L'évaluation du raisonnement clinique

Bernard CHARLIN, Georges BORDAGE**, et Cees VAN DER VLEUTEN****

Messages Clés

• Le raisonnement clinique est multidimensionnel. Son évaluation implique l'utilisation d'instruments complémentaires. • La compétence à résoudre un problème clinique ne permet pas de prédire avec confiance la capacité à résoudre un autre problème. Il convient donc d'éviter toute longue évaluation sur un même cas et de privilégier plusieurs évaluations portant sur des cas courts. • Il est souhaitable d'ancrer les évaluations sur des présentations de cas qui induisent de réelles activités de raisonnement clinique. • L'important est la tâche cognitive qu'effectue le candidat et non le format de la question. • Il est nécessaire de distinguer l'évaluation du processus de raisonnement de celle du résultat de ce raisonnement. • Il n'y a pas d'avantage notable à utiliser des méthodes complexes d'établissement des scores. • Il est souhaitable d'éviter l'effet d'indice, qui mène les candidats à répondre dans une direction. • Il est possible d'améliorer notablement la fidélité des examens en organisant des formations pour les évaluateurs. • Tout processus d'évaluation modifie les comportements d'apprentissage des étudiants. Il est important d'en tenir compte lors du choix d'une méthode.

Key Messages

• Clinical reasoning is multidimensional. Its assessment implies the use of complementary instruments. • The competency to solve a given clinical problem does not permit to predict with confidence the capacity to solve another problem. This means that, to measure competency, it is preferable using many short clinical cases instead of using a long evaluation on a unique case. • Assessments should be based on clinical scenarios and cases that will induce authentic clinical reasoning processes. • Of importance is the cognitive task carried by the candidate and not the question format. • It is necessary to make a distinction between assessment of the reasoning process and of result of this process. • There is no clear advantage of using complex methods to establish scores. • It is desirable to avoid the cueing effect which leads candidates to respond in a given direction. • The reliability of the examinations may be substantially enhanced through training of evaluators. • Every assessment process modifies the learning behaviours of the students. This fact must be taken into account when a method is to be chosen.

Pédagogie Médicale 2003 ; 4 : 42-52

Introduction

Le raisonnement clinique est constitué des processus de pensée et de prise de décision qui permettent au clinicien d'effectuer l'action jugée la plus utile dans un contexte spécifique[1,2]. Il peut être considéré comme étant l'activité intellectuelle qui synthétise l'information obtenue à partir de la situation clinique, qui l'intègre aux connaissances et aux expériences antérieures et l'utilise pour prendre des décisions de diagnostic et de prise en charge du patient[3]. Cette activité est souvent appelée résolution de problème clinique. Elle est essentielle et centrale à toute pratique professionnelle dans les sciences de la santé. Il s'agit d'un phénomène de grande complexité et il n'existe pour le moment aucun modèle des processus de raisonnement clinique, qu'il soit théorique ou issu de travaux de recherche, qui soit unanimement accepté[2]. Il est d'ailleurs probable qu'il n'y en aura

Université de Montréal, Canada - ** Université de l'Illinois à Chicago, Etats-Unis - * Université de Maastricht, Pays Bas. Correspondance : Bernard Charlin, URDESS - Faculté de médecine -Direction - Université de Montréal - CP6128 Succursale centre ville - Montréal, Québec H3C 3J7 Canada - mailto:bernard.charlin@umontreal.ca*

Pédagogie Médicale
REVUE INTERNATIONALE FRANCOPHONE D'ÉDUCATION MÉDICALE

jamais un seul, étant donné la complexité du phénomène et la multiplicité des théories existantes.
Le raisonnement clinique constitue une des trois composantes de la compétence clinique, les autres étant représentées par les connaissances (sciences de base et sciences cliniques) et par les habiletés pertinentes (cliniques, techniques et interpersonnelles). Dans une perspective d'évaluation ces trois composantes sont souvent considérées séparément, de sorte que des tests spécifiques sont conçus pour évaluer les connaissances (questions à choix multiples par exemple), d'autres pour les habiletés (l'ECOS, examen clinique objectif et structuré) et d'autres pour le raisonnement clinique et la prise de décision. En fait, ces trois composantes sont très intriquées[4] et chacune d'entre elles est influencée par tout un ensemble d'attitudes qui sont difficiles à définir et à mesurer[3].
Depuis des décennies, chercheurs et éducateurs médicaux ont été très créatifs dans la recherche d'instruments d'évaluation qui permettraient de mesurer avec efficacité le raisonnement clinique. Dans ce texte nous décrirons tout d'abord un de ces instruments, le PMP (*Patient Management Problem*), qui a fait l'objet d'une très large utilisation en Amérique du Nord, notamment dans des contextes à hauts enjeux tels que la certification en fin de formation (l'attribution du diplôme). Cette description permettra d'illustrer les principales difficultés psychométriques ou pratiques auxquelles sont confrontés les instruments d'évaluation du raisonnement clinique. Nous décrirons ensuite plusieurs instruments en précisant les forces et les faiblesses de chacun. L'article se terminera par une série de recommandations qui visent à optimiser ce type d'évaluation.

Le PMP (Patient Management Problem)

Les années 60 et 70 ont été marquées[5] par la recherche de méthodes qui permettraient de mesurer une compétence générale à raisonner adéquatement devant un problème clinique. Une telle compétence aurait constitué une capacité stable et distincte, une stratégie qui une fois acquise pouvait être appliquée à tout problème clinique, quel que soit le domaine concerné. L'idée du PMP[6] consistait à simuler sur papier, puis plus tard sur ordinateur, le processus avec lequel un médecin obtient l'histoire, collecte les informations par l'examen clinique, puis prend les décisions liées au diagnostic, à l'investigation ou à la prise en charge thérapeutique.
Un PMP typique débute par la description d'une situation clinique. Il est alors demandé à l'étudiant de recueillir des données complémentaires. Certains artifices techniques sont utilisés pour masquer les données jusqu'à ce qu'elles soient sélectionnées à l'aide d'un stylo révélateur. Après avoir obtenu les données d'interrogatoire et d'examen clinique, dans la mesure du possible conformément à la façon dont il faudrait s'y prendre en situation clinique réelle, il est demandé à l'étudiant de sélectionner les procédures d'investigation, le diagnostic et les mesures thérapeutiques appropriées. Le cheminement suivi par l'étudiant est alors comparé à celui d'un expert ou d'un groupe de référence et des scores sont attribués en fonction du caractère complet de la collecte de données, de son efficacité et de sa pertinence.
Cette méthode d'examen a suscité initialement un grand intérêt et elle devint très largement utilisée dans des buts de certification en fin de formation ou d'attribution du droit de pratique : elle était perçue comme étant une mesure valide de la capacité à résoudre les problèmes cliniques qui permettait de donner objectivement des scores tout en reflétant assez fidèlement la réalité de la vie clinique. Des préoccupations apparurent cependant progressivement, liées soit à des limites psychométriques soit à des doutes sur la réalité d'une capacité générale de résolution de problèmes cliniques[7,8]. Ces préoccupations sont présentées dans la section suivante. Elles illustrent les problèmes auxquels fait face toute démarche d'évaluation du raisonnement clinique.

Les difficultés rencontrées par l'évaluation du raisonnement clinique

L'effet d'indice (cueing effect)

Cet effet[9,10] suppose que si l'on présente une sélection limitée de réponses possibles, le candidat peut reconnaître la bonne réponse plutôt que la générer, ce qui dénature la tâche de résolution de problème et améliore artificiellement la performance. Par exemple, une étude[7] portant sur l'utilisation de formats de questions différents pour des contenus similaires, a montré un taux de réponse inférieur de 14 % et un taux de difficulté supérieur de 18 %, lorsque la réponse devait être générée et non simplement reconnue par le candidat.

La diversité des cheminements du raisonnement

On pensait à l'époque que les experts résolvaient les problèmes cliniques en suivant des cheminements de pensée optimaux, parfois même des cheminements fixes

Références

(algorithmes). L'évaluation de la compétence consistait dès lors à déterminer dans quelle mesure le candidat suivait le cheminement optimal de pensée[5]. Le développement des grilles de correction consistait habituellement à demander à un groupe d'experts (panel de référence) de décider, par consensus, quels étaient les cheminements acceptables pour le problème concerné. L'expérience a cependant montré qu'il est difficile d'atteindre un tel consensus, ce qui a conduit à une remise en question de l'utilisation de cheminements fixes par les experts[11]. Ceci venait entériner les données de la recherche sur le raisonnement clinique montrant que les experts suivent des cheminements différents dans le processus de résolution de problème[12,13].

La pondération de réponses

Les panels de référence devaient également déterminer quelle pondération (positive ou négative) devait être donnée aux options proposées dans chaque section du problème simulé. Les réponses des candidats servaient alors à calculer des scores composites tels que l'efficience, le caractère complet de la démarche, ou la compétence générale à résoudre le problème. La mise au point du dispositif de pondération provoquait d'importants débats au sein de ces panels et suscitait de nombreux biais de jugement[11]. Différents systèmes de pondération ont été développés, sans résoudre les difficultés décrites car la corrélation entre les différentes méthodes s'avérait élevée[14]. Ceci confirmait les résultats de nombreux travaux de recherche, menés dans différents domaines, indiquant qu'à peu près tous les systèmes de pondération, quelle que soit leur complexité, ont des fortes corrélations avec la simple addition des scores pour chaque item[15]. La pondération n'améliore donc pas la fidélité et la validité des scores. Par conséquent, il ne faut pas perdre de temps à discuter de pondération différentielle.

L'exhaustivité de la démarche

Beaucoup de PMP comportent un grand nombre d'items concernant l'histoire de cas, les données de l'examen physique et les tests de laboratoire, le tout réduisant les scores à une mesure d'exhaustivité de la démarche de collecte de données, mesure qui s'est avérée non reliée à la compétence clinique[16]. La recherche a montré que les experts utilisent souvent des raccourcis dans leur démarche de résolution de problème, car ils utilisent l'information disponible de façon plus stratégique que les novices[17]. Elle a également montré que l'exhaustivité n'est pas un bon prédicteur de succès diagnostique[12,18]. Par ailleurs un candidat peut compenser une omission d'acquisition de données essentielles (cotées 2 points par exemple) en cochant plusieurs données positives mais de moindre valeur (cotées 1 point). En conséquence, les systèmes de score qui donnent un poids excessif à l'exhaustivité dans la collecte de données pénalisent l'expertise et sur-cotent les cliniciens moins habiles en résolution de problème. Ce phénomène est apparu clairement dans les études sur des PMP comparant la performance d'experts à celle de cliniciens moins expérimentés. Des médecins très qualifiés obtiennent ainsi parfois des scores inférieurs à ceux d'étudiants ou de médecins moins expérimentés[10].

La spécificité de contenu (ou de cas)

La recherche a montré avec régularité que la performance dans un problème clinique prédit très mal la performance dans un autre problème, la corrélation moyenne de performance entre cas variant entre 0.1 et 0.37. Ce phénomène, décrit à propos d'études sur les PMP, a remis en question le postulat initial qui faisait de la compétence au raisonnement clinique une compétence générale, indépendante des connaissances spécifiques du domaine clinique évalué. Elle a ensuite été décrite pour les autres instruments d'évaluation du raisonnement clinique et, de ce fait, a jeté un doute sur toute procédure d'évaluation qui implique un processus long et extensif portant sur un nombre limité de cas cliniques[3].

La fidélité

À côté des caractéristiques décrites relatives à l'évaluation de la compétence clinique, les instruments doivent posséder les qualités attendues de tout instrument d'évaluation, la fidélité et la validité[19]. La fidélité d'un instrument est sa capacité à donner avec constance un même résultat pour un candidat lorsqu'il est utilisé à différentes occasions, avec des évaluateurs différents ou avec des formats différents. La fidélité peut être exprimée numériquement de plusieurs façons. La plus utilisée consiste à calculer le coefficient de cohérence interne du test, exprimé par le coefficient alpha de Cronbach. Une autre consiste à calculer le coefficient de corrélation entre deux mesures séparées dans le temps (corrélation test-retest). Une autre concerne deux évaluateurs différents pour le même test (fidélité inter-juge). Une autre enfin consiste à obtenir et comparer des mesures parallèles avec des instruments similaires mais différents (division du test en moitiés équivalentes ou tests faits parallèlement).

Alors que dans certaines études de fidélité (par exemple sur un test de connaissance fait de questions à choix

multiples), l'analyse se fait item par item, en matière d'évaluation du raisonnement clinique, le phénomène de spécificité de contenu vient compliquer l'analyse. Il est en effet indispensable de faire porter l'évaluation sur un nombre de cas suffisamment représentatif du domaine. L'analyse ne se fait donc plus item par item ; elle se fait par cas et par items. Or la relation idéale entre nombre de cas et nombre d'items reste une question non résolue en matière d'évaluation du raisonnement clinique. Un autre facteur important à considérer est le temps nécessaire pendant lequel un test doit être administré pour que soit garantie une bonne fidélité des scores. Certaines méthodes d'évaluation peuvent ainsi exiger plusieurs heures de test pour ce faire, tandis que d'autres en demanderont beaucoup moins.

La validité

La validité est la qualité d'un instrument qui mesure réellement ce qu'il prétend mesurer. Il est important d'en distinguer deux dimensions. La validité apparente est un jugement porté par un expert sur la capacité de l'instrument à mesurer ce qui lui apparaît important. Les autres formes de validité (validité concomitante*, prédictive** ou de construit***) sont des mesures empiriques qui démontrent que l'instrument mesure effectivement ce qu'il prétend mesurer[16]. Il convient de souligner que ces caractéristiques ne reposent pas sur des jugements, mais sur des études, en général statistiques, et qu'un instrument qui, à première vue est jugé valide, peut s'avérer non valide lorsqu'il est étudié empiriquement.

La faisabilité

La faisabilité est également une qualité essentielle. En effet, le meilleur instrument possible est sans utilité s'il ne peut être utilisé dans un milieu d'enseignement en raison de coûts trop importants en termes de matériel ou de personnel ou encore parce qu'il n'est pas acceptable par les étudiants ou les enseignants.

L'effet sur les apprentissages

L'évaluation a un énorme impact sur les apprentissages. Les étudiants sont stratégiques. Ils s'adaptent aux systèmes d'évaluation mis en place dans les cursus, à tel point que l'on affirme souvent que l'évaluation détermine les apprentissages[19]. C'est dire que cet impact doit être soigneusement considéré avant de choisir une méthode d'évaluation plutôt qu'une autre.

Les autres moyens d'évaluation du raisonnement clinique

Les grilles d'évaluation globale

Utilisées dans la plupart des milieux de stage, ces grilles comportent une énumération de critères à évaluer. Elles sont remplies par un ou plusieurs observateurs après une période de contact de quelques semaines ou quelques mois avec l'étudiant ou l'interne. Elles sont destinées à refléter l'ensemble des observations effectuées au cours du stage. Ces grilles ont une validité apparente, puisque la multiplicité des critères décrits suggère qu'elles sont capables de mesurer à peu près tous les aspects de la compétence clinique[16]. Cependant, les critères évaluant le raisonnement clinique ne représentent généralement qu'une fraction de la grille.

Les études empiriques montrent que ces grilles ont une fidélité très faible[20]. Un ensemble de facteurs[16] explique cela. (1). De nombreuses expériences de psychologie ont montré qu'il est pratiquement impossible d'intégrer avec fidélité un ensemble d'observations accumulées sur une longue période de temps. Toute intégration d'évènements sur-pondère soit les observations récentes soit une observation marquante qui colore le tout. Un processus d'intégration d'informations multiples recueillies pendant des périodes dépassant quelques jours est inévitablement sujet à de multiples erreurs. (2). Lorsqu'une grille comporte de multiples critères d'observation, les études ont montré que la performance sur un des critères biaise généralement l'évaluation des autres critères (effet dit de « halo »). (3). Lorsque l'observation porte sur une longue période de contacts entre l'évaluateur et l'étudiant, elle

**Validité concomitante : caractère de validité d'un test appréciant, au moyen d'un coefficient de corrélation, jusqu'à quel point les scores au test peuvent être utilisés pour estimer les scores réels effectivement obtenus par les même sujets à une autre épreuve dont la validité a déjà été établie et reconnue.*

***Validité prédictive : caractère de validité d'un test appréciant, au moyen d'un coefficient de corrélation, jusqu'à quel point les scores au test peuvent être utilisés pour estimer les scores futurs des mêmes sujets en regard d'un performance à exercer dans une situation particulière.*

****Validité de construit : validité interne qui indique le degré d'adéquation entre la structure d'un instrument de mesure et le schéma théorique illustrant les caractéristiques comportementales inter reliées du trait mesuré.*

Définitions du dictionnaire actuel de l'éducation de Renald Legendre, Éditions Larousse, Paris - Montréal, 1988

incorpore très souvent, à côté du jugement sur la performance, un jugement sur le caractère de l'étudiant. (4). sachant que l'évaluation peut être remise en cause par l'étudiant et qu'elle comporte inévitablement une subjectivité, les évaluateurs utilisent souvent une approche défensive et cotent les performances à la moyenne ou au-dessus parce qu'ils estiment qu'ils auraient du mal à défendre les autres options.
Malgré ces limites, l'instrument reste extrêmement utilisé. Il peut être amélioré en formant les évaluateurs à l'utilisation rationnelle de la grille et en réalisant une série d'observations sur des compétences bien définies et sur de courtes périodes et si possible, par des observateurs différents plutôt qu'en se basant sur le jugement global d'un seul observateur au terme d'une longue période[16].

Les questions à choix multiple (QCM)

Bien qu'elles ne mesurent que très imparfaitement la qualité du raisonnement clinique, les QCM servent à l'évaluation de certification dans de très nombreux programmes, surtout lorsque ces derniers comportent un grand nombre de candidats. Leurs principaux avantages sont leur haut degré de fidélité qui tient au fait qu'elles permettent de mesurer de larges échantillonnages de connaissances, qu'elles peuvent être administrées facilement à de grandes populations d'étudiants et que leur système d'attribution des scores est objectif. Certaines critiques sont techniques. Par exemple, l'effort requis pour bâtir un grand nombre de bonnes questions est considérable et l'élaboration de questions claires demande de l'habileté pour éviter des erreurs fréquentes telles que l'ambiguïté ou la fourniture d'indices vers la bonne réponse. D'autres critiques portent sur l'effet induit sur les apprentissages, avec une incitation à l'apprentissage superficiel et par cœur et une sur-valorisation des connaissances factuelles au détriment de la capacité de résolution de problèmes.
Les QCM représentent cependant un excellent exemple d'instruments caractérisés par une différence entre validité apparente (souvent perçue comme mauvaise si l'on prend en compte les critiques énumérées ci-dessus) et validité étudiée par des mesures empiriques. En effet, contrairement aux intuitions, il a été démontré qu'elles possèdent une excellente validité prédictive en terme de performance dans la pratique future[16]. Tout se passe comme si l'existence d'une base riche de connaissances factuelles était statistiquement associée à une bonne performance clinique actuelle et même future. Il est, par ailleurs, possible de mieux tester le raisonnement clinique en évitant les questions qui sondent des connaissances triviales et en construisant les questions autour de présentations de patients (QCM à contexte riche)[21].

L'oral

Depuis fort longtemps, l'oral représente une des composantes majeures des examens de certification. Il porte souvent sur un patient que le candidat est allé interroger et examiner avant la session orale. Il est souvent affirmé qu'il peut avantageusement mesurer à la fois l'étendue de la base de connaissances, les capacités de résolution de problème (le jugement clinique) ainsi que des attributs personnels tels que la tolérance au stress, la confiance personnelle, les valeurs et les attitudes. Il est, de plus, flexible et adaptable aux situations d'évaluation. Un ensemble de limites psychométriques ainsi que qu'un haut coût d'utilisation (en terme de temps professoral) remettent cependant en question ces avantages apparents[22,23]. Certaines caractéristiques liées aux examinateurs et aux candidats menacent particulièrement sa fidélité. Les examinateurs présentent souvent des biais en faisant preuve soit d'un excès de clémence, soit d'un excès de sévérité, soit encore d'une tendance à donner à tous un score moyen[19]. Certains facteurs personnels des candidats, tels l'anxiété, l'aisance verbale, ou l'assurance personnelle influencent les scores qu'obtiennent les candidats. Les nombreuses études qui ont porté sur la valeur des examens oraux[23] montrent avec constance que la fidélité inter-juges, (la concordance des juges à propos de la performance) est élevée, tandis que la fidélité entre les cas est typiquement faible (spécificité de contenu).
Malgré certains désavantages, l'oral reste un examen très utilisé, car il est commode à organiser, parce qu'il présente une bonne validité apparente (l'examen est crédible) et qu'il permet de se former une opinion sur des dimensions cruciales telles que la qualité du raisonnement ou le bon jugement. Il est possible d'améliorer considérablement sa valeur en évitant la pratique commune qui implique deux examinateurs dans l'évaluation d'un petit nombre de cas sur une longue période de temps (une heure ou plus). Il est beaucoup plus efficace d'évaluer les candidats sur une série de cas courts, qui durent 10 à 15 minutes chacun, avec un seul examinateur par cas. L'oral démontre alors une bonne validité prédictive des capacités cliniques futures[24,25].

La question rédactionnelle

Comme l'oral, la question rédactionnelle fait partie des traditions dans de nombreux examens de compétence, bien qu'il ait été montré avec régularité qu'elle présente d'importantes limites[26]. La fidélité inter-juges est systématiquement basse[27], même lorsque des grilles de cotation sont utilisées. Les examinateurs ne parviennent généralement pas à séparer l'appréciation de la grammaire, du style et des qualités d'écriture, de celles de la

compétence au raisonnement clinique, qui devrait être le seul objet de l'évaluation. De plus, on retrouve ici aussi le problème de spécificité de contenu lorsqu'on utilise de longues réponses pour étudier la compétence sur un nombre limité de cas cliniques[16].
La question rédactionnelle peut être améliorée en utilisant un grand nombres de questions à réponses ouvertes et courtes, ce qui permet de faire de multiples sondages dans la base de connaissances. Ce type de question permet alors d'évaluer le raisonnement qui sous-tend un processus de prise de décision, par exemple en demandant d'exposer les raisons qui conduisent à cette prise de décision. Les examens à réponses courtes sont relativement faciles à construire et ils permettent d'éviter l'effet d'indice (indices qui guident vers la bonne réponse). Il est cependant difficile d'éviter toute ambiguïté dans l'intitulé des questions et d'établir des critères de correction clairs qui font consensus entre les examinateurs. Elles sont, enfin, exigeantes en temps de réflexion pour les étudiants et en temps de correction pour les évaluateurs.

L'examen clinique objectif et structuré (ECOS)

Ce type d'examen évalue la démarche clinique par observation directe à partir de situations cliniques simulées (vrais patients, acteurs) et standardisées (tous les candidats sont soumis aux mêmes tâches cliniques). L'examen comporte des stations multiples qui évaluent chacune des comportements distincts[28]. Dans chacune des stations, à durée prédéterminée (10 à 20 minutes généralement), l'évaluateur utilise une grille d'observation prédéfinie (*check-list*). La structure de l'examen permet d'obtenir une bonne fidélité inter-juges. L'examen, « *performance-based* », est extrêmement utilisé dans les pays anglo-saxons, dans des buts de certification de compétence notamment pour mesurer les habiletés cliniques. Il ne sera que brièvement traité ici, car il pose des problèmes de validité comme outil de mesure du raisonnement clinique. En effet, par sa conception, il n'évalue que des comportements observables. Or, il est difficile de mesurer avec des grilles les subtilités d'une démarche de raisonnement exprimé à voix haute, de sorte que les grilles conduisent plus à récompenser la minutie dans la collecte de données que véritablement les qualités du raisonnement clinique.
L'ECOS n'est donc pas, malgré des avantages indéniables pour mesurer certaines composantes de la compétence clinique, la panacée attendue en matière d'évaluation du raisonnement clinique. Il est certes possible d'utiliser les stations individuelles pour réaliser des oraux courts et structurés, ou tout autre test qui permet d'évaluer des capacités cognitives supérieures, mais dès lors il ne s'agit plus d'ECOS à proprement parler et il convient de rappeler que l'ECOS, pour démontrer de bonnes qualités psychométriques, doit comporter une vingtaine de stations ou plus[16]. Cela en fait un examen coûteux en ressources (matérielles et en personnel), de sorte qu'il convient de s'assurer qu'il n'est pas possible de mesurer ces mêmes capacités par des instruments moins exigeants.

Le MEQ (Modified Essay Question)

Il s'agit d'une approche alternative au PMP[29]. Par rapport à ce dernier elle introduit dans la simulation des données amenées séquentiellement et du *feed-back*. Le MEQ a été très utilisé, sans doute en partie parce qu'il est plus facile à construire qu'un PMP. Un MEQ commence par une vignette portant sur un cas clinique. Les réponses des étudiants sont ouvertes et courtes plutôt que choisies dans une liste fixe d'options, ce qui permet d'éviter l'effet d'indice. L'information nouvelle est fournie séquentiellement en fonction des variations d'évolution qui peuvent survenir dans le cas clinique, tout en prenant garde à ne pas donner des indices sur les sections préalables ou à venir dans l'examen. Bien qu'il existe peu de travaux rapportant sa fidélité et sa validité[30], la méthode a une bonne validité apparente et apparaît réaliste.

Les questions à appariement étendu (EMQ)

Les questions à appariement étendu (EMQ, *Extended Matching Questions*) représentent une variante du principe des QCM qui est utilisée par plusieurs organismes de certification nord-américains. Elles constituent un test de reconnaissance de modèles-types (*pattern recognition*). Chaque série de questions[31] est basée sur un motif principal de consultation (difficulté respiratoire par exemple), suivi d'une longue liste de diagnostics possibles (anémie, sténose aortique, pneumonie d'aspiration, insuffisance respiratoire chronique, etc). Chaque question représente un ensemble de signes associés au motif de consultation (par exemple une femme de 55 ans, fumeuse, présente une toux productive et une difficulté de respiration progressive depuis 5 ans). Les étudiants doivent choisir au sein de la liste de diagnostics possibles ceux qui sont pertinents compte tenu du regroupement de signes présentés par le patient. Ce type de question continue sans doute à explorer les connaissances factuelles, mais il peut être adapté à la prise de décision clinique, à l'interprétation de données et à certaines activités de résolution de problèmes[21]. Les questions sont plus faciles à rédiger que les QCM et elles

semblent mieux refléter les activités cliniques que ces dernières, tout en diminuant les chances de réponse correcte par simple reconnaissance de la bonne réponse. Tout comme les QCM, elles peuvent être corrigées mécaniquement ou être présentées sur ordinateur. Elles requièrent toutefois, elles aussi, un grand travail de préparation et, de ce fait, conviennent mal aux examens comportant un nombre limité de candidats.

L'examen par éléments clés

Chaque section de cet examen[7] comporte un scénario de cas clinique, suivi par des questions conçues pour évaluer les éléments clés dans la prise en charge de ce cas. Les questions sont de format varié, avec des QCM, des questions ouvertes et courtes, ou encore des choix de réponse dans de longues listes d'options. La méthode est particulièrement utile pour mesurer la capacité de prise de décision. Un de ses avantages est l'accent mis sur les éléments clés de résolution du problème concerné, de sorte que le nombre de questions dans chaque cas est limité, ce qui permet de multiplier le nombre de problèmes évalués et donc ainsi de répondre aux contraintes de spécificité de contenu[32]. Un autre avantage est la similitude avec les tâches cliniques réelles[18]. L'instrument présente cependant certains inconvénients. La préparation du matériel d'examen demande un temps considérable (bien moindre que celui exigé par la préparation d'un PMP cependant) et un grand nombre de cas (20 à 40 cas) est nécessaire pour obtenir une bonne fidélité[21]. Enfin, la nécessité d'obtenir un consensus entre examinateurs sur la « bonne réponse » à obtenir des candidats conduit à privilégier les situations d'évaluation où le consensus est facile à obtenir et à délaisser les autres.

Le test de concordance de script (TCS)

Ce test vise à comparer l'organisation des connaissances (les scripts) des candidats à celle d'experts du domaine[33]. Chaque section débute par un scénario de cas clinique pour lequel plusieurs hypothèses sont pertinentes. Le format de question consiste à présenter une de ces hypothèses et à demander quel effet (négatif, neutre, ou positif) aurait sur le statut de cette hypothèse la découverte d'une donnée clinique complémentaire, qui n'était pas présente dans le scénario. Les questions ultérieures concernent d'autres hypothèses et d'autres données. Le crédit donné aux candidats pour chaque réponse est fonction du nombre d'experts qui ont fourni la même réponse qu'eux. Les données publiées démontrent une relative facilité de construction et d'administration du test, une bonne fidélité, une bonne validité prédictive et la capacité de détecter les personnes les plus expérimentées cliniquement, alors que les tests habituels, basés sur les consensus entre correcteurs, permettent mal cette détection[34]. L'instrument est relativement nouveau et ses qualités psychométriques demandent à être confirmées par des études portant sur de larges populations de personnes examinées. Sa structure de correction, qui permet d'incorporer la variabilité des réponses d'experts du domaine, en fait cependant un examen intéressant pour évaluer ce que l'on désigne sous le terme de problème mal défini, c'est-à-dire un problème dont les données, les buts et les solutions ne sont pas univoques. Or en médecine, comme dans les autres domaines professionnels, l'expertise repose sur la capacité à résoudre les problèmes mal définis[35]. L'examen représente par ailleurs un changement de perspective théorique. Jusqu'à maintenant la démarche, en matière d'évaluation du raisonnement clinique, a consisté à mimer la réalité le plus possible en transposant le cas sur papier ou sur ordinateur (PMP, MEQ…). Les données empiriques dues au problème de la spécificité de contenu ont ensuite amené des adaptations (examen par éléments clés, questions à appariement étendu,…), mais l'essence de ces méthodes reste une simulation de la réalité. À l'opposé, le TCS part d'une théorie du raisonnement clinique (la théorie des scripts) et vise à mesurer des processus de raisonnement jugés essentiels plutôt que l'issue d'un raisonnement devant une situation qui mime la réalité.

L'évaluation basée sur la performance : Le miniCEX

Le miniCEX (*Clinical Examination Exercice*) permet d'observer directement la compétence clinique d'un résident dans un contexte qui reflète la pratique quotidienne[36]. L'exercice dure une vingtaine de minutes pendant lesquelles le résident (l'interne) prend une histoire de cas et réalise un examen physique dans une salle d'urgence, dans une clinique externe ou sur l'étage d'hospitalisation. À la fin de l'exercice, l'observateur donne du *feed-back* au résident et complète une grille d'observation. L'outil présente certaines limites, essentiellement en termes de fidélité inter-juges et de spécificité de contenu (un seul observateur évalue la performance sur un seul cas). De plus il permet surtout d'évaluer les comportements observables et beaucoup moins les processus de raisonnement, même si on demande au résident de raisonner à haute voix. Ceci s'explique sans doute par les exigences de la situation clinique qui mobilisent les ressources cognitives du résident qui, dès lors, a du mal a faire en plus l'effort de raisonner à voix haute. Il s'agit cependant d'un outil intéressant qui donne satisfaction

aux résidents (ils sont observés de façon approfondie dans un but de *feed-back*) et qui s'inscrit dans la tendance actuelle vers une évaluation centrée sur l'observation de la performance[37].

Cette version abrégée (d'où le terme mini-CEX) tend aujourd'hui à remplacer l'ancienne version plus longue (CEX), ce qui permet de multiplier le nombre de mises en situations pour pallier les limites déjà évoquées liées à la spécificité de contenu. Ces limites sont d'ailleurs partagées par tous les tests comparables de longue durée, tels que l'OSLER (*Objective Long Structured Examination Record*) même si ces derniers conservent un intérêt théorique, notamment en évaluation formative. Dans un cas comme dans l'autre (test court ou test long), l'observation directe de l'étudiant dans son interaction avec le patient est essentielle et semble apporter une valeur ajoutée à l'évaluation du raisonnement clinique[38].

L'évaluation du raisonnement clinique : principes et recommandations

Nous nous sommes limités dans cet article à la présentation des méthodes les plus utilisées, ou les plus intéressantes conceptuellement, produites grâce aux efforts menés depuis plusieurs dizaines d'années pour développer des instruments valides et fiables. Ces recherches ont par ailleurs permis d'établir quelques principes difficiles à contourner en matière d'évaluation du raisonnement clinique.

Le raisonnement clinique est multidimensionnel. Il comporte notamment la capacité d'intégrer les données (la capacité à les obtenir fait partie de l'évaluation des habiletés cliniques, bien que les stratégies de collectes appartiennent au raisonnement), de générer les hypothèses pertinentes à la situation clinique, de décider du poids à attribuer à chaque réponse en fonction de chaque hypothèse, de décider du bon diagnostic, de prendre les décisions appropriées en matière d'investigation ou de diagnostic, etc. Aucun instrument ne permet de mesurer toutes ces dimensions. L'évaluation de chacune des dimensions du raisonnement clinique suppose donc l'utilisation d'instruments complémentaires qui mesurent chacun une ou plusieurs de ces dimensions.

Il est essentiel de tenir compte du principe de spécificité de contenu. Rappelons que ce principe implique que la mesure de la compétence à résoudre un problème ne permet pas de prédire avec confiance la performance à résoudre un autre problème. Ce principe conduit à éviter toute longue évaluation d'un même cas et à réaliser des évaluations courtes sur un éventail de cas et ce, quel que soit le format utilisé dans l'examen (oral, questions à développements, examens par éléments clés, etc).

L'évaluation doit être ancrée dans des présentations de cas qui permettent d'induire de réelles activités de raisonnement clinique. L'évaluation de la simple mémorisation factuelle n'est plus acceptable. Par ailleurs faire cheminer, dans des cas réels ou simulés, un candidat dans de laborieuses étapes de collecte de données et d'investigations multiples, est une approche inefficace si l'intention est de réellement mesurer les habiletés de raisonnement clinique. Il est en effet nécessaire, en raison de la spécificité de contenu, de mettre l'accent sur les phases de raisonnement réellement cruciales, de façon à gagner du temps d'examen et de pouvoir multiplier les cas évalués[7,8,32].

Les données de la recherche montrent qu'il est essentiel de se préoccuper de la tâche cognitive que doit effectuer l'étudiant, beaucoup plus que du format [(39)] qui recueille la réponse de l'étudiant (QCM, réponse ouverte et courte, choix dans une longue liste d'options, ou TCS). Cette tâche doit être suffisamment complexe pour requérir un réel processus de résolution de problème. Il devrait être impossible de pouvoir répondre par un simple rappel de connaissances.

En matière de résolution de problème, la psychologie cognitive distingue le processus du résultat. Faut-il mesurer le processus ou le résultat ? Peut-on ignorer totalement le processus et mettre l'accent sur la qualité de la solution ? Un haut degré de fidélité est nécessaire dans certains examens en raison des conséquences majeures des décisions prises. Or en règle générale les tests qui évaluent les solutions ont des indices de fidélité supérieurs à ceux qui évaluent les processus de raisonnement, en raison notamment des variations observées chez les experts dans ces processus. Il est donc envisageable d'utiliser ce type de test pour les examens de fin de formation, d'autant plus que ce qui est attendu en premier lieu d'un médecin qui va exercer, c'est une aptitude à donner le bon diagnostic et la bonne conduite à tenir[7]. L'utilisation systématique et exclusive de ce type de test est sans doute moins défendable en cours de formation où les étapes intermédiaires de raisonnement doivent faire l'objet de formations et d'évaluation spécifiques[3]. Le choix des instruments de mesure dépend donc des buts de l'évaluation et du niveau de formation des personnes testées.

Pour terminer ces recommandations, rappelons qu'il convient : (1) de rester simple dans les méthodes d'établissement des scores (les méthodes de pondération complexes n'apportent pas grand chose) ; (2) de prendre garde à minimiser l'effet d'indice qui guide vers la bonne réponse, présent surtout avec les QCM ; (3) d'améliorer la fidélité des examens en organisant des formations

pour les examinateurs qui permettent d'harmoniser à la fois leur utilisation des instruments d'évaluation et leurs attentes vis-à-vis du niveau de performance attendu des candidats.

Conclusion

Même si nous ne disposons pas d'instrument parfait pour évaluer le raisonnement clinique, il est crucial d'évaluer celui-ci en raison de la relation puissante qui existe entre évaluation et comportement des étudiants. Les étudiants veulent réussir et toute introduction d'un instrument d'évaluation dans un système d'apprentissage modifie le comportement d'apprentissage des étudiants[40]. Il convient donc de choisir les instruments d'évaluation en fonction des compétences attendues des étudiants au terme de leur apprentissage et non en fonction de leur facilité de conception ou d'utilisation. Ces instruments doivent donner aux étudiants des informations valides sur le niveau de performance qu'ils ont atteint afin de leur permettre de prendre, le cas échéant, les mesures de correction nécessaires[41]. Il est enfin nécessaire de créer une cohérence entre les méthodes d'enseignement et les méthodes d'évaluation. Ceci s'applique particulièrement pour le raisonnement clinique, compétence essentielle de la profession médicale.

Références

1. *Harris I. New Expectations for Professional Competence. In L Curry, JF Wegin (Eds), Educating Professionals. Responding to new expectations for competence and accountability. San Francisco: Jossey-Bass Publishers, 1993: 17-52.*

2. *Higgs J, Jones M. Clinical reasoning in the health professions. In J Higgs & M Jones, Clinical Reasoning in the Health Professions. Butterworth Heinemann: Oxford, 2000 (2nd edition): 3-14.*

3. *Newble D, Norman G, Van der Vleuten C. Assessing Clinical Reasoning. In: Higgs J, Jones M (Eds.), Clinical Reasoning in the Health Professions, Butterworth-Heinemann Ltd: Oxford, 2000 (2nd edition): 156-165.*

4. *Norman GR, Tugwell P, Feightner JW, Muzzin LJ, Jacoby LL. Knowledge and Clinical Problem Solving. Med Educ, 1985, 19: 344-536.*

5. *Schuwirth L. An approach to the assessment of medical problem solving: Computerised Case-based Testing. Ph D thesis. Maastricht, The Netherland, 1998.*

6. *McGuire CH, Babbott D. Simulation techniques in the measurement of problem solving skills. Educ Meas, 1967, 4: 1-10.*

7. *Page G and Bordage G. The Medical Council of Canada's Key Features Project: a more valid written examination of clinical decisions skills. Acad Med, 1995, 70: 104-110.*

8. *Norman G. Striking the balance. Acad Med, 1994, 69: 209-210*

9. *Norman GR, Feightner JW. A comparison of behaviour on simulated patients and patient management problems. Med Educ, 1981, 55:529-537.*

10. Newble DI, Hoare J, Baxter A. Patient Management Problems: Issues of Validity. Med Educ, 1982, 16:137-142.

11. Swanson DB, Norcini JJ, Grosso LJ. Assessment of Clinical Competence: Written and Computer-Based Simulations. Assessment and Evaluation in higher Education, 1987, 12: 220-246.

12. Elstein AS, Shulman LS, Sprafka SA. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press, 1978.

13. Gale J, Marsden P. Medical Diagnosis: From Student to Clinician. Oxford: Oxford University Press, 1983.

14. Norcini JJ, Swanson DB, Webster GD and Grosso LJ. A comparison of several methods of scoring patient management problems. In Proceedings of the 22nd Annual Conference off Research in Medical Education. Washington, DC: Association of American Medical Colleges, 1987, pp. 41-46.

15. Wainer H. Estimating coefficients in linear models: It doesn't make no nevermind. Psychological Bulletin,1976, 83: 213-217.

16. Norman GR. Theoretical and psychometric considerations. In: Report on the evaluation system for specialist certification (pp 73-80). Task force of the evaluation committee. The Royal College of Physicians and Surgeons of Canada. Ottawa, 1993.

17. Marshall J. Assessment of Problem-Solving Ability. Med Educ, 1977, 11:329-334.

18. Hatala R, Norman GR. Adapting the Key Features Examination for a clinical clerkship. Med Educ 2002, 36: 160-165.

19. Jean P, Des Marchais JE, Delorme P- Apprendre à enseigner les sciences de la santé. Guide de formation pratique. Faculté de médecine des universités de Montréal et de Sherbrooke, 1993, 4e édition.

20. Streiner DL. Global rating scales. In Neufeld VR and Norman GR, Assessing Clinical Competence. Springer: New York, 1985.

21. Jolly B and Grant J. The Good Assessment Guide. A practical Guide to Assessment and Appraisal for Higher Specialist Training. Joint Center for Education in Medicine. London: UK, 1997.

22. Levine, H.G., McGuire, C.H. (1970). The validity and reliability of oral examinations in assessing cognitive skills in medicine. J Educ Meas, 7:63-73.

23. Muzzin, LJ. Oral examinations. In Neufeld,V.R. and Norman, G.R. Assessing clinical competence. New York: Springer, 1985.

24. Solomon, D.J., Rienhart, M.A., Birdeham, R.G., Munger, B.S., Stranaman, S. An assessment of an oral examination format for evaluating clinical competence in emergency medicine. Acad Medi, 1990, (Supp) 65: S43-S44.

25. Swanson, D.B. A measurement framework for performance-based tests. In: Hart, I., Harden, R. (Eds.) Further developments in Assessing Clinical Competence. Montreal: Can-Heal publications, 1987, pp. 13 - 45.

26. Neufeld VR. Written examinations. In Neufeld,V.R. and Norman, G.R. Assessing clinical competence. New York: Springer, 1985.

27. Norcini JJ, Diserens D, Day SC et al. The scoring and reproductibility of an essay test of clinical judgement. Acad Med 1990 (Supp), 65: S41-S42.

28. Harden RM, Gleeson FA. Assessment of medical competence using an Objective Structured Clinical Examination. Med Educ, 1979 ; 13 : 39-54.

29. Hodgkin K. and Knox JDE. Problem Centred Learning: The Modified Essay Question in Medical Education Edinbburg: Churchill Livingstone, 1975.

Références

30. *Feletti GI. Reliability and validity on modified essay question. Med Educ, 1980, 55: 933-941.*

31. *Case SM, Swanson DB, and Stillman PS. Evaluating diagnostic pattern recognition: The psychometric characteristics of a new item format. In Proceedings of the 27th Conference on Research on Medical Education. Washington DC: Association of Medical Colleges, 1988, pp. 3-8.*

32. *Bordage G, Brailovsky C, Carretier H, Page G. Content Validation of Key Features on a National Examination of Clinical Decisions-making Skills. Acad Med, 70: 276-281*

33. *Charlin B, Gagnon R, Sibert L, Van der Vleuten C. Le test de concordance de script : un instrument d'évaluation du raisonnement clinique. Pédagogie Médicale, 2002, 3 : 135-144.*

34. *Charlin B., Desaulniers M, Gagnon R, Blouin D, van der Vleuten C. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. Teach Learn Med, in press, July 2002.*

35. *Schön, DA. The reflective Practicioner: How Professionals Think in Action. New York: Basic Books, 1983.*

36. *Holmboe ES, Hawkins RE. Methods for evaluating clinical competence of residents in internal medicine : a review. Ann Intern Med 1998 ; 129 : 42-48.*

37. *Dauphinee WD. Assessing clinical performance: where do we stand and what might we expect? JAMA 1995; 274: 741-743*

38. *Wass V, Jolly B. Does observation add to the validity of the long case ? Med Educ, 2001 ; 35 :729-734.*

39. *Norman, G.R., Smith, E.K.M., Powles, AA.C., Rooney, P.J., Henry, N.L. and Dodd, P.E. (1987). Factors underlying performance on written tests of knowwledge. Med Educ, 21: 297-304*

40. *Newble DI and Entwistle. Learning styles and approaches: Implications for medical education. Med Educ, 1986, 20 : 162-175.*

41. *Jouquan J. L'évaluation des apprentissages des étudiants en formation médicale initiale- Repères conceptuels et pratiques, perspectives. Pédagogie Médicale, 2002, 3 : 38-52.*