évaluation. Un tel contexte ne favorise pas la réduction de certains biais tel que l'effet de halo (Kogan, et al., 2014 ; Raj & Thorn, 2014 St-Onge et al., 2014).
- Un environnement éducatif dans lequel le processus évaluatif favorise l'observation de plusieurs dimensions de la compétence en même temps, peut occasionner une surcharge cognitive chez l'évaluateur, favorisant un jugement de type idiosyncrasique susceptible de réduire la discrimination des différents niveaux de performances. Tavares and Eva (2014) montrent que la fiabilité inter-évaluateurs diminue à mesure que le nombre de dimensions de la performance qui doivent être prises en compte par les évaluateurs augmente.
- Un environnement éducatif dans lequel l'influence d'un évaluateur trop persuasif ou ayant une forte opinion dans une réunion de programme destinée à débattre de la promotion d'un stagiaire peut constituer un facteur de distraction.
- Un environnement éducatif dans lequel il y a une contribution collective de l'équipe soignante aux tâches quotidiennes de soins aux patients peut rendre difficile l'évaluation de la performance clinique des apprenants.
- Le fait que l'apprenant se sait observé peut rendre sa performance clinique artificielle ; ainsi, sa performance peut être stimulée le temps de l'observation ou, dans certains cas, inhibée, selon le type de personnalité.
- L'émotion de l'évaluateur lors du processus évaluatif est considérée par Gomez-Garibello and Young (2018) comme un facteur social pouvant conduire l'évaluateur à une prise de décision biaisée. Ceci contribue aux bruits aléatoires des évaluations.

En somme, les facteurs environnementaux et sociaux sont susceptibles de façonner, d'influencer, de surcharger et de distraire les enseignants pendant le processus évaluatif (Gauthier et al., 2016).

6. Les perspectives conceptuelles de la variabilité du jugement des évaluateurs.

Dans le cadre de la recherche en éducation médicale, trois perspectives conceptuelles distinctes ont émergé des discussions itératives de groupes de chercheurs internationaux. Ces perspectives conceptuelles non mutuellement exclusives, portent sur les origines et les solutions possibles à la variabilité des jugements d'évaluation (Gingerich et al., 2014) :

6.1. Le postulat de l'évaluateur qui peut être formé.

Dans ce postulat, l'évaluateur est considéré comme un instrument d'évaluation (Gomez-Garibello & Young, 2018) dont la variabilité ou le manque de fidélité fait principalement référence à des lacunes de formation corrigibles. La recension de la littérature présentée par Gingerich et al. (2014) décrit comment le jugement des évaluateurs peut être influencé par des biais systématiques découlant d'une formation inadéquate.

6.2. Le postulat de l'évaluateur idiosyncrasique (Gingerich et al., 2014).

Dans ce postulat, les évaluateurs peuvent se positionner en tant que décideurs experts, susceptibles de mobiliser des informations variées qu'ils traduiront grâce à leur expertise afin de porter un jugement (Gomez-Garibello & Young, 2018). Ainsi selon Gingerich et al. (2014), la variabilité du jugement des évaluateurs découle de l'expertise qu'ils ont développé. Cette expertise unique développée par chaque évaluateur s'exprime à travers des inférences et les extrapolations que ceux-ci utilisent dans l'interprétation de la performance des apprenants.

6.3. Le postulat de l'évaluateur faillible.

Le postulat selon lequel « *l'évaluateur est faillible* » est rapporté par Gingerich et al. (2014). Il suggère la responsabilité des limites de la cognition humaine dans la variabilité du jugement des évaluateurs. Ainsi, dans ce postulat, le jugement des évaluateurs peut être facilement influencé par le contexte socio professionnel. Cela vient du fait que l'évaluateur interagit nécessairement avec ce qu'il observe, ce qui implique qu'il lui est impossible de faire des observations objectives des performances.

Par ailleurs, selon Gomez-Garibello and Young (2018), l'émotion doit être considérée comme le facteur supplémentaire intrinsèquement lié à ce processus cognitif de l'évaluateur. En

effet, l'interaction entre des considérations cognitives, relationnelles (sociales) et émotionnelles (c'est-à-dire les émotions suscitées par l'évaluation en termes de contrôle et de valeur) semble mieux représenter la manière dont le processus d'évaluation se déroule (Gomez-Garibello & Young, 2018) et contribue à biaiser le jugement des évaluateurs. Le processus cognitif régi par cette interaction est à l'origine des biais responsables de problèmes d'évaluation basés sur le jugement (tels que le manque de transparence, d'équité, de fiabilité et de validité). Wilson et Brekke (1994) considèrent ce type de biais comme des « contaminations mentales » qui se distinguent de l'incapacité à connaître ou à appliquer des règles normatives d'inférences. Ces « *contaminations mentales* » sont difficiles à éviter car elles résultent des propriétés fondamentales de la cognition humaine, dont le jugement est de nature comparative. Ainsi, l'évaluation est relative, en ce sens qu'elle fait référence à une comparaison de la cible évaluée avec une norme ou un standard pertinent. Cette relativité de l'évaluation a des implications dans la survenue de biais (Gingerich et al.,2014).

Enfin, compte tenu de ce qui précède nous nous demandons :

a) De quelles manières ces facteurs peuvent-ils influencer le processus cognitif du jugement des enseignants-cliniciens?

b) De quelles manières ces biais peuvent-ils influencer le processus cognitif des enseignants-cliniciens?

Les perspectives conceptuelles soulevées par Gingerich et al. (2014) suggèrent que la variabilité du jugement des évaluateurs est due à des erreurs de mesure provenant de l'expression de biais cognitifs, comme l'expliquent les postulats de l'évaluateur pouvant être formé et par celui de l'évaluateur faillible. Cependant, il ne semble pas y avoir d'erreur de mesure chez l'évaluateur idiosyncrasique, pour qui la mesure émane d'un prisme[14] de vérité dont le nombre de facettes dépend en grande partie du niveau d'expertise atteint par l'évaluateur.

[14] Un prisme est un polyèdre ayant deux faces isométriques et parallèles appelées bases et possédant des quadrilatères en guise de faces latérales. Chaque quadrilatère représente une parcelle de vérité dans le jugement de la performance. Le nombre de parcelles de vérité dans le jugement dépend en grande partie du niveau d'expertise de l'enseignant. Il ne s'agit donc pas ici d'une erreur de mesure de l'instrument enseignant clinicien.

Chapitre 3 – La méthodologie.

1. L'approche méthodologique.

Dans le cadre de cette étude, nous avons réalisé une recension systématique des écrits avec synthèse descriptive. Cependant, il existe plusieurs approches méthodologiques de recensions des écrits, telles que : la recension systématique, la recension intégrative ou mixte, la recension générale, narrative ou traditionnelle, la « *umbrella Review* », la méta-analyse, ou l'étude de la portée (Grant & Booth, 2009). L'approche systématique qui représente notre choix méthodologique, se définit comme « une synthèse de la littérature scientifique en réponse à une question précise. Elle utilise des méthodes explicites de recherche, de sélection et d'analyse des données » (Zaugg et al., 2014, p. 656). Autrement dit, il s'agit d'une approche rigoureuse, transparente et reproductible permettant de collecter, d'évaluer et de synthétiser systématiquement les preuves scientifiques issues des résultats de plusieurs études (Landry et al., 2008) et dont le but est de mettre le lecteur à jour au niveau des écrits publiés sur un sujet donné. Une recension systématique ne doit pas être confondue avec une recension générale « dans laquelle la recherche bibliographique n'est en général pas exhaustive et qui représente plus l'opinion d'un expert ou d'un groupe d'experts » (Zaugg et al., 2014, p. 656).

Les recensions systématiques produisent une analyse critique d'un ensemble de connaissances liées au sujet exprimé en tant que questions de recherche avec une plus grande fiabilité et une plus grande précision. De ce fait, elles permettent de réduire au minimum l'effet des biais ou des erreurs systématiques ayant pu s'immiscer dans les études primaires, puisqu'il est possible de combiner les résultats de plusieurs études (Akobeng, 2005). Les caractéristiques de la recension systématique concordent bien avec les objectifs de notre étude. Selon le type d'études incluses dans la recension, les données peuvent être synthétisées de manière quantitative, semi-quantitative ou descriptive. Or lorsque les critères de jugement dans les études ne sont pas homogènes, ou lorsque celles-ci présentent des risques de biais importants, il est plus approprié de réaliser une synthèse «semi- quantitative» ou une synthèse descriptive (Zaugg et al., 2014).

La synthèse descriptive correspond mieux aux objectifs de cette recension car elle nous permet de faire le point sur nos questions de recherche sans avoir recours à des résultats

quantitatifs : « Elle correspond à une présentation structurée des caractéristiques des études, qui commence par une analyse descriptive de chaque étude, en ce qui concerne sa méthodologie, l'intervention évaluée ou encore ses résultats. » (Zaugg et al., 2014, p. 664). Cette synthèse descriptive bien que qualitative intègrera à la fois des études qualitatives, quantitatives et mixtes.

Pour réaliser cette recension systématique de la littérature, nous avons adopté une stratégie de recherche suggérée par les lignes directrices de Dickson et al. (2014), qui décrivent comment structurer la démarche de la recension systématique. Ainsi, la littérature étudiée a été obtenue grâce à une recherche systématique des écrits en appliquant une série d'étapes rigoureuses et reproductibles, afin de recueillir, d'analyser et de synthétiser les documents selon l'approche suivante :

- La recherche bibliographique.
- Le processus de sélection des articles.
- L'extraction et l'analyse des données.

2. La recherche bibliographique.

Les mots clés : La recension de la littérature en vue de la production de la problématique et du cadre conceptuel sur notre sujet d'étude, nous a permis d'identifier les mots clés suivant : évaluation performances ; jugement évaluatif ; raisonnement évaluatif ; biais des évaluateurs ; étudiants de médecine / résidents spécialité; variabilité ; « Performance assessment; *Rater cognition; rater-based assessment; cognitive processes; judgment; rater bias; Medical students /residents speciality;* Undergraduate; Postgraduate; variability*;* assessor / Assess*; variance /Vari* ».

Les critères d'inclusion : Ont été incluses dans cette étude:

(a) les études publiées dans des revues ayant fait l'objet d'une évaluation par les pairs entre 2011 et 2021 en français et en en anglais, quel que soit le type de devis (qualitatif et quantitatif).

(b) les études dont le titre, le résumé et les mots clés contiennent les mots-clés en lien avec la variabilité du jugement des évaluateurs identifiés ci-dessus.

(c) les études dont le titre, le résumé et les mots clés contiennent les termes « étudiant de médecine à l'externat et /ou résident de toutes spécialités » ou leurs synonymes.

Les critères d'exclusion : Ont été exclues de cette étude:

(a) les études dont le titre, le résumé et les mots clés n'incluent pas les mots-clés en lien avec la variabilité du jugement des évaluateurs identifiés ci-dessus.

(b) les études dont le titre, le résumé et les mots clés, utilisent des mots-clés en lien avec la variabilité du jugement évaluateurs identifiés ci-dessus mais, n'ont pas été menées auprès d'enseignants-cliniciens en médecine.

(c) les études dont le titre, le résumé et les mots clés, utilisent des mots-clés en lien avec la variabilité du jugement évaluateurs identifiés ci-dessus mais ne semblent pas se concentrer sur les questions de la variabilité du jugement des enseignants-cliniciens.

(d) les commentaires d'articles ou autres publications non-empiriques ainsi que les articles de la littérature grise.

Les bases de données électroniques:

S'appuyant sur ces mots clés, la recherche de la littérature a été amorcée à l'aide des bases de données électroniques d'articles scientifiques (PubMed, ERIC, CINAHL, PsycINFO et Web of Science, Cochrane Library, et Scopus) en anglais et en français de l'Université d'Ottawa. Les articles publiés de 2011 à 2021 ont fait l'objet de la recherche, afin de nous focaliser sur un corpus récent tenant compte des avancées sur le plan des stratégies d'interventions récentes pour le contrôle des biais cognitifs. Notre stratégie de recherche a été confirmée et validée par un bibliothécaire (voir figure 5, 6 et 7 en annexe).

3. Le processus de sélection des articles.

Dans cette phase, les publications extraites des bases de données ont été importées dans le logiciel EndNote, ce qui a permis d'éliminer les doublons. Ces articles été ensuite sélectionnés

à partir des critères d'inclusion et d'exclusion, en fonction des informations retrouvées dans leurs titres, résumés et mots clés, ainsi :

- Dans une première étape, nous avons fait la sélection des articles à partir des titres et des mots clés.
- Dans une deuxième étape, les articles retenus lors de l'étape précédentes ont été sélectionnés à partir des résumés.

À chacune des étapes sus-mentionnées nous avons pris la décision, soit « OUI-Inclure », soit « NON-Exclure » ou soit « JE NE SAIS PAS ». Si le résumé ne nous permettait pas de décider, alors une troisième étape nous a permis d'effectuer la sélection selon une lecture intégrale des articles en question.

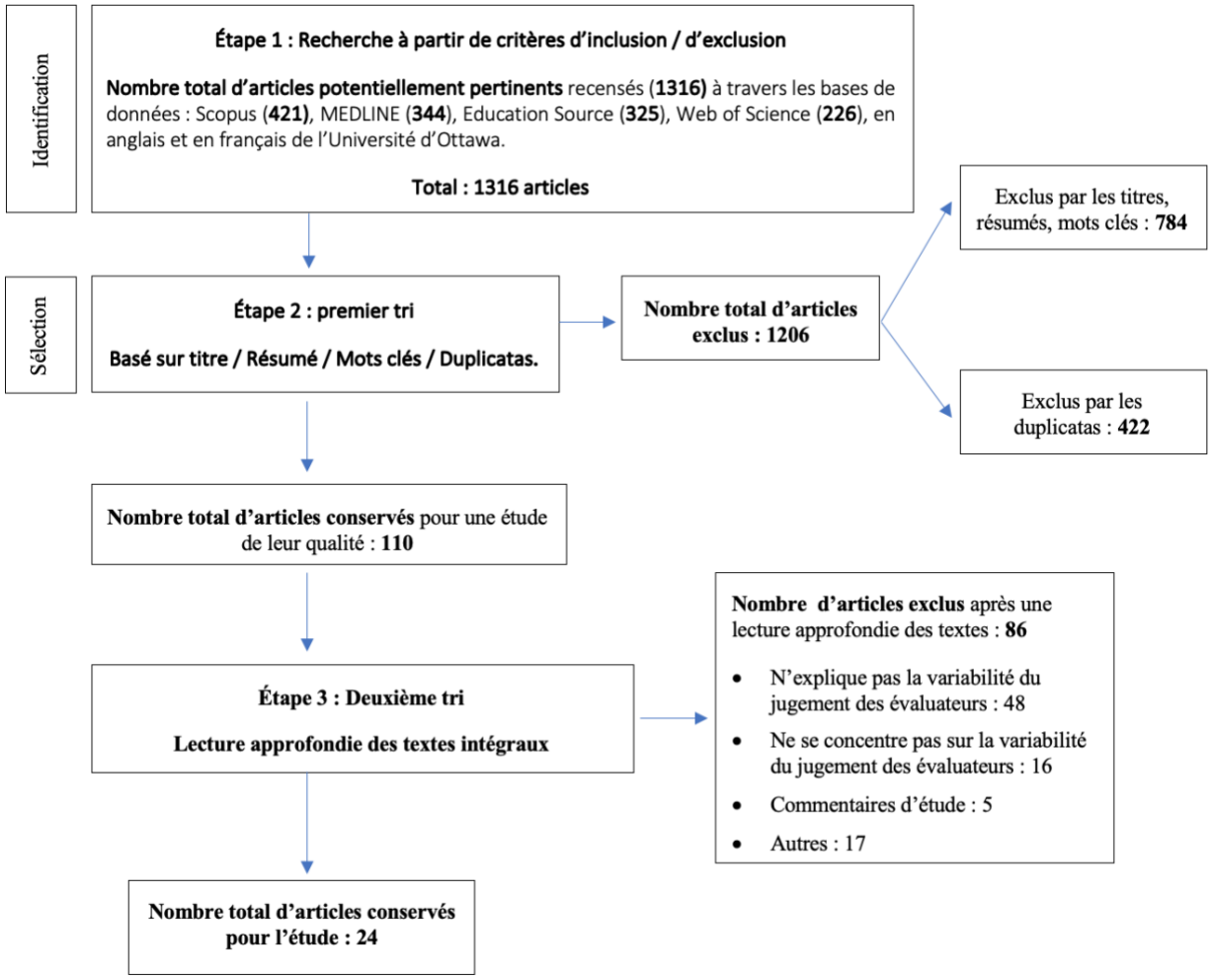
Par ailleurs, nous avons complété notre sélection en consultant la bibliographie des articles sélectionnés afin de nous assurer de ne pas manquer d'articles pertinents. La liste des études exclues a été conservée ainsi que les raisons de leur exclusion à des fins de transparence.

En pratique, le processus de sélection a été effectué par une équipe composée de deux chercheurs dont un junior, qui s'est attelé à effectuer une sélection des articles à la lumière des critères d'inclusion et d'exclusion présentés ci-dessus (voir figure 1en annexe). Le chercheur sénior a été chargé de la révision de 10% du matériel. Les décisions de sélection prises par les deux chercheurs étant accordé à plus de 80 %, le chercheur junior a pu poursuivre seul la sélection des articles. Puisque les articles ont déjà fait l'objet d'une évaluation par les pairs, nous n'avons pas procéder à une évaluation supplémentaire de leur qualité. Ainsi, tous les articles sélectionnés ont été soumis à ce processus, conformément aux critères d'inclusion établis. A la suite de cette étape 24 articles ont été retenus dont 9 études qualitatives, 13 études quantitatives et 2 études mixtes.

Figure 1

Diagramme de flux de la recension systématique de la littérature :

Organigramme illustrant les étapes de la recherche documentaire et du processus de sélection par lequel les études seront identifiées.



4. L'extraction des données.

Les 24 articles retenus (voir tableaux 4, 5 et 6 en annexe) ont fait l'objet d'une extraction des données permettant de répondre à nos questions de recherche. Cette extraction de données était particulièrement susceptible d'être une source d'erreurs et c'est la raison pour laquelle nous

avons privilégié une approche à un observateur, mais avec une discussion et validation avec un 2e observateur.

L'extraction des données qualitatives qui ont servi à l'analyse a impliqué la saisie et la classification des unités de sens présentes dans les articles retenus. Nous nous sommes servis du logiciel Covidence afin de simplifier l'extraction des caractéristiques et résultats des études et l'exportation des données et des références. Dans un premier temps, une lecture initiale du corpus nous a permis d'identifier certaines sections d'intérêt quant à leur pertinence pour notre objet d'étude et nos questions de recherche. Nous avons procédé ensuite à une seconde lecture, qui visait cette fois à extraire les « unités de sens » du corpus de recherche. Lors de cette seconde lecture, nous avons utilisé une grille d'analyse pour l'extraction des unités de sens. Cette grille était organisée selon nos questions de recherche ainsi que leurs indicateurs ou sous-thèmes. La grille d'analyse qui comprenait l'ensemble des données extraites nous a permis de grouper les informations tirées du corpus et de les répertorier dans un premier temps, et ensuite, de les synthétiser. Ce processus d'extraction de données avec la grille d'analyse nous a permis de structurer la seconde lecture de sorte qu'il a été possible de coder tout en lisant, nous évitant ainsi d'avoir une multitude de données à coder ou à étiqueter *a posteriori*. De plus, elle nous a permis de mettre de côté des informations moins pertinentes en donnant une vue d'ensemble des données recueillies. La classification des unités de sens s'est effectuée en partie lors de leur extraction, puisque chaque unité de sens a été inscrite sous une question de recherche ou un concept.

5. La synthèse des données.

L'analyse des données qualitatives extraites de notre corpus de recherche nous a permis de synthétiser nos données en codant et en catégorisant les unités de sens extraites des documents. Cette analyse de contenu a été réalisée à partir de notre grille d'analyse. Les codes ont été déterminés *a priori* grâce aux indicateurs ou dimensions retenues lors de notre recension des écrits. Ces codes ont permis de résumer les unités de sens à l'aide d'un ou de plusieurs mots clés. Les *catégories* ont été tirées du cadre conceptuel ou théorique de notre étude et ont constitué des éléments de réponse de nos questions de recherche. Ces catégories ont facilité le regroupement des codes en unités d'analyse plus vastes. Il est également à noter que la grille

d'analyse est dynamique : de nouveaux codes ou catégories qui n'avaient pas été prévus ont été générés de façon inductive au cours de l'analyse et ont été ajoutés en fonction des unités de sens extraites des documents.

Chapitre 4 – Les résultats.

Après avoir exploré les multiples facettes des facteurs influençant la variabilité du jugement des enseignants dans le domaine de l'éducation médicale, il est temps de passer à l'examen des résultats. Cette section offre un aperçu détaillé des découvertes issues de notre analyse approfondie. Nous examinerons en détail les différents facteurs et biais identifiés, ainsi que leurs implications pour la pratique éducative médicale. À travers cette exploration, nous chercherons à ouvrir des pistes de réflexion pour l'amélioration continue des processus d'évaluation en éducation médicale.

1. Le manque d'uniformité dans l'utilisation des normes de référence.

Plusieurs cadres de référence sont employés par les évaluateurs lorsque vient le moment d'évaluer les compétences des apprenants. Ces cadres semblent regrouper des connaissances composites sur les compétences attendues des stagiaires à différentes étapes de leur apprentissage. Ils sont le fruit de l'expérience accumulée par les évaluateurs au fil des années d'enseignement avec des stagiaires de niveaux de formation similaires (Yeates et al., 2013; Kogan et al., 2011; St-Onge et al., 2016; Lee et al., 2017; Forte et al., 2021). La manière dont ces évaluateurs mettent en œuvre ces cadres de référence est complexe, dynamique et très variable, ce qui peut produire des jugements différents d'un évaluateur à l'autre et d'un apprenant à l'autre (Kogan et al., 2011; Yeates et al., 2013; Lee et al., 2017). En pratique, les évaluateurs comparent des combinaisons d'observations différentes et relativement uniques des apprenants à des critères de compétence souvent incertains (voire ambigus), élaborés de manière distincte, et impliquant certains filtres individuels, tels que : (i) les expériences antérieures de l'évaluateur; (ii) la performance d'autres médecins, et (iii) l'interprétation variée des normes[15] de performance du domaine de pratique en question (Kogan et al., 2011; Yeates et al., 2013). Les critères de compétences élaborés de cette manière peuvent être classés en critères internes (les attentes des

[15] Ces normes de performance, englobent les critères du Collège royal et les objectifs fournis aux évaluateurs par le programme (St-Onge et al., 2016).

évaluateurs) et externes (les critères institutionnels), qui sont constamment comparés aux performances observées des étudiants par les enseignants au fil des années lors des évaluations. (St-Onge et al., 2016; Yeates et al., 2013). Les détails de ces critères sont exposés dans les sections qui suivent (voir tableau 1).

Tableau 1.

Sommaire des principaux résultats concernant le manque d'uniformité dans les normes de référence des évaluateurs.

Article	1.1 Les attentes des évaluateurs	1.2 Les critères institutionnels
Forte et al., 2021	Conception personnelle de bonne performance	Comparaison à la performance antérieure de l'étudiant Comparaison de la performance du résident évalué à celle de ses pairs à un stade similaire de la formation
Gingerich et al., 2018		Comparaison à la performance antérieure de l'étudiant Comparaison de la performance du résident évalué à celle de ses pairs à un stade similaire de la formation
Kogan et al., 2011	Cadre de référence fondé sur ses propres performances	Comparaison à la performance antérieure de l'étudiant Comparaison de la performance du résident évalué à celle de ses pairs à un stade similaire de la formation
Lee et al.,2017	Collection personnelle et distinctive d'expériences, de connaissances et d'instincts	
St-Onge et al., 2016	Collection personnelle et distinctive d'expériences, de connaissances et d'instincts Internalisation de critères résultant de l'assimilation de normes professionnelles	
Yeates et al., 2013	Collection personnelle et distinctive d'expériences, de connaissances et d'instincts	Comparaison à la performance antérieure de l'étudiant Comparaison de la performance du résident évalué à celle de ses pairs à un stade similaire de la formation
Yeates et al., 2015		Comparaison à la performance antérieure de l'étudiant

		Comparaison de la performance du résident évalué à celle de ses pairs à un stade similaire de la formation

1.1. Les attentes des évaluateurs

Les attentes des évaluateurs[16] sont des sources d'informations internes intégrées aux processus cognitifs de l'évaluateur. Ces sources d'informations internes sont définies comme des éléments de données[17] pertinents pour l'évaluation provenant de l'évaluateur lui-même, et non de la personne évaluée. Ces éléments émanent de leur collection personnelle et distinctive d'expériences, de connaissances et d'instincts (Forte et al., 2021; Kogan et al., 2011; Lee et al., 2017; St-Onge et al., 2016; Yeates et al., 2013), parfois qualifiés de "sentiments" ou de "réactions instinctives", d'après les commentaires de certains évaluateurs : « A lot of it is just instinct. A lot of it is when I've been a patient myself what I've looked for in a good doctor. » (Kogan et al., 2011, p. 1051)

Ces sources d'informations internes sont principalement axées sur les attentes individuelles des évaluateurs concernant ce qu'un candidat doit accomplir ou démontrer. Ces attentes sont généralement liées à la propre pratique de l'évaluateur, soit en tant que professionnel de la santé, soit en se remémorant sa propre expérience en tant que stagiaire. Ces attentes peuvent aussi être le résultat de l'internalisation de critères externes résultant de la compréhension et de l'assimilation des normes professionnelles (St-Onge et al., 2016), à en juger par le témoignage de certains évaluateurs :

It's mostly that (pause) in fact that (pause) all the information, I want her to get the information that I think I would need to make my diagnosis. (E1). I'm starting to think that she's completely not on the right track, or at least not on the same track as I am, because I (pause) I'm thinking lungs. (E7). (St-Onge et al., 2016, p.635)

Ainsi, certains évaluateurs ont développé des éléments de conception personnelle de bonne performance, de sorte qu'ils s'appuient souvent sur un cadre de référence normatif qui

[16] Voir la figure 2 dans le chapitre des annexes.

[17] Ces éléments de donnés sont des connaissances composites des compétences qu'un stagiaire devrait avoir à certains stades de leur apprentissage. (Yeates et al., 2013; Kogan et al., 2011; St-Onge et al., 2016; Lee et al., 2017; Forte et al., 2021)

leur est propre et à partir duquel ils distinguent les aspect positifs ou négatifs de la performance des stagiaires (Kogan et al., 2011). Pour ce faire, ils comparent les performances des résidents à leur propre perception, qui englobe leur appréciation de leurs forces ou limites cliniques, de leur performance passée en tant que résidents et des attentes qui leur étaient fixées à ces étapes de la formation (Kogan et al., 2011; Forte et al., 2021). Ainsi par exemple, la conception personnelle de bonne performance amène les évaluateurs à rechercher certains critères, dont : l'autonomie des stagiaires, se reflétant dans leur capacité à assumer une pratique non supervisés : « If I was going to give someone 'I did not need to be there'... then they would need to able to function as an independent physician. » (Forte et al., 2021, p.89) ; la communication interpersonnelle, qui est perçue comme cruciale par les évaluateurs et considérée comme le socle de la relation avec le patient; la façon dont les évaluateurs souhaiteraient recevoir des soins en tant que patient; la précision du diagnostic et l'observance thérapeutique du patient (Kogan et al., 2011). Dans d'autres cas, les évaluateurs se considèrent eux même comme le cadre de référence (Kogan et al., 2011; Forte et al., 2021).

1.2. Les critères institutionnels

Les critères externes[18] sont définis comme des éléments d'information que les évaluateurs prennent en compte lors des évaluations, et qui ne sont pas fondés sur leurs propres expériences personnelles. En d'autres termes, ces critères proviennent de sources extérieures à l'évaluateur, telles que des grilles d'évaluations et des attentes des institutions d'agrément ou d'accréditation. En outre, les évaluateurs prennent en considération d'autres critères externes. Cela inclut, par exemple, la recherche d'une amélioration ou d'une progression au fil du temps, en comparant les performances actuelles des résidents avec celles qu'ils ont eues précédemment avec le même évaluateur, comme l'affirment ceux-ci : «... If its someone that I've worked with before I'm looking to see whether or not I feel that they are progressing along a certain trajectory, so have they improved compared to the last time I worked with them. » (Forte et al., 2021, p. 89)

D'autres évaluateurs ont souvent tendance à comparer la performance du résident évalué à celle de ses pairs à un stade similaire de la formation (Kogan et al., 2011; Yeates et al., 2013; Yeates et al., 2015; Gingerich et al., 2018). Les évaluateurs qui se considèrent "expérimentés"

[18] Voir la figure 2 dans le chapitre des annexes.

ont particulièrement tendance à utiliser des modèles basés sur les performances d'anciens résidents (Forte et al., 2021). Ces comparaisons relatives influencent le processus d'évaluation des compétences (Yeates et al., 2013 ; Yeates et al., 2015). En effet, celles-ci génèrent souvent des effets de contraste significatifs tant pour les performances excellentes que pour celles qui sont plus faibles (Gingerich et al., 2018). Ainsi, évaluer une bonne performance après une mauvaise performance entraîne une plus grande occurrence de commentaires positifs. En revanche, lorsque la bonne performance est évaluée indépendamment, on constate une plus grande proportion de commentaires comportant des suggestions d'amélioration. Par ailleurs, lorsqu'une performance faible est précédée d'une bonne performance, la proportion de commentaires critiques est plus élevée. À l'inverse, en l'absence de préalable, la proportion de commentaires mettant en avant les aspects réussis est plus importante (Gingerich et al., 2018). Cette recherche indique que les effets de contraste peuvent conduire à une différenciation notable à la fois dans les scores attribués et dans les commentaires écrits (Gingerich et al., 2018). Ces effets de contraste peuvent affecter les évaluations même après qu'un évaluateur ait été exposé à une seule performance passée ou à la moyenne des performances antérieures (Yeates et al., 2015).

2. La génération automatique d'impressions.

L'acte de percevoir[19] d'autres personnes implique des jugements sociaux (Gingerich et al., 2011; Gingerich et al., 2014 ; Wood et al., 2014). Ces jugements initiaux sur les autres sont appelés impressions. Ils se produisent rapidement dès le premier contact avec une personne et sans que l'on réfléchisse vraiment à la manière dont ils ont été élaborés. Il s'agit de tâches de catégorisation, qui nous aident en facilitant le traitement, la structuration et l'assimilation des données concernant la personnalité et le comportement des individus que nous rencontrons (Gingerich et al., 2011; Wood et al., 2014). La première impression se réfère à tout jugement porté sur une personne sans qu'il y ait eu de rencontre préalable. Ce jugement est souvent rapide et repose sur des informations limitées. En milieu clinique, ces premières impressions façonnent les interactions initiales entre les évaluateurs et leurs stagiaires. Elles influencent les

[19] En psychologie, la perception d'autrui fait référence au processus par lequel les individus interprètent et comprennent les autres personnes, y compris leurs émotions, leurs traits de personnalité, et leurs intentions, à partir de signaux verbaux, non verbaux et contextuels.

informations retenues et conditionnent les attentes des évaluateurs quant aux comportements futurs des stagiaires (Wood et al., 2014). Les mécanismes cognitifs sous-jacents à la formation d'une première impression sont liés au fonctionnement du système 1[20], étant donné que la première impression ne semble pas être altérée par l'ajout de charges attentionnelles supplémentaires lors de l'utilisation de tests d'attention divisée[21]. En règle générale, les évaluateurs ne sont pas pleinement conscients des mécanismes qui sous-tendent la formation de leurs impressions (Wood et al., 2014). Bien qu'elles puissent être aussi précises que les impressions formées au cours d'un processus délibératif plus long, les premières impressions restent sujettes aux erreurs (Lee et al., 2017 ; Wood et al., 2014). Plusieurs facteurs peuvent influencer la précision des premières impressions, tels que : les caractéristiques des évaluateurs et des personnes évaluées ainsi que les processus cognitifs impliqués dans la formation des impressions chez les évaluateurs. Ces facteurs sont décrits en détail dans les sections qui suivent.

2.1. Les caractéristiques des personnes évaluées susceptibles d'influencer la précision des premières impressions.

Les évaluateurs semblent se baser sur certains traits physiques et psychologiques des individus évalués pour formuler une première impression qui influence ensuite leur jugement évaluatif. Parmi ces facteurs, l'on retrouve divers aspects sociaux comme la culture, l'éducation et les variations de comportement en fonction des contextes sociaux. De plus, les traits personnels inhérents des individus, tels que la race, le genre et des caractéristiques de personnalité facilement observables comme l'extraversion, l'énergie, l'amicalité et l'expressivité, le degré de chaleur/humanité et de compétence, jouent un rôle significatif dans la formation des impressions (Colson et al., 2020; Gauthier et al., 2016; Gingerich et al.,2011; Klein et al., 2019; Low et al., 2019; Wood et al., 2014). Enfin, l'apparence (tant physique que professionnelle), la gestion de l'impression[22], ainsi que les caractéristiques verbales (telles que la voix) et non-

[20] Il est généralement admis que les processus du système 1 sont rapides, sans effort, non analytiques, automatiques et/ou inconscients, tandis que les processus du système 2 sont lents, laborieux, analytiques, contrôlés et/ou conscients (Wood et al.,2014).

[21] Les tests d'attention divisés : Ce sont des tests qui sont généralement utilisés pour évaluer la coordination des processus entre le système 1 et le système 2, mais ils n'ont pas d'impact sur la première impression. (Wood et al., 2014)

[22] La gestion de l'impression : incluant l'auto-promotion, la mise en avant des aspects positifs, et la focalisation de l'attention de l'intervieweur (Wood et al., 2014).

verbales (comme le sourire et le contact visuel), peuvent toutes influencer la première impression de l'évaluateur.

Ces facteurs peuvent agir de manière intentionnelle ou non intentionnelle (Wood et al., 2014) sur la cognition de l'évaluateur lors de son jugement évaluatif. Il peut résulter de ces premières impressions des jugements dichotomiques basés sur deux dimensions. La première dimension concerne les traits socialement désirables ou indésirables qui influent directement sur les interactions avec autrui, incluant des traits positifs (comme la gentillesse ou l'honnêteté) et des traits négatifs (tels que la froideur ou la duplicité). La seconde dimension, plus variable selon les études, concerne des traits ayant un impact plus direct sur la réussite individuelle, tels que des traits positifs (tels que l'intelligence ou l'ambition) et des traits négatifs (comme l'indécision ou l'inefficacité). Bien que divers termes puissent être utilisés pour décrire chaque dimension, suggérant une divergence entre les chercheurs de divers domaines, ces derniers s'accordent sur l'existence d'un ensemble partagé de caractéristiques et de comportements. Ainsi, les deux dimensions distinctes sont catégorisées en jugements de grande valeur et de faible valeur, produisant quatre combinaisons potentielles lorsqu'elles sont croisées. Il a été suggéré que les individus et les groupes sont classés dans l'une de ces quatre catégories en fonction de leur degré de chaleur/humanité et de leur compétence, ce qui entraîne chez l'évaluateur diverses réponses émotionnelles et comportementales. Ces dimensions des jugements sociaux expliquent une grande partie de la variance dans la formation des impressions et la production de jugements évaluatifs en enseignement médical. La dimension de la compétence est associée aux connaissances de la personne, tandis que la dimension de la chaleur correspond aux compétences interpersonnelles de la personne évaluée. Ainsi, les médecins évaluateurs semblent utiliser ce processus cognitif de catégorisation pour classer les évalués dans l'une des quatre catégories décrites ci-dessus, avec les émotions, les biais et les comportements qui en découlent (Gingerich et al., 2011).

Tableau 2.

Sommaire des principaux résultats concernant les caractéristiques des personnes évaluées susceptibles d'influencer la précision des premières impressions.

Article	2.1 Les caractéristiques des personnes évaluées susceptibles d'influencer la précision des premières impressions.
Colson et al., 2020	Certains traits physiques et psychologiques des individus évalués influencent la première impression des évaluateurs.
Gauthier et al., 2016	Certains traits physiques et psychologiques des individus évalués influencent la première impression des évaluateurs.
Gingerich et al.,2011	Certains traits physiques et psychologiques des individus évalués influencent la première impression des évaluateurs. Utilisation d'un processus cognitif de catégorisation par les évaluateurs pour classer les évalués dans l'une des quatre catégories suivantes : -Les traits socialement désirables ou indésirables. -Les traits ayant un impact plus direct sur la réussite individuelle (traits positifs ou traits négatifs).
Klein et al., 2019	Certains traits physiques et psychologiques des individus évalués influencent la première impression des évaluateurs.
Low et al., 2019	Certains traits physiques et psychologiques des individus évalués influencent la première impression des évaluateurs.
Wood et al., 2014	Certains traits physiques et psychologiques des individus évalués influencent la première impression des évaluateurs. Ces caractéristiques peuvent agir de manière intentionnelle ou non intentionnelle

2.2. Les facteurs et les processus cognitifs impliqués dans la formation des premières impressions des évaluateurs et susceptibles d'en influencer la précision.

Dans certains cas, la formation des premières impressions est influencée par diverses variables en lien avec le système 1 sous-jacent dans les prises de décision. Ces variables ont

tendance à être influencées par les heuristiques, la récupération de mémoire et par d'autres biais cognitifs propre aux évaluateurs eux-mêmes. En outre, des facteurs tels que le genre, l'intelligence, l'expérience, l'humeur, l'émotion de l'évaluateur et l'accès à des informations sur les performances passées du résident peuvent aussi influencer les premières impressions (Gingerich et al., 2011; Wood et al., 2014; Wood et al., 2017).

Toutes ces variables interfèrent avec les processus cognitifs qui sous-tendent la catégorisation, qui constitue un élément crucial de la formation des impressions sur autrui. Le processus de catégorisation offre aux évaluateurs la possibilité d'utiliser leurs connaissances préalables[23] sur les comportements sociaux. En utilisant des catégories, l'évaluation ne nécessite pas de surveillance constante du comportement conforme à une catégorie spécifique. L'évaluateur met l'accent sur les comportements qui ne correspondent pas à la catégorie en question, ce qui renforce sa compréhension et sa capacité à anticiper les comportements des individus évalués, ainsi qu'à choisir la meilleure approche lors de leurs interactions avec eux (Gingerich et al., 2011; Gingerich et al., 2014; Wood et al.,2014).

Tableau 3.

Sommaire des principaux résultats concernant les facteurs et des processus cognitifs dans la formation des premières impression des évaluateurs susceptibles d'en influencer la précision.

Article	2.3 Les facteurs et les processus cognitifs dans la formation des premières impression des évaluateurs susceptibles d'en influencer la précision
Gingerich et al., 2011	Certains facteurs physiques et psychologiques de l'évaluateur ainsi que l'accès à des informations sur les performances passées du résident peuvent influencer les premières impressions. L'utilisation de la catégorisation comme moyen pour réduire la charge mentale lors de la surveillance des comportements.
Wood et al., 2014	Certains facteurs physiques et psychologiques de l'évaluateur ainsi que l'accès à des informations sur les performances passées du résident peuvent influencer les premières impressions. L'utilisation de la catégorisation comme moyen pour réduire la charge mentale lors de la surveillance des comportements.

[23] Les connaissances préalables qui peuvent être des constructions préformées qui existent dans la mémoire à long terme qui nécessite d'être activé (Gingerich et al., 2011).

Wood et al., 2017	Certains facteurs physiques et psychologiques de l'évaluateur ainsi que l'accès à des informations sur les performances passées du résident peuvent influencer les premières impressions. L'utilisation de la catégorisation comme moyen pour réduire la charge mentale lors de la surveillance des comportements.

2.3. Les conséquences de baser son jugement sur les premières impressions.

Les études sur la formation des impressions s'accordent sur le fait que les catégorisations facilitent l'application des informations concernant un individu typique d'une catégorie à une nouvelle personne. Cette approche aide à minimiser les charges cognitives requises pour observer le comportement d'une personne, anticiper ses actions et envisager les meilleures façons d'interagir avec elle. Ainsi, les premières impressions influencent nos perceptions, nos prédictions et nos théories sur les autres personnes que nous rencontrons (Gingerich et al., 2011). L'évaluation fondée sur les premières impressions comporte donc plusieurs risques, dont les suivants :

2.3.1. L'émergence de biais cognitifs de généralisation, tels que l'effet de halo.

Les risques liés à l'émergence de biais cognitifs de généralisation, tels que l'effet de halo. Dans ce contexte, les éléments sociaux ou personnels exercent une influence inconsciente sur le jugement global d'un individu (Gauthier et al., 2016). Ces biais cognitifs de généralisation peuvent affecter la façon dont nous formons nos premières impressions, en nous poussant à extrapoler rapidement des caractéristiques à partir de notre première perception. Cette première impression ainsi formée, influence alors toutes les évaluations ou tous les jugements ultérieurs concernant la personne. Cela peut conduire à des évaluations biaisées en médecine et à des prophéties auto-réalisatrices. Ce phénomène se produit lorsqu'une impression initiale influence les interactions ultérieures entre l'évaluateur et la personne évaluée (Wood et al., 2014).

2.3.2. Le développement de préjugés raciaux, ethniques ou sexistes.

Les risques liés au développement de préjugés raciaux, ethniques ou sexistes peuvent conduire à des disparités dans les évaluations de performance et dans les notes attribuées aux étudiants (Colson et al., 2020 ; Gauthier et al., 2016 ; Klein et al., 2019 ; Low et al., 2019). Ces

biais peuvent se manifester à travers les évaluations narratives, où l'on observe une différence dans l'utilisation de termes selon le genre et l'origine ethnique des étudiants en médecine. Les descriptions des étudiantes et des étudiants issus de minorités sous-représentées mettent souvent l'accent sur les traits personnels plutôt que sur les compétences, tandis que pour les étudiants masculins ou ceux n'appartenant pas à ces minorités, l'attention semble être davantage portée sur les compétences. Cela révèle une lacune dans l'application d'une évaluation basée sur les compétences (Polanco-Santana et al., 2021 ; Rojek et al., 2019 ; Ross et al., 2017). Plus précisément, les termes associés à l'agentivité[24] (comme compétent, organisé, travailleur, intelligent, motivé, autonome) sont utilisés de manière significativement plus fréquente pour décrire les résidents blancs et asiatiques, tandis que les termes associés aux compétences collaboratives (tels que sympathique, agréable, compatissant, doux, enthousiaste, amical) sont plus couramment utilisés pour décrire les étudiants en médecine noirs et hispaniques (Polanco-Santana et al., 2021; Ross et al., 2017).

2.3.3. L’humeur et les émotions des évaluateurs au moment du jugement.

L’humeur et les émotions des évaluateurs au moment du jugement peuvent influencer leur perception (Wood et al., 2014), en particulier lorsque la personne évaluée rappelle une autre personne importante dans l'esprit de l'évaluateur. De plus les émotions et l'état d'esprit des évaluateurs peuvent être affectés lorsqu'ils ont pris connaissance d'une description de la personne évaluée avant de la rencontrer (Gingerich et al., 2011).

[24] l'agentivité, adaptation de l'anglais « agency », est la faculté d’action d'un être, sa capacité à agir sur le monde, les choses, les êtres, à les transformer ou les influencer. (goal-oriented/ leader- ship/ achievement traits) (Polanco-Santana et al., 2021).

Tableau 4.

Sommaire des principaux résultats concernant les conséquences de baser son jugement sur les premières impressions.

Article	2.4 Les conséquences de baser son jugement sur les premières impressions		
	2.3.1 L'émergence de biais cognitifs de généralisation	2.3.2 Le développement de préjugés raciaux, ethniques ou sexistes.	2.3.3 L'humeur et les émotions des évaluateurs au moment du jugement.
Colson et al., 2020		Préjugés raciaux, ethniques ou sexistes	
Gauthier et al., 2016	Biais cognitifs de généralisation, tels que l'effet de halo	Préjugés raciaux, ethniques ou sexistes	
Gingerich et al., 2011			Une description de la personne évaluée avant de la rencontrer peut influencer la perception de l'évaluateur
Klein et al., 2019		Préjugés raciaux, ethniques ou sexistes	
Low et al., 2019		préjugés raciaux, ethniques ou sexistes	
Polanco-Santana et al., 2021		L'accent sur les traits personnels plutôt que sur les compétences	
Rojek et al., 2019		L'accent sur les traits personnels plutôt que sur les compétences	
Ross et al., 2017		L'accent sur les traits personnels plutôt que sur les compétences	
Wood et al., 2014	Prophéties auto-réalisatrices		L'humeur et les émotions des évaluateurs au moment du jugement peuvent influencer leur perception

3. L'adoption d'une approche déductive par la formulation d'inférences de haut niveau.

Il est généralement admis que les évaluations des étudiants en médecine ou des résidents ont tendance à être davantage marquées par l'utilisation d'inférences[25] et d'approches déductives[26]. (Colson et al., 2020 ; Gingerich et al., 2014 ; Kogan et al., 2011 ; Lee et al., 2017). Cette tendance est particulièrement prononcée chez les évaluateurs expérimentés par rapport aux novices (St-Onge et al., 2016 ; Lee et al., 2017). En d'autres mots, au lieu de se conformer strictement aux normes de référence établies, les évaluateurs, de manière consciente ou inconsciente, s'appuient davantage sur les comportements observables tels que les actions concrètes ou le langage corporel des résidents pour formuler des inférences et des déductions. Dans le cas de l'évaluation des compétences observées de façon indirecte, telles que le raisonnement clinique, les évaluateurs s'appuient largement sur des inférences découlant d'actions observables (St-Onge et al., 2016), comme le montre les commentaires de certains évaluateurs :

She appropriately identified the problems. She (pause) specified the problems, and we see that she shows clinical reasoning. That means that she is looking for, OK, weight loss, what could that be, she looks for things, she evaluates things, thus we see that she has an idea in mind, and now she is verifying it. (E2). (St-Onge et al., 2016, p.637)

Ces inférences et déductions portent notamment sur certains attributs des résidents relatifs à leur niveau de confiance, leur personnalité, leurs compétences, leur motivation à s'améliorer, leurs expériences antérieures et leur préparation (Kogan et al., 2011), tel que mentionné dans les commentaires de certains évaluateurs :

The one thing that struck me through this entire visit was he [the resident] had his arms kind of crossed. That could mean different things to different people in terms of body language.

[25] Inférences : Les évaluateurs ont tendance à collecter, à fusionner et à construire des schémas cohérents à partir de divers éléments d'information.

[26] Approche déductives : Les évaluateurs analysent les performances observées en déduisant le processus de pensée de l'individu à partir de ses actions ou de ses paroles, et ils génèrent également des explications hypothétiques pour les lacunes éventuelles dans ces performances.

It means I'm either closed to you, or I'm very comfortable with this, but it seems less likely. It could also be [representative of] an uncomfortable feeling on the part of the resident. He wants to act like he's comfortable, but internally, he's very anxious about breaking bad news, so there's a number of ways you can look at that... (Kogan et al., 2011, p.1053)

Se fier aux approches déductives pour évaluer les stagiaires comporte des risques, comme en témoigne une étude menée par Colson et ses collègues en 2020. Cette recherche met en lumière les préoccupations soulevées par l'utilisation répandue de méthodes déductives dans l'évaluation des étudiants en médecine. En effet, les étudiants participant à cette étude manifestent une préoccupation quant à la méconnaissance de leur travail par les évaluateurs, attribuable au contact limité qu'ils établissent avec les étudiants. De plus, certains étudiants signalent avoir reçu des évaluations injustifiées, puisqu'ils leur semblent que les évaluateurs ne les connaissaient pas suffisamment pour être en mesure de prononcer un jugement éclairé sur leurs compétences.

Tableau 5.

Sommaire des principaux résultats concernant l'adoption d'une approche déductive par la formulation d'inférences de haut niveau lors du processus évaluatif.

Article	3-Adoption d'une approche déductive par la formulation d'inférences de haut niveau lors de l'évaluation.
Colson et al., 2020	L'utilisation d'inférences et d'approches déductives lors de l'évaluation des apprenants en médecine. Se fier aux approches déductives pour évaluer les stagiaires comporte des risques.
Gingerich et al., 2014	L'utilisation d'inférences et d'approches déductives lors de l'évaluation des apprenants en médecine.
Kogan et al., 2011	L'utilisation d'inférences et d'approches déductives lors de l'évaluation des apprenants en médecine. Les inférences et déductions portent notamment sur certains attributs des résidents.
Lee et al., 2017	L'utilisation d'inférences et d'approches déductives lors de l'évaluation des apprenants en médecine.

St-Onge et al., 2016	Inférences découlant d'actions observables.
Lee et al., 2017	L'utilisation d'inférences et d'approches déductives lors de l'évaluation des apprenants en médecine.

4. L'accentuation de divers aspects des compétences.

Des chercheurs ont observé que lors de l'évaluation des performances des étudiants en médecine ou des résidents, les évaluateurs ont tendance à se concentrer sur certains aspects spécifiques. Cependant, ce qui est considéré comme crucial par un évaluateur peut différer de ce qui est jugé important par un autre évaluateur évaluant la même prestation (Lee et al., 2017 ; Yeates et al., 2013). Ainsi, les évaluateurs chevronnés accordent une importance accrue aux indices contextuels spécifiques, tels que la nature de la tâche ou de la situation, tandis que les évaluateurs novices mettent davantage l'accent sur la description d'aspects particuliers et distincts de la performance (Lee et al., 2017). Plusieurs éléments liés au contexte de l'interaction avec le patient et à la situation clinique semblent influencer l'évaluation. Par exemple, la complexité croissante des cas, notamment lors de consultations prolongées impliquant des situations peu définies ou des consultations résultant de la diversité linguistique des patients, surtout lorsqu'ils présentent plusieurs plaintes simultanément (Lee et al., 2017; St-Onge et al., 2016). Un autre facteur contribuant à l'accentuation de certaines dimensions de la compétence est la variation fréquente dans l'évaluation des performances des candidats due à des interprétations variables des compétences. Ainsi, chaque évaluateur semble accorder une attention différente à chaque performance, ce qui conduit à des jugements singuliers propres à chaque évaluateur (Lee et al., 2017; Yeates et al., 2013). En attribuant des niveaux d'importance variables aux différents aspects d'une même performance, les évaluateurs forment essentiellement des jugements basés sur des observations différentes, constituant ainsi une source de variabilité entre les évaluateurs, ce qui contribue aux divergences dans le processus de jugement (Yeates et al., 2013).

Tableau 6.

Sommaire des principaux résultats concernant l'accentuation de divers aspects des compétences lors du processus évaluatif.

Article	4 L'accentuation de divers aspects des compétences.
Lee et al., 2017	Ce qui est jugé crucial par un évaluateur peut différer de ce qu'un autre évaluateur considère important en évaluant la même prestation. Les éléments contextuels de l'interaction avec le patient et de la situation clinique semblent influencer l'évaluation.
St-Onge et al., 2016	Les éléments contextuels de l'interaction avec le patient et de la situation clinique semblent influencer l'évaluation.
Yeates et al., 2013	Ce qui est jugé crucial par un évaluateur peut différer de ce qu'un autre évaluateur considère important en évaluant la même prestation. Les évaluateurs basent essentiellement leurs jugements sur des observations variées.

5. La traduction du jugement narratif en chiffre.

Malgré l'utilisation de grilles d'évaluation, de nombreux enseignants ou évaluateurs se heurtent à des difficultés notables lorsqu'ils tentent de traduire leurs évaluations en notations numériques (Kogan et al., 2011 ; Yeates et al., 2013). Cette difficulté se pose particulièrement lorsque les enseignants constatent que les repères normatifs ne parviennent pas à refléter de manière adéquate ce qu'ils souhaitent communiquer. Ce défi est encore plus marqué lorsqu'il s'agit d'utiliser une échelle plus détaillée (Forte et al, 2021). Pour surmonter ces défis, les enseignants adoptent différentes approches, dont l'utilisation d'un langage narratif, avant l'assignation d'une note chiffrée (Yeates et al., 2013).

Ce langage narratif est fondé sur leur perception de l'écart entre ce qu'ils observent de la performance de l'évalué et les critères présentés dans la grille d'évaluation. Ainsi, durant cette démarche de transcription narrative, les jugements exprimés font occasionnellement appel à l'échelle utilisée dans la grille et utilisent des expressions telles que, "est en dessous des attentes » ; "est à la limite des attentes"; "conforme aux attentes"; "surpasse les attentes"(Yeates et al.,

2013) ou à travers un langage narratif et descriptif varié, allant de termes simples tels que "bon" ou "mauvais" à des expressions plus élaborées parfois teintées d'émotions, comme "agacé", "déçu", "content" ou "pas du tout content" (Yeates et al.,2013). Après avoir traduit leur évaluation en texte narratif, l'attribution de notes relatives à la performance observée implique l'utilisation des trois méthodes suivantes :

5.1. La traduction mentale des points d'ancrage.

La traduction mentale des points d'ancrage signifie qu'ils interprètent mentalement les repères normatifs qui semblent inadéquats pour exprimer leurs intentions. Les évaluateurs se servent d'échelles de notation plus familières, telles que les échelles de type Likert graduées sur 5 points, en interprétant mentalement que 5 représente une performance excellente et 1 une performance médiocre (Forte et al., 2021). En d'autres termes, il semble que les évaluateurs prennent leurs descriptions narratives personnelles et les transforment en critères mesurables. De plus, ils semblent convertir leurs impressions générales en scores attribués à la performance pour chaque compétence, ce qui signifie que la variabilité des impressions globales des évaluateurs influe sur la variabilité des scores (Lee et al., 2017 ; Yeates et al., 2013). Ainsi, la manière dont cette conversion a été effectuée peut davantage influencer la variabilité des scores (Yeates et al, 2013).

5.2. La pondération des évaluations en fonction de l'objectif de la rencontre.

La manière dont les évaluations sont pondérées peut varier en fonction des objectifs de l'évaluation du stagiaire. Par exemple, lorsqu'il s'agit d'observer une compétence spécifique chez le stagiaire, les autres compétences démontrées pendant cette observation peuvent recevoir moins d'importance. Certains évaluateurs préfèrent calculer la moyenne des compétences individuelles démontrées pour évaluer les performances globales, tandis que d'autres soutiennent une approche de notation non compensatoire où les lacunes dans une compétence ne peuvent être compensées par des performances supérieures dans d'autres domaines (Kogan et al., 2011 ; Lee et al., 2017) : « I don't think so, because competence is still, you know, absent competence – a humanistic physician who's not competent is a very dangerous person. » (Kogan et al., 2011, p.1053)

5.3. Une approche différenciée de la rigueur.

Une approche différenciée de la rigueur vise à ajuster les évaluations des compétences et à prendre en compte divers facteurs contextuels lors de la pondération des performances. La rigueur des évaluations peut être influencée par divers facteurs tels que : l'expérience de l'évaluateur, ses attentes et ses compétences cliniques personnelles, ou une combinaison de ces éléments. Par ailleurs, il apparaît dans le contexte d'études d'observations qu'une rigueur accrue est appliquée dans l'évaluation des performances, relatives aux compétences cliniques tels que : l'interrogatoire, l'examen physique et l'organisation des tâches, par rapport aux compétences non cliniques (Kogan et al., 2011 ; Lee et al., 2017). Les attentes des évaluateurs renforcent cette approche rigoureuse, car si elles ne sont pas comblées, les évaluateurs se concentrent davantage sur les lacunes ou les insuffisances découvertes (St-Onge et al., 2016).

Toutes ces circonstances engendrent une diversité considérable dans les méthodes utilisées et une incertitude quant à la conversion d'une évaluation du résident en une notation numérique (Kogan et al., 2011).

Tableau 7.

Sommaire des principaux résultats concernant la traduction du jugement narratif en chiffre lors du processus évaluatif.

Article	5 La traduction du jugement narratif en chiffre.		
	5.1 La traduction mentale des points d'ancrage.	5.2 La pondération des évaluations en fonction de l'objectif de la rencontre.	5.3 Une approche différenciée de la rigueur.
Forte et al., 2021	Les évaluateurs interprètent mentalement les repères normatifs		
Kogan et al., 2011		Les lacunes dans une compétence ne peuvent pas être compensées par de meilleurs performances dans un autres domaines.	Une rigueur accrue est appliquée dans l'évaluation des performances, relatives aux compétences cliniques par rapport aux compétences non cliniques.

Lee et al., 2017	Les variations dans les impressions globales des évaluateurs affectent la variabilité des scores.	Les lacunes dans une compétence ne peuvent pas être compensées par de meilleurs performances dans un autres domaines.	Une rigueur accrue est appliquée dans l'évaluation des performances, relatives aux compétences cliniques par rapport aux compétences non cliniques.
St-Onge et al., 2016			les évaluateurs se concentrent davantage sur les lacunes ou les insuffisances découvertes
Yeates et al., 2013	Les variations dans les impressions globales des évaluateurs affectent la variabilité des scores.		

6. L'effet de l'émotion de l'évaluateur lors du processus évaluatif.

Il semble que les décisions sont influencées par une combinaison de facteurs cognitifs, émotionnels et contextuels. Cette interconnexion détermine le processus d'évaluation comme une tétrade composée de quatre éléments distincts mais interdépendants : la cognition, l'observation, l'interprétation et les émotions, qui se mêlent tout au long de l'acte d'évaluation. Parmi ces éléments, l'émotion joue un rôle significatif dans le processus décisionnel (Gomez-Garibello & Young, 2018 ; Kogan et al., 2011). Trois types d'émotions peuvent influencer les processus de prise de décision : d'abord, les traits émotionnels ou le tempérament de la personne prenant la décision (humeur) ; ensuite, les émotions surgissant lors de la prise de décision (émotions incidentes) ; enfin, les émotions anticipées concernant les résultats possibles des décisions envisagées.

La recherche issue du domaine de la psychologie de l'éducation indique que l'évaluation des apprenants, tout comme d'autres processus de prise de décision, est influencée par les émotions. Bien que diverses émotions puissent servir de sources valables d'information lors de l'évaluation, elles peuvent également entraîner des décisions biaisées et introduire un bruit aléatoire dans le processus d'évaluation, en particulier lorsque l'évaluateur n'est pas conscient de ses propres émotions. Ceci suggère que les émotions pourraient être envisagées comme des

facteurs contribuant aux biais systématiques lors d'évaluations (Gomez-Garibello & Young, 2018). S'appuyer sur les émotions pour évaluer les performances des stagiaires peut présenter divers risques, notamment :

6.1. Les traits émotionnels de la personne qui prend la décision (humeur).

Les individus ressentant des émotions positives peuvent être davantage enclins à adopter des stratégies heuristiques dans leur processus décisionnel, ce qui peut conduire à des représentations déformées de la performance de l'apprenant en raison de l'effet de halo. En revanche, les évaluateurs éprouvant des émotions négatives, telles que l'anxiété ou la peur, ont tendance à se concentrer sur les détails du stagiaire, de la procédure ou de la situation, ce qui peut également entraîner des représentations erronées de la performance de l'apprenant (Gomez-Garibello & Young, 2018).

6.2. Les émotions suscitées lorsque la personne prend la décision (émotions incidentes).

Certains évaluateurs manifestent un biais d'indulgence en attribuant des notes plus élevées afin d'éviter la sensation désagréable de se percevoir méchant, malveillant ou démoralisant (Kogan et al., 2011; Paget et al., 2013), comme en témoigne les commentaires de certains d'entre eux :

People don't like to give low scores because it doesn't show well for them... And the answer that I've gotten, and I haven't been very satisfied with it, is that you can be a popular doctor or a good, good doctor with residents and medical students. (Kogan et al., 2011, p.1055)

6.3. Les émotions anticipées des résultats des décisions possibles (émotions attendues).

Certains évaluateurs peuvent ajuster leurs évaluations afin d'éviter des conséquences désagréables, comme des évaluations négatives de la part des résidents. En effet, il semble y avoir une relation de réciprocité entre les évaluations des résidents et celles des évaluateurs. Ainsi, les enseignants qui évaluent de manière plus positive les résidents reçoivent également des évaluations plus favorables de leur part. Cette relation significative entre les évaluations ne peut être expliquée uniquement par le nombre d'années de formation des évaluateurs ; elle dépend

plutôt du phénomène de réciprocité (Gardner & Scott., 2016). D'autres répercussions désagréables peuvent influencer l'évaluation des enseignants. Il s'agit, par exemple, des appréhensions que certains évaluateurs ont concernant les réactions d'étudiants insatisfaits de leurs rétroactions, comme le mentionne ceux-ci :

If anyone's in the satisfactory category, I tend to put them in the 6 range because I don't want to be having a conversation about why it wasn't a 6 instead of a 5. But [what] I really want to say to this resident is, you were fine, you were at your level of training, that's where I want you to be. I don't want to be negotiating about why I picked it as a 4 or 5 or 6 in that range. So, I pick it as a 6 so that takes that conversation off the table. (Kogan et al., 2011, p.1055)

Ainsi, certains évaluateurs manifestent un biais d'indulgence en préférant surévaluer les notes pour éviter d'avoir à donner des justifications aux étudiants (Kogan et al., 2011 ; Paget et al., 2013). Pour d'autres évaluateurs, les répercussions négatives proviennent plutôt de la culture institutionnelle, qui tend à favoriser les étudiants et placer les évaluateurs dans une position inconfortable d'avoir à justifier les notes négatives données aux étudiants. Ainsi, les affiliations entre l'évaluateur et l'évalué basées sur leur appartenance à la même institution d'enseignement, semble inciter les professeurs locaux à être moins enclins à évaluer négativement les résidents que les professeurs externes. Ce genre de relation peut entraîner un biais d'évaluation locale. (Kogan et al., 2011 ; Lee et al., 2017).

Tableau 8.

Sommaire des principaux résultats concernant l'effet de l'émotion de l'évaluateur lors du processus évaluatif.

Article	6 L'effet de l'émotion de l'évaluateur lors du processus évaluatif.			
	Introduction : Le rôle de l'émotion de façon globale dans la prise de décision	6.1 Les traits émotionnels de la personne qui prend la décision (humeur).	6.2 Les émotions suscitées lorsque la personne prend la décision (émotions incidentes).	6.3 Les émotions anticipées des résultats des décisions possibles (émotions attendues).
Gardner & Scott., 2016				Phénomène de réciprocité
Gomez-Garibello & Young	L'émotion joue un rôle significatif dans le processus décisionnel Émotions comme facteurs contribuant aux biais systématiques lors d'évaluations	Les traits émotionnels peuvent entraîner des représentations erronées de la performance de l'apprenant		
Kogan et al., 2011	L'émotion joue un rôle significatif dans le processus décisionnel		Biais d'indulgence de certains évaluateurs pour éviter la sensation désagréable de se percevoir comme malveillant .	Biais d'indulgence de certain évaluateur pour éviter d'avoir à donner des justifications aux étudiants Biais d'évaluation local
Lee et al., 2017				Biais d'évaluation local
Paget et al., 2013			Biais d'indulgence de certains évaluateurs pour éviter la sensation désagréable de se percevoir comme malveillant .	Biais d'indulgence de certain évaluateur pour éviter d'avoir à donner des justifications aux étudiants

Chapitre 5 – Discussion et conclusion

1. La discussion

Le but de cette étude était de déterminer les facteurs et les biais qui influencent la variabilité du jugement des enseignants-cliniciens. Plus spécifiquement, l'étude visait à examiner de quelles manières ces biais et ces facteurs influencent cette variabilité ainsi que le double processus de décision d'enseignants-cliniciens dans le contexte de l'évaluation de la performance des apprenants en médecine. La section suivante analyse les résultats principaux de cette étude en les comparant avec d'autres recherches pertinentes recensées dans la littérature. Ensuite, nous discutons des implications pour les évaluateurs intervenant auprès des apprenants en médecine. Enfin, ce chapitre se termine avec la présentation des limites de l'étude et les pistes de recherches futures.

La démarche évaluative permet de porter un jugement, à partir de normes ou de critères explicitement établis, sur la valeur de la performance d'un apprenant (Legendre et al., 2005). Au cœur de cette démarche se trouve le jugement évaluatif qui est complexe et multidimensionnel. Cette étude nous amène à considérer trois aspects déterminants de l'évaluation des performances des stagiaires. Il s'agit du jugement de l'évaluateur, des normes du domaine (Legendre et al., 2005) et de l'émotion des évaluateurs lors de ce processus décisionnel (Gomez-Garibello & Young, 2018; Kogan et al., 2011). Ces trois aspects seront détaillés dans la section suivante.

1.1. La multiplicité des cadres de référence.

Le premier aspect déterminant de l'évaluation des performances des stagiaires, fait référence à la multiplicité des cadres de références utilisés pour l'évaluation des stagiaires. Selon les résultats de notre étude, plusieurs cadres de référence sont employés par les évaluateurs lorsque vient le moment d'évaluer les compétences des apprenants (Yeates et al., 2013; Kogan et al., 2011; St-Onge et al., 2016; Lee et al., 2017; Forte et al., 2021). Ces cadres de référence peuvent être classé en deux catégories reflétant les informations utilisées par les évaluateurs. En premier lieu, on retrouve les informations externes (les critères institutionnels), et en second lieu, on retrouve le informations internes/personnelles (les attentes des évaluateurs) (St-Onge et al., 2016). Ces informations, qu'il s'agissen de critères institutionnels ou des attentes des évaluateurs, sont parfois incertaines ou ambiguës (Kogan et al., 2011; Yeates et al., 2013). En

effet, le caractère ambigu des critères institutionnels peut résider dans l'interprétation variée des normes[27] de performance du domaine de pratique en question ainsi que l'utilisation d'autres critères tels que la performance d'autres médecins et la performance de l'apprenant lors d'une séance antérieure (Kogan et al., 2011; Yeates et al., 2013). Les attentes des évaluateurs mettent en évidence un processus idiosyncrasique basé sur une construction individuelle de ce qui est considéré comme une bonne performance. Elles découlent de sources externes intériorisées, émanant d'une collection personnelle et distinctive d'expériences, de connaissances et d'instincts des évaluateurs (Forte et al., 2021; Kogan et al., 2011; Lee et al., 2017; St-Onge et al., 2016; Yeates et al., 2013), parfois qualifiés de "sentiments" ou de "réactions instinctives" (St-Onge et al., 2016). Puisque le jugement évaluatif est basé sur une comparaison de la performance observée aux critères ou normes, des normes incertaines peuvent avoir un effet sur l'évaluation. L'un des éléments clés réside dans le processus de comparaison des performances de l'apprenant. Cela peut se faire soit par rapport aux attentes des évaluateurs, qui peuvent parfois être influencées par les performances professionnelles de l'évaluateur ou par les critères institutionnels, qui sont fondés sur les performances des pairs.

Cette notion de relativité des performances a un impact sur le processus d'évaluation des compétences (Yeates et al., 2013 ; Yeates et al., 2015), en raison des effets de contraste qu'elle engendre (Gingerich et al., 2018). Ainsi, la manière dont les évaluateurs mettent en œuvre les cadres de référence est complexe, dynamique et très variable, ce qui peut produire des jugements différents d'un évaluateur à l'autre et d'un apprenant à l'autre (Kogan et al., 2011; Yeates et al., 2013; Lee et al., 2017). Nos résultats sont en accord avec les travaux de Gautier et al., (2016) qui mettent en évidence l'utilisation de critères incertains par l'évaluateur lors de la phase de traitement de l'information. Selon Gautier et ses collaborateurs, l'apprenant est catégorisé dans des schémas préétablis. Au cours de ce processus, la performance observée est comparée avec un concept personnel de compétence et un certain nombre d'exemples antérieurs stockés dans la mémoire.

En somme, il semble que les critères d'évaluation ou les cadres de compétences ne soient pas toujours connus ou correctement appliqués par les évaluateurs (Gautier et al., 2016;

[27] Ces normes de performance, englobent les critères du Collège royal et les objectifs fournis aux évaluateurs par le programme (St-Onge et al., 2016).

Gingerich et al., 2014). Pourtant, les cadres de compétences sont essentiels pour guider l'éducation médicale axée sur les compétences, en fournissant une structure organisée décrivant les compétences nécessaires pour la pratique professionnelle efficace (Harris et al., 2017). D'ailleurs, les travaux réalisés par Fontaine et Loye (2017), ainsi que Lacasse et al. (2017), indiquent l'existence d'un référentiel du Collège Royal des médecins et chirurgiens du Canada (CanMEDS), qui encadre la pratique des professionnels et fournit également des pistes précises pour l'évaluation des apprentissages des stagiaires en indiquant concrètement les compétences à développer par ces derniers.

1.2. Les étapes du jugement évaluatif.

Le deuxième aspect déterminant de l'évaluation des performances des stagiaires, fait référence au jugement évaluatif qui implique un processus à plusieurs étapes. Les résultats de notre étude sont en accord avec les travaux de Gauthier et al. (2016) qui présente trois étapes distinctes dans le processus du jugement de l'évaluateur. Ces trois étapes clés du processus de jugement, à savoir l'observation, le traitement et l'intégration de l'information, comprennent la plupart des facteurs qui contribuent à la variabilité des évaluations, comme l'ont identifié notre recherche. Ces facteurs sont divers et peuvent être liés à l'enseignant et aux stagiaires. Dans la section suivante, nous décrivons les divers facteurs qui pourraient influencer le jugement de l'évaluateur, en fonction des trois étapes du processus d'évaluation présentées par Gauthier et al. (2016).

1.2.1. **La phase d'observation** peut être définie comme la phase de perception et de sélection de l'information. Celle-ci peut être influencée notamment par trois facteurs, dont la génération automatique d'impressions, la formulation d'inférences et l'accentuation de certaines variables lors de l'évaluation.

La génération automatique d'impressions est à l'origine des premières impressions. Les premières impressions semblent jouer un rôle déterminant dans la variabilité du jugement des évaluateurs. Il ressort de notre étude que les évaluateurs se basent parfois sur certains traits physiques et psychologiques, dont la race, l'âge, le sexe et l'extraversion, pour formuler une première impression qui influence ensuite leur jugement évaluatif. De plus, la précision de cette première impression peut être influencée par certains traits de l'évaluateur. Ces deux sources

d'influence peuvent avoir des conséquences positives ou négatives sur l'évaluation d'un étudiant, , telles que l'émergence de biais cognitifs de généralisation, comme l'effet de halo (Gauthier et al., 2016) ainsi que le développement de préjugés raciaux, ethniques ou sexistes (Colson et al., 2020 ; Gauthier et al., 2016 ; Klein et al., 2019 ; Low et al., 2019). Nos constatations concordent avec les conclusions de la littérature telles qu'exposées dans l'étude de Gingerich et al., (2014) qui notent que les biais cognitifs influencent de manière significative les processus d'évaluation et de jugement dans l'éducation médicale.

Dans cette même perspective, Feinsilber et al. (2019) ont montré que les préjugés tels que l'effet de halo peuvent conduire à une diminution des critères d'évaluation des stagiaires, ce qui se traduit par une variabilité inappropriée des niveaux de supervision clinique. Ces préjugés peuvent conduire non seulement à des évaluations erronées, mais également à un retard crucial dans l'identification en temps réel d'une lacune qui pourrait être facilement corrigée. De plus, ces préjugés peuvent favoriser un népotisme institutionnel exercé par des directeurs de programme jouissant d'un pouvoir disproportionné, ainsi que la détérioration fréquente de la réputation des stagiaires au sein de certains programmes, encourageant ainsi l'établissement d'une culture du bouc émissaire.

La formulation d’inférence de haut niveau réfère au fait que les évaluations des étudiants en médecine reposent souvent sur des inférences et des approches déductives, surtout réalisées par des évaluateurs expérimentés. Ces derniers se basent sur les comportements observables pour formuler des inférences et des déductions, plutôt que de comparer strictement ces comportements observables aux normes établies (Colson et al., 2020 ; Gingerich et al., 2014; Kogan et al., 2011 ; Lee et al., 2017).

Ceci est particulièrement remarquable dans l'évaluation des compétences indirectes telles que le raisonnement clinique (St-Onge et al., 2016 ; Lee et al., 2017). Cependant, cette méthode comporte des risques, comme le souligne une étude récente menée par Colson et ses collègues en 2020. Les étudiants expriment des préoccupations concernant le fait que les évaluateurs ne comprennent pas suffisamment leur travail et signalent des évaluations injustifiées, attribuées au manque de connaissance des évaluateurs sur leurs compétences. Nos résultats confirment les observations de Gingerich et al. (2014) ainsi que de Kogan et al. (2014), mettant en lumière la tendance courante des évaluateurs à déduire les connaissances, les compétences et les attitudes

des stagiaires. Bien que ces déductions soient inévitables, l'absence de vérification ou de validation peut potentiellement altérer l'évaluation précise du stagiaire. Par conséquent, les déductions fondées sur l'observation directe influencent encore plus la validité de l'évaluation. Même si atteindre une objectivité absolue dans l'évaluation semble impossible, elle peut être améliorée par une approche méthodique qui intègre des principes de transparence, de rigueur et de cohérence, comme le soulignent Fontaine et Loye (2017).

Finalement, nos résultats montrent que les évaluateurs mettent l'accent sur divers aspects des performances des étudiants ou résidents, ce qui conduit à des jugements variables (Lee et al., 2017 ; Yeates et al., 2013). Les évaluateurs expérimentés se concentrent sur des indices contextuels spécifiques, tandis que les novices se focalisent sur des aspects particuliers de la performance (Lee et al., 2017). De plus, des éléments tels que la complexité des cas et la diversité linguistique des patients influencent les évaluations (Lee et al., 2017; St-Onge et al., 2016). La variation dans l'interprétation des compétences contribue également à des jugements singuliers et à une variabilité entre les évaluateurs, ce qui peut entraîner des divergences dans le processus d'évaluation (Lee et al., 2017; Yeates et al., 2013).

Nos découvertes divergent de la littérature existante en ce qui concerne l'impact de l'attention différentielle sur divers aspects des compétences comme facteur de la variabilité du jugement des évaluateurs. En effet, Lacasse et al. (2017) montrent que l'accentuation de l'attention de l'évaluateur n'est pas spécifique à l'évaluateur mais est plutôt orientée par un cadre de référence constitué de jalons à atteindre. Ainsi, cela implique que la surveillance active de l'évolution des compétences se fait en utilisant des jalons, qui sont des indicateurs observables de développement spécifiant les attentes à différentes étapes de la formation dans divers domaines ou contextes de pratique. Cette approche relevée par Lacasse et al. (2017) est soutenue par des recommandations du CMFC (Conseil médical du Canada) qui recommande que ces indicateurs soient utilisés par les cliniciens enseignants pour évaluer la progression des résidents.

1.2.2. **La phase de traitement de l'information** fait référence à un processus de catégorisation implicite et automatique (Gauthier et al., 2016) qui a été déjà abordé plus haut dans les cadres de références.

1.2.3. **La phase d'intégration de l'information** se caractérise par la combinaison des différentes sources d'information pour produire le jugement. Cette phase comprend une production narrative du jugement et la traduction de cette dernière en note (Gauthier et al., 2016). Les résultats de notre recherche suggèrent que, au cours de cette phase d'intégration, les évaluateurs semblent rencontrer de la difficulté lorsqu'ils tentent de traduire leurs évaluations subjectives en notations numériques (Kogan et al., 2011 ; Yeates et al., 2013). Cette difficulté est exacerbée lorsque les repères normatifs ne parviennent pas à refléter adéquatement ce que les évaluateurs veulent communiquer (Forte et al, 2021). Pour surmonter ces défis, les évaluateurs adoptent différentes approches, notamment l'utilisation d'un langage narratif pour exprimer leurs jugements avant d'attribuer une note chiffrée (Yeates et al., 2013). Les méthodes utilisées pour traduire l'évaluation narrative en notation numérique, comportent les éléments suivants:

- La traduction mentale des points d'ancrage : Les évaluateurs interprètent mentalement les repères normatifs et utilisent des échelles de notation plus familières pour exprimer leurs intentions (Forte et al., 2021 ; Yeates et al, 2013).
- La pondération des évaluations en fonction de l'objectif de la rencontre : Les évaluations sont pondérées en fonction des objectifs spécifiques de l'évaluation du stagiaire, ce qui peut entraîner des approches différentes, comme la moyenne des compétences individuelles ou une approche non compensatoire (Kogan et al., 2011 ; Lee et al., 2017).
- Une approche différenciée de la rigueur : Les évaluateurs ajustent les évaluations en tenant compte de divers facteurs contextuels, tels que leurs expériences, leurs attentes (St-Onge et al., 2016) et leurs compétences cliniques personnelles. Cette approche peut varier en fonction des compétences évaluées, avec une rigueur généralement accrue pour les compétences cliniques par rapport aux compétences non cliniques (Kogan et al., 2011 ; Lee et al., 2017).

En somme, les défis et les différentes méthodes utilisées par les enseignants pour convertir une évaluation narrative en une notation numérique soulignent la diversité des approches et l'incertitude qui en découle (Kogan et al., 2011), ce qui entraine une variabilité dans le jugement des évaluateurs. Cette ambiguïté et les conséquences néfastes de la conversion des évaluations en chiffres correspondent aux conclusions des travaux de Gautier et ses collègues (2016), qui notent que le passage à des évaluations chiffrées peut entraîner une perte de sens et altérer l'évaluation

de la performance des étudiants. La section suivante présente le troisième aspect déterminant de l'évaluation des stagiaires.

1.3. L'émotion de l'évaluateur

L'émotion de l'évaluateur lors du processus évaluatif est un autre aspect déterminant de l'évaluation des performances des stagiaires. En effet, les émotions semblent faire partie intégrante du processus d'évaluation, aux côtés de la cognition et de l'observation. Il s'agit de processus interconnectés de sorte que les émotions jouent un rôle significatif dans l'évaluation et peuvent influencer les décisions prises par les évaluateurs (Gomez-Garibello & Young, 2018; Kogan et al., 2011).

On distingue trois types d'émotions qui influencent le processus de prise de décision des évaluateurs : les traits émotionnels (humeur) de la personne prenant la décision, les émotions surgissant lors de la prise de décision (émotions incidentes), et les émotions anticipées concernant les résultats possibles des décisions (Gomez-Garibello & Young, 2018). Les émotions peuvent entraîner des biais dans le processus d'évaluation, tels que l'effet de halo, le biais d'indulgence, le biais d'évaluation local, ou inciter les enseignants à ajuster leurs évaluations pour éviter des conséquences désagréables, telles que des réactions négatives des étudiants ou des répercussions institutionnelles (Kogan et al., 2011 ; Paget et al., 2013). En somme, les résultats de notre étude suggèrent que, bien que les émotions puissent fournir des informations précieuses, leur impact sur l'évaluation doit être pris en compte pour assurer un processus juste et équitable puisque les émotions peuvent introduire des biais et du bruit dans le processus d'évaluation. Nos résultats corroborent les conclusions de plusieurs chercheurs, dont Gingerich et al. (2014). Ils notent comment l'émotion influence la modification des évaluations afin d'éviter des conséquences négatives, ainsi que l'activation des stéréotypes et leur impact sur le processus de jugement. De plus, Wilson & Brekke (1994) arrivent à la conclusion que cette émotion peut résulter d'une contamination mentale, qui se définit comme un processus par lequel une personne aboutit, par exemple, à une émotion non désirée en raison d'un traitement mental inconscient ou incontrôlable. Autrement dit, cette contamination mentale pourrait expliquer l'origine des différents types d'émotions.

1.4. Quelques pistes de solution :

En prenant en considération les données issues de cette recherche, qui suggèrent que lors de l'évaluation, les biais cognitifs sont souvent inconscients et influencés par divers facteurs tels que les émotions et les préjugés préexistants, plusieurs stratégies peuvent être envisagées pour améliorer la fiabilité du jugement des évaluateurs. Parmi celles-ci, nous proposons :

- L'utilisation de paires d'examinateurs dans les évaluations cliniques à fort enjeu, telles que les examens médicaux finaux, afin que les évaluations soient portées par deux examinateurs et que la note finale soit obtenue par consensus (Faherty et al., 2020).
- L'offre d'une formation continue ciblée aux évaluateurs, pour réduire la variabilité des jugements et assurer la validité des notes et la crédibilité des diplômes décernés (Colson et al., 2020 ; Fontaine et Loye, 2017 ; Gingerich et al., 2014 ; Hodwitz et al., 2019 ; Klein et al., 2019). Cette formation devrait viser à améliorer la mise en œuvre de démarches rigoureuses qui respectent les principes de transparence, de cohérence (interne et externe), d'égalité et d'équité (Fontaine et Loye, 2017). La cohérence des jugements devrait s'appuyer sur des lignes directrices pertinentes, des critères de performance et un cadre de référence commun pour évaluer les performances. Les évaluations détermineront si le stagiaire a atteint les objectifs en détectant le comportement observable attendu et en le comparant à des critères préétablis, idéalement fondés sur les meilleures pratiques basées sur des données probantes (Colson et al., 2020 ; Gingerich et al., 2014). En résumé, l'évaluation des apprenants devrait se faire en fonction de leurs résultats individuels plutôt que par rapport à ceux de leurs pairs (Gingerich et al., 2014).
- La mise en œuvre de stratégies visant à améliorer la diversité du corps enseignant et des instances universitaires (Colson et al., 2020).

  La promotion d'un dialogue ouvert sur le genre et les préjugés parmi les stagiaires et les professeurs dans le cadre des stratégies visant à lutter contre les biais de genre durant la formation médicale des diplômées (Klein et al., 2019).
- L'intégration d'activités professionnelles confiables (APC) dans le curriculum de formation médicale permet une évaluation objective des compétences grâce à des tâches cliniques standardisées et spécifiques. Ces APC renforcent la cohérence entre les

évaluateurs, favorisant ainsi une progression des apprentissages plus transparente et équitable (Ten Cate et al., 2015).

2. Les limites de l'étude

Une des limites de cette revue systématique réside dans le nombre restreint d'études disponibles sur le sujet de la variabilité du jugement des évaluateurs dans le contexte de l'évaluation des apprenants en médecine. Ce faible nombre d'études pourrait restreindre la robustesse et la généralisabilité des conclusions. La présente revue systématique est aussi confrontée à une limitation importante en raison de la méthodologie hétérogène observée parmi les études incluses, ce qui a compliqué l'interprétation des résultats et pourrait limiter la validité des comparaisons entre les résultats. Un autre défi rencontré lors de cette revue systématique est l'hétérogénéité des résultats rapportés dans les études incluses, ce qui a limité la possibilité d'extraire des résultats cohérents et comparables parmi les études.

3. Conclusion et pistes de recherche.

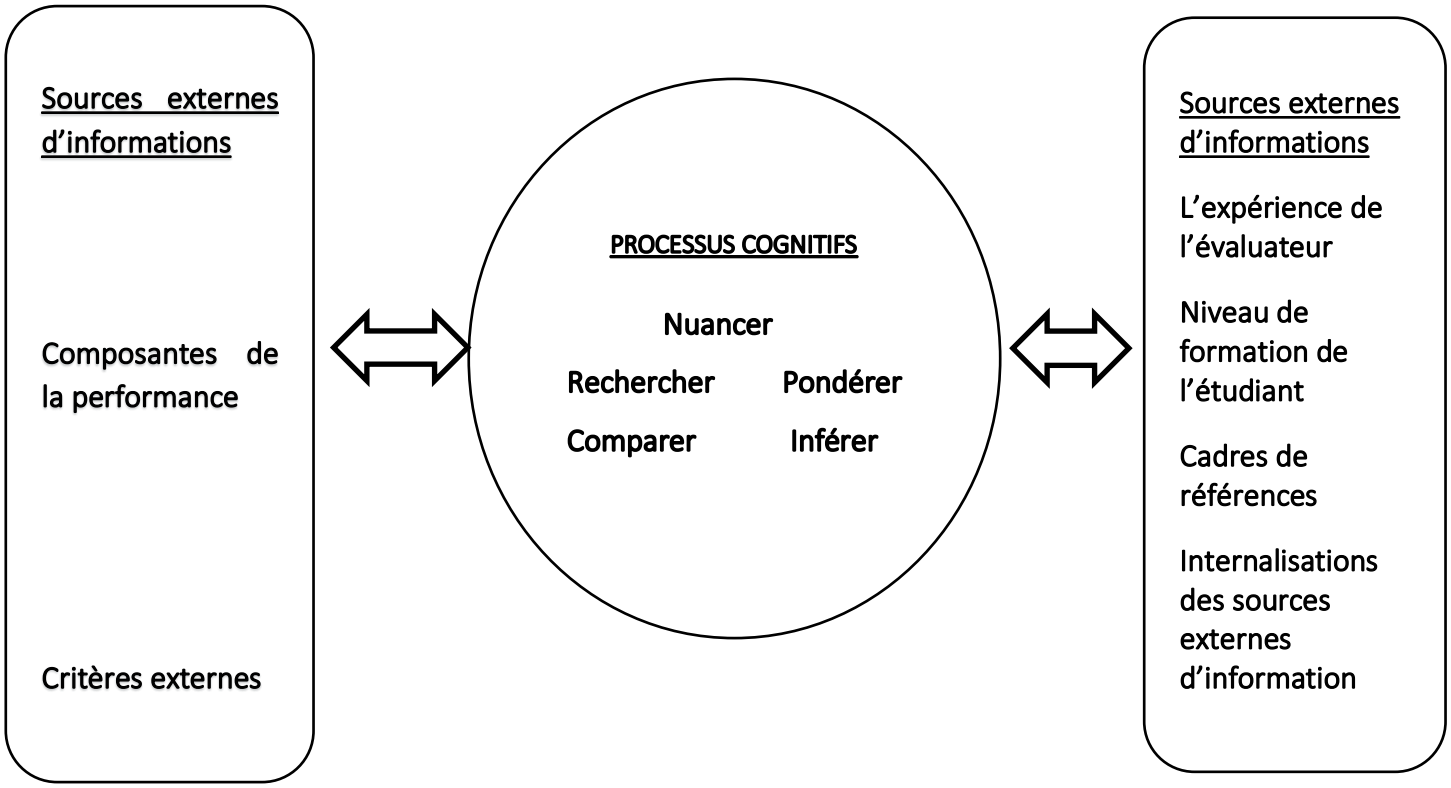
Après avoir scruté en détail les divers facteurs et biais qui influent sur la variabilité du jugement des enseignants en éducation médicale, nous avons identifié plusieurs éléments clés, notamment le manque d'uniformité dans l'utilisation des normes de référence, la génération automatique d'impressions, l'adoption d'une approche déductive pour formuler des inférences de haut niveau, la mise en avant de différentes dimensions des compétences, la translation du jugement narratif en chiffres, et l'effet de l'émotion de l'évaluateur lors du processus évaluatif. Nous notons que la variabilité du jugement des évaluateurs peut découler de diverses sources d'erreur et de biais, ainsi que de la combinaison de plusieurs sources d'information et de critères lors de l'évaluation. Les manifestations des biais cognitifs et de certains facteurs influençant le jugement des évaluateurs dans l'évaluation des résidents sont complexes et difficiles à étudier. Notre étude n'a pas réussi à identifier les mécanismes sous-jacents à l'influence des préjugés sur l'évaluation. Par conséquent, nous n'avons pas pu proposer d'explications alternatives aux résultats de la recherche de Morewedge and Kahneman (2010), qui suggèrent que l'influence des préjugés peut être expliquée par les principes de l'activation associative. Selon cette hypothèse, les jugements sont formés par une combinaison pondérée d'informations, avec une tendance à surestimer les informations fortement activées, tandis que les connaissances pertinentes mais non

activées peuvent être sous-estimées ou négligées. Étant donné l'importance de l'évaluation des résidents dans l'enseignement médical axé sur les compétences, il est nécessaire de mener des recherches futures pour mieux comprendre la génération et l'influence des biais cognitifs ainsi que des facteurs affectant le jugement des évaluateurs. Ces études devraient également examiner plus en profondeur, l'impact de ces biais sur les apprenants et explorer des interventions visant à les corriger dans le domaine de l'éducation médicale.

Chapitre 6 – Annexes.

Figure 2

Modèle conceptuel des processus cognitifs en jeu pour une évaluation significative selon St-Onge et al. (2016).



Évaluation significative

Figure 3 : Diagramme récapitulatif des principales notions présentées dans le cadre conceptuel.

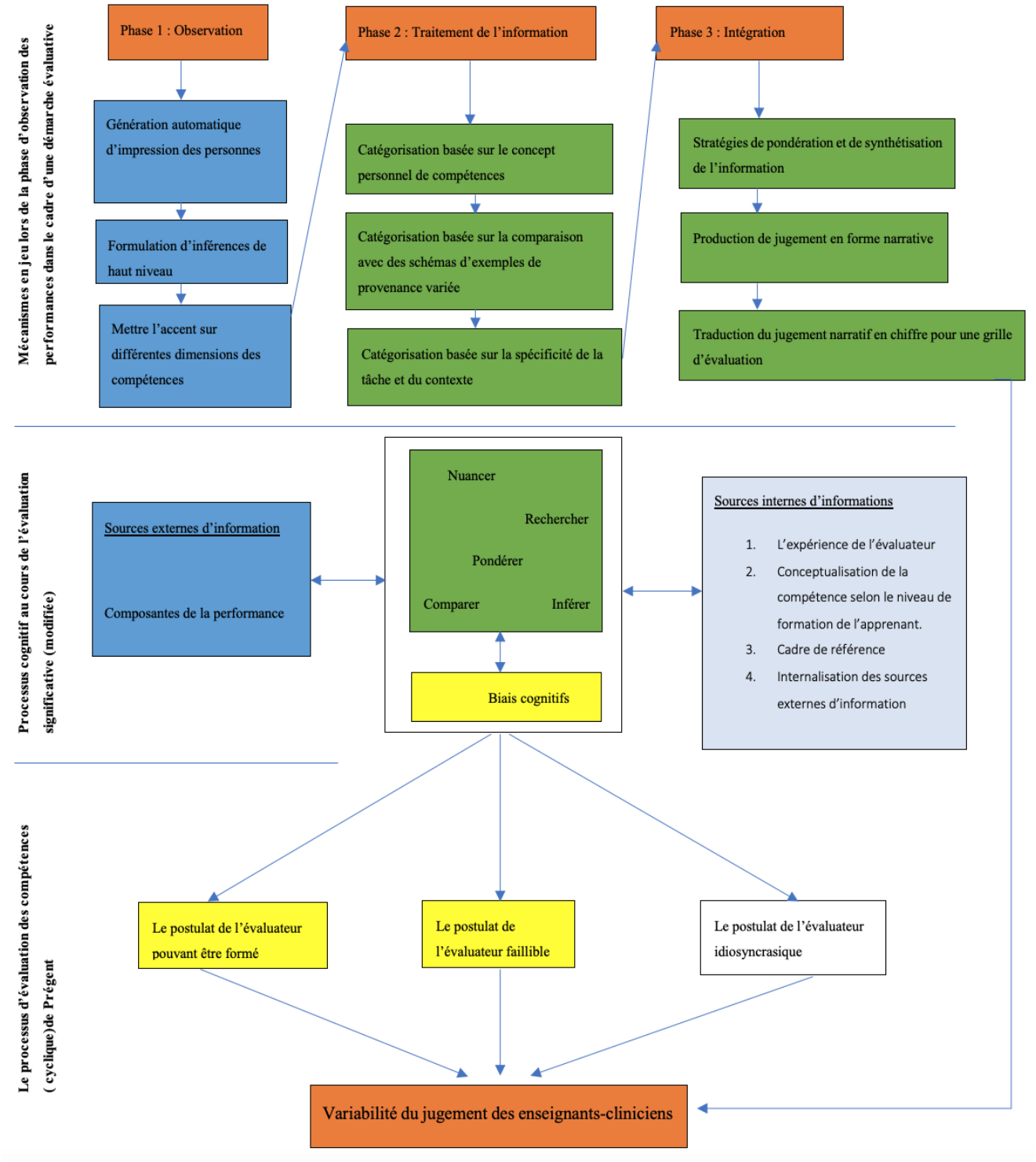


Figure 4 : Ébauche du schème de codification

Documents	De quelle manière les biais cognitifs contribuent-ils à la variabilité du jugement des évaluateurs ?												De quelle manière les facteurs environnementaux contribuent-ils à la variabilité du jugement des évaluateurs ?						
	l'évaluateur pouvant être formé			l'évaluateur faillible															
	le manque d'uniformité sur l'utilisation des normes de références pour juger des compétences des apprenants	l'évaluation qui fait fi des comportements observables au profit d'une approche par déduction	l'évaluateur modifie les jugements d'évaluation afin d'éviter des répercussions désagréables	Génération automatique d'impressions des personnes	Formulation d'inférences de haut niveau	Mettre l'accent sur différentes dimensions des compétences	Catégorisation de l'information à travers des schémas basés sur le concept personnel de compétences	Catégorisation de l'information à travers des schémas basés sur la comparaison avec des schémas d'exemples de provenance variée	Catégorisation de l'information à travers des schémas basés sur la spécificité de la tâche et du contexte	Stratégies de pondération et de synthétisation de l'information	Production de jugement en forme narrative	Traduction du jugement narratif en chiffre pour une grille d'évaluation	Effets de la pression et du manque de temps lié aux exigences des activités clinique sur la variabilité des évaluateurs	Effets d'un environnement éducatif ne suscitant pas le perfectionnement du corps professoral	Effets d'un environnement éducatif dans lequel le processus évaluatif favorise l'observation de plusieurs dimensions de la compétence en même temps (peut constituer un facteurs de distraction)	Effets d'un environnement éducatif dans lequel l'influence d'un évaluateur trop persuasif ou ayant une forte opinion (peut constituer un facteurs de	Effet de distraction d'un environnement éducatif dans lequel il y a une contribution collective de l'équipe soignante aux tâches quotidiennes de	Effet de distraction lié au fait que le stagiaire se sait observé (peut rendre sa performance clinique artificielle)	Effet de distraction de l'émotion de l'évaluateur lors du processus évaluatif
DOC 01																			
DOC 02																			
DOC 03																			

Figure 5 [28]

Exemple de stratégie de sur recherche Scopus

```
( TITLE-ABS-
KEY ( ( ( assessor* OR rater OR "Expert
assessment*" OR "Clinical
trainer" OR preceptor* OR teacher* OR ( ( rater W/2
assessment ) OR judgement* OR ( performance W/2 (
assess* OR evaluation ) ) OR "Clinical
assess*" OR ( training W/2 ( rating OR assessment ) )
OR "Educational
Measurement" ) ) W/8 ( bias OR variabilit* OR variable
* OR variance OR emotion OR "Examiner
Variability" OR "Examiner
factor*" OR discrimination ) ) ) ) AND ( ( TITLE-ABS-
KEY ( ( "Clinical
Medicine" OR ( competency W/2 education ) OR "Clini
cal
Competenc*" OR internship OR residency ) ) ) OR ( TI
TLE-ABS-
KEY ( ( ( medical OR medicine ) W/2 ( undergraduate
OR postgraduate OR student OR residen* ) ) ) ) ) AND
PUBYEAR > 2009 AND ( LIMIT-
TO ( LANGUAGE , "English" ) OR LIMIT-
TO ( LANGUAGE , "French" ) )
```

[28]Une erreur concernant la période de recherche s'est introduite dans la stratégie de recherche bibliographique. Cependant, elle a été rectifiée lors de la sélection des articles en appliquant les critères d'inclusion.

Figure 6

Exemple de stratégie de recherche sur Web-of-Science

((((assessor* OR rater OR "Expert assessment*" OR "Clinical trainer" OR preceptor* OR teacher* OR ((rater NEAR/2 assessment) OR judgement* OR (performance NEAR/2 (assess* OR evaluation)) OR "Clinical assess*" OR (training NEAR/2 (rating OR assessment)) OR "Educational Measurement"))) NEAR/8 ((bias OR variabilit* OR variable* OR variance OR emotion OR "Examiner Variability" OR "Examiner factor*" OR discrimination)))) AND ((("Clinical Medicine" OR (competency NEAR/2 education) OR "Clinical Competenc*" OR internship OR residency)) OR (((medical OR medicine) NEAR/2 (undergraduate OR postgraduate OR student OR residen*)))
) (Topic) and **French** or **English** (Languages) and **2022** or **2021** or **2020** or **2019** or **2018** or **2017** or **2016** or **2015** or **2012** or **2011** or **2010** or **2009** or **2014** or **2013** (Publication Years)

Figure 7

Exemple de stratégie de recherche sur Ovid MEDLINE

((Assessor* or Rater or Expert assessment* or Clinical trainer or Preceptor* or Teacher* or ((Rater adj2 assessment) or judgement* or (Performance adj2 (Assess* or evaluation)) or Clinical assess* or (Training adj2 (rating or Assessment)) or Educational Measurement)) adj8 (bias or variabilit* or variable* or variance or Emotion or Examiner Variability or Examiner factor* or discrimination)).ti,ab,kf.
Cognitive Biases/ or Social Bias/ or Gender Bias/ or Racial Bias/ or Cultural Differences/ or Minority Groups/ or Negative attitudes/ or Social Discrimination/ or Stereotypes/ or Observer Variation/
Cognition/ or Educational Measurement/ or Judgment/ or Mental Processes/ or Decision Making/
and/2-3
or/1,4
((medical or medicine) adj2 (undergraduate or postgraduate or student or residen*)).ti,ab,kf.
(Clinical Medicine or (Competency adj2 Education) or Clinical Competenc* or Internship or Residency).ti,ab,kf.
exp Education, Medical/
or/6-8
and/5,9
limit 10 to yr="2009 -Current"
limit 11 to (english or french)

Tableau 9.

Mécanismes en jeu lors de la phase d'observation des performances dans le cadre d'une démarche évaluative, selon Gauthier et al. (2016).

Phase 1 : Observation	**Brève description des mécanismes**
Génération automatique d'impressions des personnes	Biais cognitif de généralisation (effet de halo) où des aspects sociaux ou des traits personnels influencent inconsciemment le jugement
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
Mettre l'accent sur différentes dimensions des compétences	Attention portée à différents éléments d'une performance en fonction d'une compréhension différente des composantes de compétences

Tableau 10.

Mécanismes en jeu lors de la phase de traitement de l'information dans le cadre d'une démarche évaluative, selon Gauthier et al. (2016).

Phase 2 : Traitement de l'information	**Brève description des mécanismes**
Catégorisation de l'information à travers des schémas basés sur:	
Le concept personnel de compétences	Englobe : 1) une composante expérientielle ; 2) des éléments de conception personnelle de bonne performance ; 3) des références externes intériorisées et : 4) des définitions opérationnelles variées des compétences ciblées
La comparaison avec des schémas d'exemples de provenance variée	Utilisation d'exemples en lien avec les apprenants précédents, soi-même, des collègues, d'anciens apprenants, pour établir des points de référence et des standards pour comparer des composantes de la performance
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 laquelle s'ajoutent les buts et le contexte de l'évaluation qui, eux, jouent un rôle d'agents médiateurs sur l'évaluation

Tableau 11.

Mécanismes en jeu lors de la phase d'intégration de l'information dans le cadre d'une démarche évaluative, selon Gauthier et al. (2016).

Phase 3 : Intégration	**Brève description des mécanismes**
Stratégies de pondération et de synthétisation de l'information	Ces stratégies sont diverses et variées ; elles incluent une pondération égale de toutes les sources d'information dont les évaluateurs sont conscients, une priorisation par rapport à l'importance relative des composantes d'une compétence donnée ou encore des aspects d'un comportement qui équivalent à un échec, ou d'une perte de confiance face au raisonnement de l'apprenant.
Production de jugement en forme narrative	Le jugement se développe en forme narrative et les demandes de comparaison et de catégorisation des performances complexes créent beaucoup d'incertitude chez les évaluateurs
Traduction du jugement narratif en chiffre pour une grille d'évaluation	La traduction d'un jugement global en description narrative qui est subséquemment traduite en pointage pour les différentes échelles d'une grille d' évaluation

Tableau 12.

Tableau d'extraction des données de la recherche 1/4

AUTEUR	TITRE DE L'ARTICLE	LANGUE DE PUBLICATION	PAYS DE PUBLICATION	REVUE DE PUBLICATION	TYPE D'ARTICLE
Colson (2020)	Washington University School of Medicine in St. Louis Case Study_ A Process for Understanding.	Anglais	USA	Academic Medicine	Recherche, QI
Forte (2021)	How Teachers Adapt Their Cognitive Strategies When Using Entrustment Scales.	Anglais	CANADA	Academic Medicine	Recherche
Gardner (2016)	Repaying in Kind_ Examination of the Reciprocity Effect in Faculty and Resident Evaluations.	Anglais	USA	Journal of Surgical Education	Recherche (retrospective) retrospective review
Gingerich (2011)	Repaying in Kind_ Examination of the Reciprocity Effect in Faculty and Resident Evaluations.	Anglais	CANADA	Academic Medicine	Revue non systématique de la littérature
Gingerich (2018)	Comparatively salient/ examining the influence of preceding performances on assessors' focus and interpretations in written assessment comments.	Anglais	UK	Advances in Health Sciences Education / Home-spinger	Recherche
Gomez-Garibello (2018)	Emotions and assessment: considerations for rater-based judgements of entrustment.	Anglais	CANADA	The cross-cutting edge - Medical Education	Revue non systématique de la littérature (narrative)
Klein (2019)	Gender Bias in Resident Assessment in Graduate Medical Education	Anglais	USA	Journal of General Internal Medicine	Revue systématique de la littérature
Klein(2020)	Association of Gender with Learner Assessment in Graduate Medical Education.	Anglais	USA	JAMA Network Open | Medical Education	Recherche
Kogan (2021)	Opening the black box of clinical skills assessment via observation_ a conceptual model.	Anglais	USA	Medical Education	Recherche
Lee (2017)	Factors Influencing Mini-CEX Rater Judgments and Their Practical Implications_ A Systematic Literature Review.pdf	Anglais	AUSTRALIE	Academic Medicine	Revue systématique de la littérature
Low (2019)	Racial Ethnic Disparities in Clinical Grading in Medical School.pdf	Anglais	USA	Teaching and Learning in Medicine An International Journal /Routledge	Recherche: Étude rétrospective des évaluations
Paget (2013)	Rater variables associated with ITER ratings.	Anglais	CANADA	Advances in Health Sciences Education / Home-spinger	Recherche
Polanco-Santana (2021)	Ethnic_Racial Bias in Medical School Performance Evaluation of General Surgery Residency Applicants.	Anglais	USA	Journal of Surgical Education	Recherche (retrospective)
Rojek (2019)	Differences in Narrative Language in Evaluations of Medical Students by Gender and Under-represented Minority Statu	Anglais	USA	Journal of General Internal Medicine.	Recherche
Ross (2017)	Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations.	Anglais	USA	PLOS ONE	Recherche
Shaw (2021)	How biased are you_ The effect of prior performance information on attending physician ratings and implications for learner handover.	Anglais	CANADA	Advances in Health Sciences Education / Home-spinger	Recherche
Sherbino (2013)	The reliability of encounter cards to assess the CanMEDS roles.	Anglais	CANADA	Advances in Health Sciences Education / Home-spinger	Recherche
St-Onge (2016)	Expectations, observations, and the cognitive processes that bind them_ expert assessment of examinee performance.	Anglais	CANADA	Advances in Health Sciences Education / Home-spinger	Recherche
Wood (2014)	Exploring the role of first impressions in rater-based assessments.	Anglais	CANADA	Advances in Health Sciences Education / Home-spinger	Revue non systématique de la littérature (narrative)
Wood (2017)	The influence of first impressions on subsequent ratings within an OSCE station	Anglais	CANADA	Advances in Health Sciences Education / Home-spinger	Recherche
Yeates (2013)	Seeing the same thing differently-Mechanisms that contribute to assessor differences in directly.	Anglais	UK	Medical Education	Recherche
Yeates (2013)	'You're certainly relatively competent'_ assessor bias due to recent experiences pdf.	Anglais	CANADA	Medical Education	Recherche
Yeates (2015)	Are Examiners' Judgments in OSCE-Style Assessments Influenced by Contrast Effects_.	Anglais	UK/CANADA	Academic medicine	Recherche
Yeates (2015)	Relatively speaking_ contrast effects influence assessors' scores and narrative feedback - Yeates - 2015 - Medical Education - Wiley Online Library.	Anglais	UK	Medical Education	Recherche

Tableau 13.

Tableau d'extraction des données de la recherche. 2/4

PREMIER AUTEUR	TITRE DE L'ARTICLE	OBJECTIFS DE RECHERCHE (But, QR, HR)	APPROCHE MÉTHODOLOGIQUE Méthode de recherche à la collecte de données: Qualitative, quantitative, mixte
Colson (2020)	Washington University School of Medicine in St. Louis Case Study_ A Process for Understanding.	WUSM (Washington University School of Medicine in St. Louis) leadership evaluated whether students' race, ethnicity, and gender were associated with their receipt of honors in the 6 core clerkships, key determinants of AOA selection. This case study describes WUSM's process to understand and address bias in clerkship grading and AOA nomination so that other medical schools might benefit from what has been learned.	Mixed methods (Qualitative and Quantitative)
Forte (2021)	How Teachers Adapt Their Cognitive Strategies When Using Entrustment Scales.	we aimed to provide a conceptualization of the variable ways in which teachers' cognitive strategies underpinned how they formulated. and recorded assessments of trainees.	Qualitative
Gardner (2016)	Repaying in Kind_ Examination of the Reciprocity Effect in Faculty and Resident Evaluations.	The purpose of this study was to investigate the extent to which lenient-grading faculty receive higher evaluations from surgery residents.	Quantitative
Gingerich (2011)	Repaying in Kind_ Examination of the Reciprocity Effect in Faculty and Resident Evaluations.	This critical review examines investigations of rater idiosyncrasy from impression formation literatures to ask new questions for the parallel problem in rater-based assessments. This paper represents a synthesis of related research domains focused on understanding the source of variance in social judgments. Although the measurement limitations in rater-based assessments undoubtedly stem from many complex factors, this paper explores the perplexing origins of rater variance when raters observe the same act. This paper is necessarily nonsystematic and non-exhaustive in order to present a preliminary understanding of vast literatures investigating problems analogous to those with rater-based assessments. Accordingly, the intent is to stimulate different ways of asking questions about the limitations of rater- based assessments prior to negotiating potential solutions.	Qualitative
Gingerich (2018)	Comparatively salient/ examining the influence of preceding performances on assessors' focus and interpretations in written assessment comments.	we investigated assessors' cognition by using the insight provided by "clusters of consensus" to determine whether any change in the salience of performance aspects was induced by contrast effects.	Mixed methods
Gomez-Garibello (2018)	Emotions and assessment: considerations for rater-based judgements of entrustment.	In this narrative review, we explore the influence of raters' emotions in the assessment of learners.	Qualitative
Klein (2019)	Gender Bias in Resident Assessment in Graduate Medical Education	We sought to examine the available evidence of the potential for and impact of gender bias in resident assessment in graduate medical education.	Qualitative
Klein (2020)	Association of Gender with Learner Assessment in Graduate Medical Education.	How is gender associated with faculty assessment of internal medicine resident performance?	Quantitative
Kogan (2011)	Opening the black box of clinical skills assessment via observation_ a conceptual model.	This study was intended to develop a conceptual framework of the factors impacting on faculty members' judgements and ratings of resident doctors (residents) after direct observation with patients.	Qualitative
Lee (2017)	Factors Influencing Mini-CEX Rater Judgments and Their Practical Implications_ A Systematic Literature Review.pdf	The authors of this systematic literature review aim both to identify the factors influencing mini-CEX rater judgments in the medical education setting and to translate these findings into practical implications for clinician assessors.	Qualitative
Low (2019)	Racial Ethnic Disparities in Clinical Grading in Medical School.pdf	To evaluate the association between race/eth- nicity and clinical grading, we examined Medical Student Performance Evaluation (MSPE) summary words (Outstanding, Excellent, Very Good, Good) and 3rd-year clerkship grades among medical students at the University of Washington School of Medicine. The analysis included data from July 2010 to June 2015.	Quantitative
Paget (2013)	Rater variables associated with ITER ratings.	Here our objective was to study the association between rater variables and ITER ratings.	Quantitative
Polanco-Santana (2021)	Ethnic_Racial Bias in Medical School Performance Evaluation of General Surgery Residency Applicants.	We evaluated bias in medical student performance evaluations (MSPE) of general surgery residency applicants	Quantitative

Rojek (2019)	Differences in Narrative Language in Evaluations of Medical Students by Gender and Under-represented Minority Statu	To identify and enumerate text descriptors in a database of medical student evaluations using natu- ral language processing and identify differences by gen- der and URM status in descriptions.	Quantitative
Ross (2017)	Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations.	he transitions from medical school to residency is a critical step in the careers of physicians. Because of the standardized application process–wherein schools submit summative Medical Student Performance Evaluations (MSPE's)–it also represents a unique opportunity to assess the possible prevalence of racial and gender disparities, as shown elsewhere in medicine.	Quantitative
Shaw (2021)	How biased are you_ The effect of prior performance information on attending physician ratings and implications for learner handover.	The purpose of this study is thus to determine whether exposure to positive or nega- tive PPI in the form of LH influences scoring of subsequent performances in the medi- cal education context and if so in what direction. We will also aim to ascertain how participants use LH in their assessments.	Quantitative
Sherbino (2013)	The reliability of encounter cards to assess the CanMEDS roles.	The purpose of this study was to determine the reliability of a computer-based encounter card (EC) to assess medical students during an emergency medicine rotation.	Quantitative
St-Onge (2016)	Expectations, observations, and the cognitive processes that bind them_ expert assessment of examinee performance.	the purpose of this study was to qualitatively investigate the cognitive processes of raters, and to create a framework informed by the Theory of Expertise and the Dual-Process Theories of Rea- soning that conceptualizes those processes when raters assess a complex performance.	Qualitative
Wood (2014)	Exploring the role of first impressions in rater-based assessments.	The goal of this paper, therefore, is to contribute to a better understanding of the cognitive processes used by raters.	Qualitative
Wood (2017)	The influence of first impressions on subsequent ratings within an OSCE station	This study explores the influence of first impressions in an OSCE. Specifically, the purpose is to begin to examine the accuracy of a first impression and its influence on subsequent ratings.	Quantitative
Yeates (2013)	Seeing the same thing differently- Mechanisms that contribute to assessor differences in directly.	To explore assessors' cognitive processes whilst undertaking performance assessments, and thereby to illuminate the sources of variability in assessor judgements.	Qualitative
Yeates (2013)	'You're certainly relatively competent'_ assessor bias due to recent experiences pdf.	we investigated whether confidence in the rating assigned predicts susceptibility to manipulation and whether prompting consideration of typical performance lessens the influence of recent experience.	Quantitative
Yeates (2015)	Are Examiners' Judgments in OSCE-Style Assessments Influenced by Contrast Effects_.	Laboratory studies have shown that performance assessment judgments. can be biased by "contrast effects." Assessors' scores become more positive, for example, when the assessed performance is preceded by relatively weak candidates. The authors queried whether this effect occurs in real, high- stakes performance assessments despite increased formality and behavioral descriptors.	Quantitative
Yeates (2015)	Relatively speaking_ contrast effects influence assessors' scores and narrative feedback - Yeates - 2015 - Medical Education - Wiley Online Library.	This study examines the mechanism and robustness of these findings to advance under- standing of assessor cognition. We test the influence of the immediately preceding performance relative to that of a series of prior performances. Further, we examine whether assessors' narrative comments are similarly influenced by contrast effects.	Quantitative

Tableau 14.

Tableau d'extraction des données de la recherche. 3/4

PREMIER AUTEUR	TITRE DE L'ARTICLE	MÉTHODE DE RECUEIL DES DONNÉES par ex., sondages, entretiens, études de cas,analyse d'échelle de mesure
Colson (2020)	Washington University School of Medicine in St. Louis Case Study_ A Process for Understanding.	1-Sondage anonyme sur le web: To start, students were invited to participate in 2 main activities: an anonymous web- based survey and/or focus groups. 2) Groupes de discussion (focus groupe.) 3) Questionnaire structuré: We also examined student comments from the AAMC Graduation Questionnaire administered to our graduating students in 2017 and from the most recent Year 2 Questionnaire. From these qualitative data, we observed 2 main overarching themes: assessment and the learning environment.
Forte (2021)	How Teachers Adapt Their Cognitive Strategies When Using Entrustment Scales.	1) un entretien semi structuré : Through cognitive interviews and retrospective verbal protocol analysis, we explored how faculty used the new ES to assess learners.25,26 Specifically, we used a "think aloud" protocol to explore what the anchors meant to them and how they decided when to use them. This allowed us to gain an understanding of the cognitive processes teachers employed when making rating decisions by asking them to describe rather than explain their thoughts.27 .
Gardner (2016)	Repaying in Kind_ Examination of the Reciprocity Effect in Faculty and Resident Evaluations.	A-L'analyse de contenu: A retrospective review of 2 years of faculty evaluations of residents and resident evaluations of faculty was conducted. The 2012 to 2014 timeline of this review was chosen to provide the most recent data that coincided with a recent departmental change in evaluation forms and information management systems.
Gingerich (2011)	Repaying in Kind_ Examination of the Reciprocity Effect in Faculty and Resident Evaluations.	A-L'analyse de contenu: This paper represents a synthesis of related research domains focused on understanding the source of variance in social judgments. (MEDLINE, ERIC, and PsycINFO were used to search for articles investigating social judgment processes including impression formation and associated sociocognitive processes.)
Gingerich (2018)	Comparatively salient/ examining the influence of preceding performances on assessors' focus and interpretations in written assessment comments.	1) -Observation structurée participante avec intervention: We used existing data from an experimental study designed to investigate contrast effects (Yeates et al. 2015a) as the stimulus material for further data collection. The dataset contained scores and written comments collected in response to video-recorded clinical encounters that (a) had been exposed to conditions that induced contrast effects resulting in significant differences; (b) had been exposed to conditions that did not induce contrast effects, therefore producing no significant differences; (c) had not been exposed to conditions intended to induce contrast effects (i.e. an "unbiased" or control condition).
Gomez-Garibello (2018)	Emotions and assessment: considerations for rater-based judgements of entrustment.	A-L'analyse de contenu: We summarise existing literature that describes the role of emotions in assessment broadly, and rater-based assessment specifically, across a variety of fields.
Klein (2019)	Gender Bias in Resident Assessment in Graduate Medical Education	A-L'analyse de contenu: A comprehensive literature review was performed to capture relevant primary studies for inclusion into this review.
Klein (2020)	Association of Gender with Learner Assessment in Graduate Medical Education.	A-L'analyse de contenu (Étude rétrospective et transversale des évaluations des résidents par les professeurs) Data included faculty assessments of internal medicine resident performance during general medicine inpatient rotations from July 1, 2016, to June 30, 2017. Assessment data included 20 of 22 internal medicine–specific reporting Milestones and 6 core competencies (patient care, medical knowledge, systems-based practice [SBP], practice-based learning and improvement [PBLI], professionalism, and interpersonal and communication skills [ICS]). 1) Questionnaire structuré Each site used a unique assessment tool, which, in aggregate, included 130 quantitative questions, 45 of which used exact wording of the ACGME's reporting Milestones and 85 of which used variations of the Milestones wording. We also collected resident and faculty demographic data as well as rotation setting and date.
Kogan (2011)	Opening the black box of clinical skills assessment via observation_ a conceptual model.	I-Questionnaire structuré sur les caractéristiques démographiques: Prior to their assigned study day, faculty members completed a web-based demographic questionnaire that has been previously described.20 II-Protocole A: (Repeat sequence with other video encounters) 1) -Observation structurée participante avec intervention: Faculty watch video of SP and SR, list SR's strength/weaknesses, then rates SR using mini-CEX (15 minutes) 2)- Une échelle de Likert à 9 points: After watching each of four video encounters (Fig. 1a), faculty staff completed a mini-clinical evaluation exercise (mini-CEX). The mini-CEX, developed by the American Board of Internal Medicine (ABIM) to provide residents with feed- back about their history-taking, physical examina- tion, counselling and interpersonal skills, details seven competencies that are rated on a 9-point scale (1–3 = unsatisfactory, 4–6 = satisfactory, 7–9 = superior).25,26 3) Entrevue semi-structurée: Study investigator interviews faculty about their observations, ratings, and feedback (15 minutes) III-Protocole B: (Repeat sequence with 2nd live encounter) 1) -Observation structurée participante avec intervention: Faculty observes SR with SP live (15 minutes) ==> SP and SR leave room ==> Faculty completes mini-CEX (5 minutes)==> SR returns.Faculty gives SR feedback (10 minutes)

		2)- Une échelle de Likert à 9 points: After watching each of four video encounters (Fig. 1a), faculty staff completed a mini-clinical evaluation exercise (mini-CEX). The mini-CEX, developed by the American Board of Internal Medicine (ABIM) to provide residents with feed- back about their history-taking, physical examina- tion, counselling and interpersonal skills, details seven competencies that are rated on a 9-point scale (1–3 = unsatisfactory, 4–6 = satisfactory, 7–9 = superior).25,26 3) Entrevue semi-structurée: SR leaves room ==> Faculty interviewed by study investigator about observations, ratings, feedback (10 minutes)
Lee (2017)	Factors Influencing Mini-CEX Rater Judgments and Their Practical Implications_ A Systematic Literature Review.pdf	A-L'analyse de contenu: The authors searched for internal and external factors influencing mini-CEX rater judgments in the medical education setting from 1980 to 2015 using the Ovid MEDLINE, PsycINFO, ERIC, PubMed, and Scopus databases. They extracted the following information from each study: country of origin, educational level, study design and setting, type of observation, occurrence of rater training, provision of feedback to the trainee, research question, and identified factors influencing rater judgments. The authors also conducted a quality assessment for each study.
Low (2019)	Racial Ethnic Disparities in Clinical Grading in Medical School.pdf	A-L'analyse de contenu: we examined Medical Student Performance Evaluation (MSPE) summary words (Outstanding, Excellent, Very Good, Good) and 3rd-year clerkship grades among medical students at the University of Washington School of Medicine. The analysis included data from July 2010 to June 2015. B-Échelle numérique:This value is used to create a summary word that groups students into four designations: Outstanding, Excellent, Very Good, and Good. One of these four descriptors is noted in the final paragraph of each student's MSPE to summarize their overall 3rd-year performance.
Paget (2013)	Rater variables associated with ITER ratings.	1) -Observation structurée transversale participante avec intervention From our online assessment program we recorded the rating for the student on the global rating scale and calculated the time interval in days between the last day of the rotation and the date that the preceptor submitted the completed ITER online. 2) -Une échelle de Likert à 5 points: Each of our clerkship rotation uses the same ITER that includes a global rating scale for the overall assessment of performance. On this five point Likert scale the descriptors are: 1 = unsatisfactory, 2 = borderline, 3 = at expected level, 4 = above expected level, and 5 = outstanding.
Polanco-Santana (2021)	Ethnic_Racial Bias in Medical School Performance Evaluation of General Surgery Residency Applicants.	A-L'analyse de contenu Étude rétrospective évaluant le biais ethnique / race, mesuré par l'utilisation différentielle des termes agentiques et communaux, dans les MSPE (medical student performance evaluations) des candidats à la résidence.
Rojek (2019)	Differences in Narrative Language in Evaluations of Medical Students by Gender and Under-represented Minority Statu	A-L'analyse de contenu Revue de littérature Étude observationnelle des évaluations narratives
Ross (2017)	Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations.	A-L'analyse de contenu Revue de littérature
Shaw (2021)	How biased are you_ The effect of prior performance information on attending physician ratings and implications for learner handover.	1) -Observation structurée participante avec intervention, Les participants ont été instruits d'évaluer les performances cliniques d'un apprenant interagissant avec un patient dans une série de vidéos après avoir reçu des commentaires négatifs, positifs ou aucun commentaire. Les performances ont été évaluées immédiatement après chaque visionnage de vidéo à l'aide de l'outil mini Clinical Evaluation Exercise (mini-CEX) 2) Ils devaient répondre sur deux échelles de Likert à 9 points L'outil mini-CEX (« Annexe 2 ») comprend six échelles de notation à 9 points représentant différentes compétences, ainsi qu'une échelle à 9 points représentant la compétence clinique globale. 3) Questionnaire : Un court questionnaire a été administré pour évaluer comment les évaluateurs ont utilisé la LH fournie, leurs impressions générales sur la LH et sa crédibilité (« Annexe 5 »). Comme l'utilisation formelle de LH n'est pas la pratique habituelle dans cette université, le questionnaire a également été utilisé pour déterminer si les participants étaient capables de déduire la véritable finalité de l'étude.
Sherbino (2013)	The reliability of encounter cards to assess the CanMEDS roles.	1) -Observation structurée participante avec intervention évaluer les étudiants en médecine lors d'une rotation en médecine d'urgence. 2) -Une échelle de Likert à dix points: À la fin d'une garde au service des urgences, une EC a été évaluée (de 1 à 10) pour chaque étudiant sur la compétence d'Expert Médical, 2 autres rôles supplémentaires et une note globale. L'analyse de 1 819 ECs (155 sur 186 étudiants) a révélé ce qui suit : Collaborateur, Manager, Défenseur de la santé et Érudit ont été évalués dans moins de 25 % des ECs.
St-Onge (2016)	Expectations, observations, and the cognitive processes that bind them_ expert assessment of examinee performance.	A) Entrevue semi-structurée: Nous avons mené des entretiens semi-structurés avec des participants identifiés comme d'excellents évaluateurs et nous leur avons demandé d'exprimer verbalement leurs réflexions sur la performance clinique d'un stagiaire afin d'explorer leur processus cognitif.
Wood (2014)	Exploring the role of first impressions in rater-based assessments.	A-L'analyse de contenu Une revue narrative synthétisera la recherche dans ces trois domaines (jugement social et prise de décision, psychologie éducative et psychologie cognitive) et se concentrera sur les processus cognitifs sous-jacents, la précision et l'impact des premières impressions sur les évaluations basées sur les évaluateurs.
Wood (2017)	The influence of first impressions on subsequent	1) -Observation structurée participante avec intervention, évalué et noté des vidéos de performances standardisées, scénarisées , scriptées et réalisé par des résidents

	ratings within an OSCE station	2) Ils devaient répondre sur une échelle de Likert à 6 points:Un total de 23 évaluateurs (c'est-à-dire des examinateurs médecins) ont visionné chaque vidéo et ont été invités à évaluer globalement les compétences cliniques de l'examinateur après 60 secondes (First Impression GR) en fournissant une note sur une échelle d'évaluation globale à six points, puis à évaluer leur confiance dans la précision de ce jugement en donnant une note sur une échelle d'évaluation à cinq points (Confidence GR) 3) Ils devaient répondre sur une échelle de Likert à 5 points: La deuxième série d'évaluations (voir la section "Annexe") a été réalisée à la fin de chaque vidéo. Cette série comprenait une échelle d'évaluation en cinq points (OSCE Scale) inspirée de l'échelle d'évaluation OSCE utilisée dans la Collaboration nationale d'évaluation du Conseil médical du Canada OSCE
Yeates (2013)	Seeing the same thing differently- Mechanisms that contribute to assessor differences in directly.	1) -Observation structurée participante et sans intervention, évalué et noté des vidéos de performances standardisées, scénarisées , scriptées et réalisées par des médecins de première année de fondation (année PG1). 2) les évaluateurs ont commenté leurs réflexions à la fois pendant le visionnage et de manière rétrospective. (Les participants ont ensuite participé à des entretiens ciblés de suivi). les commentaires des participants ont été analysés en utilisant les principes de la théorie ancrée (Strauss et Corbin 1998). 3) 2) Ils devaient répondre sur une échelle de Likert à 6 points: Les évaluateurs notent 7 domaines de la performance (interrogatoire; examen physique; compétences en communication; jugement critique; professionnalisme; organisation/efficacité; et prise en charge clinique globale) en utilisant une échelle de Likert à 6 points ancrée au point 4 par rapport au critère de "répond aux attentes pour l'achèvement de la première année F1". Le point 3 représente "limite pour l'achèvement de la F1", les autres points comprenant "bien en dessous", "en dessous", "au-dessus" et "bien au-dessus" de ce critère, ainsi que "incapable de commenter
Yeates (2013)	'You're certainly relatively competent'_ assessor bias due to recent experiences pdf.	1) -Observation structurée participante avec intervention de, performances scénarisées par de véritables médecins de première année de fondation (F1) lors de consultations avec des patients simulés 2) Ils devaient répondre sur une échelle de Likert à 7 points: Cette échelle était ancrée au point 1 avec le texte "Je n'ai aucune confiance du tout dans les scores que j'ai donnés", "J'ai de sérieux doutes sur mes scores, les scores "corrects" pourraient facilement être différents des scores que j'ai donnés". Au point 7, l'échelle était ancrée avec "Je suis très confiant dans les scores que j'ai donnés". 3) Questionnaire structuré fermé à choix multiple sur "Le pourcentage de F1 qui ferait mieux que cette performance (5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95%)", "Le pourcentage de F1 qui ferait moins bien que cette performance (95, 90, 80, 70, 60, 50, 40, 30, 20, 10, 5%)" et "Le pourcentage de F1 qui ferait aussi bien que cette performance (de 0 à 95% par incréments de 5%)".
Yeates (2015)	Are Examiners' Judgments in OSCE-Style Assessments Influenced by Contrast Effects_.	A-L'analyse de contenu Premier ensemble de donné a été extrait de l'Évaluation Clinique du Programme de Fondation du Royaume-Uni (UKFPO) de 2011.La réussite à cette évaluation est une exigence pour tout diplômé ayant quitté l'école de médecine depuis plus de deux ans avant de commencer le Programme de Fondation. 1-Observation structurée 2-Une échelle de Likert à cinq points allant de "échec clair" à "excellent" Le deuxième ensemble de données provenait du Multiple Mini Interview (MMI) de 2008 utilisé pour la sélection à la Faculté de Médecine de l'Université de l'Alberta. 1-Multiple Mini Interview (MMI) 2-Une échelle de Likert à cinq points allant de "Insatisfaisant, En dessous de la moyenne, Moyen, Au-dessus de la moyenne, et Excellent."
Yeates (2015)	Relatively speaking_ contrast effects influence assessors' scores and narrative feedback - Yeates - 2015 - Medical Education - Wiley Online Library.	1) -Observation structurée participante avec intervention de performances par des médecins en première année de fondation (FY-1) en consultation avec un patient simulé. 2) -Une échelle de Likert à six points allant de 1 = bien en dessous des attentes ; 2 = en dessous des attentes ; 3 = limite ; 4 = atteint les attentes ; 5 = au-dessus des attentes ; et 6 = largement au-dessus des attentes. 3)-Questionnaire non structuré (questions ouvertes) sur la performance de l'étudiant

Tableau 15.

Tableau d'extraction des données de la recherche. 4/4

PREMIER AUTEUR	TITRE DE L'ARTICLE	PARTICIPANTS À LA RECHERCHE étudiant en médecine, précepteurs

Colson (2020)	Washington University School of Medicine in St. Louis Case Study_ A Process for Understanding.	Medical student: In these revised analyses, using data for students who matriculated between academic years (AYs) 2008-2009 and 2015-2016 and took the shelf exam between AYs 2011-2012 and 2017-2018 (n = 840)
Forte (2021)	How Teachers Adapt Their Cognitive Strategies When Using Entrustment Scales.	family medicine residents Medical teachers (FM staff.)
Gardner (2016)	Repaying in Kind_ Examination of the Reciprocity Effect in Faculty and Resident Evaluations.	A total of 1480 resident assessments and 2274 faculty assessments were included in this study
Gingerich (2011)	Repaying in Kind_ Examination of the Reciprocity Effect in Faculty and Resident Evaluations.	Medical students
Gingerich (2018)	Comparatively salient/ examining the influence of preceding performances on assessors' focus and interpretations in written assessment comments.	La population étudiée était des professionnels de la santé, un professionnel de la santé et des stagiaires et des coordinateurs d'évaluation.
Gomez-Garibello (2018)	Emotions and assessment: considerations for rater-based judgements of entrustment.	Étudiant en medecine et enseignant
Klein (2019)	Gender Bias in Resident Assessment in Graduate Medical Education	Residents et enseignants
Klein (2020)	Association of Gender with Learner Assessment in Graduate Medical Education.	Résidents et enseignants
Kogan (2011)	Opening the black box of clinical skills assessment via observation_ a conceptual model.	Résidents et enseignants Of the 48 faculty staff who agreed to participate, 44 (92%) completed the study; Data collection occurred between March and August 2009 with three to six faculty staff participating each study day. On their study day, faculty members individually watched four videos and two live scenarios of a standardised postgraduate year 2 (PGY2) resident
Lee (2017)	Factors Influencing Mini-CEX Rater Judgments and Their Practical Implications_ A Systematic Literature Review.pdf	Enseignants: mini-CEX assessors (Population)
Low (2019)	Racial Ethnic Disparities in Clinical Grading in Medical School.pdf	Résidents et enseignants
Paget (2013)	Rater variables associated with ITER ratings.	Stagiaire en médecine et Précepteurs Participants in our study were clinical clerks from the graduating classes of 2008 and 2009 at the University of Calgary, and preceptors who completed online ITERs for the man- datory rotations during the 18 month time period from February 2008 to July 2009. During the study period 234 preceptors completed 1050 on-line ITERS for 257 students.
Polanco-Santana (2021)	Ethnic_Racial Bias in Medical School Performance Evaluation of	Étudiant en Médecine

	General Surgery Residency Applicants.	
Rojek (2019)	Differences in Narrative Language in Evaluations of Medical Students by Gender and Under-represented Minority Statu	Étudiant en médecine et précepteurs :A total of 87,922 clerkship evaluations from core clinical rotations at two medical schools in dif- ferent geographic areas.
Ross (2017)	Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations.	Étudiant en médecine et précepteurs
Shaw (2021)	How biased are you_ The effect of prior performance information on attending physician ratings and implications for learner handover.	Enseignants et résidents : Les participants étaient composés de 58 membres du corps professoral et de résidents de dernière année du département de médecine de l'Université d'Ottawa.
Sherbino (2013)	The reliability of encounter cards to assess the CanMEDS roles.	Étudiant en médecine et précepteurs
St-Onge (2016)	Expectations, observations, and the cognitive processes that bind them_ expert assessment of examinee performance.	Précepteurs (ou évaluateurs ou Enseignants): In total, 18 faculty members were invited to participate as ''expert raters''; 11 accepted (7 men and 4 women). They had, on average, 15 ± 7 years of experience as an assessor and 17.64 ± 9.5 years of clinical experience.
Wood (2014)	Exploring the role of first impressions in rater-based assessments.	
Wood (2017)	The influence of first impressions on subsequent ratings within an OSCE station	Un e-mail a été envoyé à tous les médecins de la Faculté de médecine de l'Université d'Ottawa qui avaient été examinateurs d'OSCE pour l'une des épreuves administrées à l'école de médecine (n = 325). Au total, 24 évaluateurs se sont portés volontaires pour cette étude dans le délai imparti. Un participant n'a pas pu terminer l'étude, laissant ainsi les données de 23 évaluateurs pour l'analyse.
Yeates (2013)	Seeing the same thing differently- Mechanisms that contribute to assessor differences in directly.	Précepteurs All participants were consultant physicians from the North West of England. Inclusion criteria were: being a consultant for 2 years or more, assessing 5 or more Mini-CEXs per year, and comfort in assessing a general medical topic.
Yeates (2013)	'You're certainly relatively competent'_ assessor bias due to recent experiences pdf.	précepteurs The study population was consultant doctors from the UK (senior doctors working in specialities associated with general [internal] medicine, who are overseen by one of the UK Royal Colleges of Physicians). We also included emergency medicine doctors who had trained via the Royal College of Physicians' examination system, as they are fre- quently involved in supervising trainees in the management of acute presentations of general internal medicine patients. To meet inclusion cri- teria, the participants had to be comfortable assessing general internal medicine case material and estimate that they assessed at least five mini- CEXs per year.
Yeates (2015)	Are Examiners' Judgments in OSCE-Style Assessments Influenced by Contrast Effects_.	1-Étudiant en médecine (dans un programme de fondation) Le Programme de Fondation est un programme obligatoire et généraliste de deux ans visant à faciliter la transition entre l'école de médecine et la formation spécialisée. 2-Étudiants de premier cycle pour la formation médicale 3-Les évaluateurs
Yeates (2015)	Relatively speaking_ contrast effects influence assessors' scores and narrative feedback - Yeates - 2015 - Medical Education - Wiley Online Library.	1-consultants ou des médecins en formation spécialisée en médecine interne générale ou en médecine d'urgence de tout le Royaume-Uni

Chapitre 7– Bibliographie

Abernot, Y. (1993). *Les méthodes d'évaluation scolaire: techniques actuelles et innovations*: Dunod.

AFMC. (2012). *L'avenir de l'éd́ucation médicale au Canada projet postdoctoral une vision collective pour les études médicales postdoctorales au Canada*. Ottawa, Ont: Association des facultés de médecine du Canada.

Akobeng, A. K. (2005). Understanding systematic reviews and meta-analysis. *Archives of disease in childhood, 90*(8), 845-848. doi:10.1136/adc.2004.058230

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in education : principles, policy & practice*, *25*(6), 551-575. doi:10.1080/0969594X.2018.1441807

Chasseigne, G. r. (2007). *Cognition, santé et vie quotidienne*. Paris: Publibook.

Colson, E. R., Perez, M., Blaylock, L., Jeffe, D. B., Lawrence, S. J., Wilson, S. A., & Aagaard, E. M. (2020). Washington University School of Medicine in St. Louis Case Study: A Process for Understanding and Addressing Bias in Clerkship Grading. Acad Med, 95(12S Addressing Harmful Bias and Eliminating Discrimination in Health Professions Learning Environments), S131-S135. doi:10.1097/ACM.0000000000003702

Creswell, J. W. (2014). *Research design : qualitative, quantitative, and mixed methods approaches* (4th ed. ed.). Thousand Oaks: SAGE Publications.

CRMCC. (2015). *Projet de terminologie en éducation médicale*: Collège royal des médecins et chirurgiens du Canada.

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 1: origins of bias and theory of debiasing. *BMJ Quality & Safety*, *22*(Suppl 2), ii58-ii64. doi:10.1136/bmjqs-2012-001712

Dickson, R., Cherry, M. G., & Boland, A. (2014). Carrying out a systematic review as a master's thesis. *Doing a systematic review: A student's guide*, 1-16.

Evans, J. S. B. T., & Frankish, K. (2009). *In two minds dual processes and beyond*. Oxford: Oxford University Press.

Faherty, A., Counihan, T., Kropmans, T., & Finn, Y. (2020). Inter-rater reliability in clinical assessments: do examiner pairings influence candidate ratings? *BMC medical education*, *20*(1), 147-147. doi:10.1186/s12909-020-02009-4

Feinsilber, D., Siripala, D. S., & Mears, K. A. (2019). Review of Cognitive Biases in ACGME Milestones Training Assessments in Post-graduate Medical Education Programs. *Curēus (Palo Alto, CA)*, *11*(8), e5518-e5518. doi:10.7759/cureus.5518

Fontaine, S., & Loye, N. (2017). L'évaluation des apprentissages : une démarche rigoureuse. *Pédagogie médicale*, *18*(4), 189-198. doi:10.1051/pmed/2018013

Forte, M., Morson, N., Mirchandani, N., Grundland, B., Fernando, O., & Rubenstein, W. (2021). How Teachers Adapt Their Cognitive Strategies When Using Entrustment Scales. Acad Med, 96(11S), S87-S92. doi:10.1097/ACM.0000000000004287

Frank, J. R., Snell, L. S., Cate, O. T., Holmboe, E. S., Carraccio, C., Swing, S. R., . . . Harris, K. A. (2010). Competency-based medical education: theory to practice. *Medical teacher, 32*(8), 638-645. doi:10.3109/0142159X.2010.501190

Gardner, A. K., & Scott, D. J. (2016). Repaying in Kind: Examination of the Reciprocity Effect in Faculty and Resident Evaluations. J Surg Educ, 73(6), e91-e94. doi:10.1016/j.jsurg.2016.04.015

Gauthier, G., Christina, S.-O., & Dory, V. (2016). Synthèse et conceptualisation des processus cognitifs du jugement évaluatif de l'enseignant clinicien. *Pédagogie médicale, 17*(4), 261-267.

Gauthier, G., St-Onge, C., & Tavares, W. (2016). Rater cognition: review and integration of research findings. *Medical education, 50*(5), 511-522.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the 'black box' differently: assessor cognition from three research perspectives. *Medical education, 48*(11), 1055-1068. doi:10.1111/medu.12546.

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-based assessments as social judgments: rethinking the etiology of rater errors. Acad Med, 86(10 Suppl), S1-7. doi:10.1097/ACM.0b013e31822a6cf8.

Gingerich, A., Schokking, E., & Yeates, P. (2018). Comparatively salient: Examining the influence of preceding performances on assessors' focus and interpretations in written assessment comments. Advances in Health Sciences Education, 23, 937-959.

Gomez-Garibello, C., & Young, M. (2018). Emotions and assessment: considerations for rater-based judgements of entrustment. *Medical education*, *52*(3), 254-262. doi:10.1111/medu.13476

Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health information and libraries journal, 26*(2), 91-108. doi:10.1111/j.1471-1842.2009.00848.x

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological review, 71*(6), 438-456. doi:10.1037/h0040736

Harris, P., Bhanji, F., Topps, M., Ross, S., Lieberman, S., Frank, J. R., . . . Sherbino, J. (2017). Evolving concepts of assessment in a competency-based world. *Medical teacher, 39*(6), 603-608. doi:10.1080/0142159X.2017.1315071

Hodwitz, K., Kuper, A., & Brydges, R. (2019). Realizing one's own subjectivity: assessors' perceptions of the influence of training on their conduct of workplace-based assessments. Academic Medicine, 94(12), 1970-1979.

Irby, D. M., Cooke, M., & O'Brien, B. C. (2010). Calls for reform of medical education by the Carnegie Foundation for the Advancement of Teaching: 1910 and 2010. *Academic medicine, 85*(2), 220-227. doi:10.1097/ACM.0b013e3181c88449

Jenicek, M. (2010). *Medical error and harm: Understanding, prevention, and control*: CRC Press.

Jouquan, J., & Bail, P. (2003). A quoi s'engage-t-on en basculant du paradigme d'enseignement vers le paradigme d'apprentissage ? *Pédagogie médicale, 4*(3), 163-175. doi:10.1051/pmed:2003006

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed. ed.). New York: Farrar, Straus and Giroux.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment, 49*, 81.

Klein, R., Julian, K. A., Snyder, E. D., Koch, J., Ufere, N. N., Volerman, A., . . . From the Gender Equity in Medicine, w. (2019). Gender Bias in Resident Assessment in Graduate Medical Education: Review of the Literature. J Gen Intern Med, 34(5), 712-719. doi:10.1007/s11606-019-04884-0

Klein, R., Ufere, N. N., Rao, S. R., Koch, J., Volerman, A., Snyder, E. D., . . . Gender Equity in Medicine, w. (2020). Association of Gender With Learner Assessment in Graduate Medical Education. JAMA Netw Open, 3(7), e2010888. doi:10.1001/jamanetworkopen.2020.10888

Kocovski, S. (2009). *Ergonomie et management : optimisez vos produits et vos processus*. Liège: Edipro.

Kogan, J. R., Conforti, L. N., Iobst, W. F., & Holmboe, E. S. (2014). Reconceptualizing Variable Rater Assessments as Both an Educational and Clinical Care Problem. *Academic medicine, 89*(5), 721-727. doi:10.1097/ACM.0000000000000221

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: a conceptual model. Med Educ, 45(10), 1048-1060. doi:10.1111/j.1365-2923.2011.04025.x

Lacasse, M., Renaud, J.-S., Cantat, A., & Saucier, D. (2016). Développement de compétences avancées dans la formation des futurs médecins : l'exemple de la médecine familiale au Canada. *Éducation et francophonie*, *44*(2), 126-151. doi:10.7202/1039025ar

Landry, R., Becheikh, N., Amara, N., Ziam, S., Idrissi, O., & Castonguay, Y. (2008). La recherche, comment s'y retrouver? revue systématique des écrits sur le transfert de connaissances en éducation.

Laurin, S., Audetat Voirol, M.-C., & Sanche, G. (2013). L'approche par compétences lubie pédagogique ou réel progrès? *Le médecin du Québec*, *48*(3), 87-90.

Lee, V., Brain, K., & Martin, J. (2017). Factors Influencing Mini-CEX Rater Judgments and Their Practical Implications: A Systematic Literature Review. Acad Med, 92(6), 880-887. doi:10.1097/ACM.0000000000001537.

Long, H. A., French, D. P., & Brooks, J. M. (2020). Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis. *Research Methods in Medicine & Health Sciences*, *1*(1), 31-42.

Low, D., Pollack, S. W., Liao, Z. C., Maestas, R., Kirven, L. E., Eacker, A. M., & Morales, L. S. (2019). Racial/Ethnic Disparities in Clinical Grading in Medical School. Teach Learn Med, 31(5), 487-496. doi:10.1080/10401334.2019.1597724

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in cognitive sciences, 14*(10), 435-440. doi:10.1016/j.tics.2010.07.004

Mussweiler, T. (2007). Assimilation and contrast as comparison effects: A selective accessibility model.

Naik, V. N., Wong, A. K., & Hamstra, S. J. (2012). Review article: Leading the future: guiding two predominant paradigm shifts in medical education through scholarship. *Canadian Journal of Anesthesia/Journal canadien d'anesthésie*, *59*(2), 213-223. doi:10.1007/s12630-011-9640-1

Nguyen, Q., & Blais, J.-G. (2007). Approche par objectifs ou approche par compétences ? Repères conceptuels et implications pour les activités d'enseignement, d'apprentissage et d'évaluation au cours de la formation clinique. *Pédagogie médicale, 8*(4), 232-251. doi:10.1051/pmed:2007026

Paget, M., Wu, C., McIlwrick, J., Woloschuk, W., Wright, B., & McLaughlin, K. (2013). Rater variables associated with ITER ratings. Adv Health Sci Educ Theory Pract, 18(4), 551-557. doi:10.1007/s10459-012-9391-y.

Polanco-Santana, J. C., Storino, A., Souza-Mota, L., Gangadharan, S. P., & Kent, T. S. (2021). Ethnic/Racial Bias in Medical School Performance Evaluation of General Surgery Residency Applicants. J Surg Educ, 78(5), 1524-1534. doi:10.1016/j.jsurg.2021.02.005.

Prégent, R., Bernard, H., & Kozanitis, A. (2009). *Enseigner à l'université dans une approche-programme : guide à l'intention des nouveaux professeurs et chargés de cours*. Montréal: Presses internationales Polytechnique.

Raj, J. M., & Thorn, P. M. (2014). A Faculty Development Program to Reduce Rater Error on Milestone-Based Assessments. *Journal of graduate medical education, 6*(4), 680-685. doi:10.4300/JGME-D-14-00161.1

Rey, B. (2014). *La notion de compétence en éducation et formation : enjeux et problèmes*. Louvain-la-Neuve: De Boeck.

Rojek, A. E., Khanna, R., Yim, J. W. L., Gardner, R., Lisker, S., Hauer, K. E., . . . Sarkar, U. (2019). Differences in Narrative Language in Evaluations of Medical Students by Gender and Under-represented Minority Status. J Gen Intern Med, 34(5), 684-691. doi:10.1007/s11606-019-04889-9

Romainville, M. (2011). Objectivité versus subjectivité dans l'évaluation des acquis des étudiants. *Revue internationale de pédagogie de l'enseignement supérieur, 27*(2). doi:10.4000/ripes.499.

Ross, D. A., Boatright, D., Nunez-Smith, M., Jordan, A., Chekroud, A., & Moore, E. Z. (2017). Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations. PloS one, 12(8), e0181659. doi:10.1371/journal.pone.0181659.

Sherbino, J., Kulasegaram, K., Worster, A., & Norman, G. R. (2013). The reliability of encounter cards to assess the CanMEDS roles. Adv Health Sci Educ Theory Pract, 18(5), 987-996. doi:10.1007/s10459-012-9440-6

Shaw, T., Wood, T. J., Touchie, C., Pugh, D., & Humphrey-Murto, S. M. (2021). How biased are you? The effect of prior performance information on attending physician ratings and implications for learner handover. Adv Health Sci Educ Theory Pract, 26(1), 199-214. doi:10.1007/s10459-020-09979-6.

St-Onge, C., Chamberland, M., Lévesque, A., & Varpio, L. (2014). The role of the assessor: exploring the clinical supervisor's skill set. *The clinical teacher, 11*(3), 209-213. doi:10.1111/tct.12126

St-Onge, C., Chamberland, M., Lévesque, A., & Varpio, L. (2016). Expectations, observations, and the cognitive processes that bind them: expert assessment of examinee performance. *Advances in health sciences education : theory and practice, 21*(3), 627-642. doi:10.1007/s10459-015-9656-3

Tardif, J. (2006). *L'évaluation des compétences : documenter le parcours de développement.* Montréal: Chenelière-éducation.

Tavares, W., & Eva, K. W. (2014). Impact of rating demands on rater-based assessments of clinical competence. *Education for primary care, 25*(6), 308-318. doi:10.1080/14739879.2014.11730760

Ten Cate, Th. J. Olle, et al. "Medical Competence: The Interplay between Individual Ability and the Health Care Environment." *Medical Teacher*, vol. 32, no. 8, 2010, pp. 669–75, https://doi.org/10.3109/0142159X.2010.500897.

Weber, E. U., & Johnson, E. J. (2006). Constructing Preferences From Memory. In (pp. 397-410): Cambridge University Press.

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and learning in medicine, 15*(4), 270-292. doi:10.1207/S15328015TLM1504_11

Wilson, T. D., & Brekke, N. (1994). Mental Contamination and Mental Correction: Unwanted Influences on Judgments and Evaluations. *Psychological bulletin, 116*(1), 117-142. doi:10.1037/0033-2909.116.1.117

Wood, T. J. (2014). Exploring the role of first impressions in rater-based assessments. *Advances in health sciences education : theory and practice, 19*(3), 409-427. doi:10.1007/s10459-013-9453-9.

Wood, T. J., Chan, J., Humphrey-Murto, S., Pugh, D., & Touchie, C. (2017). The influence of first impressions on subsequent ratings within an OSCE station. Adv Health Sci Educ Theory Pract, 22(4), 969-983. doi:10.1007/s10459-016-9736-z

Yeates, P., O'Neill, P., Mann, K., & Eva, K. (2013). Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. Adv Health Sci Educ Theory Pract, 18(3), 325-341. doi:10.1007/s10459-012-9372-1.

Yeates, P., Moreau, M., & Eva, K. (2015). Are Examiners' Judgments in OSCE-Style Assessments Influenced by Contrast Effects? Acad Med, 90(7), 975-980. doi:10.1097/ACM.0000000000000650.

Yeates, P., O'Neill, P., Mann, K., & K, W. E. (2013). 'You're certainly relatively competent': assessor bias due to recent experiences. Med Educ, 47(9), 910-922. doi:10.1111/medu.12254.

Yeates, P., Cardell, J., Byrne, G., & Eva, K. W. (2015). Relatively speaking: contrast effects influence assessors' scores and narrative feedback. Med Educ, 49(9), 909-919. doi:10.1111/medu.12777.

Zaugg, V., Savoldelli, V., Sabatier, B., & Durieux, P. (2014). Améliorer les pratiques et l'organisation des soins : méthodologie des revues systématiques. *Santé publique (Vandoeuvre-lès-Nancy, France)*, *26*(5), 655-667. doi:10.3917/spub.145.0655