La Revue de médecine interne 39 (2018) 4–9

Disponible en ligne sur ScienceDirect www.sciencedirect.com

Elsevier Masson France EM|consulte www.em-consulte.com

Article original

Validité et reproductibilité de deux grilles d'observation des compétences cliniques des internes en DES de médecine interne

Validity and reproducibility of two direct observation assessment forms for evaluation of internal medicine residents' clinical skills

P. Pottier[a,*,d], F. Cohen Aubart[b], O. Steichen[c], M. Desprets[a], M. Pha[b], A. Espitia[a], S. Georgin-Lavialle[c], A. Morel[d], J.B. Hardouin[d]

[a] *Service de médecine interne, Hôtel-Dieu, CHU de Nantes, place Alexis-Ricordeau, 44093 Nantes, France*
[b] *Service de médecine interne 2, hôpital de la Pitié-Salpêtrière, université Paris-VI – Pierre-et-Marie-Curie, Assistance publique–Hôpitaux de Paris, 75013 Paris, France*
[c] *Service de médecine interne, hôpital Tenon, UPMC université Paris 06, Sorbonne universités, AP–HP, 75970 Paris, France*
[d] *SPHERE U1246, Inserm, université de Nantes–université de Tours, 44000 Nantes, France*

INFO ARTICLE

Historique de l'article :
Disponible sur Internet le 20 novembre 2017

Mots clés :
Compétence
Évaluation
Médecine interne
Étude psychométrique

RÉSUMÉ

Contexte. – La réforme du troisième cycle des études médicales se veut centrée sur l'acquisition de compétences. En médecine interne, une évaluation des connaissances théoriques par *e-learning* et des connaissances pratiques par un e-carnet de stage sont en cours de création. Parallèlement, une réflexion se met en place sur des grilles d'évaluation des compétences cliniques. Dans ce cadre, nous avons voulu évaluer la reproductibilité et la validité de deux grilles d'évaluation par observation directe.
Méthode. – Une étude multicentrique et prospective a été menée de novembre 2015 à octobre 2016 pour évaluer les traductions françaises du MINI-Clinical Examination Exercice (MINI-CEX) et du Standardized Patient Satisfaction Questionnaire (SPSQ). La passation était réalisée 2 fois au cours du semestre, par le même binôme d'observateur pour chaque interne.
Résultats. – Dix-neuf internes ont été inclus. La reproductibilité inter-juge était satisfaisante pour le MINI-CEX : coefficients de corrélation intraclasse (CCI) entre 0,4 et 0,8 et moyenne pour le SPSQ : CCI entre 0,2 et 0,7 avec une bonne cohérence interne pour les deux grilles (Cronbach entre 0,92 et 0,94). Des différences significatives entre les distributions des scores donnés par les deux observateurs ont été constatées ainsi qu'une variabilité importante des scores selon le centre.
Conclusion. – La valeur absolue des scores, trop variable, ne doit pas être prise en compte dans l'évaluation mais peut avoir un intérêt pour le suivi de la progression des compétences. Ces grilles pourraient servir de support pour le débriefing en s'appuyant sur les tendances générales données par les scores.

Keywords:
Skills
Assessment
Internal medicine
Psychometric study

ABSTRACT

Introduction. – The revision of the French medical studies' third cycle ought to be competency-based. In internal medicine, theoretical and practical knowledge will be assessed online with e-learning and e-portfolio. In parallel, a reflection about clinical skills assessment forms is currently ongoing. In this context, our aim was to assess the reproducibility and validity of two assessment forms based on direct clinical observation.
Method. – A prospective and multicentric study has been conducted from November 2015 to October 2016 aiming at evaluating the French translations of the MINI-Clinical Examination Exercice (MINI-CEX) and the Standardized Patient Satisfaction Questionnaire (SPSQ). Included residents have been assessed 2 times over a period of 6 months by the same binoma of judges.

* Auteur correspondant.
Adresse e-mail : pierre.pottier@univ-nantes.fr (P. Pottier).

https://doi.org/10.1016/j.revmed.2017.10.424
0248-8663/

Results. – Nineteen residents have been included. The inter-judge reproducibility was satisfactory for the MINI-CEX: intraclass coefficients (ICC) between 0.4 and 0.8 and moderate for the SPSQ: ICC between 0.2 and 0.7 with a good internal coherence for both questionnaires (Cronbach between 0.92 and 0.94). Significant differences between the distributions of the scores given by the judges and a significant inter-center variability have been found.
Conclusion. – If the absolute value of the scores should not be taken into account in the evaluation process given its high variability, it could be of interest for the follow-up of the progression in the competencies. These forms could support the residents' debriefing based on the general trends given by the scores.

1. Contexte

La réforme du troisième cycle des études médicales françaises (TCEM) qui sera mise en œuvre à partir de novembre 2017 se veut centrée sur l'acquisition de compétences. Une compétence est définie comme un « savoir agir en situation ». Elle mobilise, pour un contexte donné, de nombreuses ressources externes et de multiples savoirs d'ordre théoriques (savoirs déclaratifs) et pratiques (savoir-faire et savoir-être) [1]. Une approche par compétence implique de définir à la fois les compétences et les outils pour les évaluer [2]. En effet, il est indispensable que l'évaluation des apprentissages soit alignée sur les principes pédagogiques invoqués : une évaluation axée sur les connaissances théoriques orientera inévitablement les comportements d'apprentissage vers l'acquisition de connaissances théoriques qui ne sont qu'une des composantes de la compétence et aura possiblement une influence négative sur la motivation des étudiants. Par définition, les compétences ne s'observent que dans l'action. Plus précisément, elles s'infèrent à partir des performances réalisées dans l'action. Elles ne peuvent donc s'apprécier qu'à partir de mises en situation. Cela conduit certains collèges d'enseignants à définir des familles de situations au cours desquelles un certain nombre de compétences données sont susceptibles d'être observées [3].

Parallèlement, il est important d'évaluer la fiabilité et la reproductibilité des outils choisis pour évaluer les compétences. De nombreux outils cliniques ont été développés pour observer et évaluer les compétences en milieu professionnel. Certains sont destinés à évaluer l'examen clinique, soit en supervision directe en présence de l'étudiant, en situation réelle [4] ou simulées [5] comme les examens cliniques objectifs structurés [6], soit en supervision indirecte sur la base de portfolio [7–9] ou de récits de situations authentiques [10]. Ces évaluations peuvent être suivies d'une rétroaction (ou feed-back) par un enseignant, par un pair ou par les patients [11].

La présence physique d'un observateur peut gêner l'étudiant et biaiser l'évaluation. Des évaluations indirectes d'après des enregistrements vidéos ont été validées en milieu simulé [12] et en milieu réel [13]. Les évaluations peuvent être réalisées par les patients simulés eux-mêmes ou par des enseignants. La plupart de ces dispositifs d'évaluations des compétences sont coûteux et consommateurs de temps pour les enseignants.

Des outils d'évaluation des compétences dites relationnelles, visant à appréhender la qualité de la relation médecin–patient instaurée au cours de l'examen clinique, ont également été développés. De la même façon que pour les outils d'évaluation de l'examen clinique, ces grilles ont été utilisées en observation directe ou indirecte à partir d'enregistrements audio ou vidéo et remplies de façon reproductible par des patients réels, simulés ou par des enseignants [14].

En France, à notre connaissance, aucun outil n'a été validé pour évaluer les compétences des internes en médecine interne en situation professionnelle.

Le collège des enseignants de médecine interne a mené en 2015 une réflexion sur la nature des compétences à acquérir au cours du diplôme d'enseignement spécialisé (DES) de médecine interne ce qui a conduit à la définition de 287 compétences par méthode Delphi. Une évaluation des connaissances théoriques par *e-learning* et un e-carnet de stage sont actuellement en cours de création. Parallèlement, il a été mis en place une réflexion sur des grilles d'évaluation des compétences cliniques des étudiants inscrits en DES de médecine interne. Nous avons donc mené une étude prospective pour évaluer la pertinence et la reproductibilité de deux grilles d'évaluation par observation directe des internes en situation professionnelle.

2. Méthodes

2.1. Objectifs et critères d'inclusion

L'objectif de l'étude était de déterminer la faisabilité, la reproductibilité et la validité de deux grilles d'observation des compétences cliniques « au lit du malade » chez des étudiants de TCEM inscrits en DES de médecine interne en France.

Il s'agissait d'une étude multicentrique et prospective menée dans les services de médecine interne, de maladies infectieuses et de médecine vasculaire du CHU de Nantes et les services de médecine interne des hôpitaux de Tenon et de la Pitié-Salpêtrière de novembre 2015 à octobre 2016. Les critères d'inclusion des étudiants étaient :

- l'inscription en DES de médecine interne ;
- la signature d'un consentement éclairé.

Le critère d'exclusion était le refus de l'interne de participer.

Il s'agissait d'une étude pilote expérimentale réalisée dans le cadre de la réforme du 3e cycle des études médicales. Dans ce contexte, aucune limite d'inclusion n'a été définie et il n'a pas été réalisé de calcul d'un nombre de sujets nécessaires. Les objectifs de faisabilité et de reproductibilité ont motivé le choix d'une étude multicentrique.

Il était précisé à l'étudiant que les résultats n'étaient pas pris en compte dans l'évaluation du stage et que l'étude était destinée uniquement à évaluer la pertinence des grilles et non les compétences des étudiants, ce d'autant que la présence de deux observateurs était susceptible de diminuer les performances de l'interne.

Une information était délivrée au patient sur les objectifs de l'étude et son consentement signé était recueilli.

Cette étude a obtenu l'avis favorable du Groupe nantais d'éthique dans le domaine de la santé (GNEDS), référence RC15_0452.

Les deux grilles étudiées étaient les suivantes (Annexe, Figs. 1 et 2) :

- la traduction française (non validée) du MINI-CEX, grille américaine utilisée pour évaluer la conduite de l'examen clinique par les résidents (internes), notamment dans les services de médecine interne [4] ;
- la traduction française (validée) du SPSQ, grille américaine utilisée pour évaluer les compétences relationnelles dans un contexte

de consultations simulées par des comédiens professionnels, ayant démontré de bons critères psychométriques avec des étudiants en médecine français et une bonne reproductibilité entre observateurs externes et comédiens [15].

Toutes les grilles étaient anonymisées pour la collection des résultats.

3. Déroulement de l'étude

Lors de l'examen clinique d'entrée d'un patient, les deux grilles étaient remplies par un binôme d'observateurs (professeur des universités–praticien hospitalier, maître de conférences des universités–praticien hospitalier ou assistant-chef de clinique) présents dans la chambre au moment de l'examen clinique de l'interne. Le SPSQ était également rempli par le patient à la fin de l'examen clinique.

Une rétroaction a été faite par au moins l'un des deux observateurs à la fin de l'observation de l'interne à partir des éléments des grilles sans dévoiler ceux-ci.

La passation était réalisée deux fois au cours du semestre, la première dans les deux premiers mois et la seconde dans les deux derniers mois, par le même binôme d'observateurs pour chaque étudiant.

Outre les scores obtenus aux grilles, les éléments suivants étaient recueillis : le sexe de l'étudiant, l'ancienneté d'inscription dans le DES de médecine interne, la complexité du problème posé par le patient, le caractère nouveau ou anciennement suivi sur le site du patient et le temps de passation de l'épreuve (rétroaction non comprise).

4. Analyse statistique

4.1. Fiabilité des grilles

La reproductibilité inter-observateur a été évaluée par le coefficient de corrélation de intraclasse (CCI) pour le score total de chaque grille, en considérant que les observateurs ont été répartis au hasard (car les binômes d'observateurs n'ont pas été constamment les mêmes dans les centres 2 et 3) et par le coefficient de corrélation de Pearson pour chaque item de chaque grille.

La cohérence interne des questionnaires a été évaluée en estimant le coefficient Alpha de Cronbach.

Les comparaisons des distributions des variables quantitatives ont été réalisées par le test des rangs signés de Wilcoxon, compte tenu des petits effectifs.

4.2. Validité des grilles

La validité des grilles a été appréciée en étudiant l'effet en analyse univariée puis multivariée de plusieurs paramètres sur la variation des scores MINI-CEX et SPSQ :

- l'effet du service de recrutement (appelé effet centre). Pour des internes de même niveau mais provenant de centres différents, les scores de ces grilles ne devraient pas varier ;
- l'effet du moment de l'évaluation dans le stage (appelé début/fin de stage). Pour une utilisation répétée dans le même stage, ces grilles devraient dépister une amélioration des compétences dans le temps, quel que soit le centre et quel que soit le niveau de l'interne ;
- l'effet de l'ancienneté de l'interne (appelé année du DES). Les scores devraient être meilleurs pour les internes plus expérimentés pour des niveaux de difficulté identique, quel que soit le centre, quel que soit le moment de l'observation ;
- l'effet de la complexité de la situation.
- À niveau de DES identique, les scores obtenus lors des situations complexes devraient être moins élevés que lors des situations simples, quel que soit le centre, quel que soit le moment de l'observation.
- De plus, la validité des grilles a été testée par l'étude de la validité convergente positive :

- entre les items relationnels du MINI-CEX et le SPSQ remplis par les patients (validité concourante négative). Ces deux questionnaires mesurant des construits proches (compétences cliniques et compétences relationnelles), on s'attend à ce qu'ils soient corrélés entre eux ;
- pour le SPSQ, par la corrélation entre le SPSQ rempli par les observateurs et le SPSQ rempli par les patients (coefficients de corrélation intraclasse). On s'attend à une corrélation positive forte entre ces deux scores.

5. Résultats

5.1. Population de l'étude et données générales

Sur la période de l'étude, 20 internes inscrits en DES de médecine interne étaient présents sur les sites participant à l'étude. Dix-neuf internes ont été inclus (9 au CHU de Nantes [centre 1], 7 à la Pitié-Salpétrière [centre 3] et 3 à Tenon [centre 2]). Un interne n'a pas pu être inclus en raison d'une impossibilité de passation, aucun n'a refusé de participer.

Les internes inclus se répartissaient de la façon suivante en fonction de leur niveau d'expérience : 1re année : 3 internes, 2e année : 3 internes, 3e année : 4 internes, 4e année : 6 internes, 5e année : 3 internes.

Le temps moyen pour réaliser l'évaluation (rétroaction non comprise) a été de 25,7 minutes ± 7,1 (min : 15, max : 45).

Parmi les 38 patients pris en charge, 11 était des patients déjà connus, 27 des nouveaux patients. Les objectifs de l'entrevue avec l'interne étaient les suivants : diagnostic ($n = 26$), suivi ($n = 16$), traitement ($n = 10$), information du patient ($n = 7$).

La répartition des niveaux de complexité des problèmes médicaux pris en charge par les internes était la suivante : 6 problèmes de niveau faible, 27 de niveau modéré, 5 de niveau élevé. La satisfaction globale des observateurs à l'usage du MINI-CEX (item 8 du questionnaire MINI-CEX, Fig. 2) était de 6,7 ± 1,5 pour sur une valeur maximale de 9 ($n = 76$).

Les items des deux grilles sont apparus pertinents pour les observateurs avec un taux de coche « non applicable » inférieur à 13 % pour tous les items (Tableau 3) sauf deux : l'item no 5 du MINI-CEX (taux de non applicable : 30 %) et l'item no 6 du SPSQ (taux de « non applicable : 38 %).

5.2. Reproductibilité

5.2.1. Reproductibilité inter-juge

Concernant la reproductibilité inter-observateur, les CCI étaient compris entre 0,39 et 0,81 pour le MINI-CEX et entre 0,22 et 0,68 pour le SPSQ. Concernant le MINI-CEX, les distributions des scores donnés par les deux observateurs différaient significativement entre les centres 1 et 3. Pour le SPSQ, les distributions des scores ne différaient pas significativement (Tableau 1).

Concernant la reproductibilité entre les observateurs et le patient (Tableau 2), les comparaisons entre les score de SPSQ donnés par les observateurs et ceux donnés par le patient font apparaître des discordances selon les centres :

- pour le centre 1, les distributions étaient significativement différentes alors que les CCI étaient faibles et négatifs (scores inversement corrélés) ;

Tableau 1
Comparaisons entre les deux observateurs par centre.

Grille	*n*	Moyenne	Écart-type	*p*[a]	CCI[b]
Centre 1					
CEX_observ_1	18	42,3	8,4	0,002	0,87
CEX_observ_2	18	38,1	9,5		
SPSQ_observ_1	18	22,4	6	0,08	0,72
SPSQ_observ_2	18	24,3	5,8		
Centre 2					
CEX_observ_1	6	48,2	6,7	0,13	0,86
CEX_observ_2	6	45,5	7,2		
SPSQ_observ_1	6	25,7	4,7	0,75	0,5
SPSQ_observ_2	6	25,5	5,1		
Centre 3					
CEX_observ_1	14	47,6	7,5	0,03	0,52
CEX_observ_2	14	51,7	4,7		
SPSQ_observ_1	14	38,1	6,2	0,44	0,22
SPSQ_observ_2	14	39,7	3,5		

CEX : Mini Clinical Exercise ; SPSQ : Standardized Patient Satisfaction Questionnaire ; observ : observateur.
[a] Test de Wilcoxon.
[b] Coefficient de corrélation intraclasse.

Tableau 2
Comparaisons entre le score SPSQ patient et les scores SPSQ observateurs.

Grille	*n*	Moyenne	Écart-type	*p*[a]	CCI[b]
Centre 1					
SPSQ_patient	18	32,4	7,8		
SPSQ_observ_1	18	22,4	6	0,005	−0,24
SPSQ_observ_2	18	24,3	5,8	0,01	−0,08
Centre 2					
SPSQ_patient	6	27,7	6,1		
SPSQ_observ_1	6	25,7	4,7	0,59	0,44
SPSQ_observ_2	6	25,5	5,1	0,29	0,55
Centre 3					
SPSQ_patient	14	40,5	8,6		
SPSQ_observ_1	14	38,1	6,2	0,37	0,19
SPSQ_observ_2	14	39,7	3,5	0,77	−0,03

SPSQ : Standardized Patient Satisfaction Questionnaire ; observ : observateur.
[a] Test de Wilcoxon.
[b] Coefficient de corrélation intraclasse.

- pour le centre 2, les distributions n'étaient pas significativement différentes alors que les CCI étaient positifs et relativement élevés ;
- pour le centre 3, les distributions n'étaient pas significativement différentes alors que les CCI étaient discordants (l'un positif, l'autre négatif) et faibles.

L'analyse des corrélations inter-observateur par item (Tableau 3) montrait des items mieux corrélés pour les MINI-CEX (*R* entre 0,59 et 0,75) que pour le SPSQ (de 0,43 à 0,6). Il y avait plus de désaccord entre les deux observateurs pour l'item 5 de la grille MINI-CEX (compétences pédagogiques) et pour l'item 6 de la grille SPSQ (décision partagée) avec des coefficients de Pearson respectivement de 0,25 et 0,36.

5.2.2. *Cohérence interne des grilles*

Les coefficients de Cronbach calculés pour chaque observateur étaient de 0,94 et 0,93 pour le MINI-CEX et 0,93 et 0,92 pour la grille SPSQ.

5.3. *Validité*

5.3.1. *Analyse des variances*

L'analyse univariée montrait que tous centres confondus, les observateurs donnaient des scores comparables ($F = 0{,}39$; $p = 0{,}53$ et $F = 0{,}52$; $p = 0{,}47$ pour le MINI-CEX et le SPSQ, respectivement) (Tableau 4).

Tableau 3
Coefficient de corrélation inter-juge par item.

Items	*R*	% NA
CEX		
Habileté à conduire un entretien médical	0,63	0
Habileté à conduire un examen physique	0,59	0
Qualités humaines/professionnalisme	0,61	1
Raisonnement clinique	0,75	13
Compétences pédagogiques	**0,25**	38
Organisation/efficacité	0,71	0
Compétence clinique globale	0,73	0
SPSQ		
L'interne a donné tous les éléments à son patient. Il a été avenant et franc, et ne lui a pas caché d'éléments le concernant	0,6	0
Il a accueilli son patient de manière chaleureuse. Il l'a appelé par une dénomination qui lui convenait. Il a toujours été courtois, jamais désagréable ou grossier	0,6	0
Il s'est mis à son niveau. Il ne s'est pas adressé à lui comme s'il était supérieur ou ne l'a pas traité comme un enfant	0,6	0
Il a laissé son patient raconter son histoire. Il a été à son écoute de manière attentive et lui a posé des questions attentionnées. Il ne l'a pas systématiquement interrompu pendant qu'il parlait	0,49	1
Il a montré à son patient de l'intérêt vis à vis de sa personne et n'a pas agi comme s'il l'ennuyait. Il n'a pas ignoré ce qu'il avait à lui dire	0,43	0
Il a discuté avec son patient de différentes options, lui a demandé son avis, lui a offert la possibilité de faire des choix et laissé participer à la décision	**0,36**	30
Il a encouragé son patient à poser des questions et y a répondu clairement. Il ne les a pas évité ni lui a fait de leçon	0,6	5
Il a expliqué à son patient ce qu'il avait besoin de savoir concernant son état de santé	0,57	8
En utilisant des mots que le patient pouvait comprendre, il lui a expliqué la ou les causes de son état de santé et les raisons qui justifient ses traitements. Il lui a expliqué les termes médicaux techniques dans un langage clair	0,49	1
Il a cherché à comprendre les sentiments du patient au sujet de son état de santé. Il a reconnu l'impact de son état de santé	0,48	13

CEX : Mini Clinical Exercise ; SPSQ : Standardized Patient Satisfaction Questionnaire ; NA : non applicable.
En gras, les items ayant un coefficient de corrélation inter-juge inférieur à 0,4.

Tableau 4
Analyses de variances uni- et multivariées.

Variables	CEX		SPSQ	
	F	*p*	*F*	*p*
Univariées				
Centre	8,87	0,0004	69,1	< 0,0001
Sexe	0,39	0,53	1,22	0,27
Année du DES	3,6	0,001	8,2	< 0,0001
Début/fin stage	1,18	0,28	2,01	0,16
Complexité	1,48	0,23	1,47	0,23
Observateur	0,39	0,53	0,52	0,47
Multivariées				
Centre	4,4	0,02	42,5	< 0,0001
Année du DES	2,3	0,08	3,3	0,04
Interaction	3,0	0,06	0,05	0,95
Début/fin stage	2,3	0,13	6,32	0,01

CEX : Mini Clinical Exercise ; SPSQ : Standardized Patient Satisfaction Questionnaire.

On note en revanche un effet centre et un effet de l'ancienneté significatifs pour les deux grilles.

Le sexe de l'interne, la complexité de la situation clinique et le moment de l'observation (début ou fin de stage) n'avaient aucune part dans la variance des scores des deux grilles.

L'analyse multivariée confirmait l'effet du centre pour les deux scores et de l'ancienneté pour le score SPSQ mais pas pour le MINI-CEX.

Pour le SPSQ, on notait également un effet du moment de l'observation (début vs fin de stage).

5.3.2. Validité convergente positive

Le coefficient de corrélation de Pearson entre le MINI-CEX et le SPSQ, tous centres confondus et les deux observateurs confondus était de 0,63.

Le MINI-CEX n'était pas corrélé au SPSQ patient ($R = -0{,}15$ et $R = 0{,}04$ pour les deux observateurs).

6. Discussion

Cette étude préliminaire tricentrique avait pour objectif d'évaluer la reproductibilité et la validité de deux outils d'évaluation des compétences des étudiants en TCEM traduits de l'anglais, le MINI-CEX et le SPSQ.

Pour un centre donné, la reproductibilité inter-juge semble satisfaisante pour le MINI-CEX (CCI entre 0,5 et 0,9) et moyenne pour le SPSQ (CCI entre 0,2 et 0,7). Dans cette étude préliminaire, les observateurs n'avaient pas eu de formation sur les items des grilles. Les items de la grille SPSQ sont plus riches en concepts, ce qui nécessite probablement un temps de formation pour les clarifier et s'assurer que tous sont compris de la même manière par les observateurs. Les corrélations inter-juges par items (Tableau 3) étaient satisfaisantes pour le MINI-CEX (coefficients variant entre 0,6 et 0,7) à l'exception de l'item « compétences pédagogiques » pour lequel les évaluations entre les deux observateurs semblaient peu concordantes (coefficient de corrélation à 0,25) et qui a recueilli le pourcentage de « non applicable » le plus élevé (38 %). Cet item peut sembler flou et nécessite une précision qui n'a pas été donnée aux observateurs. Il fallait comprendre la compétence de l'étudiant à expliquer au patient son raisonnement et à justifier ses décisions de prescription d'examens complémentaires ou de traitement. Les observateurs étaient moins concordants sur les items du SPSQ dont les coefficients de corrélation varient entre 0,4 et 0,6. L'item évaluant la décision partagée (6e item) était le moins concordant et celui qui a recueilli le pourcentage de « non applicable » le plus élevé (30 %).

Pour le MINI-CEX, malgré des CCI supérieurs à 0,5, il a été souvent constaté des différences significatives entre les distributions des scores donnés par les deux observateurs. Ces résultats n'incitent pas à utiliser le score brut à un temps *t* comme un marqueur quantitatif fiable de la compétence.

Les deux questionnaires ont démontré une bonne cohérence interne, ce qui prouve que les items de chaque questionnaire mesurent bien chacun un même construit théorique à savoir la compétence à mener un examen clinique et le savoir être lors de l'entrevue d'entrée du patient en milieu hospitalier. Les valeurs des coefficients alpha de Cronbach étaient supérieures à 0,9 ce qui est cohérent avec les données de la littérature [4,14]. En outre, il est apparu que ces deux construits sont liés lorsqu'ils sont évalués par un observateur externe : plus un interne apparaît compétent dans la conduite d'un examen clinique, plus il semble démontrer de bonnes capacités relationnelles (coefficient de corrélation à 0,63). Ce lien n'est plus vrai si les compétences relationnelles sont évaluées par le patient (coefficients de corrélation entre le MINI-CEX de chaque observateur et le SPSQ patient à −0,15 et 0,04). Les discordances observées entre les scores de SPSQ donnés par le patient et ceux donnés par les observateurs externes suggèrent que le patient semble avoir une perception différente des compétences relationnelles de l'interne. L'absence de reproductibilité des évaluations des compétences relationnelles entre observateur externe et patient est bien connue dans la littérature [16,17], l'évaluation d'un observateur ayant tendance à être plus sévère que celle du patient, comme si la relation médecin–patient était mieux perçue de l'intérieur de la relation que de l'extérieur. Cela ne remet pas en cause l'utilisation de ces grilles par l'un ou l'autre à partir du moment où l'enseignant clinicien en est informé et qu'il en tient compte pour la rétroaction qu'il proposera lors du débriefing.

Les deux questionnaires ont démontré en analyse univariée leur capacité à déceler des différences en fonction de l'ancienneté de l'interne. En multivariée, seul le SPSQ a varié significativement et de façon indépendante des autres variables étudiées en fonction de l'ancienneté et entre le début et la fin du stage.

Les résultats montrent une variabilité importante des scores d'évaluation selon le centre, indépendamment de l'ancienneté d'inscription de l'étudiant en DES de médecine interne. En effet, en analyse multivariée, comprenant dans le modèle l'ancienneté de l'étudiant, l'effet du centre est resté significatif tant pour le MINI-CEX que pour le SPSQ. L'utilisation des scores pour comparer des étudiants évalués dans des sites différents ne semble donc pas pertinente.

L'ensemble de ces résultats suggèrent que la valeur absolue du score n'a que peu d'intérêt et ne doit pas être prise en compte dans l'évaluation compte tenu de sa variabilité inter-centre et des différences significatives de distributions entre les observateurs attestées par les tests de Wilcoxon. En revanche, ces grilles pourraient servir de support à un débriefing visant à cibler les points à améliorer, en s'appuyant sur l'évaluation de tendances générales données par les scores (bon/moyen/à améliorer).

Dans la perspective d'une évaluation formative (tout au long de la formation), nos résultats montrent que les scores du MINI-CEX et du SPSQ peuvent avoir un intérêt pour le suivi de la progression, ce qui est parfaitement compatible avec la notion de développement continu des compétences tout au long de la vie professionnelle [1].

Pour améliorer la reproductibilité des questionnaires, il semble nécessaire de modifier le libellé de l'item 5 du MINI-CEX dont la traduction de *counselling skills* par « compétences pédagogiques » peut avoir été mal comprise. Le libellé « compétences en éducation thérapeutique » pourrait se révéler plus fidèle et mieux interprété par les observateurs. L'item 6 du SPSQ, plus adapté au contexte de la consultation ambulatoire pourrait, quant à lui, être supprimé. La reproductibilité des grilles devrait vraisemblablement être améliorée par la planification d'un temps de formation des observateurs ou par la rédaction d'une fiche explicative sur les items des questionnaires.

Pour augmenter la validité de l'évaluation, notamment celle du raisonnement clinique, il serait utile, en complément de l'observation de l'examen clinique, de demander à l'interne de faire la synthèse de la situation après l'entrevue. La qualité du raisonnement clinique semble en effet difficile à juger par une simple observation de l'examen clinique.

La limite principale de l'étude est le petit nombre d'internes inclus comparé à l'étude de validation du MINI-CEX de Norcini et al. [4] réalisée sur 421 internes et plus de 1200 rencontres interne-patient aux États-Unis. Notre travail s'inscrit dans un contexte institutionnel pédagogique fort différent. En effet, l'évaluation des compétences des internes sur la base d'une observation directe à partir d'une grille étant encore très peu développée en France, nous avons délibérément choisi de réaliser une étude pilote multicentrique expérimentale sur un nombre limité d'internes. Ce choix doit nous inciter à rester prudent quant à la validité des analyses statistiques. Notre étude cherchait en premier lieu à évaluer la faisabilité et la reproductibilité des grilles choisies entre deux observateurs. Concernant leur validité, c'est-à-dire, leur capacité à réellement mesurer ce qu'elles sont censées mesurer, les données de la littérature montrent qu'elle varie en fonction du nombre d'évaluations réalisées. Norcini et al. ont ainsi montré qu'il fallait au moins 10 à 12 évaluations avec le MINI-CEX pour une précision satisfaisante. D'autres études devront être menées pour montrer leur validité dans le cadre d'évaluations sanctionnantes ou à forts

enjeux, car ce n'était pas l'objectif de ce travail. Une étude multicentrique de plus grande envergure est en cours pour juger de la faisabilité de ces deux outils d'évaluation des compétences cliniques reposant sur l'observation directe des internes en activité professionnelle. Pour mieux documenter l'acquisition de compétences, d'autres outils d'évaluation in situ des compétences (ou *work-based assessment* [18]) seront évalués dans cette étude tels qu'un questionnaire d'évaluation d'un geste technique dérivée de la grille Direct Observation of Procedural Skills (DOPS) [19] et un questionnaire d'évaluation multi-source (Multisource Feedback) [17]. La diversification des outils permettra de disposer d'une évaluation composite documentant les différentes dimensions de la compétence [16].

Déclaration de liens d'intérêts

Les auteurs déclarent ne pas avoir de liens d'intérêts.

Annexe 1. Matériel complémentaire

Le matériel complémentaire (Fig. S1, S2) accompagnant la version en ligne de cet article est disponible sur https://doi.org/10.1016/j.revmed.2017.10.424.

Références

[1] Tardif J. Chapitre 1. In: L'évaluation des compétences. Documenter le parcours de développement. Canada: Chenelière Éducation; 2006. p. 17–51.
[2] van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programs. Med Educ 2005;39:309–17.
[3] Attali C, Huez JF, Valette T. Lehr-Drylewicz. Les grandes familles de situations cliniques. Exercer 2013;108:165.
[4] Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. Ann Intern Med 2003;138:476–81.
[5] Swanson DB, van der Vleuten CP. Assessment of clinical skills with standardized patients: state of the art revisited. Teach Learn Med 2013;25:S17–25.
[6] Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979;13:41–54.
[7] Gadbury-Amyot CC, McCracken MS, Woldt JL, Brennan RL. Validity and reliability of portfolio assessment of student competence in two dental school populations: a four-year study. J Dent Educ 2014;78:657–67.
[8] Naccache N, Samson L, Jouquan J. Le portfolio en éducation des sciences de la santé : un outil d'apprentissage, de développement professionnel et d'évaluation. Pedagogie Med 2006;7:110–27.
[9] Moonen-van Loon JMW, Overeem K, Donkers HH, van der Vleuten CP, Driessen EW. Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. Adv Health Sci Educ 2013;18:1087–102.
[10] Le Mauff P, Pottier P, Goronflot L, Barrier J. Évaluation d'un dispositif expérimental d'évaluation certificative des étudiants en fin de troisième cycle de médecine générale. Pedagogie Med 2006;7:142–54.
[11] Donnon T, Al Ansari A, Al Alawi S, Violato C. The reliability, validity, and feasibility of multisource feedback physician assessment: a systematic review. Acad Med 2014;89:511–6.
[12] Swartz MH, Colliver JA, Bardes CL, Charon R, Fried ED, Moroff S. Global ratings of videotaped performance versus global ratings of actions recorded on checklists: a criterion for performance assessment with standardized patients. Acad Med 1999;74:1028–32.
[13] Reinders ME, Blankenstein AH, van Marwijk HW, Knol DK, Ram P, van der Horst HE, et al. Reliability of consultation skills assessments using standardised versus real patients. Med Educ 2011;45:578–84.
[14] Schirmer JM, Mauksch L, Lang F, Marvel MK, Zoppi K, Epstein RM, et al. Assessing communication competence: a review of current tools. Fam Med 2005;37:184–92.
[15] Morel A, Hardouin JB, Pottier P. Validation of the french version of the standardised patient satisfaction questionnaire (SPSQ). In: Abstract Congress of the Association of medical education in europe (AMEE). 2015.
[16] Lie D, Boker J, Bereknyei S, Ahearn S, Fesko C, Lenahan P. Validating measures of third year medical students'use of interpreters by standardized patients and faculty observers. J Gen Intern Med 2007;22(Suppl 2):336–40.
[17] Burt J, Abel G, Elmore N, Newbould J, Davey A, Llanwarne N, et al. Rating communication in GP consultations: the association between ratings made by patients and trained clinical raters. Med Care Res Rev 2016 [pii: 1077558716671217].
[18] Nair BK, Moonen-van Loon JM, Parvathy MS, van der Vleuten CP. Composite reliability of workplace-based assessment for international medical graduates. Med J Aust 2016;205:212–6.
[19] Wilkinson JR, Crossley JG, Wragg A, Mills P, 1 Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. Med Educ 2008;42:364–73.