Pédagogie Médicale **19**, 37-50 (2018)

https://doi.org/10.1051/pmed/2019001

Pédagogie Médicale

Disponible en ligne :
www.pedagogie-medicale.org

CONCEPTS ET INNOVATIONS

Sélection des futurs médecins : sur quelles bases empiriques ?

Selection of future medical practitioners: on which empirical basis?

Pascal DETROZ[1,*] et Nathalie LOYE[2]

[1] Université de Liège, Liège, Belgique
[2] Université de Montréal, Montréal, Québec, Canada

Manuscrit soumis pour publication le 20 décembre 2016 ; commentaires éditoriaux formulés aux auteurs le 26 octobre et le 18 décembre 2018 ; accepté pour publication le 14 janvier 2019

Résumé – Contexte et problématique : Le processus de sélection des étudiants à l'entrée des études de médecine revêt un très fort enjeu sociétal, en ce sens qu'il conditionne le profil des étudiants entamant des études de médecine et donc, en partie, le profil des futurs professionnels. Il n'y a en fait pas de consensus concernant les meilleures méthodes de sélection. Les examens et concours à l'entrée des études de médecine reposent sur une diversité de manières de faire à travers le monde, dont, pour la plupart, la qualité n'a pu être empiriquement prouvée. **Exégèse :** Le présent article propose un recensement des différents processus et outils de contingentement des étudiants dans le domaine de la santé, pour en proposer une lecture actualisée et critique. Nous puisons également dans la littérature définissant le concept de validité pour nous questionner sur la qualité de ces outils, mais aussi sur la qualité méthodologique des études qui y réfèrent. **Conclusion :** Les données probantes justifiant la qualité de l'un ou l'autre dispositif de sélection des étudiants à l'entrée des études de médecine sont rares, soit parce que les dispositifs sont effectivement de faible qualité, soit parce que les études qui en justifient la qualité ne tiennent pas compte d'une vision moderne de la validité.

Mots clés : sélection, études médicales, validité, psychométrie

Abstract. Context and background: The selection process of students at the outset of medical studies represents a strong societal issue because it determines the profile of students who study medicine and, consequently, it influences the profile of future medical practitioners. Up to now, there is no consensus regarding best selection methods. Indeed, competitive examinations at the beginning of medical studies are organized worldwide in many ways. **Analysis:** The present contribution takes an inventory of the various procedures and tools implemented for the purpose of selecting future students in the domain of health and reviews them critically. We shall also use the literature dealing with the concept of validity in order to question the quality of existing procedures and tools as well as the quality of related studies. **Conclusion:** There is little evidence to support the quality of any of the selection instruments, either because they are actually of low quality, or because the studies that justify their quality do not take into account a modern vision of validity.

Keywords: selection, medical education, validity, psychometry

Introduction

Le processus de sélection des étudiants à l'entrée des études de médecine revêt un très fort enjeu sociétal. Comme le souligne Morrison [1], « *Admission to medical school is effectively admission to the medical profession because dropout rates during basic medical education are extremely low.* (p. 3) » (L'admission aux études de médecine est, en réalité, l'admission à la profession médicale parce que le taux d'abandon durant les études médicales de base est extrêmement faible. – traduction libre –). Lorsqu'il écrit ces propos, Morrison s'appuie sur des données provenant du Royaume-Uni (le rapport du *Higher Education Funding Council for England*) [2] et des États-Unis (*Association of American Medical Colleges*, 2014) [3]. Sans présager du taux d'abandon ailleurs dans le monde, il est cependant évident que les candidats qui ne seront pas lauréats d'un processus de sélection donnant accès aux études médicales ne seront pas médecins.

*Correspondance et offprints : Pascal DETROZ, Traverse des architectes 4, B63B, 4000 Liège, Belgique.
Mailto : p.detroz@uliege.be

Il est dès lors important de se poser la question de la qualité des processus et des instruments qui permettent de filtrer les profils d'étudiants qui seront autorisés à entamer ou poursuivre des études de médecine. Une lecture approfondie des principaux articles sur le sujet révèle que l'ensemble des procédures de sélection présente des défauts. Ces imperfections ont bien sûr des conséquences. Elles signifient qu'il existe probablement des postulants « meilleurs » qu'une partie de ceux qui ont été choisis et, par voie de conséquence, que le groupe sélectionné comprend vraisemblablement des individus qui n'auraient pas dû l'être. Des enjeux d'égalité et d'équité, de coûts liés à l'échec scolaire, mais aussi, ultimement, de santé publique sont donc au cœur des processus de sélection. Il est alors essentiel de pouvoir argumenter la pertinence de ces derniers, ce qui relève de la responsabilité sociale des facultés de médecine [4].

Problématique

De nombreux articles ont été produits à propos des outils de sélection donnant accès aux études de médecine. Ainsi, récemment, Patterson *et al.* [5], dans un article de synthèse en anglais, ont répertorié 194 articles pertinents écrits sur le sujet ces 18 dernières années. Au même moment, Fayolle *et al.* [6] ont réalisé en français une revue systématique de 1116 publications dont, au final, 22 ont satisfait aux critères d'inclusion de leur étude.

Malgré une littérature abondante dédiée aux processus de sélection, les auteurs précités constatent l'absence d'un consensus clair et définitif concernant les avantages et inconvénients des différentes méthodes utilisées. C'est aussi ce que soutiennent Wright et Bradeley [7] en soulignant que la sélection des étudiants est un champ fertile à la controverse : « *The selection of medical students has, over the years, raised a number of controversial issues* (p. 1070) » (La sélection des étudiants de médecine a, au fil des années, soulevé un certain nombre de questions de controverse. – traduction libre –).

Une des raisons de ce manque de consensus tient au concept même de sélection. Celui-ci cache en effet des réalités diverses et variées. Ce terme recouvre à la fois des examens et des concours, des filtres à l'entrée des études ou des pratiques de contingentement plus tardif, ainsi qu'une multitude d'outils allant des tests recourant à des questions à choix multiple (QCM) aux entretiens de motivation, en passant par le tirage au sort.

Même les visées de la sélection varient d'un pays à l'autre. S'agit-il de sélectionner les étudiants qui ont les prérequis pour entamer des études exigeantes et coûteuses ? D'octroyer l'accès à un nombre prédéterminé d'étudiants les mieux armés cognitivement ? De ne garder que les étudiants ayant des aptitudes (ou des prédispositions concernant ces aptitudes) souhaitables en milieu médical, telles que la communication et l'empathie ? De viser la sélection d'une élite digne d'effectuer un métier réputé prestigieux ou, au contraire, de veiller à une mixité sociale nécessaire à la pratique du métier, notamment face à un public défavorisé ou en milieu rural ?

Une littérature opaque, parce que non consensuelle, un concept polysémique et des visées de sélection divergentes sont des conditions peu propices à des prises de décision rationnelles et étayées. En l'absence de consensus clair, les décisions quant à l'organisation d'un processus de sélection sont arbitrées par des jugements d'ordre politique, reposant essentiellement sur les assomptions ou les idéologies des décideurs et ce, au détriment d'une réflexion basée sur une étude rigoureuse des données à prendre en compte.

Prenons l'exemple de l'examen d'entrée mis en place récemment en Fédération Wallonie Bruxelles de Belgique. Celui-ci comprend une section éliminatoire centrée sur l'empathie (si l'étudiant n'obtient pas 8/20 pour cette section de l'examen, il sera éliminé d'office). Ce principe de sélection signifie en fait que l'aptitude (ou une certaine disposition) à l'empathie chez les futurs étudiants de médecine est jugée essentielle par les organisateurs de l'examen. Cependant, afin de ne pas trop alourdir le processus de sélection, seuls dix items de type « test de jugement situationnel » sont consacrés à l'évaluation de l'empathie. Et pour ne pas déstabiliser les étudiants, le barème de notation choisi est le même que pour le reste de l'épreuve ; en l'occurrence des points sont retranchés en cas d'erreur, en lien avec le souci que la correction prenne en compte les réponses faites au hasard – pseudo-chance ou *guessing*. En réalité, chaque réponse correcte vaut deux points et chaque erreur entraîne une pénalité de 2/3 de points.

Or, que nous dit la recherche à propos de l'évaluation d'une aptitude telle que l'empathie ?

Premièrement, celle-ci nécessite un certain nombre d'items pour être évaluée. Si l'on considère les principaux tests d'empathie validés scientifiquement, on s'aperçoit qu'ils comprennent au minimum 20 items. Ainsi, le test dénommé *Mehrabian Measure of Emotional Empathy* compte 33 items et celui intitulé *Hogan's Empathy test* en compte 66. En lien direct avec le contexte médical, le *Jefferson Scale of Physician Empathy* compte, quant à lui, 20 items. Or, le nombre d'item est essentiel pour le psychométricien. Si celui-ci est insuffisant pour mesurer un construit, la fidélité de la mesure n'est pas assurée.

Deuxièmement, mesurer l'empathie n'est pas simple et semble peu opportun dans le cadre d'un test de sélection précoce. Hemmerdinger *et al.* [8] ont recensé et analysé 59 instruments de mesure de l'empathie en milieu médical, dont certains ont été utilisés à des fins de sélection des étudiants. Ils concluent : « *No empathy measures were found with sufficient evidence of predictive validity for use as selection measures for medical schools.* » (Aucune mesure de l'empathie n'a démontré suffisamment de validité prédictive pour être utilisée comme mesure de sélection dans les écoles de médecine. – traduction libre –).

Troisièmement, la correction pour la pseudo-chance (*guessing*) postule que les réponses erronées des étudiants ont été données au hasard. Si ce postulat peut être soutenu dans le cadre d'un test cognitif, il pose problème lorsqu'il est appliqué à un test de jugement situationnel où c'est l'attitude qu'adopterait le répondant face à une situation

professionnelle qui est visée. Peut-on, dans ce cas, considérer que l'erreur est due au hasard ? N'est-elle pas plutôt l'expression d'une erreur de jugement de la part de l'individu ?

Une intention louable telle que celle d'inclure des considérations attitudinales dans un test de sélection peut donc mener à des décisions et des conséquences potentiellement délétères puisqu'il paraît douteux que le recours à seulement 10 questions puisse donner lieu à un jugement correct sur l'empathie des répondants. Pourtant, les étudiants qui obtiennent un score inférieur à 8/20 pour l'empathie se voient automatiquement refuser l'accès aux études de médecine en Belgique francophone, quelle que soit la qualité de leurs réponses aux autres questions de l'examen.

Compte tenu des lectures effectuées et du cas belge décrit ci-dessus, notre objectif, à travers cet article, est de réaliser un état des lieux des diverses méthodes qui guident les processus d'admission aux études de médecine puis de porter un regard critique sur ces méthodes, essentiellement du point de vue de leur validité.

Par le biais de cet article, nous aimerions aussi contribuer à donner aux débats, concernant les procédures de sélection, une inflexion légèrement différente. Actuellement, le débat public, particulièrement en France et en Belgique, porte essentiellement sur la question : « Faut-il sélectionner les étudiants qui souhaitent entamer des études de médecine ? » Nous voudrions que cette question soit complétée par d'autres questions tout aussi importantes, à savoir : « Dispose-t-on de moyens psychométriques valides pour sélectionner les étudiants susceptibles d'entreprendre des études de médecine ? Quels sont ces moyens ? Est-il raisonnable de les utiliser ? Sous quelles conditions et à quelles fins ? ».

Dans un premier temps, nous décrirons notre cadre de référence. Celui-ci repose sur la validité en tant que critère définitoire de la qualité en évaluation. Nous décrirons ensuite les principaux outils utilisés dans les tests de sélection en médecine et nous soulignerons leurs forces et leurs faiblesses respectives, puis nous considérerons ces outils du point de vue de leur validité. Enfin, nous conclurons par quelques propositions qui nous semblent utiles pour poursuivre la réflexion.

Le cadre conceptuel de la validité

Le terme « validité » est présent dans la plupart des articles qui traitent de la sélection des étudiants accédant aux études de médecine. Cependant, ce concept y est généralement défini comme la relation entre la mesure résultant de la procédure de sélection et le score obtenu lors d'une épreuve de fin d'année (en premier ou deuxième cycle des études médicales). Si la relation entre deux variables est bien un élément qui plaide en faveur de la validité d'une épreuve d'évaluation, considérer ce seul lien pour juger de la validité de l'épreuve est aujourd'hui reconnu comme insuffisant. Précisons notre cadre de référence en adoptant une perspective historique.

Les années 1921 à 1951 sont décrites par Newton et Shaw [9] comme une période de « cristallisation de la validité ». Son étude repose alors soit sur des calculs de corrélation entre deux scores issus de tests ayant la même visée (validité concomitante), soit sur des calculs de corrélation entre le score au test et une autre mesure et ce, dans une perspective causale (la première mesure conditionnant la deuxième). C'est le cas de la corrélation recherchée entre le résultat d'un étudiant au test de sélection en médecine et celui obtenu par ce même étudiant à la fin du premier ou du deuxième cycles des études médicales. On parle alors de validité prédictive. La notion de fidélité date également de cette époque et réfère à un indicateur de la stabilité des scores obtenus à l'aide d'un même instrument d'évaluation [10].

Dès les années 1950, d'autres formes de validité apparaissent et complètent les approches initiales. Il s'agit de la validité de contenu [11] puis de la validité de construit [12]. À cette époque, la validité est d'abord vue comme une propriété de l'instrument de mesure (l'instrument est valide ou non). Puis, dans les années 1970, elle devient une propriété des données que l'instrument permet de recueillir [13] (les données générées par l'instrument sont valides ou non).

Dans les années 1980, Messick [14] propose une approche unifiée de la validité comprise comme l'intégration d'un ensemble de preuves en lien avec l'usage et l'interprétation de données (y compris celles relatives à l'étude de la fidélité) et ce – il s'agit là d'une nouveauté – dans une perspective sociale. Ce paramètre « social » permet de renforcer le concept de validité dans la mesure où il s'agit aussi de s'intéresser aux conséquences (individuelles et sociales) des résultats générés par le test.

Aujourd'hui, pour beaucoup d'auteurs, la validité fait référence au degré avec lequel les preuves et la théorie soutiennent l'interprétation des résultats d'un test pour un usage précis de celui-ci (AERA, 2014, p. 11) [15]. La validation, qui historiquement reposait exclusivement sur des analyses psychométriques effectuées en aval de la conception du test, devient donc une procédure dynamique continue visant à rechercher des preuves de qualité tout au long du processus d'évaluation, de la création du test jusqu'à l'interprétation des résultats.

Cette approche « globale » est défendue par plusieurs auteurs, dont Kane [16] qui tente de la rendre plus facilement applicable. Ce dernier [16,17] propose d'établir un argumentaire de validité en passant par quatre types d'inférences (pour plus de détails relativement à l'application de ce modèle de validité au contexte des sciences de la santé, voir Pennaforte et Loye [18] et Loye [19]) qui sont :
– les inférences de notation, qui consistent à définir :
- les conditions de construction de l'instrument, notamment en ce qui a trait au contenu à évaluer, à la formulation et au caractère crédible des questions posées ;
- les modalités de l'évaluation qui ont trait à ce qui est demandé à l'évaluateur, au mode de fabrication du résultat, à l'étude théorique du lien entre le comportement observé et le résultat ;

- les modes de collecte et d'analyse des données dans une perspective de contrôle qualité ;

– les inférences de généralisation, qui visent à soutenir :
- la représentativité de l'échantillon des observations, notamment par l'étude des différentes variables qui peuvent nuire au processus (la tâche, les correcteurs, l'échelle, les caractéristiques des candidats) ;
- la reproductibilité de l'évaluation pour que les résultats soient constants au fil des tâches, des évaluateurs, des occasions. C'est à ce niveau que se situe la validité psychométrique (études de fidélité, généralisabilité, théorie de réponse à l'item – TRI-, fonctionnement différentiel des items – FDI-, etc.) ;

– les inférences d'extrapolation, qui visent à :
- expliquer le fonctionnement de la tâche (lien avec le construit visé *via* un modèle plus large, capacité à discriminer) ;
- s'assurer que la tâche est réaliste ;
- définir dans quelle mesure le score est corrélé à d'autres sources d'informations concernant le candidat ;

– les inférences d'implication, qui visent à soutenir la crédibilité et les conséquences des résultats de l'évaluation. Même si ce type d'inférences reste peu documenté, nous pensons qu'elles sont de toute première importance dans le contexte de la sélection à l'entrée des études en médecine et nécessiteraient en conséquence d'être mieux prises en considération.

Dans la suite de cet article, après avoir passé en revue, décrit et analysé de manière critique les principaux outils de sélection, nous les analyserons sous l'angle spécifique des preuves de validité auxquelles ils peuvent prétendre (*cf.* partie discussion).

Les instruments de sélection utilisés

Traditionnellement, deux approches sont utilisées pour la sélection des futurs étudiants en médecine [20]. La première se centre sur les acquis scolaires antérieurs des candidats à ce type d'études, la seconde s'intéresse à des aptitudes autres que cognitives de ces mêmes candidats.

Les acquis scolaires antérieurs

La première approche table, pour sélectionner les meilleurs candidats, sur leurs acquis scolaires antérieurs. Deux types d'indicateurs sont alors classiquement utilisés : 1) les indicateurs de performance scolaire produits par le système éducatif et ; 2) les résultats à un test intégré au processus de sélection.

Les indicateurs de performance scolaire produits par le système éducatif

Les indicateurs de performance scolaire produits par le système éducatif (par exemple, le *A level* au Royaume-Uni, le *Grade Point Average* (GPA) aux États-Unis, la cote R au Québec ou encore le résultat au Baccalauréat en France) possèdent la meilleure validité prédictive concernant la réussite dans les études de premier cycle (*undergraduate*) en médecine [21]. Ils constituent également des prédicteurs fiables – bien que moins puissants – de la réussite des études de deuxième cycle (*postgraduate*), voire même de la carrière de médecin [22].

L'utilisation de ce type d'indicateurs présente toutefois trois points faibles et non des moindres :

– s'ils possèdent la meilleure validité prédictive, ils sont bien loin d'être parfaits. Ainsi, Ferguson *et al.* [21] ont réalisé une méta-analyse à partir de 62 études qui visaient à mettre en lien les résultats antérieurs et les résultats des étudiants à l'issue du premier cycle des études de médecine. Dans cette méta-analyse, le score antérieur explique 23 % de la variance de la réussite en premier cycle. Ces auteurs ont également effectué une méta-analyse sur la base de cinq études faisant le lien entre ces résultats antérieurs et la réussite au deuxième cycle. Le score antérieur n'explique plus alors que 6 % de la variance des résultats obtenus. Ces données laissent présager que les indicateurs de performance scolaire produits par le système éducatif ont certes un certain pouvoir de prédiction concernant la réussite des étudiants en médecine mais que ce pouvoir est relativement limité ;

– de plus, Esmail *et al.* [23] ont souligné très tôt que ce mode de sélection défavorisait les minorités ethniques, ce qui a été confirmé plus récemment [24]. D'autres études [25,26] ont par ailleurs montré que cette façon de procéder n'était pas non plus socialement neutre et qu'elle favorisait les étudiants provenant de milieux socio-économiques favorisés. La dimension éthique liée au problème des disparités ethniques et sociales provoquées par des modes de sélection reposant sur les acquis scolaires antérieurs ne concerne pas seulement l'équité due à chaque candidat à titre individuel mais comporte une dimension professionnelle et sociétale. Comme le soulignent Lievens *et al.* [27], en donnant accès aux études de médecine à des étudiants issus de minorités économiques ou ethniques, c'est l'ensemble des futurs médecins qui améliorent leur expérience éducative en travaillant leurs compétences culturelles. Au niveau sociétal, cela permet aussi de diversifier le profil des médecins et de faire en sorte que cette profession soit plus représentative de la population globale. Les auteurs ajoutent qu'il s'agit là d'un facteur important de santé publique dans la mesure où l'accès des populations issues de minorités aux soins de santé s'en trouve favorisé. Il y aurait donc clairement un biais de sélection lié à l'origine ethnique et sociale des candidats lorsque les critères d'admission aux études de médecine reposent exclusivement sur leurs performances scolaires antérieures. Notons que les Pays-Bas ont tenté de remédier à ce problème en organisant un tirage au sort des futurs étudiants et ce, en pondérant les chances d'être tirés au sort par les résultats scolaires antérieurs [28] ;

– au Royaume-Uni, l'utilisation des résultats antérieurs des candidats aux études de médecine pose un problème technique. L'inflation des grades sur une échelle de mesure ne comptant que quelques échelons a créé un

effet plafond. Beaucoup d'étudiants obtenant le maximum, les mesures ont perdu de leur pouvoir discriminant [29].

Les tests de mesure des acquis d'apprentissage scolaire antérieurs

De nombreux pays ont fait le choix de développer des tests d'aptitudes spécifiques appelés à mesurer les acquis d'apprentissage scolaire antérieurs. Dans certains cas, ces tests ont été mis en place pour pallier l'absence d'indicateurs de performance scolaire produits par le système d'éducation obligatoire (par exemple, aucun indicateur de ce type n'existe en Fédération Wallonie-Bruxelles). Les structures de ces tests sont assez variables car constituées de sections différentes comme en témoigne le tableau I.

Le *Medical College Admission Test* (MCAT) est un test utilisé essentiellement aux États-Unis et au Canada. L'*Undergraduate Medicine and Health Admission Test* (UMAT) et le *Graduate Medical School Admission* (GAMSAT) sont les deux outils de sélection mis en œuvre et pilotés par l'*Australian Council for Educational Research* (ACER). Le *BioMedical Admissions Test* (BMAT) est utilisé par quelques universités au Royaume-Uni et s'adresse également aux candidats aux études vétérinaires. Enfin, le *United Kingdom Clinical Aptitude Test* (UKCAT) est utilisé par la plupart des universités du Royaume-Uni. Il s'agit d'un test d'aptitudes cognitives, qui ne fait pas appel aux acquis scolaires antérieurs, et qui porte plutôt sur quatre dimensions cognitives générales.

Ce relevé non exhaustif des tests d'aptitudes montre la place prépondérante que tiennent les acquis d'apprentissage en sciences puisque, hormis le UKCAT et l'UMAT, ces tests évaluent tous des connaissances scientifiques. C'est toutefois le seul point commun que nous avons pu observer.

Quelles que soient leurs faiblesses et qualités respectives (une analyse au cas par cas serait nécessaire pour en faire état), tous ces tests peuvent se prévaloir d'études qui les accréditent sur la base de leur validité prédictive. Tous ont également fait l'objet, dans le même temps, d'études plus nuancées, voire clairement négatives concernant cet aspect spécifique (voir Patterson *et al.* [5] pour un relevé complet). Par ailleurs, McManus *et al.* [22] mettent en doute la qualité scientifique de certaines de ces études en faveur de la validité prédictive. Ils leur reprochent de porter sur l'ensemble du test et non sur chacune de ses sections. Ils émettent l'hypothèse – en menant une étude sur un article portant sur la validité prédictive du BMAT [30] – que la validité prédictive des tests d'aptitudes est « tirée vers le haut » par les sections du test qui mesurent les acquis scolaires antérieurs des étudiants. Quoi qu'il en soit, seules quelques inférences d'extrapolation sont explorées dans cette littérature, et de surcroît avec des résultats variables.

D'autres pays n'ont pas recours à des tests d'aptitudes, mais élaborent cependant, d'année en année, des questions d'examen ou de concours à l'adresse des futurs étudiants en sciences médicales.

C'est le cas de la Fédération Wallonie-Bruxelles (Belgique) pour laquelle depuis 2017–2018, une commission est chargée de rédiger chaque année des questions d'examen portant sur les matières définies par le décret du 29 mars 2017 relatif aux sciences médicales et dentaires. Selon ce décret, l'examen d'entrée comprend deux parties. La première porte sur la connaissance et la compréhension des matières scientifiques et concerne explicitement les matières de chimie, physique, biologie et mathématiques. La première partie de cet examen vise donc bien à mesurer des acquis d'apprentissage antérieur. Ce n'est par contre pas le cas de la deuxième partie qui vise à évaluer des aptitudes de l'ordre de la communication et de l'analyse critique de l'information. Cette deuxième partie s'articule en fait en quatre sous-sections portant respectivement sur : 1) le raisonnement ; 2) la communication ; 3) l'éthique et 4) l'empathie.

Le cas de la France est, lui aussi, particulier. Les compétences antérieures des étudiants ne sont pas évaluées dans le cadre des concours de sélection appliqués aux études médicales. La sélection a lieu durant la première année commune aux études de santé (PACES) et peut se dérouler soit à la fin du premier quadrimestre, soit à la fin de l'année. Chacune des huit unités d'enseignement qui composent la première année (sept unités communes et une spécifique à l'orientation choisie) fait l'objet d'une évaluation. Les questions d'examen sont rédigées par chaque université, de manière indépendante. Ces questions peuvent être de type QCM ou faire appel à d'autres modalités de questionnement comme, par exemple, des questions à réponse ouverte. Notons que le gouvernement français a annoncé le 18 septembre 2018 que ce concours serait abrogé dès la rentrée 2020. En effet, à partir de cette date, les étudiants pourraient prendre part librement à un premier cycle commun d'une durée de trois ans, durant lequel ils s'orienteraient progressivement vers les différentes filières de santé en fonction de leurs choix et des résultats qu'ils auront obtenus au terme de ce cycle.

Ces tests d'admission et/ou de sélection élaborés à un niveau national, fédéral ou local ne donnent en général pas lieu à des publications scientifiques. Nous ne pouvons, dès lors, nous prononcer sur leur validité. Nous n'avons toutefois pas de raison de penser que ces tests soient plus valides que les tests de sélection dont il est question dans la littérature passée en revue ci-dessus.

Évaluation des dimensions non cognitives

Jusqu'à la fin du siècle passé, l'évaluation des dimensions non cognitives se pratiquait sous la forme d'un entretien face à un jury, *via* l'analyse de lettres de recommandation ou encore en analysant le dossier personnel du candidat (*personal statements*). Ces méthodes posaient une variété de problèmes et la littérature pointe un manque de validité (notamment prédictive), mais aussi de fidélité de ce type de procédures [31–34]. À propos des procédures de sélection axées sur les dimensions non cognitives, Norman [35] écrit en 2004 : « *It is not too big a stretch to suggest that, once we go beyond marks,*

Tableau I. Les contenus des tests d'aptitudes utilisés dans différents pays pour la sélection des étudiants en médecine.

Nom du test	Contenus évalués organisés en sections	Pays
Medical College Admission Test (MCAT)	Les fondations chimique et physique des systèmes biologiques Les fondations biologique et biochimique des systèmes vivants Les fondations psychologique, sociale et biologique du comportement L'analyse critique et le raisonnement	États-Unis Canada
Undergraduate Medicine and Health Admission Test (UMAT)	Le raisonnement logique et la résolution de problème La compréhension des individus Le raisonnement non verbal	Australie
Graduate Medical School Admission (GAMSAT)	Raisonner dans les sciences humaines et sociales La communication écrite Le raisonnement en biologie et sciences physiques (40 % des questions en chimie, 40 % en biologie et 20 % en physique)	
BioMedical Admissions Test (BMAT)	Les aptitudes et skills (aptitudes générales en résolution de problème, compréhension d'arguments, analyse de données et inférence) Les connaissances scientifiques et leurs applications (couvrant les matières scientifiques et mathématiques) Les tâches écrites évaluant la capacité à sectionner, développer et organiser des idées et à les communiquer par écrit avec concision	Royaume-Uni
United Kingdom Clinical Aptitude Test (UKCAT)	Le raisonnement verbal Le raisonnement quantitatif Le raisonnement abstrait L'analyse de décision	
Pas de test d'aptitude, mais un examen	Chimie Biologie Physique Mathématique Évaluation des capacités de raisonnement, d'analyse, d'intégration, de synthèse, d'argumentation, de critique et de conceptualisation ; Évaluation de la capacité à communiquer et à percevoir les situations de conflit ou potentiellement conflictuelles ; Évaluation de la capacité à percevoir la dimension éthique des décisions à prendre et de leurs conséquences pour les individus et la société ; Évaluation de la capacité à faire preuve d'empathie, de compassion, d'équité et de respect.	Fédération Wallonie Bruxelles de Belgique
Pas de concours à l'entrée	Examens en fin de première portant sur la matière vue lors de la première année commune des études de santé.	France

many of our schools are engaged in the process of conducting a very elaborate, labour-intensive, and expensive lottery ». (Il n'est pas exagéré de suggérer qu'une fois que nous allons au-delà des notes scolaires, beaucoup de nos écoles sont engagées dans le processus de réalisation d'une loterie très élaborée, qui demande beaucoup de travail et qui coûte cher. – traduction libre –).

Néanmoins, la volonté de ne pas limiter la décision d'admission aux études médicales à une mesure des savoirs scolaires perdure, essentiellement pour les deux raisons suivantes : 1) les dimensions non cognitives ont trait à des aptitudes jugées essentielles pour exercer une profession médicale et 2) elles sont supposées diminuer la reproductibilité sociale inhérente à l'évaluation des acquis scolaires antérieurs. Aujourd'hui, les instruments conçus pour évaluer ces dimensions se sont améliorés et diversifiés. Nous en présentons une liste non exhaustive.

Les tests de personnalité

Des recherches sont menées depuis au moins cent ans dans le but de développer et valider des tests de personnalité. La raison de leur utilisation dans le cadre des processus de sélection des étudiants de médecine est liée au fait que certains traits de personnalité sont particulièrement favorables à la pratique médicale. Ainsi, quelques études sur le sujet ont mis en évidence une corrélation entre les facteurs du Big 5 (ouverture, conscience professionnelle, extraversion, agréabilité et stabilité émotionnelle) et la réussite des études de médecine [36–38]. Cependant, comme le soulignent Schripsema *et al.* [39] en s'appuyant sur des études de Le *et al.* [40], il semble que le lien entre traits de personnalité et réussite ne soit pas linéaire, mais plutôt curvilinéaire. Par exemple, si avoir une conscience professionnelle est un facteur de réussite dans les premières années d'études, en avoir de trop peut devenir un facteur de contre-performance dans le domaine clinique [5]. Cela rend les résultats de ces tests difficilement interprétables. Par ailleurs, les tests de personnalité pêchent généralement par un défaut de fidélité, étant donné que les réponses des étudiants sont souvent entachées de désirabilité sociale [41], ce qui, en termes de validité, se traduit par des inférences problématiques de généralisation et d'extrapolation.

Les tests de jugements situationnels (situational judgement tests – SJT)

Ces tests [42] sont construits à partir de situations professionnelles scénarisées et présentées par écrit ou sous la forme d'une vidéo. Le répondant est donc confronté à une situation professionnelle et il doit choisir, parmi un ensemble de solutions qui lui sont proposées, celle(s) qui lui semble(nt) la/les plus pertinente(s) pour faire face à la situation. Des dispositions telles que l'intégrité, l'empathie, la capacité à gérer des relations interpersonnelles, la centration sur le bien-être du patient ou encore le sens du leadership peuvent être évaluées efficacement à l'aide de cette méthode [43,44]. Ce type de tests est par ailleurs assez praticable, car il est à réponse standardisée et peut être présenté à un grand nombre d'étudiants sans entraîner de coûts trop élevés. Il est considéré comme ayant parfois une bonne validité prédictive, essentiellement par rapport à la réussite des études cliniques [45]. Par ailleurs, le test ne désavantage pas les étudiants issus d'un niveau socio-économique faible [27]. C'est un atout important que ne possèdent pas les tests qui mesurent des compétences académiques. Par contre, il semble ne pas être complètement neutre quant à l'origine ethnique [23].

La validité de ces tests dépend toutefois de la manière dont ils ont été conçus. Par exemple, l'étude de Lievens *et al.* [27], en 2016, porte sur l'intégration d'une épreuve de jugement situationnel dans le UKCAT. Les situations qui y sont proposées sont le fruit d'un travail rigoureux qui impliquait notamment des psychologues en organisation accompagnés de 27 médecins ayant une expertise liée aux aptitudes visées. Les 27 médecins avaient été choisis de façon à respecter une certaine diversité ethnique ainsi qu'une représentation équilibrée des genres. Il est d'ailleurs possible que les auteurs de l'étude seraient parvenus à d'autres conclusions si ces préalables méthodologiques n'avaient pas été respectés. Par exemple, nous savons que les consignes liées à la réponse (donner un choix ou juger d'un comportement proposé suite à une situation) influencent de manière importante la validité [46]. Il en va de même pour le support utilisé pour présenter le cas. Selon Sharma et Nagle [47], le support vidéo augmente la qualité de l'épreuve en comparaison à une version papier-crayon, notamment parce que ce support diminue le poids de la compétence en littératie, compétence défavorable aux minorités. Ces instruments récents font donc l'objet d'études qui visent à produire des inférences de notation en examinant de façon critique le soin apporté à leur élaboration.

Dans certains cas, et cela nous semble être une dérive, les tests de jugements situationnels (TJS) sont conçus au niveau local ou national par des personnes qui ne sont pas expertes de ce type de dispositifs. Ainsi, des TJS peuvent être créés par des psychologues spécialistes de l'empathie mais n'ayant aucune expertise concernant la création de TJS, voire la création de tests en général. Même si les tests élaborés dans de telles conditions ne font en général pas l'objet d'études de validité, nous pouvons raisonnablement douter de leur qualité psychométrique.

Les mini entrevues multiples (MEM)

Ces tests reposent sur une méthodologie proche des examens cliniques objectifs structurés (ECOS). Ils sont composés d'une succession de stations (en général 8 à 12) qui durent chacune entre 8 à 10 minutes, chacune associée à un évaluateur différent, ce qui minimise les biais liés au correcteur. À chaque station, un scénario standardisé est proposé au candidat [48]. Ce scénario vise à mesurer une série de compétences, d'aptitudes ou de traits de personnalité, par exemple la communication, la tolérance à l'ambiguïté, l'ouverture d'esprit ou encore la résolution

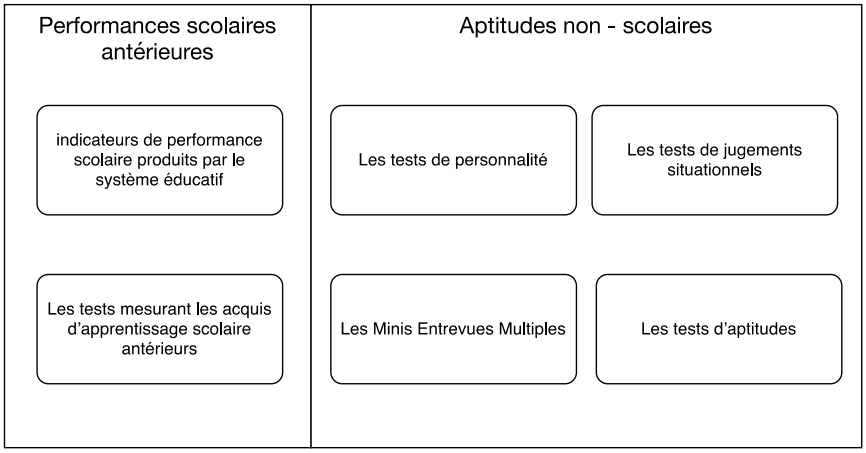

Figure 1. Les différentes épreuves employées pour la sélection des étudiants en médecine.

de problème. Plusieurs études se sont penchées sur la fidélité ou la validité de ce type d'épreuve. Ainsi, celle d'Eva *et al.* [49] montre que les MEM ont une fidélité juste satisfaisante (0,73 quand 12 stations de 10 minutes étaient utilisées) et une validité prédictive modérée (corrélation entre les MEM et des ECOS de 0,35 lors des études de premier cycle et de 0,43 pour les études de second cycle). Plus tôt, Reiter *et al.* [50] et Harris et Owen [51] avaient déjà apporté des preuves de la validité prédictive des MEM. Récemment, Oluwasanjo *et al.* [52] ont obtenu une corrélation positive, significative mais modérée (R = 0,384) entre le score global aux MEM et des ECOS proposés à des internes en début de formation. Finalement, Bergeron *et al.* [53] mettent en évidence une bonne valeur prédictive des résultats lors de l'externat.

Cependant, comme le soulignent Eva *et al.* [49], le MEM est plus un processus d'évaluation qu'un test (p. 768). Cela signifie notamment que la qualité du MEM est assez dépendante du processus de construction. Ainsi, Knor et Hissbach [54], qui ont répertorié 66 études portant sur le MEM, dressent comme constat que le design des MEM varie largement en termes de compétences visées, du type de stations, de détails organisationnels (nombre de stations, temps de préparation, durée...) et de système de notation, et ce d'une institution à l'autre ou même dans la même institution en fonction des années (p. 1164). Or, certaines de ces variables peuvent avoir des effets sur la fidélité ou la validité du dispositif [55]. Par ailleurs, le fait de devoir recourir à des correcteurs et les problèmes de fidélité inter-juges qui en découlent constituent une difficulté majeure non encore résolue à ce jour [56], si ce n'est en augmentant le nombre de stations, ce qui paraît une solution peu satisfaisante, notamment pour des raisons de praticabilité et de coûts.

Les tests d'aptitudes pour évaluer les dimensions (non)-cognitives

Certains tests d'aptitudes ont également leur place dans cette section, lorsqu'ils visent à évaluer des dimensions cognitives ou non, mais dans tous les cas indépendantes des acquis scolaires antérieurs. C'est, par exemple, le cas de l'UKCAT déjà mentionné précédemment (voir Tableau I). Ce test a été conçu par un *consortium* d'une vingtaine d'universités anglaises en collaboration avec la société de *testing* Pearson Vue [57]. Il s'agit d'un test d'aptitude cognitive qui porte sur quatre dimensions qui sont : 1) le raisonnement verbal ; 2) le raisonnement quantitatif ; 3) le raisonnement abstrait et 4) l'analyse de décision (un test de jugement situationnel lui a été ajouté en 2012 pour compléter le processus d'admission). Composée d'items de type QCM, son administration est électronique. Ce qui est étonnant, au vu des enjeux importants de la sélection, c'est qu'il a été utilisé pour sélectionner de futurs médecins dans 23 universités dès 2006 [58] alors qu'il n'avait pas encore livré de réels gages de validité. Les premières études *a posteriori* n'allaient d'ailleurs pas être très rassurantes sur son caractère prédictif. En effet, Lynch *et al.* [59], dans une des premières études sur le sujet portant sur 341 étudiants issus de deux écoles de médecine écossaises (Dundee et Aberdeen) ne trouvent aucun lien entre les résultats à l'UKCAT et la réussite du premier cycle en médecine. Yates et James [58] qui s'intéressent au caractère prédictif de ce test sur la réussite aux deux premières années de cursus des étudiants à Nottingham (N = 209) n'identifient que des liens très faibles et qui ne s'appliquent qu'à certaines épreuves. Deux études sont légèrement plus positives en ce qui concerne les caractéristiques du UKCAT. Wright et Bradley [57], s'intéressant à 304 étudiants de Newcastle, font état d'une validité prédictive modérée sur les tests de connaissances en première et deuxième année de médecine. Ce caractère prédictif semble toutefois s'estomper avec le temps et disparaît quand il porte sur les ECOS. McManus *et al.* [60] font mention, dans leur étude réalisée dans 12 écoles de médecine et auprès de 4811 étudiants, d'une valeur prédictive significative, mais faible, de ce test. Leurs résultats montrent cependant qu'utilisé conjointement avec une autre méthode de sélection classique – les résultats scolaires antérieurs (*A score*) – l'UKCAT amène

une validité incrémentielle. Une dernière étude [61], plus positive encore, montre des corrélations positives et significatives (de 0,24 à 0,36) entre l'UKCAT et des examens classiques, d'une part, et des ECOS de quatrième année, d'autre part. Constatons que malgré une validité qui reste encore largement à démontrer, la plupart des étudiants du Royaume-Uni sont soumis depuis une décennie à cette épreuve et le seront cette année encore.

La figure 1 résume les analyses que nous venons de livrer sur les divers outils d'évaluation.

On y perçoit bien la vision limitée des auteurs quant au concept de validité, qui devrait être objet d'étude pour élaborer un argumentaire complet et cohérent soutenant la validité des démarches entreprises. Là encore quelques intérêts pour les inférences de notation sont relevés, mais les études se focalisent clairement sur les inférences d'extrapolation.

Discussion

Des instruments imparfaits

Même si la validité semble une préoccupation annoncée par les nombreux auteurs consultés, notre constat est que leur vision en est bien imparfaite et incomplète. Nous notons aussi que les résultats publiés sont bien peu stables dans le temps ou dans l'espace. Ainsi, si des décideurs ayant pour mission d'opter pour un processus de sélection valide se penchaient sur cette littérature, il leur serait difficile d'opter avec sérénité pour l'une ou l'autre procédure. Cela explique sans doute pourquoi les choix sont souvent faits sur base d'enjeux politiques ou même simplement des préférences des principaux décideurs [5]. Évidemment, plusieurs outils sont souvent combinés, ce qui nous semble une bonne chose. Les tenants des sciences sociales nous apprennent en effet que quand une information est incertaine, l'inclure dans un processus de triangulation en la confrontant à d'autres données issues d'autres sources et méthodes permet d'améliorer le jugement. Comme le soulignent Denzin et Lincoln [62], cette méthode favorise par ailleurs la rigueur et la profondeur de l'analyse, à la condition toutefois que la dynamique de comparaison des informations se fasse elle-même avec rigueur. Or, nous savons peu de choses sur la manière dont les jurys traitent ces différentes sources d'information. De plus, cela ne saurait suffire à légitimer l'utilisation de données à la qualité plus qu'incertaine ni justifier un manque de rigueur dans la construction des outils.

Or, tant les procédures de sélection axées sur les acquis scolaires antérieurs que celles évaluant les dimensions non cognitives posent des problèmes de validité. À tout le moins, nous défendons l'idée que les études de validation qui tentent de les légitimer souffrent de problèmes méthodologiques importants. En conséquence, en 2018, nous ne pouvons sans doute toujours pas rejeter comme faux les propos que Benbassat et Baumal [20] tenaient en 2007 : « *the use of non-cognitive criteria is costly in terms of time and manpower, and their reliability and validity is a matter of controversy* (p. 8) » (L'utilisation de critères non-cognitifs est coûteuse en termes de temps et de main d'œuvre et leur fidélité ainsi que leur validité sont objet de controverse. – traduction libre –).

Pour appuyer cette affirmation, la figure 2 propose un regard global sur ce que nous avons observé dans les études liées aux diverses méthodes de sélection que nous avons citées. Il met en évidence le fait que les études de validité dans le domaine de la sélection des futurs médecins sont essentiellement de nature corrélationnelle et, pour celles qui les abordent, réduisent les inférences de généralisation au seul calcul d'un coefficient de fidélité, avec de surcroît un résultat souvent peu satisfaisant. Il est assez rare que soient mentionnées dans la littérature des inférences de notation précisant par le menu le processus avec lequel le test a été construit ou les vérifications éventuelles qui ont porté sur le processus de réponse. Quant aux inférences d'implication, il n'en est tout simplement jamais question, probablement parce qu'il est difficile à l'heure actuelle de savoir comment s'y prendre pour en témoigner.

Nous ne prétendons pas que la figure 2 reflète la totalité de la recherche sur le sujet. Nous osons affirmer, par contre, que les études portant sur la validité des modalités de sélection à l'entrée en médecine sont insuffisantes.

En effet, de manière globale, notre jugement sur les méthodes de validation de ces instruments est sévère, ce qui limite d'après nous grandement la portée des études citées dans cet article. Nous nous attardons donc dans les paragraphes qui suivent à réfléchir aux types de preuves qui pourraient être proposées dans de telles études, notamment en mettant l'accent sur la définition des contenus à évaluer.

Preuves liées au contenu dans les épreuves de sélection

Pour envisager des inférences de généralisation basées sur des approches psychométriques plus appropriées qu'un simple regard sur la cohérence interne de l'épreuve (fidélité), il est absolument nécessaire de disposer d'un construit à mesurer, de définir ce que l'on cherche à mesurer. Or, comme le soulignent Patterson *et al.* [5], ceci nécessiterait, dans le cas qui nous occupe, de disposer d'une taxonomie claire, soutenue théoriquement, qui expliquerait quelles sont les compétences, cognitives ou non, d'un bon médecin. Les auteurs soulignent (p. 50) que, pour réaliser un tel travail, la conduite de recherches visant à déterminer ce qu'est un praticien compétent, et ce à divers moments de sa formation puis de sa pratique, est une étape obligatoire. À ce jour, nous ne disposons pas d'une telle taxonomie de référence, même si au Royaume-Uni, le *General Medical Council* (voir www.gmc-uk.org/guidance/) a donné une définition des bonnes pratiques médicales et des compétences y afférentes, que le cadre de référence CanMEDS du Canada tente aussi de définir. Le problème de ces taxonomies est qu'elles se limitent à l'identification des compétences, sans spécifier, pour chacune d'elles, un *continuum* de développement [63] ni de pistes pour les acquérir.

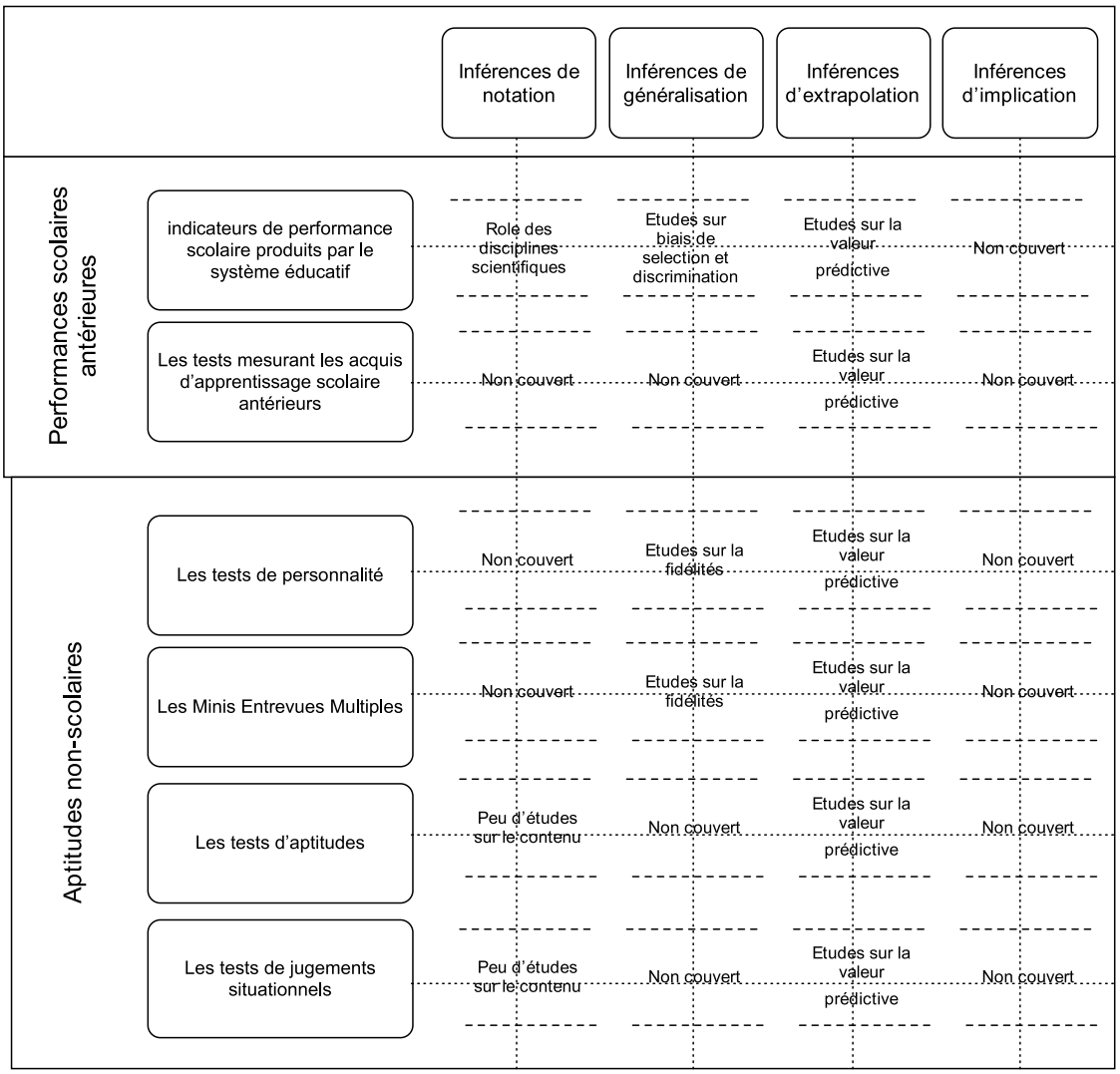

Figure 2. Panorama de synthèse des modalités de validation des méthodes de sélection des étudiants pour les études médicales.

Même si nous développions et disposions d'une telle taxonomie, son utilisation dans une procédure très précoce de sélection nous semble en outre poser deux types de problèmes.

1. En termes de développement se pose la question de savoir si une telle taxonomie peut être unifiée. En effet, hormis les compétences médicales, quelles sont les habiletés non cognitives communes à un médecin généraliste, un pédiatre, un neurochirurgien et un anesthésiste ? Cette question nous semble cruciale si la perspective est de les utiliser pour la sélection. Albanese *et al.* [32] notent à cet effet « *The literature identifies up to 87 different personal qualities relevant to the practice of medicine and selecting the most salient of these that can be practically measured is a challenging task* (p. 1) ». (La littérature identifie jusqu'à 87 qualités personnelles pertinentes pour la pratique médicale et sélectionner les plus importantes d'entre elles pour en faire des objets de mesure réalisable est une tâche difficile. – traduction libre –). Malgré les difficultés d'une telle entreprise, nous plaidons, dans ce contexte, pour l'émergence de référentiels de compétences pour chacune des professions médicales, créés sur la base d'une analyse soigneuse de l'activité. Il s'agit là d'une condition nécessaire pour identifier, ensuite, les compétences qui sont transversales et communes à la plupart des profils de médecins, quelle que soit leur spécialité. Chacune de ces compétences devrait ensuite être pensée de manière développementale. Cela permettrait de pouvoir estimer le niveau de développement nécessaire à l'entrée en médecine.

À défaut, les procédures de sélection axées sur les qualités personnelles pêcheront toujours par manque de validité. Que permettent-elles de mesurer réellement et surtout comment être assuré que ce sont bien les qualités les plus importantes pour exercer la profession de médecin ? Cette réflexion nous semble trop embryonnaire dans la littérature. Par exemple, aujourd'hui, est-il possible d'affirmer que les médecins anglais ont un besoin supérieur de raisonnement verbal (c'est l'une des sections de l'UKCAT) par rapport aux étudiants américains (le MCAT en est dépourvu) ? Et si oui, qui en a décidé et en fonction de quels arguments ?

Dans le cas des tests de jugements situationnels ou des MEM, plusieurs auteurs décrivent de manière un peu péremptoire ces procédures comme valides ou non valides. Cette catégorisation nous semble éminemment abusive. En effet, ces études passent le plus souvent sous silence les principes qui ont été à la base du choix et du développement des situations (SJT) ou des stations (MEM), mais également la manière dont les évaluateurs attribuent un résultat aux candidats. Elles passent donc sous silence l'argumentaire montrant en quoi ces choix sont pertinents quant à ce qui est à évaluer. Il est donc impossible de porter un jugement sur la validité dans de telles conditions.

2. Quand bien même nous disposerions de telles taxonomies, cela amène à une seconde question : que doit-on mesurer à l'entrée des études de médecine ? À ce stade, pour des études exigeantes et coûteuses comme celles de médecine, il semble assez raisonnable d'évaluer les prérequis nécessaires. C'est ce que vise l'évaluation des acquis scolaires antérieurs.

Pour les compétences non cognitives, le problème se révèle plus ardu. S'agit-il d'évaluer des compétences en germe, des prédispositions, des prérequis attitudinaux qui seront, par la suite, développés dans les cursus de formation ? Ou s'agit-il d'évaluer, déjà, des aptitudes bien ancrées qui sont valorisées dans la profession médicale ?

S'il s'agit d'évaluer des prédispositions, on peut se demander s'il est possible de déterminer le niveau minimalement acceptable de celles-ci, niveau minimum sur lequel s'appuieraient les facultés de médecine pour poursuivre les développements de ces aptitudes professionnelles jusqu'alors en germe chez les étudiants ?

Nous pensons que fixer un tel niveau n'est pas évident pour deux raisons principales. D'abord, pour déterminer si les étudiants possèdent ces prérequis nécessaires, il faudrait pouvoir disposer, pour chacun des domaines non cognitifs envisagés, d'un modèle de développement. Ainsi, nous pourrions identifier où se situe l'individu sur un *continuum* et refuser l'accès aux études aux individus n'ayant pas atteint un certain stade jugé nécessaire aux acquisitions subséquentes. Or, nous l'avons souligné, autant il semble difficile de développer une taxonomie précise de ces habiletés non cognitives, autant c'est une gageure d'imaginer que nous pourrions disposer pour chacune d'elles d'un modèle cognitif élaboré.

Ensuite, dans la logique ainsi développée, ces prérequis constitueraient le socle nécessaire aux apprentissages visés par le cursus de formation. Il est évident que des dimensions non cognitives sont aujourd'hui prises en compte dans la plupart des formations médicales. Mais est-ce à dire, toutefois, qu'il s'agit d'objectifs de formation clairement énoncés, soutenus pédagogiquement et entraînés à travers des dispositifs efficients par des professionnels compétents ? Pour toutes les dimensions ? Dans toutes les universités ? Et si c'était le cas, à partir du moment où aucun modèle développemental n'est disponible, comment sont-elles soutenues par des activités d'enseignement évolutives et comment sont-elles évaluées ?

S'il s'agit d'évaluer précocement des aptitudes bien ancrées afin de sélectionner ceux qui les ont déjà développées, on peut se demander s'il est légitime de sélectionner précocement des candidats sur base d'aptitudes professionnelles qui, pour certaines, ne seront pas ou prou enseignées de manière explicite lors des études. La question de la responsabilité des établissements ne formant pas les étudiants à des dimensions pourtant jugées nécessaires à la pratique du métier ne se poserait-elle pas de manière criante ?

Tant que ces questions ne seront pas résolues, il nous semble difficile d'imaginer que les processus de sélection offrent la rigueur nécessaire à cet enjeu sociétal d'importance. Ainsi, nous pensons que les inférences de notation sont trop peu développées actuellement dans la littérature et sont à traiter en priorité pour étudier la validité des instruments utilisés pour la sélection des étudiants à l'entrée en médecine, notamment en ce qui a trait à toutes les dimensions non cognitives.

Preuves axées sur la relation à d'autres variables

La plupart des études voient dans le lien – corrélationnel – entre une mesure issue du test de sélection et une mesure d'évaluation en cours d'étude (au premier ou au deuxième cycle, dans la formation théorique ou lors de l'internat et de l'externat) des preuves de la validité du processus de sélection. Nous pourrions pourtant nous demander à rebours si de telles corrélations ne devraient pas conduire à questionner les preuves de la validité des épreuves et examens en cours d'études. En effet, l'un des problèmes majeurs de la vérification de la validité en référence à un critère est le doute possible quant à la validité de l'autre mesure utilisée dans la corrélation. Dès lors, par exemple, quelles preuves les ECOS réalisés dans une université donnent-ils quant à la suffisance de leur qualité pour être utilisés dans de telles études corrélationnelles ? Le même constat peut d'ailleurs être dressé pour les tests QCM souvent proposés aux étudiants en fin de premier cycle.

En outre, il semble légitime de se demander si les liens corrélationnels décrits entre les MEM et les ECOS pourraient être induits par une compétence transversale – permettant en l'occurrence de réussir ce type d'épreuve quel que soit le construit mesuré –, telle que, par exemple, une compétence de gestion de stress teintée d'assertivité favorable à la réussite.

Dans le même esprit, la corrélation entre les tests d'aptitudes visant des savoirs scolaires et la performance en premier cycle – qui est supérieure à celle entre ces mêmes tests et la performance en second cycle – ne pourrait-elle pas s'expliquer par la capacité à répondre efficacement à des QCM (qui constituent le plus souvent la mesure de performance en premier cycle) ? Nous n'en

savons évidemment rien, mais ces hypothèses ne peuvent sans doute pas être balayées d'un simple revers de la main, la corrélation pouvant peut-être s'expliquer par d'autres raisons que celles proposées.

À ce titre, les études qui mettent en lien une variable prédictive (ici le score obtenu au processus de sélection) avec la réussite des étudiants ont été largement critiquées dans la littérature pédagogique [64]. La compréhension des facteurs influençant la réussite des études nécessite en effet un appareillage complexe prenant en compte plusieurs variables, mais aussi leurs interactions. De telles études ne sont pas simplement corrélationnelles et impliquent des modèles statistiques plus complexes, par exemple multivariés et même multiniveaux. Récemment, des auteurs [65] préconisent également d'abandonner les études centrées sur les variables au profit d'études favorisant l'émergence de profils d'étudiants, par exemple à partir d'analyses en clusters. En outre, la psychométrie offre une grande variété de modèles qui peuvent soutenir des inférences de généralisation et également offrir des preuves de validité sous forme de liens avec d'autres variables (par exemple les analyses de fonctionnement différentiel d'items).

Conclusion

Les processus de sélection des étudiants pour l'entrée ou la poursuite dans les études de médecine ont ouvert un champ théorique dense et complexe. De nombreuses procédures de sélection y sont détaillées. Nous invitons toutefois les différents auteurs à prendre en compte les principes de qualité définis dans la littérature en termes de validité et de fidélité. Sans doute cela devrait-il passer par des études plus fondamentales visant à déterminer ce qui caractérise les médecins compétents à divers stades de leur apprentissage puis de leur carrière, caractéristiques qu'il ne nous semble pas possible d'unifier pour l'ensemble des métiers auxquels ouvrent les études de médecine.

Il nous semble également nécessaire de considérer avec plus d'acuité et de finesse une vision moderne de la validité telle que définie en ce début de XXI^{e} siècle pour s'atteler à la tâche. En effet, les auteurs qui traitent de validité s'inscrivent majoritairement dans la visée de produire des inférences d'extrapolation. Nous pensons que cette vision de la validité est non seulement limitée, mais également souvent erronée par le caractère restrictif des études corrélationnelles qui sont présentées. Au-delà de travaux approfondis pour définir les contenus à évaluer, des études visant à produire plus largement des inférences de généralisation et d'extrapolation sont nécessaires. Il sera alors peut-être possible de penser à des manières de prendre en compte des questionnements de nature plus largement politique et de réfléchir aux conséquences des choix dans le cadre des inférences d'implication.

Cette dernière proposition est tout à fait en lien avec la première recommandation du consensus développé par le groupe de la conférence d'Ottawa sur le sujet. Les auteurs qui la signent, parmi les plus informés sur le sujet, plaident pour que les comités de sélection adoptent les principes de « bonnes pratiques » en évaluation. À travers cet article, nous donnons du corps à cette proposition et offrons au lecteur un cadre conceptuel, celui de la validité en évaluation, pour concrétiser cette démarche qualité.

Nous voudrions conclure sur un autre élément qui nous paraît essentiel. Les commissions qui organisent la sélection des étudiants ont de lourdes responsabilités, respectivement vis-à-vis des facultés de médecine (faire en sorte que les étudiants qui y entrent soient les plus aptes possible à suivre ce type d'études) ou vis-à-vis de la société (favoriser la formation de médecins outillés des compétences adéquates au regard de l'exigence de responsabilité sociale, en faisant en sorte, par exemple que les médecins de demain couvrent les besoins en termes de santé publique). Cela se concrétise notamment par la préoccupation de varier les profils sociaux et ethniques des étudiants sélectionnés. Mais elles ont également un autre type de responsabilité, qui est la responsabilité individuelle vis-à-vis de chacun des candidats, parce que celle-ci est immédiate et qu'une erreur de sélection, si elle est de faible ampleur, sera sans conséquence majeure au niveau des facultés et de la santé publique, mais qu'elle pourra mettre fin de manière douloureuse, parce qu'injuste, à des espoirs vocationnels portés par des jeunes que nous devons aussi protéger. Cela devrait être la responsabilité première, parce que mère des deux autres, qui devrait obnubiler les membres de ces commissions. En plus de la mise en œuvre de démarches favorisant la validité, et sachant le caractère irrémédiable des erreurs de mesure, il est opportun d'inclure systématiquement dans le processus de sélection une réflexion démocratique et éthique, en amont, pendant et en aval des épreuves.

Contributions

Les deux auteurs ont contribué solidairement à l'élaboration du travail, une contribution majoritaire ayant été apportée par Pascal Detroz.

Approbation éthique

Sans objet.

Conflits d'intérêts

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

Références

1. Morrison J. Selecting for medical education. Med Educ 2016;50:3-5.
2. Higher Education Funding Council for England. Non-continuation rates: Trends and profiles. 2016 [On-line] Disponible sur https://webarchive.nationalarchives.gov.uk/20160106192551/http://www.hefce.ac.uk/analysis/ncr/nc/.

3. Association of American Medical Colleges. US Medical School Applicants and Matriculants by school, state of legal residence and sex 2014. AAMC 2014 [On-line] Disponible sur https://www.aamc.org/data/facts/.
4. Boelen C. Global consensus on social accountability of medical schools. Santé Publique 2011;23:247-50.
5. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Med Educ 2016;50:36-60.
6. Fayolle A-V, Passirani C, Letertre E, Ramond A, Perrotin D, Saint-André J-P, Richard I. Sélection des étudiants en médicine : facteurs prédictifs de réussite ; une revue systématique de la littérature. Presse Med 2016;45:483-94.
7. Wright SR, Bradley PM. Has the UK Clinical Aptitude Test improved medical student selection? Med Educ 2010;44 (11):1069-76.
8. Hemmerdinger JM, Stoddart SD, Lilford RJ. A systematic review of tests of empathy in medicine. BMC Med Educ 2007;7:24.
9. Newton PE, Shaw SD. Validity in educational and psychological assessment. Thousand Oaks (CA): Sage Publications, 2014.
10. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;6:297-335.
11. Cureton EE. Validity, in Educational measurement, Lindquist EF, Editor. American Council on Education: Washington DC, 1951, p. 621-94.
12. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin 1955;52(4):281-302.
13. Cronbach LJ. Test validation, in Educational measurement (2nd ed.), Thorndike RL, Editor. American Council on Education: Washington DC, 1971, p. 443-507.
14. Messick S. Validity, in Educational measurement (3rd ed.), Linn RL, Editor. American Council on Education and Macmillan: New York (NY), 1989, p. 13-104.
15. AERA, APA, NCME. Standards for educational and psychological testing. Washington DC: AERA, APA, & NCME, 2014.
16. Kane, MT. Validation, in Educational measurement (4th ed.), Brennan RL, Editor. American Council on Education and Praeger: Westport, CT, 2006, p. 17-64.
17. Kane MT. Validation as a pragmatic, scientific activity. Journal of Educational Measurement 2013;50:115-22.
18. Pennaforte T, Loye N. Une approche pragmatique de validation en éducation médicale : l'application du modèle de Kane à un outil d'évaluation du raisonnement clinique, in Mesure et évaluation des compétences en éducation médicale : regards actuels et prospectifs, Dionne E, Raiche I, Editors. Presses de l'Université du Québec : Québec, 2017, p. 143-176.
19. Loye N. Et si la validation était plus qu'une suite de procédures techniques ? Mesure et Évaluation en Éducation 2018;41:97-123.
20. Benbassat J, Baumal R. Uncertainties in the selection of applicants for medical school. Adv Health Sci Educ Theory Pract 2017;12:509-21.
21. Ferguson E, James D, Madeley L. Factors associated with success in medical school: Systematic review of the literature. BMJ 2002;324:952-7.
22. McManus IC, Ferguson E, Wakeford R, Powis D, James D. Predictive validity of the BioMedical Admissions Test: An evaluation and case study. Med Teach 2011;33:53-7.
23. Esmail A, Nelson P, Primarolo D, Torna T. Acceptance into medical school and racial discrimination. BMJ 1995;310:501-2.
24. Stegers-Jager KM, Steyerberg EW, Lucieer SM, Themmen APN. Ethnic and social disparities in performance on medical school selection criteria. Med Educ 2015;49:124-133.
25. Simmenroth-Nayda A, Görlich Y. Medical school admission test: advantages for students whose parents are medical doctors? BMC Med Educ 2015;15:81.
26. Girotti JA, Park YS, Tekian A. Ensuring a fair and equitable selection of students to serve society's health care needs. Med Educ 2015;49:84-92.
27. Lievens F, Patterson F, Corstjens J, Martin S, Nicholson S. Widening access in selection using situational judgement tests: evidence from the UKCAT. Med Educ 2016;50: 624-36.
28. Cohen-Schotanus J, Muijtjens AMM, Reinders JJ, Agsteribbe J, van Rossum HJM, van der Vleuten CPM. The predictive validity of grade point average scores in a partial lottery medical school admission system. Med Educ 2006;40:1012-19.
29. McManus I, Dewberry C, Nicholson S, Dowell JS, Woolf K, Potts HW *et al.* Construct-level predictive validity of educational attainment and intellectual aptitude tests in medical student selection: Meta-regression of six UK longitudinal studies. BMC Medecine 2013;11:243.
30. Emery JL, Bell JF. The predictive validity of the BioMedical Admissions Test for pre-clinical examination performance. Med Educ 2009;43:557-64.
31. Ferguson E, James D, O'Hehir F, Sanders A, McManus IC. Pilot study of the roles of personality, references, and personal statements in relation to performance over the five years of a medical degree. BMJ 2003;326:429-32.
32. Albanese MA, Snow MH, Skochelak SE, Huggett KN, Farrell PM. Assessing personal qualities in medical school admissions. Acad Med 2003;78:313-21.
33. Kreiter CD, Yin P, Solow C, Brennan RL. Investigating the reliability of the medical school admissions interview. Adv Health Sci Educ Theory Pract 2004;9:147-59.
34. Siu E, Reiter HI. Overview: What's worked and what hasn't as a guide towards predictive admissions tool development. Adv Health Sci Educ Theory Pract 2009;14:759-75.
35. Norman G. The morality of Medical School Admissions. Adv Health Sci Educ Theory Pract 2004;9:79-82.
36. Lievens F, Coetsier P, De Fruyt F, De Maeseneer J. Medical students' personality characteristics and academic performance: a five-factor model perspective. Med Educ 2002;36:1050-6.
37. Lievens F, Ones DS, Dilchert S. Personality scale validities increase throughout medical school. J Appl Psychol 2009;94:1514-35.
38. Hojat M, Erdmann JB, Gonnella JS. Personality assessments and outcomes in medical education and the practice of medicine: AMEE Guide No. 79. Med Teach 2013;35: e1267-e1301.
39. Schripsema NR, van Trigt AM, van der Wal MA, Cohen-Schotanus J. How different medical school selection processes call upon different personality characteristics. PloS One 2016;11:e0150645.
40. Le H, Oh I-S, Robbins SB, Ilies R, Holland E, Westrick P. Too much of a good thing: Curvilinear relationships between personality traits and job performance. J Appl Psychol 2011;96:113-33.
41. Griffin B, Wilson IG. Faking good: Self-enhancement in medical school applicants. Med Educ 2012;46:485-90.
42. Motowidlo SJ, Dunnette MD, Carter GW. An alternative selection procedure: The low-fidelity simulation. J Appl Psychol 1990;75:640-47.
43. Patterson F, Ashworth V, Kerrin M, O'Neill P. Situational judgement tests represent a measurement method and can be designed to minimise coaching effects. Med Educ 2013;47:220-1.

44. Lievens F. Adjusting medical school admission: Assessing interpersonal skills using situational judgement tests. Med Educ 2013;47182-9.
45. Schmitt N, Keeney J, Oswald FL, Pleskac TJ, Billington AQ, Sinha R *et al.* Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. J Appl Psychol 2009;94:1479-97.
46. McDaniel MA, Hartman NS, Whetzel DL, Grubb WL. Situational judgement tests, response instructions, and validity: A meta-analysis. Pers Psychol 2007;60:63-91.
47. Sharma N, Nagle YK. Development of pictorial situational judgement test of affect. Psychology development of pictorial situational judgement test of affect. Psychology 2015;6:400-408.
48. Liao S-C, Hsiue T-R, Lin C-H, Huang A-M. Multiple mini-interviews combined with group interviews in medical student selection. Med Educ 2014;48:1104
49. Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR. Predictive validity of the multiple mini-interview for selecting medical trainees. Med Educ 2009;43:767-75.
50. Reiter HI, Eva KW, Rosenfeld J, Norman GR. Multiple mini-interviews predict clerkship and licensing examination performance. Med Educ 2007;41:378-84.
51. Harris S, Owen C. Discerning quality: Using the multiple mini-interview in student selection for the Australian National University Medical School. Med Educ 2007;41: 234-41.
52. Oluwasanjo A, Wasser T, Alweis R. Correlation between MMI performance and OSCE performance – a pilot study. J Community Hosp Intern Med Perspect 2015;5:27808.
53. Bergeron L, Saint-Onge C, Martel S, Hanna D. Évaluation édumétrique d'un dispositif d'entrevues structurées multiples pour la sélection de candidats dans un programme postgradué de dermatologie. Pédagogie Médicale 2011;12: 17-27.
54. Knorr M, Hissbach J. Multiple mini-interviews: same concept, different approaches. Med Educ 2014;48:1157-75.
55. Uijtdehaage S, Doyle L, Parker N. Enhancing the reliability of the multiple mini-interview for selecting prospective health care leaders. Acad Med 2011;86:1032-9.
56. Roberts C, Clark T, Burgess A, Frommer M, Grant M, Mossman K. The validity of a behavioural multiple-mini-interview within an assessment centre for selection into specialty training. BMC Med Educ 2014;14:169.
57. Wright SR, Bradley PM. Has the UK clinical aptitude test improved medical student selection? Med Educ 2010;44: 1069-76.
58. Yates J, James D. The UK clinical aptitude test and clinical course performance at Nottingham: A prospective cohort study. BMC Med Educ 2013;13(1):32.
59. Lynch B, Mackenzie R, Dowell J, Cleland J, Prescott G. Does the UKCAT predict Year 1 performance in medical school? Med Educ 2009;43:1203-9.
60. McManus IC, Dewberry C, Nicholson S, Dowell JS, McManus I, Richards P, Reiter H. The UKCAT-12 study: Educational attainment, aptitude test performance, demographic and socio-economic contextual factors as predictors of first year outcome in a cross-sectional collaborative study of 12 UK medical schools. BMC Med 2013;11:244.
61. Husbands A, Mathieson A, Dowell J, Cleland J, MacKenzie R, McManus I, Norman G. Predictive validity of the UK clinical aptitude test in the final years of medical school: A prospective cohort study. BMC Med Educ 2014;14:88.
62. Denzin N, Lincoln Y. Handbook of qualitative research (2nd ed.). London, Thousand Oaks (CA) and New Delhi: Sage, 2000.
63. Morrison J. Selecting for medical education. Med Educ 2016;50:3-5.
64. Allen J, Robbins SB, Sawyer R. Can measuring psychosocial factors promote college success? Appl Meas Educ 2009;23: 1-22.
65. De Clercq M, Galand B, Frenay M. Transition from high school to university: A person-centered approach to academic achievement. Eur J Psychol Educ 2017;32:39-59.

Citation de l'article : Detroz P., Loye N., Sélection des futurs médecins : sur quelles bases empiriques ? Pédagogie Médicale 2018:19;37-50