UNIVERSITÉ DE STRASBOURG

École doctorale Sciences Humaines et Sociales
-
Perspectives européennes

Laboratoire interuniversitaire des sciences de l'éducation et de la communication

Thèse présentée par

Anne-Laure PHILIPPON

Soutenue le 2 mars 2022

Pour obtenir le grade de : Docteur de l'université de Strasbourg

Discipline/Spécialité : Sciences de l'éducation

Evaluer les futurs médecins en situation simulée dans le contexte de l'urgence vitale : développement d'un outil et réflexions sur son apport dans le cadre d'une approche par compétence(s).

Directeur de thèse :	M. Triby Emmanuel	Professeur, Université de Strasbourg
Co-directeur de thèse :	M. Freund Yonathan	Professeur, Sorbonne Université, Paris
Rapporteurs :	M. Ardouin Thierry	Professeur, Université de Rouen
	Mme Charpentier Sandrine	Professeure, Université de Toulouse
Examinateurs :	M. Baron Georges-Louis	Professeur émérite, Université de Paris
	M. Oriot Denis	Professeur, Université de Poitiers
	Mme Pacurar Ecatarina	Professeure, Université de Lille
	M. Pelaccia Thierry	Professeur, Université de Strasbourg

”All assessments are imperfect measures of knowledge, skills, and performance. The critical question in the development and administration of an assessment is “How imperfect is it?”

(Howley, 2004)

A mon époux, Jacques-Marie,
pour ton amour, ta patience, ton soutien sans limites
et ta compréhension de ce qui me meut,
probablement mieux que moi.

A mon fils, Gustave, né pendant cette aventure doctorale,
et à son frère ou à sa sœur, qui s'invite avec joie à la toute fin du parcours.
Je vous souhaite de vivre des expériences professionnelles aussi riches
d'enseignements, d'émotions et d'émerveillements,
dans vos domaines de prédilection.

REMERCIEMENTS

Je souhaite remercier profondément le Pr Triby, qui a accepté de diriger cette thèse et les contraintes qui allaient avec : celles de mon métier, celles d'un positionnement disciplinaire et méthodologique à la croisée des Sciences de l'éducation et de la pédagogie médicale et celles de ma vie privée, qui ont pu parfois retarder quelques échéances. Merci pour votre soutien et pour nos entretiens strasbourgeois qui ont toujours été source de motivation, d'apprentissage, ainsi qu'un réel plaisir aidant à la réalisation de la thèse. Merci de votre disponibilité, jusqu'aux derniers moments, et merci de votre accompagnement rigoureux.

Au Pr Freund, qui dès le premier jour de notre rencontre autour d'un mannequin de simulation et d'une session de simulation improvisée, a su voir en moi une urgentiste et une enseignante, puis une chercheuse en herbe. Grâce à ton intuition et à ta proposition de faire un Master en Sciences de l'Education, j'ai découvert le domaine qui me permettrait d'allier recherche et enseignement. Merci de m'avoir fait découvrir les joies, les difficultés et les exigences de la recherche, et d'avoir partagé avec moi ta passion de la médecine d'urgence, métier que je suis fière d'exercer au sein de notre beau service. Service dans lequel je ne serai probablement jamais venue, sans cette simulation improvisée. Merci au hasard de la vie. Merci, enfin, d'avoir accepté de me suivre, de me soutenir et de m'encadrer dans ce travail de thèse, d'avoir écouté mes doutes et errements (le plus dur !) et d'avoir cru que je pouvais le faire. Et bien sûr, merci pour le Plastic, le Rochelle, Londres et les virées états-uniennes !

Je souhaite ensuite remercier les Pr Ardouin et Charpentier, qui ont accepté d'être les rapporteurs de cette thèse. Merci pour votre « jugement évaluatif » porté par le regard des Sciences de l'Education et par celui de la médecine d'urgence. Merci Sandrine de nous soutenir dans le développement national de la simulation en médecine d'urgence.

Au Pr Baron, qui avez suivi mes débuts en Sciences de l'Education, je suis très honorée et fière de votre participation à ce jury de thèse. Merci de m'avoir appris que la confusion fait toujours partie d'un processus initial de recherche et qu'il faut s'en servir plutôt que de la fuir.

Je souhaite également remercier le Pr Pelaccia, qui contribue au développement de la pédagogie médicale, par ses livres, références en la matière, et par son implication dans les différentes formations et réformes indispensables à la formation des futurs médecins et professionnels de santé. Merci également au rôle que tu as joué dans ce parcours doctoral, en

mettant sur ma route le Pr Triby et en me permettant ainsi de trouver un directeur de recherche qui a su m'encadrer et m'accompagner. Merci également de ton accueil lors de mon année strasbourgeoise et pour les discussions que nous avons pu avoir, m'aidant ainsi à avancer.

Je tiens également à remercier les Pr Pacurar et Oriot, qui ont accepté d'être évaluatrice et évaluateur de cette thèse. Au Pr Oriot, merci d'être le pionnier en France dans la recherche sur l'évaluation par la simulation et d'avoir inspiré mon travail.

***

Ce travail de thèse a été riche en rencontres, découvertes et je souhaite ici remercier tous les enseignants qui ont participé à la recherche, pour leur enthousiasme et leur volonté de participer à un travail décalé de leurs pratiques habituelles. Merci particulièrement à Jennifer Truchot, Margaux Dumont, Aurélien Baud et Carine Zumstein pour leur implication, pour les discussions enrichissantes que nous avons pu avoir et pour leur soutien dans les (quelques) moments de démotivation. Merci également à toute l'équipe de l'UNISIMES, pour leur accueil tout au long de mon année strasbourgeoise, ma bulle d'air et de sociabilité malgré le confinement ! J'espère que ce travail dans votre équipe sera à l'origine d'autres projets. Enfin, grâce à ce travail et cette année à Strasbourg, j'ai pu rencontrer de nouveaux collègues urgentistes, que je souhaite vivement remercier pour leur accueil chaleureux et amical au sein d'une équipe soudée malgré les défis quotidiens. Merci à vous, urgentistes de Hôpitaux Universitaires de Strasbourg et bravo pour votre travail qui vise toujours le « juste soin », dans un environnement parfois hostile.

Aux étudiants et internes en médecine, que j'ai plaisir à accompagner au quotidien, merci de votre implication, de vos remarques qui aident à avancer. Vous êtes ce qui me motive à continuer ces travaux de recherche et d'enseignement et pour cela je ne vous remercierais jamais assez !

Un grand merci au Pr Riou, d'avoir été un chef de service exemplaire, montrant la voie aux plus jeunes d'entre nous, disponible et à l'écoute de ses équipes en tout premier lieu. Merci d'être un doyen tout aussi exemplaire, ambitieux pour sa faculté, soutenant l'enseignement par simulation et les innovations pédagogiques. Merci de tout le travail accompli pour la naissance et la défense de la spécialité de médecine d'urgence, nous vous devons beaucoup.

Au Pr Duguet, pionnier de la simulation dans notre faculté, merci d'avoir été présent et accompagnant dans mon parcours facultaire et pour les différents projets d'enseignements menés à bien grâce à votre soutien. Merci de voir la vie et la médecine différemment, et de le transmettre inlassablement.

Merci enfin à mes collègues des urgences de La Pitié-Salpêtrière, autrement appelées les urgences de « La-Mitié », ce qui illustre bien la représentation que nous nous en faisons. Je ne peux pas tous vous citer, mais le travail que nous accomplissons ensemble tous les jours, dans une certaine adversité, parle pour moi. Un merci spécial au Pr Hausfater qui a permis que ce travail de recherche se fasse, par son soutien et sa compréhension de mes différents besoins. Pour conclure, un merci infini à mes « collègues-devenus-amis », ce qui veut tout dire : Adeline, Julie, Jérôme, Pauline (que serait ce bureau sans toi ?) et Yonathan. Sans vous, la vie quotidienne aurait moins de saveur.

***

A mes chers amis, Edouard et Olivier, Claire et Amandine, Kathleen et Aymeric, Alex et Gleise, Joyce et Mathilde, pour votre oreille attentive, vos paroles attentionnées, vos bons petits (et grands !) plats, votre soutien sportif (!), vos feux de cheminée légendaires et pour avoir toujours su faire une place à mon travail, « quelque peu » envahissant ces derniers temps, et tout cela sans reproches !

Enfin, à ma famille et tout particulièrement à mes parents, à ma sœur et à mon beau-frère, sans oublier mes grands-parents. Merci d'avoir largement contribué à ce que je suis, merci d'être un exemple et merci d'avoir toujours compris, accepté et soutenu mes choix, avec les conséquences qu'ils impliquaient et notamment votre grande aide tant matérielle qu'affectueuse tout au long de mon parcours médical. Je vous sais toujours à mes côtés et je vous en remercie. Que cette thèse, qui est la fin d'une étape, mais pas la dernière (!) soit la marque de ma reconnaissance et de mon amour.

TABLE DES MATIERES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

LISTE DES ABREVIATIONS, DES SIGLES ET DES ACRONYMES

ABMS	American Board of Medical Specialties
ACAT	Acute Care Assessment Tool
ACGME	Accreditation Council for Graduate Medical Education
ANTS	Anaesthetists' non-technical skills
APC	Approche Par Compétences
CBME	Competence-Based Medical Education
CHU	Centre Hospitalo-Universitaire
CMU	Capacité de médecine d'urgence
CNUMU	Collège National des Universitaires de Médecine d'Urgence
DES	Diplôme d'étude spécialisé
DESMU	Diplôme d'études spécialisées en Médecine d'Urgence
DESCMU	Diplôme d'études spécialisées complémentaires de médecine d'urgence
DFASM	Diplôme de Formation Approfondie en Science médicale
DFGSM	Diplôme de Formation Générale en Science Médicale
DOPS	Direct observation of procedural skills
EPA	Entrustable Professional Activities
ECN	Examen National Classant
ECOS	Examens Cliniques Objectifs Structurés
LCME	Liaison Committee on Medical Education
mini-CEX	Mini-clinical evaluation exercise
OSATS	Objective Structured Assessment of Technical Skills
PROFILES	Principal Relevant Objectives and a Framework for Integrative Learning and Education in Switzerland
QCM	Question à choix multiples
QROC	Question à réponse ouverte courte
SBA	Simulation-based assessment
SoFraSimS	Société Française de Simulation en Santé

INTRODUCTION

Le travail doctoral que nous avons mené s'inscrit dans une démarche de recherche débutée il y a quelques années lors d'un master de Sciences de l'Education. Il visait à étudier la place et le rôle de l'évaluation par la simulation au sein de d'un enseignement des urgences vitales destiné à des étudiants de deuxième cycle des études médicales, qui était donc à un stade relativement novice de leur apprentissage. En s'intéressant à leur perception de deux situations d'évaluation par la simulation, il apparaissait que les étudiants reconnaissaient à cette modalité d'évaluation une plus-value, en comparaison aux évaluations facultaires. Ils rapportaient en effet un changement dans leurs pratiques de travail, décrivant une préparation différente des sessions de simulation, ainsi qu'un atout pour leur pratique et leur place en stage hospitalier. De plus, ils remarquaient que l'évaluation par la simulation avait l'intérêt de placer au centre de la formation des connaissances et habiletés plus « basiques et généralistes » que celles habituellement demandées dans les évaluations facultaires, mais qui, selon eux, n'étaient pas adaptées à leur futur métier d'interne. Cependant, ils relevaient également un manque d'équité et d'objectivité des évaluations, ainsi qu'une mise en situation stressante pour eux (Philippon et al., 2021, Annexe 1).

Dans le cadre de ce travail doctoral, qui a suivi les règles de la thèse par articles, et pour questionner ce manque d'équité et d'objectivité relevé par les étudiants, nous avons souhaité nous intéresser à la place et au rôle de l'évaluation par la simulation en décrivant et analysant le développement d'un outil d'évaluation par la simulation, ainsi que la perception qu'en avait les enseignants.

Comme nombre de programmes universitaires et scolaires, l'enseignement et l'apprentissage basés sur l'approche par compétence sont de plus en plus répandu au sein de la formation des professionnels de santé, même si leur formalisation et utilisation est rarement globale (Lemenu & Heinen, 2015; Nguyen & Blais, 2007). Dans le domaine médical, l'apprentissage par

compétence a été formalisé en 1999 aux Etats-Unis avec les recommandations de l'Accreditation Council for Graduate Medical Education (ACGME). Il établit six domaines de compétences que devrait pouvoir maîtriser un étudiant en médecine à la fin de la formation commune à tous les médecins. L'ACGME précise également que ces compétences doivent être maîtrisées et que les facultés doivent pouvoir s'en assurer en les évaluant (Batalden et al., 2002).

Nommées compétences « génériques » en France, elles ont été redéfinies en 2013, au cours de d'une réforme du deuxième cycle des études médicales[1] (Arrêté du 8 avril 2013). Il s'agit des compétences de clinicien, de communicateur, de membre d'une équipe soignante pluriprofessionnelle qui coopère avec elle, et d'acteur de santé publique. L'étudiant en médecine doit également avoir des compétences scientifiques, afin de pouvoir actualiser ses connaissances, mais aussi de formuler une problématique de recherche. Enfin, l'étudiant doit avoir des compétences dans le domaine de l'éthique et de la déontologie et il doit apprendre à s'autoévaluer, à être réflexif pour connaitre ses limites et argumenter ses décisions.

Bien qu'ancré dans le système de formation actuel, l'apprentissage par compétence et notamment l'évaluation de toutes les facettes qui les composent pour former un médecin compétent restent un défi pour les facultés de médecine, en raison des difficultés de sa mise en place et des différents obstacles qu'elle peut rencontrer (Epstein, 2007; Farrell, 2005; Holmboe et al., 2011).

La simulation est un outil mais également une méthode d'enseignement basée sur l'apprentissage en situation et sur l'apprentissage expérientiel (Kolb, 1984). Elle a pour objectif de « remplacer les expériences de la "vraie" vie par des expériences encadrées, évoquant ou reproduisant les aspects fondamentaux du monde réel, d'une manière interactive » (Gaba, 2004).

[1] Les compétences génériques et objectifs de formation sont détaillées dans l'annexe de l'arrêté du 8 avril 2013 et sont accessibles Bulletin officiel du ministère de l'enseignement supérieur et de la recherche en date du 16 mai 2013 :
http://www.enseignementsup-recherche.gouv.fr/pid20536/bulletin-officiel.html?cid_bo=71544&cbo=1

S’appuyant sur différentes théories éducatives et notamment sur des théories socio-cognitives, elle place l’apprenant au centre de sa formation, dans un contexte spécifique, entouré de ses pairs dans un système d’apprentissage collaboratif, pendant la mise en situation mais également après, lors du débriefing (Bleakley, 2006; Pottier, 2013). Grâce à cette mise en situation, elle a la capacité d’aborder des thématiques d’enseignement complexes, variées, en faisant appel à plusieurs capacités des étudiants.

Elle permet également d’appréhender la notion de compétence, en tant que savoir-agir complexe, mais également d’en exposer les différentes composantes. En effet, une compétence est composée de plusieurs éléments distincts, qui sont des capacités ou habiletés qui permettent à un individu d’être qualifié de compétent (Scallon, 2007). La notion de contexte est majeure dans sa définition, et on ne peut exercer une compétence en dehors d’une mise en situation de l’activité (Pastré et al., 2006). Par ailleurs, la compétence fait appel à la notion de mobilisation de ressources internes et externes, et aux habiletés de l’apprenant : cognitives, techniques et psychomotrices (Scallon, 2015). Dans le cadre de l’urgence vitale par exemple, faire le diagnostic de méningite repose sur un savoir théorique, sur une capacité à interroger un patient pour en tirer des informations pertinentes, puis sur celle à faire une ponction lombaire, au sein d’une équipe pluridisciplinaire et après avoir obtenu l’accord du patient. Il faut également avoir la capacité de hiérarchiser les différentes actions à entreprendre afin de les prioriser et de les anticiper. Dans ces situations complexes, on voit l’apport que peut être un enseignement par la simulation, qui tient compte du contexte et de la situation.

Pour l’enseignement des urgences vitales, qu’elles soient basiques (la prise en charge des premières minutes de l’arrêt cardiaque) ou plus complexes (un travail en équipe avec un patient polytraumatisé), elle apparait être un outil adapté et qui a déjà fait ses preuves (Gaba, 2010; Garcia et al., 2015; Kessler et al., 2011; Mundell et al., 2013). Elle réunit ainsi la possibilité d’explorer et de mettre en œuvre les capacités cognitives, techniques et « non-techniques » que

doit mobiliser un étudiant face à une détresse vitale, ce qui n'est pas le cas dans les enseignements « habituels », qu'ils soient facultaires (enseignements dirigés ou cours magistraux) ou même hospitaliers. En effet, la particularité de l'urgence et la réactivité dont doit faire preuve l'étudiant pour initier une telle prise en charge, font qu'ils n'ont pas la possibilité de réaliser des gestes pratiques au cours de leurs stages et qu'ils restent la plupart du temps observateurs, ce qui créé un décalage entre les objectifs du stage et les acquis réalisés (Langevin & Hivon, 2007). Plusieurs auteurs ont ainsi démontré que les étudiants en médecine ne savent pas mettre en pratique ce qu'ils ont pu apprendre pendant leur formation initiale, et qu'ils ne sont pas préparé à gérer les premiers instants, pourtant cruciaux d'un arrêt cardiaque ou d'une urgence vitale (McEvoy et al., 2014; Xi et al., 2015). Ainsi, la simulation semble être une solution à ce manque de pratique et son utilisation pour évaluer les étudiants et certifier leurs compétences pourrait remédier à ce manque de préparation.

Ces différentes notions sont abordées en profondeur dans le premier chapitre de notre travail, qui sera suivi, dans le deuxième chapitre, de l'exposé du cadre de développement d'un outil d'évaluation et des méthodes employées dans les différentes études pour réaliser le recueil et l'analyse des données.

Dans le troisième chapitre, nous rapporterons les résultats obtenus au cours des différentes recherches et les mettrons en lien les uns avec les autres.

Enfin, dans le quatrième chapitre, seront discutés les différents apports de la recherche, en rapport avec la validité des scores étudiés, mais également avec leur place dans un programme de formation en médecine d'urgence et, plus généralement sera abordée la place de l'évaluation dans une activité d'enseignement par simulation en médecine d'urgence.

CHAPITRE 1 : CADRE THEORIQUE : L'EVALUATION DES APPRENTISSAGES AU SEIN DE L'APPROCHE PAR COMPETENCES

Ce chapitre aura pour objectif l'approche du concept d'évaluation, dans une approche « pédagogie médicale ». Il s'intéressera à l'apport de l'évaluation pour l'enseignement et l'apprentissage, aux différentes modalités d'évaluation des compétences puis à ce qui fait l'objet plus spécifique de notre recherche : la création d'un score d'évaluation.

1. L'APPROCHE PAR COMPETENCES : EMERGENCE ET ETAT DES LIEUX

Avant d'approfondir la notion de compétence, il nous paraît fondamental de décrire brièvement le cadre de développement de l'enseignement médical dans le monde occidental et notamment en France, puis le contexte de l'émergence de l'approche par compétence puis décrire l'approche et enfin analyser son déploiement actuel pour y situer notre travail.

1. 1 Bref rappel du cadre de développement de l'enseignement médical au 20ème siècle

1.1.1 Organisation des études médicales en France et en Amérique du Nord

Bien avant que la « révolution flexnérienne » n'atteigne les facultés françaises, l'enseignement médical, au début du XXème, est délivré au sein de cinq facultés de médecine (dont les facultés de Paris et Strasbourg), et est centré principalement sur un enseignement hospitalier « au lit du malade », par une médecine « installée » à l'hôpital (Picard et Mouchet, 2009). La clinique ou « art médical » y tient une place prépondérante et bien que le débat sur la place des sciences « auxiliaires » que sont la chimie, la physique, la biologie et autres sciences fondamentales, apparaissent dans les discussions sur le contenu de l'enseignement, elles sont reléguées au second plan ainsi, que l'illustrent les propos d'Armand Trousseau lors d'une séance d'ouverture

à la faculté de médecine : « de grâce, un peu moins de science, un peu plus d'art, messieurs ! » (L'expérience, 1842, p158).

Au même moment outre-Atlantique, Abraham Flexner, un spécialiste de l'Education, est missionné en 1908 par Andrew Carnegie pour établir un rapport sur les conditions de l'enseignement de la médecine aux Etats-Unis, mais également au Canada. Visitant les 148 écoles de médecine des deux pays, il soulève cinq faits marquants : une production massive de praticiens sans instruction réelle, des écoles de médecine nombreuses, mais sans objectif de formation établi, l'absence d'enseignement des sciences fondamentales, des écoles qui forment beaucoup d'étudiants au rabais et en suant d'une piètre méthode pédagogique et enfin des hôpitaux totalement indépendants des écoles de médecine, ne recevant pas les étudiants (Flexner, 1910, p10-11; Moll, 1968).

Flexner émet un rapport et préconise trois axes de développement pour l'enseignement « moderne » de la médecine aux Etats-Unis : un enseignement multidisciplinaire qui contient les sciences fondamentales et se déroule à l'université, une initiation à la recherche médicale et enfin une formation partagée entre l'université et les hôpitaux pour donner une place à la pratique. En France, quelques quarante années plus tard, sous l'impulsion de jeunes médecins hospitaliers, qualifiés de « néo-cliniciens », la réforme de l'hôpital public fait son chemin, avec au centre des réflexions une volonté d'améliorer la formation des médecins, les soins aux malades et la recherche médicale (Picard p100, Chevandier, p 306). Leurs réflexions aboutissent en 1958 à la réforme du système hospitalier avec la création des centres hospitaliers et universitaires, menée de de front avec la réforme de l'enseignement médical et de la recherche (Ordonnance n°58-1373, 1958). A partir des dix écoles et des douze facultés de médecine, vingt-deux centres hospitalo-universitaires (CHU) voient jour et ont pour principales missions : le soin, l'enseignement et la recherche. L'organisation actuelle des études médicales puise toujours ses

racines dans la réforme de 1958, les étudiants partagent leur temps entre stages hospitaliers et enseignement facultaire, et les enseignants y ajoutent leur activité de recherche.

Les modèles d'enseignement se ressemblent alors et une de leurs caractéristiques repose sur un enseignement théorique délivré à l'université pendant la première partie des études médicales, puis un enseignement pratique, fait d'observations pendant la première partie des études (qui correspond à l'externat en France) puis de mise en pratique lors de la dernière phase (l'internat ou résidanat). Peu à peu, et devant l'augmentation importante des connaissances à acquérir, la place de la théorie devient prépondérante pendant le premier cycle des études puis les étudiants obtiennent une certification leur permettant de mettre en œuvre ce qu'ils ont appris à l'université et de devenir internes.

Ainsi, regarder une petite partie de l'histoire de l'enseignement de la médecine illustre une de ses constantes : la recherche permanente d'un équilibre entre enseignement théorique et pratique, recherche qui est au cœur des réformes actuelles.

1.1.2 Behaviorisme et pratiques pédagogiques au 20ème siècle

Flexner, considéré comme le père de l'enseignement moderne de la médecine, a donc permis l'introduction des sciences dans le curriculum médical, et a imaginé le concept d'hôpital universitaire (Albano & d'Ivernois, 2001). Les années cinquante et soixante voient ensuite la naissance des premiers départements de pédagogie médicale, devenus nécessaires pour mener une réflexion pédagogique, organiser l'enseignement des vastes savoirs issus du développement des connaissances scientifiques et choisir parmi eux, ceux qui sont utiles et essentiels à la pratique médicale (Pelaccia & Triby, 2011).

S'inspirant de la pensée et des travaux de Bloom et Gagné, les pédagogues médicaux tels que Miller et Abrahamson (1962) prônent un enseignement basé sur les objectifs pédagogiques, au sein duquel le contenu et son évaluation ont une place majeure. Le courant behaviouriste domine

alors le monde de l'éducation et base son raisonnement sur la conception que la connaissance se transmet, étape par étape, tout en étant décomposée du plus simple vers le plus complexe. L'apprentissage est centré sur le contenu de l'enseignement et sur l'action l'enseignant : il détermine l'environnement d'apprentissage, les stimuli qui déclenchent l'apprentissage et délivre enfin un renforcement positif ou négatif sur l'action de l'apprenant. Le renforcement, ou feedback, a pour but de modifier le comportement de l'étudiant, mais il ne porte pas de réflexion ni d'analyse réelle de l'action (Kay & Kibble, 2016; Pottier, 2013). La note reçue après un examen, sans autre commentaire que l'évaluation chiffrée en est l'exemple parfait. Les taxonomies d'objectifs voient le jour et décrivent des objectifs pédagogiques à obtenir pour valider une formation, divisant les objectifs sont divisés en trois domaines : cognitif (knowledge), psychomoteur (skills) et affectif (attitudes), autrement appelés savoir, savoir-faire et savoir-être. Ils ont été et sont toujours utilisés dans le domaine de la pédagogie médicale (Bloom B, et al, 1956; Ten Cate, 2017) . Un des objectifs de la division des objectifs est de pouvoir les décrire pour mieux les évaluer, puisqu'ils deviennent morcelables et donc mesurables. La taxonomie de Bloom permet ainsi, dans le domaine cognitif, d'évaluer des connaissances, une capacité à analyser des données (lors de la résolution d'un cas clinique) puis au niveau le plus élevé une capacité à résoudre des problèmes complexes, par exemple via l'utilisation de tests de concordances de scripts ou via l'apprentissage par problème (Charlin et al., 2000; Schmidt et al., 2011).

Ainsi, dans l'approche par objectifs, en partant de l'hypothèse que la somme des morceaux d'apprentissage serait égale au tout, on observe une « décomposabilité des apprentissages, un morcellement du programme » et la logique d'enseignement repose sur l'échelonnement des apprentissages (Albano & d'Ivernois, 2001; Aouni, 2012; Englander & Carraccio, 2018).

1.1.3 Emergence de l'approche par compétences

Une des conséquences de l'utilisation de la pédagogie basée sur les objectifs est la création de formations centrées sur le contenu, dans une logique certificative de l'enseignement (Aouni,

2012). Les programmes deviennent « la force motrice de la pédagogie » quitte à oublier l'objectif final de la formation : le soin des patients (Englander 2018). Englander note qu'il existe deux certitudes avec la formation par objectifs. La première est que les étudiants en médecine éprouvent de grandes difficultés pour débuter leur internat, car ils n'y sont pas préparés à cause du manque de cohérence entre les programmes et de la séparation profonde entre théorie et pratique. La deuxième est que le seul résultat garanti de cette approche est une grande variabilité des résultats de la formation qui ne s'adapte pas aux besoins réels des étudiants. Même si la théorie et la pratique sont au cœur de la formation, elles restent encore trop indépendantes et leur enseignement n'est pas cohérent.

L'approche par objectif se heurte à un deuxième obstacle, représenté par nombre croissant de connaissances à enseigner, bien loin du volume de connaissances décrit par Flexner et qui submerge à la fois les enseignants et les étudiants (Carracio, 2002, p 362). Batalden illustre ainsi la situation: "Medical educators these days are bombarded with teaching requirements—genetics, ethics, communication skills, molecular medicine, geriatrics, sexual health, and computer literacy, to mention a few." "These demands reflect the continued growth in scientific knowledge coupled with society's expectation that physicians minister to social and psychological as well as physical infirmities"[2] (Batalden et al., 2002). L'augmentation des connaissances scientifiques entraine alors une augmentation de la place de l'enseignement théorique, mais sans assurer une préparation adéquate à la pratique de la médecine.

C'est dans ce contexte que la notion d'approche par compétence « Competency-based education » émerge. D'abord évoquée dans les années 50-60 aux Etats-Unis, elle se développe dans les années 80 pour devenir le nouveau mode recommandé d'enseignement au début des

[2] « Les enseignants de médecine sont actuellement bombardés par des demandes d'enseignements croissantes :la génétique, l'éthique, la médecine moléculaire, la gériatrie, la sexologie, et les connaissances en informatique, pour n'en citer que très peu. » « Ces demandes reflète la croissance continue des connaissances scientifiques, couplées avec une société qui attend des médecins qu'ils s'occupent également infirmités sociales, psychologiques et physiques. » (traduction libre)

années 2000. Un des premiers auteurs anglo-saxons à avoir posé les jalons de l'enseignement guidé par les compétences est Ralph Tyler, qui en 1949, posait 4 questions que, selon lui, tout établissement de formation devrait pouvoir poser et résoudre : "1. What purposes should a school seek to attain? 2. What educational experiences can be provided to attain the purposes? 3. How can these be organized? 4. How can one determine whether the purposes are being attained?"[3] (Ten Cate, 2017; Fontaine et Loye 2017). Sans citer le terme de compétence, un des points clé de la future approche pédagogique figure dans les questions posées : il s'agit de l'objectif final de l'enseignement.

Dans le courant médical, les pédiatres sont les premiers, en 1972, à éditer un document décrivant les étapes de mise en place d'un enseignement se basant la notion de compétence (Burg et al. 1976 ; Carraccio et al. 2002). Cependant, les consignes et directives pour rédiger les programmes basés sur la compétence resteront trop superficielles, sans réelle définition des compétences attendues, ce qui ne permettra pas à cette approche de se déployer suite à ces réflexions. En 1978, l'OMS suggère qu'un système d'enseignement basé sur l'emploi des compétences pourrait aider à diriger les efforts en matière de pédagogie médicale vers les besoins de santé spécifiques de chaque pays : " the intended output of a competency-based program is a health professional who can practice medicine at a defined level of proficiency, in accord with local conditions, to meet local needs" [4] (McGaghie et al. 1978 ; Powell et Carraccio 2018). Dans les années 80, devant la constatation d'une insuffisance de l'approche par objectif, et avec le développement des théories de l'éducation socioconstructivistes, une tentative de mise en place de l'approche par compétences émerge, notamment aux Etats-Unis, au Québec puis en Europe (Aouni, 2011 p 121). Elle échouera et ne sera pas adoptée par les communautés éducatives, probablement à

[3] 1. Quels objectifs l'école devrait-elle viser ? 2. Quelles expériences éducatives permettront l'atteinte de ces objectifs ? 3. Comment ces expériences éducatives seront-elles organisées ? 4. Comment déterminer si ces objectifs ont été atteints ? D'après Fontaine, 2017, traduction libre

[4] "Le résultat escompté d'un programme basé sur les compétences est un professionnel de la santé capable d'exercer la médecine à un niveau de compétence défini, en accord avec les conditions locales, pour répondre aux besoins locaux" (traduction libre)

cause d'un manque d'aiguillage des enseignants quant à l'évaluation des étudiants, à la fois à cause de l'absence d'outils spécifiques mais également à cause de l'insuffisance de description des différents niveaux attendus de l'étudiant à un instant précis de son parcours (Carraccio, 2002, Powell, 2018).

Enfin, alors que le 20ème siècle se termine, la notion de compétence s'impose dans différents champs de la société, et devient centrale, non seulement pour les formations, mais également dans le cadre politique, professionnel et privé. Une formation se définit alors par les compétences nécessaires à maîtriser à son issue, un poste de travail est décrit par les compétences requises pour y prétendre et, dans la vie quotidienne, les citoyens doivent avoir acquis des compétences « clés » pour évoluer dans la société actuelle (Coulet, 2016). A l'origine de l'adoption du paradigme des compétences, on trouve plusieurs éléments parmi lesquels une volonté de se recentrer sur l'étudiant, de lui apporter un parcours plus professionnalisant, mais également une volonté d'évaluer des programmes plus standardisés qui suivent une approche précise permettant de définir les objectifs d'une formation et de les mesurer (à la fois pour en apprécier leur qualité mais également pour les comparer à l'aide d'enquêtes nationales et internationales portant sur les matières principales et les compétences clés).

En Europe, on soutient l'adoption de l'approche par compétence pour toute la société. Tout d'abord via le *Livre blanc sur l'éducation et la formation*, publié en 1996 par la Commission européenne et qui propose « la création d'un processus européen permettant de confronter et de diffuser largement les définitions des compétences clés et de trouver les meilleurs moyens de les acquérir, de les évaluer et de les certifier » et qui insiste sur l'importance de développer ses compétences « tout au long de la vie » (Commission des Communautés Européennes 1995; Coulet, 2016). Puis, en 1999, via le Processus de Bologne dont la déclaration a été signée par les ministres de 29 pays européens qui s'engagent à créer un espace commun européen de l'enseignement supérieur grâce à une harmonisation des structures universitaires et avec un

processus d'assurance qualité, dans le but de contribuer à l'accroissement de la compétitivité et de l'attractivité de l'enseignement supérieur européen (Déclaration commune des ministres européens de l'éducation 1999; Musselin, et al, 2007). Le processus de Bologne repose sur une approche par compétence, et plus précisément par « acquis d'apprentissage » avec une philosophie d'enseignement centré sur l'étudiant, au sein d'un enseignement de masse. Les principes philosophiques qui sous-tendent le processus de Bologne sont les suivants : mobilité, employabilité, dimension sociale et apprentissage tout au long de la vie, et les quatre objectifs de l'enseignement supérieur européens deviennent alors : la préparation à la vie comme citoyens actifs dans des sociétés démocratiques, la préparation au marché du travail, le développement personnel et le développement et le maintien d'une large base de connaissances de pointes (Lemenu & Heinen, 2015).

Dans le domaine de la pédagogie médicale, un autre élément favorise l'adoption de l'approche par compétence. Il s'agit d'un rapport intitulé « To Err is Human » et qui pointe le manque de sécurité et de qualité des soins prodigués aux Etats-Unis avec la mise en évidence de nombreux effets indésirables liés à des erreurs humaines et systémiques (Institute of Medicine (US) Committee on Quality of Health Care in America, 1999). L'élément marquant de ce rapport est que la moitié des évènements indésirables pourrait être évitée. Le public prend alors conscience que la qualité des soins n'est pas à la hauteur de ses attentes. Une des causes émises par le rapport et par les contestations du public est un problème de formation des futurs professionnels de santé et notamment un manque de cohérence entre théorie et pratique. Les pédagogues, cette fois appuyés par un soutien institutionnel, identifient et proposent une solution : les futurs médecins seront formés par une approche guidée sur la finalité de la formation et non plus par une approche guidée par un contenu principalement théorique et qui ne se préoccupe pas des compétences finales nécessaires à acquérir avant de pouvoir pratiquer. Le projet se nomme

« The Outcome project », est publié en 2002 et a pour ambition de former des internes « compétents au premier jour de leur internat » (Batalden et al., 2002).

Les canadiens, de leur côté, rédigent leur premier référentiel de compétences en 1996, puis les USA en 1999 suivis par la Chine, l'Iraq et l'Europe, au travers de la conférence de Bologne. Au début des années 2000, il semble qu'un mouvement vers l'approche par compétence, guidé par les finalités de la formation se forme, et qu'il devient mondial. Deux écueils sont alors identifiés et vont devoir être traités selon ses défenseurs : il faut régler le problème de l'évaluation en créant des outils valides, reproductibles et prédictifs, et le mouvement devra apporter la preuve de son efficacité pour devenir pérenne et accepté (Carraccio et al. 2002, p5).

1.2 L'approche par compétences : principes et définitions

L'approche par compétence s'enracine en partie dans le courant constructiviste qui, initié par Piaget (1896-1980) au début du XXème siècle, remet en cause le principe de décomposition des apprentissages. Le constructivisme, contrairement au behaviorisme, prend en compte les connaissances antérieures de l'apprenant et s'intéresse à la manière avec laquelle les apprenants non pas reçoivent la connaissance, mais la construisent (Aouni, 2012; Pottier, 2013).

Selon le courant constructiviste, les interactions entre les différentes composantes d'une connaissance, et plus largement d'un programme d'enseignement (le tout) sont plus importantes que les composantes elles-mêmes, puisque c'est grâce à la manière dont elles se réalisent qu'elles donnent du sens à l'apprentissage et créent de nouvelles connaissances (Aouni, 2012). Le constructivisme s'intéresse à la manière dont l'apprenant, en confrontant ses connaissances et à partir d'outils d'apprentissage et d'expériences va en créer de nouvelles. L'individu, actif, construit lui-même ses nouvelles connaissances grâce aux matériaux fournis par l'environnement, par l'enseignant. En médecine, une des méthodes utilisant ces concepts est l'apprentissage par problème, apparu dès les années 60 et qui se répand à la fin du XXème siècle (Schmidt et al 2011; Shrivastava, et Ramasamy 2013). L'approche par compétence se place

également dans une approche socio-constructiviste de l'apprentissage développées notamment par Vygotski (Lemenu & Heinen, 2015, p 24). Dans cette logique, l'étudiant est acteur de la construction de ses savoirs, en lien avec ceux qu'il a déjà acquis, en interaction avec l'enseignant qui crée des situations d'apprentissage, et avec ses pairs qui sont acteurs de la co-construction de savoirs. L'acquisition se fait sans ignorer le contexte socio-culturel dans lequel les acteurs du savoir évoluent.

1.2.1 Définitions de la compétence

Bien que qualifié par différents auteurs de « mot valise », « concept étendard », « terme polysémique, polymorphe », « nouveau paradigme », le terme de compétence « a le vent en poupe » puisque, comme nous l'avons exposé ci-dessus, l'ambition actuelle est de l'implémenter, de le développer et de le pérenniser dans de nombreux champs de la société (Crahay, 2006; Lemaitre & Hatano, M, 2007; Postiaux et al., 2010). Notre travail n'étant pas centré sur une discussion du terme en lui-même, nous allons parcourir rapidement les différentes définitions de la compétence, principalement en fonction de leur utilisation dans le monde de la pédagogie médicale. Deux ressources principales sont à l'origine de ce travail de définition : les auteurs américains et canadiens car leur travail influence le développement de l'APC au sein de la pédagogie médicale, mais également les auteurs qui s'inscrivent dans la discipline des Sciences de l'Education et de l'activité car leur regard est complémentaire et a permis de définir certains éléments de la recherche présentée.

En 2002, Epstein et Hundert proposent une définition de la notion de compétence en se basant sur une revue de la littérature entre 1966 et 2001. Elle serait " the habitual and judicious use of

communication, knowledge, technical skills, clinical reasonning, emotions, values, and reflection in daily practice for the benefit of the individual and community being served"[5]

En 2010, afin de fixer un langage commun pour les pédagogues, les responsables de programme et les étudiants, Jason Frank, pour le Royal College of Physicians and Surgeons of Canada et associé à des chercheurs et représentants de différentes institutions et collèges d'enseignants canadiens, américains, néerlandais, écossais, australiens et gallois publie une conférence de consensus autour de l'approche par compétence. Ils définissent ses bases théoriques, différentes définitions et concepts ainsi que les différents défis qui sont identifiés pour le futur, avec la volonté forte d'en faire un paradigme d'enseignement qui transformera la pédagogie médicale (Frank et al., 2010).

Ils utilisent alors deux termes pour parler de compétence : « the competence » qui est « the array of abilities (knowledge, skills, and attitudes) across multiple domains or aspects of performance in a certain context. Statements about competence require descriptive qualifiers to define the relevant abilities, context, and stage of training. Competence is multi-dimensional and dynamic. It changes with time, experience, and setting" [6] et "the competency" définie par "an observable ability of a health professional, integrating multiple components such as knowledge, skills, values, and attitudes. Since competencies are observable, they can be measured and assessed to

[5] « L'utilisation habituelle et judicieuse de communication, connaissances, savoir-faire techniques, raisonnement clinique, émotions, valeurs et réflexion dans la pratique de tous les jours pour le bénéfice des individus et de la collectivité ». (traduction libre)

[6] "L'ensemble des capacités (savoir, savoir-faire, savoirs-être) dans plusieurs domaines ou aspects de la performance, dans un certain contexte. Les prises de position à propos de la compétence nécessitent des qualificatifs descriptifs pour définir les capacités, le contexte et le stade de formation pertinents. La compétence est multidimensionnelle et dynamique. Elle évolue avec le temps, l'expérience et le contexte". (traduction libre)

ensure their acquisition. Competencies can be assembled like building blocks to facilitate progressive development" [7] (Frank et al. 2010).

Dans ce cadre, la notion de compétence représente un ensemble d'habiletés, au sein de plusieurs domaines de compétence ou même par certains aspects, une certaine forme de performance. Ces habiletés sont dépendantes du contexte dans lequel elles s'expriment et surtout, du niveau d'apprentissage au cours duquel elles sont utilisées. Elles sont donc par essence multidimensionnelles, dynamiques et variables au cours du temps, de l'expérience et du contexte.

Enfin, les auteurs considèrent que ce qu'ils nomment « competency » et plus souvent « competencies » car le terme est fréquemment employé au pluriel, représentent les « ingrédients » ou « éléments de la compétence, ce que certains auteurs francophones appellent « composantes des compétences » (Scallon 2015, p 97). Pour Albanese, les « competencies » sont les habiletés ou capacités qui constituent des unités organisationnelles des compétences (Albanese et al., 2008).

Dans le champ de l'éducation, de la formation des adultes et de la didactique professionnelle, de nombreuses définitions existent, qui varient en fonction de la discipline de travail ou du cadre théorique dans lequel les auteurs travail. Ainsi, une impression de « nébuleuse », voire même de « brouillard sémantique » existe (Jonnaert, 2017). Cependant, il existe des points ou dynamiques communes ces définitions qu'il est possible d'identifier pour « cerner » la notion de compétence.

Une des composantes importantes de la notion de compétence est sa « confrontation majeure à l'activité » comme élément « moteur » de la compétence (Baudouin, 2002). En effet, elle est

[7] " une capacité observable d'un professionnel de la santé, intégrant des composantes multiples telles que les savoirs, savoir-faire, savoirs-être. Puisque les composantes des compétences sont observables, elles peuvent être mesurées et évaluées pour garantir leur acquisition. Les composantes des compétences peuvent être assemblées comme des blocs de construction pour faciliter le développement progressif". (traduction libre)

décrite comme étant une capacité à résoudre une tâche-problème par Gillet, ou encore un savoir-agir (Tardif et al, 2006), un ajustement de l'action. Pour Leplat elle est toujours une compétence « pour » quelque chose et donc la logique de l'action est également présente (Leplat, 1997, in Baudoin 2002, p 151). Perrenoud note ainsi sa composante « synergique », dans un souci « d'orchestration » de l'action, et illustre ainsi le caractère organisé, fonctionnel d'une compétence (Perrenoud, 2002). La compétence est opératoire, au sein de l'action et organisée. Le verbe « mobiliser » est également fréquemment employé dans les différentes définitions de la notion de compétence (Le Boterf 1994; Postiaux et al. 2010; Scallon 2007; Tardif et al, 2006).

Le lien indispensable de la compétence avec l'activité impose que la compétence se révèle « en » activité, et en situation. La compétence est située et de ce fait, a un lien fort avec l'expérience. Ainsi, la compétence au sein de ce que certains auteurs appellent des « familles de situation », qui serait des situations professionnelles au sein desquels, grâce à des schéma opératoires, la compétence de l'apprenant s'exerce (Allal 2013; Scallon 2015; Tardif et al, 2006). Un point majeur dans la nature de l'activité est qu'elle doit être complexe et contextualisée (ce que permet la famille de situation et donc la « mise en » situation »). En effet, et cela rejoint une des composantes des définitions anglo-saxonnes, la compétence est dynamique, complexe et elle varie selon les contextes.

Un élément majeur de la notion de compétence est son lien avec les savoirs. Dans le sens très large du terme savoir puisqu'ils peuvent être savoirs cognitifs, procéduraux, raisonnement, mémoire etc. (Jonnaert, 2017). Postiaux parle de « savoirs en acte » et Jonnaert appelle à se méfier de l'opposition parfois trop simpliste entre savoirs et compétence. En effet, la tentation est grande d'opposer ces notions qui peuvent paraitre banales alors que, bien loin d'être opposées, elles interagissent sans cesse pour « produire des compétences » (Jonnaert, 2017; Postiaux et al., 2010).

Dans la suite de ce travail et pour les réflexions ultérieures, nous utiliserons la définition de Tardif, la plus souvent employée dans les sciences de la santé, et qui regroupe les composants essentiels de la compétence : « un savoir-agir complexe prenant appui sur la mobilisation et la combinaison efficaces d'une variété de ressources internes et externes à l'intérieur d'une famille de situations » (Tardif et al, 2006).

1.2.2 Objectifs et principes de l'approche par compétence

L'approche par compétence a d'abord été développée avec un objectif à l'esprit : améliorer la sécurité des soins apportés aux patients. Dans cette optique, le projet initial de l'ACGME se nommait « The Outcome Project », afin d'insister sur la finalité des programmes de formation au sein de l'APC : le devenir des patients (Batalden et al., 2002; C. Carraccio et al., 2002). Le but de l'utilisation de l'APC est non plus de délivrer un contenu qui serait maitrisé à la fin d'un programme de formation, mais avant tout, de former un professionnel de santé compétent, fiable pour les patients et pour le système de santé dans lequel il va exercer. Le programme est guidé par les besoins de la population et le formateur doit être guidé par ce que devra être le futur professionnel de santé. Une des manières de réfléchir au sein de l'APC est d'aborder le programme de formation avec l'image du professionnel de santé compétent en tête : « starting with the end in mind»[8] (Covey 1989, dans Englander et Carraccio 2018).

Pour atteindre l'objectif fixé, le deuxième but de l'APC est de standardiser non pas le contenu, mais les résultats des formations, en décrivant les compétences à développer et à acquérir tout au long de la formation. L'objectif est donc de former un professionnel qui est capable et s'engage à fournir un service fiable et sûr, en temps utile, avec efficacité, efficience et en délivrant des soins centrés sur le patient : "physicians must not only master a body of knowledge but also possess the ability to apply that knowledge in service to others, conduct themselves as

[8] « débuter avec la fin en tête » (traduction libre)

professionals, work effectively in teams, communicate compassionately with patients and respectfully with colleagues, collaborate to improve systems of care, and engage in critical reflection and lifelong learning" [9] (Holmboe 2015; Lucey et al, 2018).

En 2010, Frank définit ainsi l'approche par compétence dans le champ de la médecine : "approach to preparing physicians for practice that is fundamentally oriented to graduate outcome abilities and organized around competencies derived from an analysis of societal and patient needs" [10] (Frank et al., 2010). Afin d'atteindre les objectifs de standardisation de résultats d'une formation qui est guidée par les besoins des patients, plusieurs principes sous-tendent le développement de l'APC, ils sont résumés dans la figure 1.

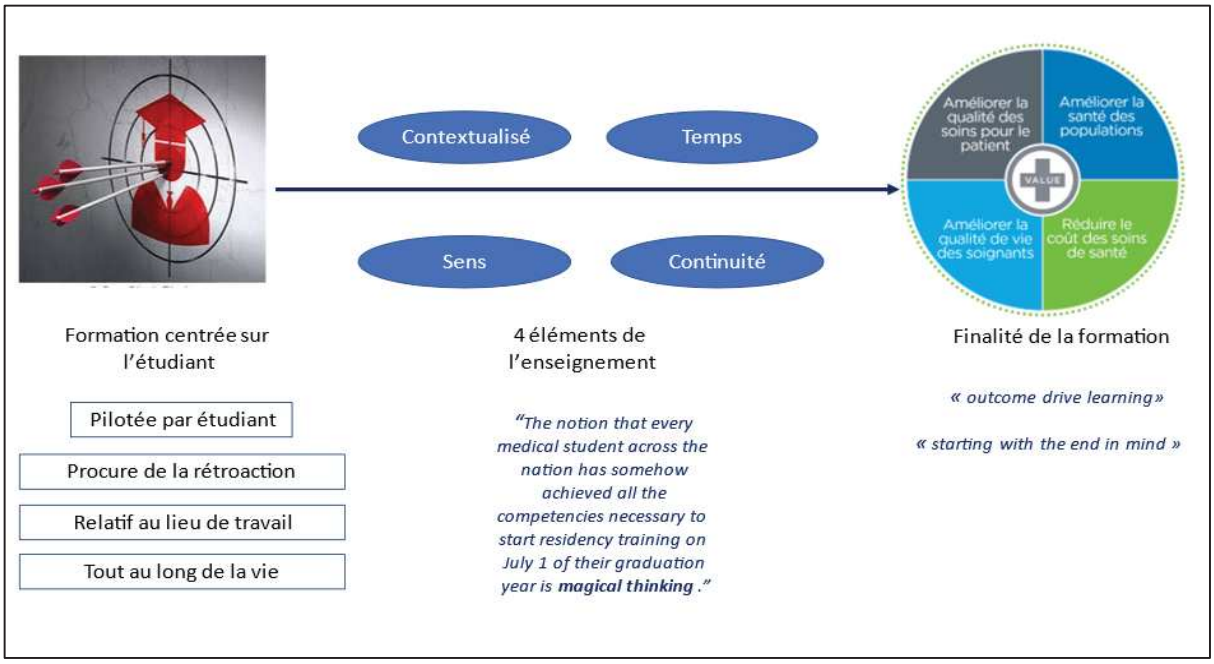


Figure 1 – Conditions pédagogiques de l'approche par compétences

[9] « les médecins doivent non seulement maîtriser un ensemble de connaissances, mais aussi être capables d'appliquer ces connaissances au service des autres, de se comporter comme des professionnels, de travailler efficacement en équipe, de communiquer avec compassion avec les patients et respectueusement avec leurs collègues, de collaborer pour améliorer les systèmes de soins, et de s'engager dans une réflexion critique et un apprentissage tout au long de la vie professionnelle. (traduction libre)

[10] « une approche de la préparation des médecins à la pratique qui est fondamentalement orientée vers les capacités des diplômés et organisée autour de compétences, dérivées d'une analyse des besoins de la société et des patients »

Le premier point fondamental de l'APC est l'utilisation d'un l'enseignement centré sur l'étudiant, dans une perspective constructiviste qui prend en compte ses connaissances antérieures, son rythme d'apprentissage et son potentiel à agir lui-même sur la formation. Il existe une grande responsabilisation des étudiants, que l'on appelle plus volontiers « apprenants » dans l'APC (Aouni, 2012; Fazio et al., 2018).

Pour mettre en place l'apprentissage centré sur l'apprenant et sur l'acquisition de compétences, deux autres principes d'enseignements sont majeurs : la continuité entre les enseignements et la variabilité dans le temps des acquisitions des apprenants, nommée en anglais « competency-based, time variable education ». La continuité s'entend au sens de continuité entre les différents enseignements : entre matières fondamentales et cliniques, entre apprentissage des connaissances cliniques et stages hospitaliers, au sein du stage entre connaissances cliniques et travail au sein d'une équipe, puis avec le patient.

L'idée est de ne pas morceler les différentes composantes des compétences, pour les aborder dans leur globalité, au plus proche de ce qu'elles sont dans la profession. Par exemple, enseigner la microbiologie sans lien avec la clinique, avec la pris en charge réelle du patient n'a pas dans sens dans le cadre de l'APC (Englander & Carraccio, 2018). Selon les auteurs, le manque de continuité dans les enseignements entraine un manque de sens, à la fois car les étudiants apprennent mieux quand ils peuvent faire des liens, mais également un manque de sens pour la pratique et le vécu en stage qui ne permet pas aux étudiants de mettre en application ce qu'ils apprennent à l'université. Cela conduit à une mise en difficulté pour développer des compétences telles que la communication ou le travail en équipe (Englander, 2018). A ce titre, les auteurs recommandent de développer des durées de stages longues et ce dès l'externat (Bernabeo et al., 2011; Hirsh et al., 2012; Norris et al., 2009).

Dans la logique de la continuité et du sens donné aux apprentissages, la possibilité de faire varier le temps de formation est également centrale dans un système d'approche par compétences, au sein duquel on reconnait que les étudiants n'apprennent pas tous au même rythme.

Dans un programme idéal, ils devraient pouvoir franchir différentes étapes de formation, préalablement définies, à leur rythme et non pas au rythme des semestres et des évaluations pas toujours adaptées à l'évaluation du développement de leur compétence. Les étudiants devraient pouvoir adapter le rythme, la durée de leur entrainement en fonction de leurs capacités à acquérir des compétences, variables selon les compétences elles-mêmes, mais également le moment de la formation (parfois perturbé par des raisons personnelles par exemple). Englander résume bien la situation actuelle, et qui n'est pas adaptée à l'APC : "The notion that every medical student across the nation has somehow achieved all the competencies necessary to start residency training on July 1 of their graduation year is magical thinking." [11] (Englander & Carraccio, 2018).

En effet, les programmes de formation actuels sont basés sur « une durée d'exposition » aux connaissances, ressources, stages mais pas sur la démonstration d'une maitrise ces compétences indispensables à acquérir pour pouvoir débuter l'internat puis sa vie professionnelle.

Comme le décrit bien Carraccio en 2002, l'expérience de la formation, auparavant principalement définie par une durée d'exposition à un lieu de stage devient alors définie par l'objectif final de la formation (devenir un médecin compétent), qui doit alors guider le processus pédagogique (Carraccio et al., 2002). Ainsi l'APC "de-emphasizes time-based training and promises greater accountability, flexibility, and learner-centeredness" [12] (Frank et al., 2010).

[11] "L'idée que chaque étudiant en médecine à travers le pays a, d'une manière ou d'une autre, atteint toutes les compétences nécessaires pour commencer son internat le 1er juillet de l'année de leur diplôme est une pensée magique."

[12] « ne met pas l'accent sur la formation basée sur le temps et promet une plus grande responsabilité, une plus grande flexibilité et une plus grande attention à l'apprenant »

Puisqu'une compétence est contextualisée, alors son enseignement ne doit pas être éloigné de la réalité de la pratique et devrait avoir du sens et un rapport permanent avec le lieu de travail (Fazio et al., 2018). En effet, la compétence possède deux caractéristiques qui la rendent indissociable de l'environnement de travail puis qu'elle est intégratrice (située) et combinatoire (dans le sens de ma mobilisation et combinaison des ressources) (Demeester, 2020). Par essence, la compétence est associée à l'expérience puisqu'elle ne s'exprime qu'en situation.

Enfin, les méthodes d'enseignement utilisées au sein de l'APC devraient permettre de procurer de la rétroaction fréquente aux apprenants, en les aidant à développer une pratique réflexive et ce, tout au long de leur vie professionnelle. Les étudiants devraient donc avoir les outils pour identifier eux-mêmes les compétences qu'ils acquièrent. Le propre même d'une compétence est de ne pas être figée dans le temps, et elle se développe ou s'appauvrit tout au long de la vie professionnelle (Lucey et al., 2018).

Ainsi, l'APC se veut une approche basée sur la pertinence des connaissances plutôt que sur leur exhaustivité, dans une approche globale plutôt que fragmentée et avec une vision de la formation centrée sur l'apprenant et guidé par les besoins des patients, au sein d'un système de santé spécifique. L'évaluation des compétences, qui représente un des défis majeurs de l'APC devrait donc suivre les principes fondamentaux de l'APC.

1.2.3 Etapes de développement de l'approche par compétence : l'exemple de l'Amérique du Nord

Entre 1999 et 2002, l'Accreditation Council for Graduate Medical Education (ACGME) et la Liaison Committee on Medical Education (LCME), associées à l'ABMS (American Board of Medical Specialties) décident que la « competency-based medical education » ou CBME deviendra l'approche pédagogique de l'enseignement médical puis, ils prescrivent des recommandations pour les déployer, afin d'éviter l'échec dans les années 80. Au sein de

« l'Outcome Project », ils définissent les différentes étapes à franchir pour développer et installer cette nouvelle approche de l'enseignement médical (Batalden et al., 2002).

La première étape requise est la définition de compétences indispensables pour devenir un médecin compétent, responsable envers la société et qui délivre des soins pertinents. Initialement recensés par des experts (200 médecins et 18 acteurs du système de santé tels que des administrateurs, des infirmiers, des patients ou des employés administratifs) et par une revue de la littérature, les experts identifient en 2002 six domaines de compétences indispensables à développer et acquérir pour devenir médecin : « patient care; medical knowledge; practice-based learning and improvement; professionalism; interpersonal skills and communication; and systems-based practice »[13] (Batalden et al. 2002). En 2013, afin de créer une taxonomie commune de ces domaines de compétence ces six domaines sont repris, et deux leurs sont ajoutés : « Interprofessional collaboration » et « Personal and Professional Development»[14]. Les huit domaines de compétence sont ensuite déclinés en 58 composantes, encore appelées « sous-compétences, subcomptetencies » (Englander et al., 2013).

En Suisse, dans le programme PROFILES (Principal Relevant Objectives and a Framework for Integrative Learning and Education in Switzerland) et au Canada au sein d'un cadre nommé CanMeds, 7 domaines de compétences « générales » sont attendues de la part de tous les médecins pour répondre de façon efficace aux besoins de ceux qu'ils soignent. Ils sont utilisés sous le terme de « rôles » dans l'activité clinique du médecin et sont les suivantes : expert médical, communicateur, leader, collaborateur, promoteur de santé, érudit et professionnel (Figure 2).

[13] le soin aux patients, les connaissances médicales, ; apprentissage et amélioration fondés sur la pratique, compétences de communication, professionnalisme et pratique au sein d'un système de santé.
[14] Collaboration interprofessionnelle et développement personnel et professionnel

En France, les domaines de compétences ont été définis dans l'arrêté du 8 avril 2013, qui réformait les deux premiers cycles des études médicales. Elles sont au nombre de sept et sont semblables aux domaines de compétences suisses et canadiens, même s'ils ne sont pas énoncés de manière identique. Le futur médecin devra acquérir des compétences de clinicien, de communicateur (avec le patient, son entourage), de membre d'une équipe soignante pluriprofessionnelle et qui coopère avec elle. L'étudiant en médecine doit également acquérir des compétences de santé publique afin de pouvoir en être acteur, des compétences scientifiques afin de réactualiser ses connaissances professionnelles pendant toute sa carrière et pour être capable de formuler une problématique de recherche. L'étudiant devra enfin avoir une attitude responsable aux plans éthique et déontologique, et il doit apprendre à développer sa réflexivité afin de démontrer sa capacité de remise en question, d'auto-évaluation qui lui permette de connaitre ses limites et d'argumenter ses décisions (Ministère des affaires sociales et de la santé, ministère de la défense et ministère de l'enseignement supérieur et de la recherche, 2013). La réforme du 2ème cycle actuellement en cours a pour ambition de d'utiliser l'APC pour former les futurs médecins et va donc s'appuyer sur les concepts et principes énoncés ci-dessus (Conférence des Doyens des facultés de Médecine, Collège National des Enseignants en Médecine, 2020; Décret n° 2021-1156, 2021).

Les domaines de compétences identifiés ont en commun de pouvoir être utilisés pour toutes les spécialités médicales et sont destinés à tous les niveaux d'apprentissage.

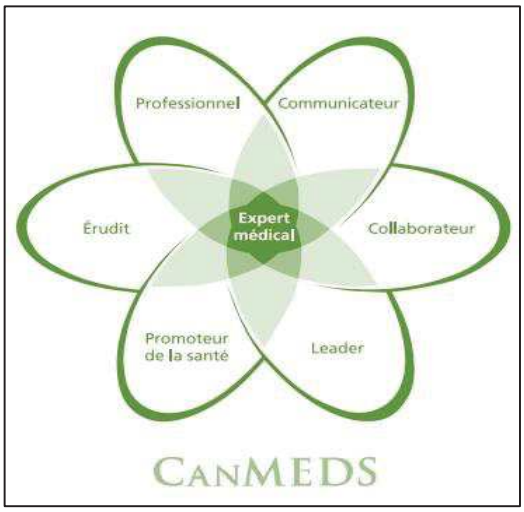


Figure 2 – Les 7 « rôles » du CanMeds (Frank et al, 2015)

La deuxième étape pour développer l'APC a pour objectif de définir des niveaux de performance pour chaque compétence. Elle s'intitule « The Milestone Project ». Les enseignants doivent décrire, pour chaque compétence, des « observable marker of an individual ability along a development continuum » [15] (Englander et al., 2017; Frank et al., 2010). Pour chaque niveau d'un domaine de compétence et d'une composante de ce domaine, une description doit être établie, afin de guider les acquisitions de l'apprenant, mais également les évaluations.

La troisième étape, celle qui relève le défi initialement perdu dans les années 1980, est celle de l'évaluation, qui doit se faire au sein d'un programme dans une approche globale des compétences. Dans les années 2000, l'étape de l'évaluation est décrite comme une des principales barrières à la pérennisation de l'approche par compétence. Nous détaillerons la mise en place d'une évaluation des compétences dans le chapitre suivant puisqu'il s'agit de notre thématique de recherche.

[15] « marqueur observable d'une capacité individuelle le long d'un continuum de développement »

Enfin la dernière étape doit constituer en une évaluation des programmes basés sur la CBME. Ils doivent faire preuve de leur efficacité, à la fois sur le contenu de la formation, mais également sur la qualité et la sécurité des soins et donc sur l'impact de l'APC sur les patients, à l'origine de la mise en place de de la CBME (Batalden et al., 2002).

1.2.4 20 ans après : implémentation et développement de l'APC

Actuellement, il existe peu d'institutions qui ont adopté une approche par compétence dans son entièreté, mais la plupart des pays ont défini les compétences requises pour devenir un professionnel de santé responsable, comme nous avons pu le voir avec les Etats-Unis en 2002 et 2013, le Canada en 2007, la Suisse en 2016 et la France. Les facultés se heurtent à de nombreux obstacles notamment en terme d'adaptabilité des programmes et en terme d'évaluation (Lucey et al., 2018). En 2009, l'université de Toronto a instauré un programme entièrement guidé par l'APC, pour 14 internes de chirurgie orthopédique. Les étudiants rapportaient une grande satisfaction d'avoir participé au programme. Trois d'entre eux ont effectué leur internat en 3 ans au lieu de 4 et deux autres en cinq ans, les six autres ayant terminé leur internat en quatre ans (Ferguson et al., 2013). Aux Etats-Unis, un programme de pédiatrie propose, dès l'externat, d'intégrer une approche basée intégralement sur les compétences avec une évolution conditionnée non pas par le temps, mais par la réussite à des évaluations qualitatives et quantitatives, variées, nombreuses et qui déterminent le passage dans le niveau supérieur. Parmi les douze étudiants inclus dans le programme, huit étaient prêts à devenir internes plus tôt, à la fin de la 4ème année de médecine (Andrews et al., 2018).

En France, à notre connaissance, seule l'université d'Aix-Marseille a mis en place à la rentrée 2018 une approche par compétence pour plusieurs filières disciplinaires tels que le droit, les Sciences du Sport, une licence intitulée « Santé Visuelle » et à l'Ecole Supérieure du Professorat et de l'Education. Dans certaines filières les notes ont disparu, avec mise en place d'évaluations formatives, d'auto-évaluations, la mise en place d'un livret de compétences, d'un portfolio et une

mise en place de travaux de groupe en Master 2. Les auteurs du programme ont constaté une grande implication, accompagnée d'une grande rigueur et d'une profondeur de réflexion de la part des enseignants. Il n'existe pas encore d'évaluation du programme, mais l'expérience est à suivre (Demeester, 2020, p 181-197).

Parmi les difficultés actuellement rencontrées pour une mise en place globale de l'APC le manque d'uniformisation du langage utilisé a pu être identifié, même s'il tend à être homogénéisé avec les efforts permanents réalisés par les pédagogues entre 2000 et 2015 pour définir la compétence, les composantes des compétences (ou facettes), les jalons, puis les activités professionnelles confiables (Ten Cate, 2017; Frank et al., 2010 ; Frank & Danoff, 2007). Le manque de continuité dans les différents enseignements est également un obstacle majeur à l'application de l'APC (entre les matières, entre les matières et les stages, entre les stages et les objectifs pédagogiques ou encore avec ce qu'on demande aux étudiants de réaliser avec les patients) (Englander & Carraccio, 2018). De même, la variabilité attribuée aux temps d'apprentissage est très peu appliquée, car difficile à mettre en place. De tels programmes de formation existent aux Etats-Unis, au Canada ou au Pays-Bas. Leur utilisation permet d'adapter les enseignements et les évaluations à la progression des apprenants. Dans la majeure partie des cas, cela permet aux apprenants d'être en avance sur le déroulement habituel des études, et pour une minorité, d'être plus « en retard », mais pour des raisons personnelles qui permettent l'épanouissement des apprenants (Cangiarella et al., 2017; Ferguson et al., 2013; Lucey et al., 2018).

Les principaux obstacles de la mise en place de l'APC sont la nécessité de certifier les étudiants pendant une durée prédéterminée, avec des notes, qui, nous allons le voir, vont à l'encontre de l'APC. L'idéal serait en effet de développer des mesures de la performance valides, dans le sens où elles sont la mesure indirecte mais fiable de la compétence. Aux USA, le programme national

de Matching, qui requiert une préparation synchrone des étudiants et une évaluation de leurs connaissances semble être un obstacle pour développer l'APC dans sa totalité.

Les défis de l'APC sont encore nombreux : adopter la notion de variabilité du temps et avec elle, la possibilité pour un étudiant d'acquérir et de développer de nouvelles compétences au sein de stages hospitalier avec des rotations bien plus longues que celles qui existent actuellement. Les rotations courtes (de 1 à 3 mois) pour les étudiants en médecine, ne permettent en effet pas de mettre en place des exercices d'évaluation formative comme l'observation directe et régulière des étudiants, qui serait une solution intéressante pour l'évaluation des compétences, en milieu authentique. Le temps est également parfois employé plus à des révisions de connaissances non pertinentes qu'à un travail autour de l'amélioration des compétences (Lucey et al., 2018).

Un dernier défi à relever pour les universités, les hôpitaux, les enseignants et les étudiants qui est celui d'aligner activités pédagogiques, activés d'apprentissages et activités évaluatives et l'objectif du programme : le patient, et instaurer au programme de formation des activités en lien avec le raisonnement clinique, le travail en équipe, à la fois pour l'apprentissage mais également pour l'évaluation (Carraccio & Englander, 2013; Kogan et al., 2018).

Même si des pistes ont été trouvées pour l'évaluation des compétences et que l'échec des années 80 semble avoir été dépassé, il persiste une lutte pour l'évaluation qui requiert l'invention et la redéfinition des rôles des enseignants, des apprenants et des universités. L'observation directe devrait y être centrale, sur le lieu de travail et dans le temps de la formation. Lucey propose ainsi que les enseignants se fondent sur l'activité d'un professionnel pour observer et décrire encore les habiletés nécessaires à acquérir pour le futur professionnel, mais également que les apprenants aient des clés pour naviguer dans l'APC : avec une connaissance des objectifs de la formation, une incitation à être responsable et actif dans sa formation et enfin à imaginer des outils qui leur permettraient d'obtenir une rétroaction fréquente sur leur apprentissage (Carraccio & Englander, 2013; Lucey et al., 2018; Powell & Carraccio, 2018).

Un défi majeur reste cependant l'évaluation des compétences, et nous allons l'aborder dans le chapitre suivant, à l'aune de l'évaluation par la simulation qui reste notre principal sujet de réflexion.

2. Evaluer avec la simulation dans une approche par competences

Alors que l'évaluation tient une place centrale dans de nombreux systèmes humains, elle est sujette à de nombreux débats, semblant parfois être une « menace », une « folie » qui « envahit tout » (Hadji, dans Demeester, 2020, p 5). Les pratiques évaluatives ont en effet conquis de nombreux champs de la société, que ce soit dans les pratiques professionnelles, la politique, le spectacle, la gastronomie, ou encore les transports dont le terme même d'ubérisation de la société est issue. Dans le champ de la pédagogie scolaire ou universitaire, elle a également une place de choix, puisqu'elle constitue une des trois activités au cœur de toute formation que sont l'enseignement, l'apprentissage et l'évaluation (Romainville, 2013, p 11 et 38). Puisque notre recherche a pour objectif premier de développer un outil d'évaluation, le chapitre actuel va décrire les buts et effets d'une évaluation, et va tenter de situer notre démarche dans le cadre actuel de l'évaluation des professionnels de santé, au sein d'une approche par compétence, avec l'exemple de la simulation qui est l'outil que nous avons choisi d'étudier.

2.1 Pourquoi évaluer ?

Après l'enseignement et l'apprentissage, l'évaluation fait partie intégrante du processus éducatif, puisqu'elle a des impacts sur l'activité des apprenants, sur celle des enseignants et, dans le cadre des sciences de la santé, sur les patients et le système de soins (Allal, 2013; Brailovsky et al., 1998; Holmboe et al., 2011). La question du pourquoi pourrait donc ne pas sembler légitime, mais elle permet de rappeler quelques définitions, d'exposer les effets de l'évaluation, et enfin objectifs qu'elle devrait réaliser.

2.1.1 Définitions et approches évaluatives

L'évaluation est définie par Nadeau en 1978 comme un jugement de valeur porté sur une mesure, dans le but de prendre une décision. D'un point de vue étymologique, Gérard nous rappelle que sa racine indo-européenne « wal » signifie « exprimer sa force ». Sa forme latine « evaluatio » est composée du substantif « e » (hors de) et d'un dérivé du verbe « valuere » (être fort, bien portant, puissant ». L'intérêt d'analyser le terme est de voir en l'évaluation, une démarche qui consiste donc à « faire sortir la valeur de ce qu'on évalue, à en montrer la force » (Gérard, 2013).

En anglais, il existe deux termes : « assessment » et « evaluation ». Dans la littérature, le terme « assessment » est toujours employé lorsqu'il s'agit de l'évaluation des acquis des apprenants, du processus de collecte des données et de ce que les apprenants ont acquis, et comment ils l'ont acquis (Scallon, 1999). Il s'agit ici de s'intéresser processus de récupération des données de l'évaluation ainsi qu'à la manière et à la raison pour lesquelles les apprenants ont réussi ou pas. Le terme « evaluation » se rapporte plus à un processus de jugement terminal, se concentrant sur les notes, leur distribution et finalement s'intéressant plus à un groupe d'apprenant ou à un programme de formation qu'à l'étudiant lui-même. Le terme « evaluation » est employé également pour aborder le processus évaluation des institutions, des organisations. Il fait plus facilement référence à une forme sommative de l'évaluation, même si les termes « formative assessment » et « sommative assessment » restent employés lorsqu'il s'agit d'analyser les formes d'évaluation des apprenants.

Les différences sémantiques des termes anglais, permettent de faire le lien avec deux approches évaluatives, définies par leurs finalités et par le moment où elles sont utilisées : l'évaluation formative et l'évaluation sommative, même si ces termes connaissent actuellement une évolution que nous exposerons rapidement après les avoir présentés, car leur utilisation reste la plus fréquemment employée (Fontaine & Loye, 2017).

L'évaluation formative se déroule au cours de l'apprentissage, en relation étroite avec l'enseignement. Elle a pour objectif de fournir des données qui caractérisent l'état de l'apprentissage des étudiants à cet instant précis de l'évaluation. Elle permet donc à la fois à l'enseignant et à l'apprenant de se situer pour adapter le contenu pédagogique pour l'un, les activités pédagogiques pour l'autre. Elle fournit donc la rétroaction essentielle à la progression de l'étudiant et permet de guider, de réassurer l'étudiant en promouvant la pratique réflexive (Epstein, 2007; Fontaine & Loye, 2017). Avec l'utilisation de l'évaluation formative, la relation apprenant/enseignant s'apparente à celle qui peut exister entre un entraineur et un athlète. Elle est constructive et fait progresser l'un et l'autre avec un même objectif : la compétence professionnelle (Schartel & Metro, 2010). La simulation utilise en permanence ce mode d'évaluation, puisque le temps du débriefing, majeur dans une séance de simulation, est en partie dédié à fournir un retour sur la performance observée (Dieckmann et al., 2009; Flanagan, 2008).

Quant à l'évaluation sommative, elle se déroule à la fin d'un apprentissage et elle a pour mission de vérifier si les compétences requises sont acquises, si les objectifs d'apprentissage sont atteints. Elle est dite sommative car elle fait référence à la somme des mesures utilisées pour évaluer et pour prendre une décision concernant le niveau de l'apprenant (Vial, 2012, p 197). Lorsqu'elle permet de délivrer un diplôme, elle devient certificative, car l'autorité qui a permis l'obtention du diplôme certifie à la société que le futur professionnel est compétent. Dans ce cadre, l'évaluation décide donc de la réussite d'un étudiant au sein d'un programme de formation (Jouquan, 2002). Si l'évaluation statue sur la manière dont la réussite de l'étudiant est déterminée, par rapport à une norme (celle du groupe, celle fixée par une courbe gaussienne), ou par rapport à des critères de performances, alors elle se nomme respectivement normative ou critériée (Bertrand, in Pelaccia, 2016, p 348). Les objectifs de l'évaluation sommative sont l'attribution de notes, la prédiction de la réussite ou de l'échec d'un cours ou d'une formation, le point d'accès à une formation, basé sur les résultats obtenus à l'évaluation et enfin, la certification.

Une certaine forme de rétroaction est fournie par l'évaluation sommative, mais qui, sous la forme de note est d'une nature totalement différente de celle fournie par l'évaluation formative (Scallon, 1999). Dans leur volonté de définir un cadre à l'évaluation des individus, Norcini et al. ont défini l'évaluation sommative comme une évaluation qui doit rendre compte du niveau de responsabilité que peut prendre l'évalué. Les évaluations sommatives composent donc les évaluations dites à « enjeu élevé » puisqu'elles doivent permettre de prendre une décision sur le niveau de l'apprenant (Norcini et al., 2011; 2018).

Une des approches actuelles de l'évaluation mérite une attention particulière lorsque l'on s'intéresse à l'évaluation. Il s'agit d'une approche qui distingue plus précisément trois fonctions de l'évaluation. La première fonction, « assessment as learning », « évaluation en tant qu'apprentissage » permet d'impliquer l'étudiant dans le processus évaluatif et lui donne l'occasion de gérer ses apprentissages par l'évaluation, cette dernière devenant elle-même un apprentissage et pas une fin en soi (Hayward, 2015; Krupat & Dienstag, 2009). La deuxième fonction, « assessment for learning », « évaluation pour l'apprentissage » donne à l'évaluation un rôle de support pour planifier et ajuster l'enseignement. Elle s'apparente à l'évaluation formative puisqu'elle fournit des informations à la fois à l'enseignant et à l'apprenant, quant à la poursuite des activités d'enseignement. Enfin, « l'assessment of learning », « évaluation de l'apprentissage » s'inscrit dans une perspective sommative de l'évaluation (Fontaine & Loye, 2017; Schuwirth & Van der Vleuten, 2011).

2.1.2 Effets de l'évaluation sur les apprentissages

« L'évaluation guide les apprentissages » ou « Assessment drives Learning » est un adage fréquemment énoncé et retrouvé dans les travaux qui s'intéressent à l'activité évaluative. Il illustre les effets que peut avoir l'évaluation sur l'apprentissage, à savoir sur les pratiques des apprenants, sur leur motivation, et sur leurs différentes adaptations à l'évaluation. Plusieurs effets

sont ainsi décrits donnant à l'évaluation un rôle moteur dans les apprentissages (Cilliers et al., 2010; van der Vleuten et al., 2012).

L'effet éducatif de l'évaluation souligne la manière avec laquelle les apprenants adoptent leur mode d'apprentissage au mode avec lequel il leur sera demander de restituer des connaissances, habiletés ou attitudes. L'évaluation pilote les efforts des étudiants, leurs méthodes de travail et le temps consacré à l'apprentissage en fonction de la nature des épreuves qu'ils auront à passer (Romainville et al, 2013, p 11 et 38). Cela influence donc la manière de penser l'évaluation, puisqu'elle aura un retentissement sur la manière d'apprendre des apprenants. Il s'agit ici de réfléchir à l'alignement pédagogique d'une formation, qui, dès le départ doit avoir déterminé les compétences à certifier et doit donc avoir pensé l'évaluation en amont du dispositif d'apprentissage (Demeester, 2020, p 174). De ces compétences et de leurs différentes composantes, découlent des modalités d'évaluation spécifiques (Pangaro & Ten Cate, 2013). Dans la recherche que nous avons menée en amont de la thèse, avec des étudiants de 4ème année de médecine, l'effet éducatif est mis en exergue par un changement de pratiques des étudiants. Se sachant évalués en situation simulée, ils modifiaient leur préparation aux sessions de simulation, de même que leurs outils de révision : de la simple fiche résumée pour un cours donné, ils avaient modifié le contenu de celle-ci, en hiérarchisant les connaissances, les actions à entreprendre et l''organisation optimale pour la prise en charge du mannequin (Philippon et al., 2021).

L'évaluation a un impact sur la motivation des apprenants, par son contenu, par le sens qu'elle donne à l'apprentissage. Pour avoir un impact motivationnel positif, il est nécessaire de créer des évaluations en lien avec ce qui fait sens pour l'apprenant, ce qui enclenchera des comportements d'apprentissage basées sur la motivation intrinsèque, dont on sait qu'elle a un meilleur impact sur la rétention et sur la compréhension des apprentissages. Il faut donc fournir aux étudiants les

objectifs de l'évaluation, ses modalités et les contenus à maitriser (T. Pelaccia et al., 2008; Van Der Vleuten, 1996).

L'évaluation a également un effet catalytique, dans le sens où la rétroaction qu'elle fournit aux apprenants, si elle est pertinente et si elle a du sens, permet de guider leurs apprentissages ultérieurs, en les adaptant à leur niveau de progression. On retrouve ici l'intérêt d'utiliser des évaluations régulières, dans le cadre d'une approche formative (Norcini et al., 2011).

Enfin, un des effets notables de l'évaluation est l'effet testing. Par la mobilisation de leurs connaissances en situation évaluative, les étudiants les réorganisent et les restructurent. L'effet testing participe donc au renforcement des apprentissages, surtout si l'évaluation demande de fabriquer la réponse plutôt que de restituer des connaissances apprises par cœur. Ainsi, cela permet également d'adapter les activités d'apprentissage, en vue de l'évaluation, puisque les apprenants devraient s'entrainer à fabriquer des réponses, en fonction des situations potentielles fournies par l'examen, plutôt que de restituer un chapitre de cours (Larsen et al., 2008).

Inversement, l'évaluation peut avoir un impact motivationnel négatif si elle ne fournit pas assez de rétroaction, si elle semble injuste aux apprenants et si elle ne fait pas sens dans leur apprentissage. Deux études intéressantes sur l'utilisation des notes ou d'une validation binaire ont également mis en évidence cet état de fait : les étudiants préfèrent être validés ou pas, plutôt qu'être classés. Cela diminue le stress lié aux examens et renforce les liens du groupe d'étudiant (Bloodgood et al., 2009; Rohe et al., 2006). Un étude a suggéré que le fait de recevoir des prix, des mentions, était approuvé par les étudiants mais que, plus ils avançaient dans le cursus, plus il s'agissait d'un facteur de démotivation (O'Neill et al., 1999). Ainsi, hormis le stress que l'impact motivationnel négatif de l'évaluation engendre chez les apprenants, elle peut déclencher des conduites d'apprentissage qui ne sont pas constructives ni efficaces sur la rétention des connaissances, en mobilisant par exemple la motivation extrinsèque, qui permet une restitution immédiate des connaissances, mais pas leur rétention à long terme (T. Pelaccia et al., 2008).

2.1.3 Buts de l'évaluation

Puisque l'évaluation a de nombreux effets sur l'apprentissage, outre les objectifs traditionnels d'une évaluation qui sont la certification des acquis et des compétences ainsi que la sélection des candidats pour intégrer une filière de formation, les objectifs en lien avec les effets sur l'apprentissage son également à considérer, notamment dans une perspective formative de l'évaluation.

Un de ses premiers objectifs est donc d'optimiser les capacités des évalués en leur procurant une direction à suivre pour leurs apprentissages futurs et en leur donnant la motivation de le faire (Epstein, 2007). L'évaluation permet en effet d'émettre un jugement sur les capacités maîtrisées ou pas des apprenants, et ainsi de leur donner, quand cela est réalisé notamment en produisant de la rétroaction, des possibilités et moyens d'amélioration.

L'évaluation peut également avoir un objectif de sélection des apprenants, en leur permettant d'atteindre un niveau supérieur dans la formation, ou encore d'intégrer un cursus spécifique de formation (telle qu'une spécialisation, un diplôme universitaire). Ici, l'activité évaluative va permettre de classer, trier les apprenants, dans le but de retenir ceux qui sont apparemment aptes, d'après leurs résultats aux examens. L'approche docimologique, qui vise à analyser la validité d'une évaluation (c'est-à-dire le degré de conformité avec lequel la mesure rend compte ce qu'il a pour but de mesurer) prend ici une place importante puisque l'utilisation d'outils docimologiques permet d'assurer que les examens sont pertinents, fiables et qu'ils n'induisent pas de différences entre les apprenants (Schuwirth & Van der Vleuten, 2011). On parle ici de fonction diagnostique de l'évaluation (Bertrand, in Pelaccia 2016, p 348).

La fonction sélective de l'évaluation est l'objet d'attentions fortes, tant de la part du public, que de celle des communautés éducatives et des apprenants. En effet, la sélection des apprenants est souvent réalisée grâce à l'attribution de notes, parfois exprimées en lettre (plutôt dans les systèmes anglo-saxons). Une étude récente en France, réalisée auprès du grand public, rapporte

que 79% des français sont attachés aux notes et ne souhaitent pas les voir disparaitre (sondage BVA-Presse régionale – Foncia, 2017, in Demeester, 2020, p 195). De même, une enquête sur l'évaluation par la simulation que nous avons réalisée au moyen de focus groupes auprès d'étudiants de 4ème année de médecine mettait en évidence un attachement aux notes, pour des raisons de tri, d'équité et de classement (Philippon, 2017).

Le but ultime de l'évaluation, qui est en fait le but premier de l'approche par compétence, est d'assurer que les futurs professionnels sont compétents et donc qu'ils apporteront aux patients la sécurité des soins, au sein d'un système de santé dans lequel ils agissent de manière juste et appropriée. Les universités, chargées de la certification des futurs médecins ont ainsi une responsabilité sociale majeure à assumer tant dans l'identification des apprenants compétents que de ceux qui ne le sont pas et ne sont donc pas prêts à agir en professionnel (Boelen et al., 2000; Carraccio et al., 2016; Epstein, 2007). Pour Epstein, l'évaluation « protège le public » et, pour répondre à cet objectif ambitieux, les évaluations pratiquées dans les universités doivent donc être pertinentes, valides et fiables (Epstein, 2007).

2.1.4 Critères de qualité d'une évaluation

Deux principaux états évaluatifs existent au sein des institutions : l'évaluation « isolée », « unique » avec un seul but évaluatif, et les systèmes d'évaluation. Dans l'approche par compétence, il est clairement identifié que les évaluations doivent être liées les unes aux autres afin de représenter de manière fiable « la » compétence d'un futur professionnel, dans le but de mesurer ses différentes facettes, dans différents contextes (Van Der Vleuten, 1996). En 2010, Norcini et al. ont défini un cadre et des recommandations qui identifiaient des critères « *for good assessment* », que l'on traduira dans notre travail par « *évaluation de qualité* » (traduction libre), dans le contexte d'une évaluation unique. Au nombre de sept, et inspirés du travail de Van der Vleuten sur l'utilité d'une évaluation, qu'il définit par l'équation représentée ci-dessous, ils sont repris sans modification dans les dernières recommandations de 2018 (Tableau 1).

Tableau 1 – Cadre pour une évaluation de qualité, d'après Norcini et al (2010, 2018). Equation de l'utilité d'une évaluation (Van der Vleuten, 1996)	
1. Validité ou cohérence	Les résultats de l'évaluation sont adaptés à son but spécifique et reposent sur différentes preuves cohérentes.
2. Reproductibilité, Fiabilité, Consistance	Les résultats de l'évaluation seront les mêmes quelles que soient les circonstances. Désigne la constance d'un test
3. Equivalence/objectivité	La même évaluation conduit à un score ou à des décisions équivalentes selon les institutions ou moments d'évaluation
4. Faisabilité	L'évaluation est pratique, réaliste et sensible, selon les circonstances et le contexte.
5. Impact pédagogique	L'évaluation motive les futurs évalués à le préparer de telle sorte que cela aura un bénéfice pédagogique
6. Effet catalytique	L'évaluation fournit de la rétroaction de telle sorte que cela motive les parties prenantes à créer, améliorer et soutenir la pédagogie : elle conduit le futur apprentissage améliore la qualité globale du programme de formation.
7. Acceptabilité	Les différentes parties prenantes (apprenants, enseignants, institutions...) trouvent crédible les processus et résultats de l'évaluation.
Utilité = Fiabilité * Validité * Acceptabilité * Impact Pédagogique * Coût (Van der Vleuten, 1996)	

Les systèmes d'évaluation intègrent une série de mesures pour un même individu et permettent leur assemblage qui donne une représentation plus complète des différents habilités, aptitudes, connaissances des apprenants. En effet, il est impossible à un examen seul d'évaluer et de mesurer toutes les composantes des compétences autrement représentées par les habiletés cognitives, psychomotrices, relationnelles (Norcini et al., 2018). Même si le cadre des recommandations qui s'intéresse aux systèmes d'évaluation dépasse notre objet et notre question de recherche, nous le décrivons rapidement car il permet de situer notre travail.

Norcini décrit quatre situations évaluatives que l'on peut rencontrer au sein des différentes institutions. L'illustration qui en découle est tout à fait parlante puisqu'elle permet de représenter la situation de l'évaluation dans les études médicales en France et donc de représenter également le contexte de notre travail de recherche. Les quatre états qui sont représentés par

la figure sont les suivants : l'absence d'évaluation, une évaluation unique, une évaluation avec de multiples outils qui ne sont pas nécessairement liés les uns aux autres et une évaluation au sein d'un système d'évaluation qui crée des liens et du sens entre les différents outils d'évaluation (Figure 3).

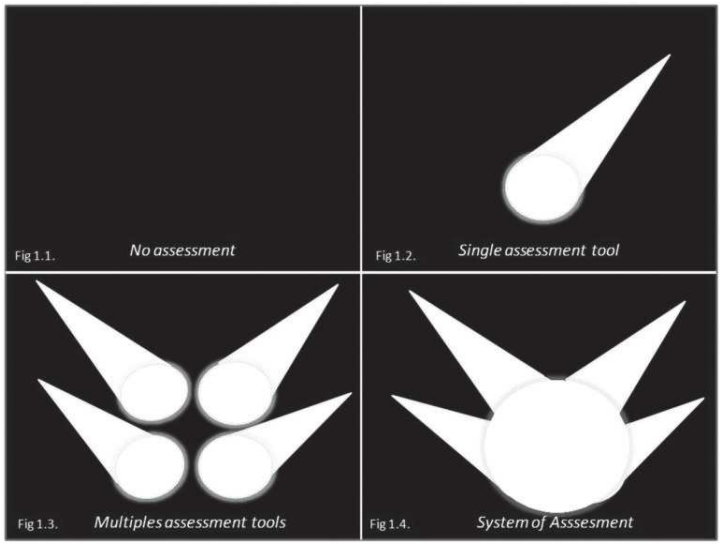


Figure 3 – Illustration des différents états évaluatifs, tirée de Norcini et al (2018)

En France, en ce qui concerne l'évaluation des compétences médicales et particulièrement de la compétence de clinicien, nous étions jusqu'à présent dans la première situation de la Figure 3, puisque sans évaluation valide des habiletés cliniques. La réforme du deuxième cycle va permettre une modification de la situation en créant des modalités d'évaluations certificatives basées sur les Examens Cliniques Objectifs Structurés (ECOS), qui permettent une mise en situation semi-authentique des apprenants. Nous serons donc en situation d'évaluations multiples (troisième situation) et si les différentes évaluations offrent une vision globale des capacités et habiletés des apprenants, en les triangulant, alors nous atteindrons la quatrième situation. Notre travail se situe dans la deuxième situation puisqu'il a pour objectif de développer un outil d'évaluation unique. Cependant, si cet outil s'avère valide, fiable, utilisable et acceptable, alors il

pourrait être intégré dans une perspective plus large d'évaluation, en complément d'autres outils, au sein d'un système spécifique.

Les critères de qualité d'un système d'évaluation sont les suivants : cohérence des évaluations, continuité entre les évaluations, composants des évaluations compréhensifs, faisable, guidé par un objectif final clair, acceptable par toutes les parties prenantes et transparents (Norcini et al., 2018). D'après les auteurs des recommandations, on identifie quatre types de systèmes d'évaluation, en fonction de leur objectif : les systèmes d'admission des apprenants, les systèmes qui délivrent des diplômes, ceux qui testent les progrès des apprenants et ceux qui s'intègrent dans le concept d'évaluation programmatique, décrit par Van der Vleuten en 2012 (Daniels & Pugh, 2018; Norcini et al., 2018; van der Vleuten et al., 2012).

2.2 Comment évaluer des compétences ?

Alors que dans les années 1980, les difficultés de mise en place de l'APC ont été attribuées à la difficulté de générer une évaluation de qualité, celle-ci reste le « talon d'Achille » de l'approche par compétence même si de nombreuses avancées pédagogiques ont permis de la développer (Carraccio & Englander, 2013).

Pour Tardif, dans le domaine d'évaluation des compétences, « l'évaluation semble correspondre à un processus social assez complexe où le jugement professionnel des enseignants est confronté à une multitude de critères, d'attentes, de besoins, de normes et de difficultés. En ce sens, l'évaluation est une opération difficile » (Tardif & Lessard, 1999, p 134).

Pour Englander, il faut remédier au caractère abstrait des compétences, à leur dépendance au contexte et à la difficulté d'harmonisation de leur enseignement en les reliant à l'activité via ce qui a été nommé par les pédagogues américains les « Entrustable Professional Activités » traduites par « Activités Professionnelles Confiables » (Englander & Carraccio, 2018), que nous décrirons ci-après.

Dans ce chapitre, nous aborderons les spécificités d'une évaluation dans l'approche par compétence, même si elle reste guidée par les principes de l'évaluation décrits ci-dessus, puis nous décrirons la place de l'évaluation par la simulation au sein d'une APC et dans le domaine de la médecine d'urgence.

2.2.1 Particularités de l'évaluation des compétences

Les particularités d'une évaluation de compétences découlent de la nature et des objectifs de l'approche par compétence, exposés ci-dessus. Afin d'établir des évaluations de qualité, il faut donc commencer la réflexion en s'appuyant sur les différents critères qui définissent une compétence. L'évaluation des compétences est en lien avec les critères de qualité d'une évaluation mais chacun de ces critères peut être éclairé par l'utilisation de l'approche par compétences. Ainsi, une évaluation de compétences sera logiquement guidée par l'objectif de la formation et donc par le patient. Pour s'assurer du transfert et de la maitrise des compétences, il faudra donc déployer des évaluations complexes, contextualisées et qui ciblent le soin et le patient (Schuwirth & van der Vleuten, 2011).

La compétence, de par son lien avec l'activité, se révèle en situation et grâce à la mobilisation de différentes habiletés, aptitudes, connaissances. Ces deux caractéristiques permettent de dire qu'elle est située et intégratrice (Demeester, 2020). Rappelons enfin que l'APC vise la pertinence plutôt que l'exhaustivité des savoirs, une approche globale plutôt que fragmentée des différentes composantes à acquérir pour devenir professionnel (Demeester, 2020, p 172; van der Vleuten et al., 2012). Ainsi, plusieurs principes sous-tendent l'évaluation des compétences professionnelles. Parent et Jouquan ont résumé en 2015 les différents principes directeurs d'une évaluation de compétence. Nous les avons regroupés en trois catégories : le contexte, la temporalité des évaluations et la formulation du jugement. Dans les paragraphes suivant, l'objectif est de rapporter ces principes directeurs et de les mettre en lumière avec les critères de qualité d'une évaluation (Jouquan & Parent, 2015, p 121).

Le contexte

Puisqu'une compétence ne peut pas s'affranchir de son lien à l'activité, son évaluation ne le peut pas non plus. Par conséquent, une évaluation de compétence se doit d'être authentique ou semi-authentique, et ne peut pas être coupée de l'activité à laquelle elle se réfère. Par exemple, une évaluation qui demanderait à un apprenant de « réciter » son cours sans ne le contextualiser ni mobiliser et organiser ses connaissances ne peut pas être considérée comme une évaluation authentique. L'objectif sera ici d'évaluer la maîtrise des ressources et leur mobilisation contextualisée en se concentrant sur les apprentissages importants et complexes plutôt que sur des apprentissages faciles tels que la mémorisation par exemple.

En 1990, Wiggins a décrit neuf principes qui déterminent l'authenticité d'une situation d'apprentissage : elle doit porter sur des problèmes et des défis stimulants, intégrer les caractéristiques réelles du contexte professionnel, en demandant à l'apprenant de réaliser une production ou une performance concrète, qui exigent la réalisation de tâches non routinières et complexes (pour éviter l'effet de mémorisation). Les situations à résoudre contiennent des indices suffisant pour les comprendre et les ancrer dans le réel, sans pour autant fournir des indications facilitant trop sa résolution. Elle permet de valoriser les apprenants qui font preuve de créativité et arrivent à montrer leurs habiletés personnelles. L'évaluation se base sur des critères clairs qui sont fournis et compris par les étudiants. Elle peut comporter des interactions entre l'évaluateur et l'évalué pour l'aider à la résolution du problème. Enfin, l'évaluation authentique demande que l'évaluation de la tâche s'intéresse au processus de résolution du problème (et donc aux stratégies cognitives et métacognitives utilisées par les apprenants) autant que du produit (la performance de l'apprenant) (Wiggins, 1990).

L'évaluation des compétences a donc pour point de départ des situations professionnelles, réunies sous le terme de familles de situations que chaque discipline médicale peut adopter. Il en existe 11 dans la spécialité de médecine générale. Selon Lemenu et al., une famille de

situation est « un ensemble de situations complexes représentatives de situations professionnelles courantes qui doivent être gérées par le futur diplômé novice et qui représentent suffisamment de caractéristiques communes pour mobiliser les mêmes compétences et capacités, dans les mêmes conditions de difficultés » (Lemenu & Heinen, 2015, p 84).

Une famille de situations est issue de situations professionnelles et pour les résoudre, l'apprenant doit mobiliser plusieurs compétences grâce à des ressources variées. Selon Raynal et Rieunier, une situation professionnelle de référence se caractérise par « une activité individuelle ou collective liée à la profession, un lieu géographique dans lequel s'exerce cette activité, l'objet sur lequel s'exerce l'activité et un produit caractérisant le résultat de cette activité ». (Raynal & Rieunier, 2012, cités par Lemenu & Heinen, 2015, p 85). Les familles de situation sont créées à partir du référentiel métier, en identifiant des situations professionnelles que l'on peut regrouper en famille (Lemenu & Heinen, 2015, p 89). En médecine, par exemple, les urgences vitales sont une famille de situation, de même que l'annonce d'une maladie grave ou encore la prise en charge et la gestion en aigüe de maladies chroniques. Une fois que les familles de situation sont identifiées par les formateurs, il faut créer des situations d'intégration qui constitueront une activité d'apprentissage en lien avec des situations professionnelles et les savoirs acquis pendant la formation. Elles permettent à la fois d'exercer des compétences puis de les évaluer (Roegiers, 2012, p 280). Chaque situation est composée d'invariants qui sont les suivants :

- l'habillage (la mise en scène : tenue médicale, salle d'accueil des urgences vitales, Service Mobile d'Urgence et de Réanimation : camion de secours médicalisé),
- le contexte (patient, rôle de l'étudiant, équipe pluridisciplinaire),
- la tâche à réaliser (prendre en charge une détresse respiratoire aigüe),
- les informations délivrées à l'apprenant (une partie de l'histoire du patient),
- les ressources à dispositions de l'apprenant (un téléphone, matériel médical, ordinateur).

A l'issue de la situation, l'apprenant aura mis à jour les acquis d'apprentissage dont il avait besoin pour résoudre la situation complexe (Lemenu & Heinen, 2015, p 100).

Ainsi l'évaluation des compétences, qui doit tenir compte du contexte et des situations, ne peut se faire que dans la mise en place ou l'observation de situations complexes. Par exemple, grâce à l'emploi de supports tels que le carnet de bord, le portfolio, les examens cliniques objectifs structuré (ECOS), le travail en groupe, un exposé ou encore la situation simulée qui est la situation à laquelle nous nous intéressons (Lemenu & Heinen, 2015, p 107). La situation d'évaluation a ainsi du sens, puisqu'elle se réfère directement à une situation professionnelle et qu'elle permet d'évaluer des compétences dans leur globalité.

Idéalement, une évaluation dans l'approche par compétence se déroule sur le lieu de travail, mais cela ne suffit pas. Van Der Vleuten rappelle en effet que le contenu de l'évaluation participe à sa fiabilité. En effet, il est impossible de conclure à la compétence d'un apprenant si on l'a observé dans une seule situation. Il est question ici de fiabilité de l'évaluation qu'il ne faut pas réduire, ainsi que Van Der Vleuten le souligne à la seule reproductibilité des situations ou des évaluateurs. Une évaluation est valide et fiable si elle reflète l'ensemble des compétences requises par le futur professionnel. Ainsi, lors d'une évaluation à enjeu élevée (certification ou à la fin d'un enseignement), il faut s'assurer que le contenu de l'enseignement est représenté non pas de manière exhaustive, mais pertinente. De même, la validité qui s'intéresse à la signification et au sens de la mesure obtenue ne peut pas être obtenue si l'évaluation n'est pas située et n'est pas intégratrice.

Une évaluation qui repose sur l'analyse de la progression des apprenants

En 2013, Pangaro et Ten Cate ont identifié trois cadres évaluatifs existants, dans le but de guider les évaluateurs. Il s'agit du cadre analytique qui évalue en analysant chaque composant du construit à évaluer et ne peut pas être utilisé dans l'APC. Le deuxième cadre est le cadre

systémique qui consiste en une vision holistique de l'évaluation qui doit se situer dans le monde réel (professionnel) et s'attache à l'évaluation de performance. Le dernier cadre, que les auteurs appellent développemental s'intéresse au développement progressif de l'apprenant, son objectif étant de permettre à l'évaluateur et à l'apprenant de définir le niveau de compétence observé lors de plusieurs évaluations (Pangaro & Ten Cate, 2013).

Dans une approche par compétence, il faut en effet permettre à l'apprenant de se situer lui-même par rapport à l'objectif final de maîtrise et lui proposer des évaluations régulières en lui fournissant une rétroaction permanente. On se situe ici dans le domaine de l'évaluation formative, indispensable aux évaluations sommatives ou finales. Certains systèmes offrent cette liberté aux apprenants et leurs permettent de se soumettre à l'évaluation finale lorsqu'ils se sentent prêts, ce qui leur permet d'évoluer à leur rythme (il s'agit des systèmes de variabilité des temps d'apprentissage, décrits plus hauts) (Lucey et al., 2018). Ainsi, d'après Jouquan et Parent, les évaluations sur le degré de maîtrise des ressources devraient être fréquentes et les évaluations qui portent sur le niveau de développement des apprenants devraient être périodiques (moins fréquentes) et porter sur des situations-problèmes (Jouquan & Parent, 2015) .

Dans une évaluation de compétences, il est tout d'abord fondamental de suivre la trajectoire de développement des compétences grâce à des niveaux de développement de compétences de complexité croissante. Dans cet objectif ont été pensé les « milestones », que l'on peut traduire par « jalons », dès la création de l'approche par compétence (Batalden et al., 2002; Carraccio et al., 2016). Les milestones correspondent à la description de différents niveaux de développement des compétences, et sont des indicateurs qualitatifs qui permettent de circonscrire les objectifs qu'un étudiant doit atteindre pour chaque niveau de compétence. Les milestones décrivent les sous-compétences et les niveaux de progressions différents des apprenants. Le tableau 2 illustre un exemple d'échelle d'évaluation basée sur l'appréciation d'une sous-compétence (apprendre et s'améliorer grâce à la rétroaction), décrite dans le référentiel de médecine interne aux USA.

Tableau 2 – Exemple de Milestone pour les internes de médecine interne, (traduction libre), (ACGME 2015, Yudlowski et al, p 9)

Exemple d'évaluation à l'aide de milestones, pour la sous-compétence « apprendre grâce à la rétroaction (RA) »				
Lacunes critiques			Prêt pour la pratique en autonomie	Objectif atteint
- Ne sollicite jamais de RA - Résiste à la RA qui lui est proposée	-Demande rarement une RA - est sur la défense si reçoit RA non sollicitée - ajustement trop superficiel ou temporaire suite à la RA	- sollicite uniquement la RA des superviseurs - ouvert à la rétroaction non sollicitée - incorpore la RA de manière inconstante	- demande RA de la part de toute l'équipe et des patients - accueille volontiers la RA non sollicitée - Incorpore la RA de manière constante à sa pratique	- la performance reflète continuellement l'incorporation de RA sollicitée ou non - capable de concilier des RA variables ou contradictoires
☐	☐	☐	☐	☐
Commentaires de l'évaluateur :				

Traditionnellement, les niveaux de développement des compétences sont décrits grâce à l'échelle de Dreyfus et Dreyfus, qui identifie cinq niveaux de développement : novice, débutant avancé, compétent, efficient, expert. L'objectif final étant de former des apprenants compétents au premier jour d'une prise de responsabilités, mais, par la suite, de poursuivre le développement personnel pour atteindre un niveau expert (Dreyfus & Dreyfus, 1987). Ces différents niveaux de développement ont souvent été intégrés dans les outils qui utilisaient des échelles d'évaluation globale (global rating scales, en anglais). Une fois les différents niveaux de développement identifiés, il faut pouvoir porter un jugement et sa nature déclenche de nombreuses réflexions dans le cadre de l'approche par compétence.

Par ailleurs, Schuwirth souligne que la compétence n'est pas le niveau « terminal » à atteindre pour un professionnel. En effet, de compétent, un professionnel deviendra expert. Il est important d'avoir cet élément en tête dans le cadre de l'évaluation car un professionnel expert ne raisonne pas de la même manière qu'un professionnel compétent. Schuwirth cite en exemple un expert

qui fait un diagnostic exact en posant seulement les trois questions pertinentes au patient, alors que le professionnel compétent, qui fera le même diagnostic aura eu besoin de poser plus de questions, d'être plus systématique dans sa recherche d'éléments contributifs au diagnostic. Il en est ainsi pour les différents domaines de compétence en médecine et dans les sciences de la santé (Schuwirth & van der Vleuten, 2011).

Le jugement évaluatif

Le processus de jugement est primordial dans celui de l'évaluation puisqu'il en est l'étape centrale, qu'il soit représenté par une mesure quantitative ou qualitative (Fontaine & Loye, 2017). Dans une approche par compétence, la question de la nature des habiletés évaluées est une des discussions centrales. En effet, en situation clinique, en situation simulée ou lors d'un oral, on évalue la performance d'un apprenant, définie par « ce que l'apprenant peut faire ». Il s'agit donc d'observer un comportement, des attitudes et la mobilisation de connaissances, mais sans avoir accès aux processus ou aux ressources internes qui ont conduit à la performance observée (Houzé-Cerfon et al., 2020; Yudkowsky et al., 2019, p 6).

Il est important d'avoir à l'esprit que bien souvent, les mesures dites de « compétences » sont en réalité des mesures de performances, c'est-à-dire de comportement observé, mais que la globalité (et pas l'addition) de ces mesures, si elles sont valides, constitue des inférences et représente une certaine forme de compétence. De nombreux auteurs utilisent l'un ou l'autre terme puisqu'ils se situent dans l'APC, mais le terme juste lorsqu'il s'agit d'évaluer un comportement est celui de performance. Ce point précis explique en grande partie les difficultés d'évaluation des compétences, qui ne devraient pas être morcelées, pas être uniquement observées et qui restent abstraites.

Gilbert, dans un livre intitulé l'évaluation de la performance individuelle, paru en 2017, rappelle que la notion de performance, tout comme celle de compétence, reste confuse et mal définie.

Même si sa réflexion est située dans le domaine des sciences de la gestion, il s'adresse aux enseignants et aux chercheurs qui se questionnent sur l'évaluation de la performance individuelle. Il rappelle les trois principaux sens donnés à la performance par Bourguignon en 1999. Le premier sens est celui du succès, dans le sens d'un jugement de valeur positif et qui appelle un système normatif, de notation. Le second sens donné à la performance est celui du résultat d'une action par rapport à des degrés d'acceptation de buts, d'objectifs d'un programme. Le dernier définit la performance comme une action, elle EST action, mise en acte d'une compétence et s'illustre dans le processus de l'action plutôt que dans son résultat (Gilbert & Yalenios, 2017, p 11-28).

Gilbert émet l'idée que le lien entre compétence et performance, puis potentiel d'un individu est linéaire dans le sens d'une certaine temporalité. Pour lui, « la compétence est une hypothèse. Son intérêt est de contribuer à expliquer, et éventuellement à modifier par des actions adéquates (en particulier, la formation), les conduites des individus conduisant à une performance. Quant au potentiel, il est en quelque sorte une virtualité : la capacité à acquérir des compétences nouvelles. Ces deux notions sont à la fois en amont et en aval de la performance. En amont, parce qu'elles sont supposées à l'origine d'une performance passée ou à venir. Mais aussi en aval, car que dire de la performance individuelle, et *a fortiori* du potentiel, si l'on n'a aucune idée de la performance de l'individu considéré ? » (Gilbert et Yalenios, 2017, p 13-14). La performance serait donc « située » entre deux autres notions. Tout comme la compétence, elle ne peut s'exprimer qu'en action, qu'en lien avec l'activité.

Afin de faire le lien entre pratique clinique observée et compétences, sous-compétences et milestones, les activités professionnelles confiables ont vu le jour au milieu des années 2000, amenant avec elles à la fois une solution pour évaluer les compétences, mais également l'introduction de la notion de confiance dans la compétence de l'apprenant. L'idée des activités professionnelles confiables est qu'elles sont issues du terrain via l'observation, mais plus souvent via des consensus d'experts, l'observation étant jugée trop complexe (Englander & Carraccio,

2018). Elles sont définies en 2005 par Ten Cate qui les décrit comme les: « critical activities in professional life of physicians that the specialty community agrees must be assessed and approved at some point in the ongoing formation of physicians »[16] (Ten Cate, 2005). En 2018, Englander précise cette notion de supervision puis d'autonomie au sein de la définition de l'activité professionnelle confiable : "essential task of a discipline (prof/spe/subspe) that a learner can be tested to perform without direct supervision and an individual entering practice can perform unsupervised in a given healthcare context, once sufficient competence has been demonstrated"[17].

L'avantage des Entrustable Professional Activities (EPA) est qu'elles sont les mêmes pour toute la formation et au-delà, de l'externat à la vie professionnelle. Issues de situations cliniques de départ, définies par les différentes instances pédagogiques (il en existe 265 en France et elles ont été définies en 2020, à partir des situations cliniques du programme de formation suisse, lui-même inspiré du programme CanMeds), elles sont observables et mesurables, en lien avec le terrain. Elles constituent le reflet indirect de la compétence d'un apprenant et constituent l'objectif « concret » de la formation médicale, représenté par les compétences que doit mobiliser un externe le premier jour de son internat ou un interne le premier jour de sa vie professionnelle (Ten Cate, 2005; Englander & Carraccio, 2018). La notion de confiance qui est ajoutée dans leur définition est illustrée ainsi par Ten Cate : "performing well in a profession could be defined as being entrusted to carry out all its critical EPAs"[18].

Ainsi, à partir des compétences « générales » décrites en 2002 mais difficilement mesurables, ont été créés les jalons pour décrire les niveaux de développements, et les activités

[16] « Activités critiques » de la vie professionnelle médicale, reconnues par la communauté médicale comme étant évaluée et certifiées tout au long de la formation médicale » (traduction libre)
[17] « Tâche essentielle d'une discipline pour laquelle un apprenant peut être évalué sans supervision directe et qu'un individu entrant en pratique peut exécuter sans supervision dans un contexte donné de soins de santé, une fois qu'une compétence suffisante a été démontrée » (traduction libre)
[18] « Être performant au sein d'une profession pourrait être défini comme le fait de se voir confier toutes les activités professionnelles confiables indispensables » (traduction libre)

professionnelles confiables pour faire le lien avec l'activité. En faisant le parallèle avec l'apport d'auteurs comme Gérard et Scallon, on peut les comparer aux acquis d'apprentissage qui seraient la part visible, représentative de la compétence d'un apprenant (De Ketele J.-M. & Gérard, 2005; Lemenu & Heinen, 2015, p 9-10; Scallon, 2007, p 100). Demeester et Nguyen rappellent tout de même qu'il faut faire attention à ne pas trop morceler la compétence avec de multiples indicateurs ou avec des indicateurs trop simples, sous peine de menacer la validité de sa mesure (Demeester, 2020, p 174-177 ; Nguyen & Blais, 2007).

Une fois que les activités professionnelles confiables peuvent être considérées comme des marqueurs de performance, se pose la question de l'expression de la mesure et donc la question de la nature de la rétroaction fournie à l'apprenant. Dans un système d'évaluation de compétences « idéal », une note, qu'elle soit sous forme de lettre ou de chiffre n'a pas de sens si elle ne permet pas à l'apprenant de se situer par rapport à ses objectifs ou par rapport à l'objectif final de la formation. La question du seuil de validation d'une évaluation se pose alors en ces termes : « à partir de combien suis-je compétent et est-ce qu'avoir 10/20 représente une performance qui reflète la compétence ou bien est-ce juste une performance « intermédiaire ? ». On peut ainsi parler de niveau « acceptable » et s'accorder sur le fait que 10/20 ne semble pas être un objectif pour certifier un futur professionnel de santé autonome (Demeester, 2020, p175).

Un jugement évaluatif pertinent dans une approche par compétence est plutôt caractérisé par des appréciations en rapport avec le niveau d'acquisition et c'est ainsi que les activités professionnelles confiables sont évaluées : ne peut pas faire seul, peut faire avec une supervision, peut faire en autonomie. Ainsi, l'APC engage apprenants, enseignants et institutions à modifier l'utilisation des notes pour les remplacer petit à petit par des marqueurs plus « pertinents ». Il s'agit d'un changement majeur dans le fonctionnement des systèmes évaluatifs au sein desquels la majeure partie des usagers reste attachée à la note. Cela fait partie du critère d'acceptabilité d'une évaluation et reste un défi pour la majeure partie des universités.

Une autre conséquence de l'APC dans le mécanisme d'évaluation et de jugement évaluatif est la constatation du fait que les compétences ne se compensent pas entre elles. En effet, on ne peut pas accepter qu'un futur professionnel, tout excellent clinicien qu'il soit, n'ait aucune compétence communicationnelle ou scientifique, ou bien qu'un futur médecin urgentiste n'ait pas obtenu un niveau de développement satisfaisant en matière de pédiatrie. Dans un système par compétence « la somme des parties n'est pas égale au tout », et cela s'applique pour les notes qui en morcelant les acquis d'apprentissage ne représentent pas de manière valide la compétence, mais également pour les systèmes d'évaluation au sein des Unités d'Enseignement (UE) qui permettent la compensation des compétences entre elles (Demeester, 2020, p176). Il en découle donc un changement de fonctionnement dans les systèmes évaluatifs, qui pouvaient permettre jusqu'à présent à certaines notes de se compenser entre elles pour pouvoir franchir les différentes étapes de la formation.

En France, l'université d'Aix-Marseille est une des seules universités à avoir adapté une approche par compétences lors de la rentrée 2018 et pour quelques-unes de ses composantes : la Faculté des Sciences et du Sport, l'Ecole Supérieure du Professorat et de l'Education (ESPE), la licence Santé Visuelle et la Faculté de Droit et de Sciences Politiques. Lorsqu'il a fallu revenir sur la possibilité offerte par les UE de se compenser entre elles, cela n'a pas été possible, les syndicats étudiants considérant que cette possibilité de compensation était un droit acquis et qu'une modification de ce droit constituait un retour en arrière (Demeester, 2020, p 195).

Ainsi, les deux principes « révolutionnaires » d'une évaluation de compétence qui sont l'absence de compensation et de hiérarchisation entre les compétences, associée à la difficulté d'utiliser des notes pour évaluer les compétences ne sont pas encore admis, ni dans la société (cf sondage BVA rapporté ci-dessus) ni par les acteurs universitaires (au moins les étudiants dans ce cas). Cependant, une composante de l'Université d'Aix-Marseille (l'ESPE) a finalement adopté l'abandon des notes, au profit de l'expression de quatre niveaux : débutant, intermédiaire,

compétent, expert. L'expérimentation est en cours jusqu'en 2023 et bien qu'il soit tôt pour en tirer des conclusions, l'observation réalisée après un an de mise en place semble conclure à la faisabilité d'une telle approche (Demeester, 2020, p 197).

Ainsi, l'évaluation au sein de l'APC présente plusieurs spécificités qu'il faut appréhender pour développer les bons outils d'enseignement puis d'évaluation. Une partie des éléments ci-dessus a été pris en compte dans la réforme du deuxième cycle des études médicales, en cours de réflexion et de mise en place puisque le dernier décret date de septembre 2021 et doit être mis en application à partir de l'année universitaire 2022-23 (Décret n° 2021-1156, 2021). Le texte précise une réduction des connaissances à maîtriser à la fin du deuxième cycle pour donner plus de place à l'apprentissage d'habiletés techniques, de communication ou encore du professionnalisme. Même si l'application de la réforme reste à faire, cela représente une petite révolution dans le monde médical, chaque spécialité ayant dû se résoudre à diminuer le volume de connaissances à maîtriser au premier jour de l'internat. La démarche n'est pas sans rappeler les critiques des étudiants, relevées dans l'étude préliminaire à cette thèse. Ils pointaient en effet des études surspécialisées, qui ne les préparaient pas à devenir médecin, mais plutôt à être classés à l'issue du deuxième cycle pour être en capacité de choisir la spécialité qui leur conviendrait le mieux. Ainsi, à la fois les évaluations et l'enseignement qui en découlaient étaient jugés inadaptés, laissant de côté les compétences nécessaires pour la pratique de leur futur métier et ne les aidant pas lors des stages hospitaliers (Philippon et al., 2021).

Le challenge de l'évaluation au sein de l'APC réside donc dans l'apprentissage de nouveaux domaines (tels que le professionnalisme, le travail en équipe qui était autrefois appris « sur le tas », lors des stages hospitaliers), mais également dans la nécessité d'utiliser de nombreux outils d'évaluation qui « cernent » la performance des apprenants pour mieux décrire leur compétence. Se jouent alors des questions de standardisation des évaluations, qui reste nécessaire pour respecter l'équité de la formation et la distribution des étudiants dans les différents programmes

de formation et également des questions d'alignement pédagogique entre enseignement, activités d'apprentissage et d'évaluation.

2.2.2. Outils d'évaluation des compétences

De nombreux outils d'évaluation de compétences existent et nous exposerons ici ceux qui sont utilisés le plus couramment afin d'avoir des moyens de comparaison avec ce que propose la simulation. Avant de se poser la question de la description d'un outil, il s'agit de savoir repérer les compétences et habiletés que l'on souhaite étudier. Se pose donc la question de la validité de l'outil, dans le sens du degré de conformité avec lequel il évalue la compétence visée. Dans une approche par compétence, il convient également de déterminer le niveau visé (avec l'approche de Dreyfus ou celle des activités professionnelles confiables). Afin de décrire les types d'apprentissages et les méthodes d'évaluation correspondantes, il est possible d'avoir recours aux taxonomies qui recensent les objectifs d'apprentissages, les différents niveaux visés par l'évaluation et qui font le lien avec les outils et méthodes disponibles.

Différents « niveaux d'évaluation » : les taxonomies de Bloom et de Miller

Puisque les compétences ont plusieurs facettes, ou composantes, il est important d'identifier lesquelles sont identifiées par les différents outils à disposition des évaluateurs. Plusieurs taxonomies existent, mais les deux les plus utilisées sont celles de Bloom, créée en 1956, et celle de Miller, créée en 1990.

La première identifie trois registres possibles d'un objectif pédagogique : le cognitif, l'affectif et le psychomoteur, appelés Knowledge, Attitudes ans Skills en anglais. Les registres sont en lien avec les trois dimensions d'un objectif pédagogique : le savoir, savoir-être et savoir-faire. Pour chaque registre, plusieurs niveaux de performance sont décrits, et pour chacun d'eux, les outils d'évaluation qui correspondent, un outil étant capable évaluer plusieurs niveaux (Bloom B et al., 1956). Cette taxonomie, bien qu'utile pour identifier différents éléments présents dans

l'apprentissage de compétences et dans les objectifs d'apprentissage, est plus en lien avec une pédagogie fondée sur les objectifs plutôt que sur les compétences, car, au lieu de rassembler les différentes composantes des compétences, elle les morcelle, pour les réduire à des objectifs observables et évaluables (Jouquan, 2007; Nguyen & Blais, 2007) .

La pyramide de Miller, décrite en 1990, comporte quatre niveaux :

- le « *knows »* : l'apprenant « sait », il a des connaissances et il est capable de les restituer. L'évaluation adaptée à ce niveau est une évaluation de connaissances, telle que des questions à choix multiples (QCM), des questions à réponse ouverte courte (QROC) ou des questions rédactionnelles.
- le « *knows how »* : l'apprenant « sait comment » il peut appliquer ses connaissances, les utiliser pour interpréter des données professionnelles. Un outil tel que la résolution de problème, ou problem-based learning permet d'évaluer ce niveau intermédiaire.
- le « *shows how* » : l'apprenant est capable de « démontrer » de quelle manière il mobilise ses connaissances, dans un environnement semi-authentique, tel qu'avec les ECOS ou la simulation. Les tests de concordance de script, qui s'intéressent au raisonnement clinique entrent également dans cette catégorie, de même que les portfolios avec des récits de situation complexes et authentiques, analysées par l'apprenant. Il s'agit donc d'une évaluation de performances.
- le « *does* » : l'apprenant a la capacité d'agir de manière autonome, en situation réelle. L'évaluation se déroule donc sur le lieu de travail, avec mise en œuvre de la pratique clinique à l'aide d'outils variés tels que la supervision directe, l'analyse rétrospective de dossiers médicaux ou de courriers, la présentation de cas ou encore l'évaluation d'une ou de plusieurs situations cliniques, grâce au Mini CEx (Miller, 1990).

On peut enfin citer le modèle de Cambridge qui s'intéresse au sommet de la pyramide de Miller et à la complexité de l'évaluation en situation réelle, en rappelant qu'il faut prendre en compte, dans une telle évaluation, deux types d'influence qui se situent au-delà de la compétence de l'apprenant. Il s'agit du système dans lequel évolue les acteurs de l'évaluation (l'apprenant et le patient, ainsi que l'équipe hospitalière ou ambulatoire). En effet, il ne faut pas omettre les règles émises par les différents systèmes (gouvernement, faculté, hôpital, recommandations professionnelles etc.) et l'influence qu'elles ont dans l'interaction entre l'apprenant et le patient. La deuxième influence de la compétence d'un individu est l'individu lui-même, son état d'esprit, de santé physique et mentale ainsi que les éléments de sa vie privée. Ces éléments peuvent finalement être rapportés à une partie du « contexte » au sein duquel il est impératif d'évaluer les compétences, mais qui complexifie leur appréciation (Holmboe & Iobst, 2020, p 6).

Outils à disposition des formateurs pour évaluer des compétences

Dans la seconde édition du livre intitulé « Assessment in Health Professions Education », paru en 2020, les auteurs identifient quatre grandes méthodes d'évaluation au sein des sciences de la santé : les évaluations écrites, les évaluations orales, les évaluations de performance et les évaluations sur le lieu de travail (workplace-based assessment). Deux autres méthodes, plus transversales et qui se classent dans plusieurs des catégories précédentes, sont également décrites : l'évaluation narrative et les portfolios (Yudkowsky et al., 2019).

Nous décrirons ici succinctement les différentes méthodes et leur capacité à évaluer les compétences, afin d'y placer l'évaluation par la simulation pour ensuite la décrire en détail dans le chapitre suivant. Pour décrire ces différentes méthodes, nous nous sommes appuyés sur des ouvrages paru entre 2018 et 2020 à savoir, le livre de Yudkowsky, Soo Park et Downing « Assessment in Health Profession Education », sur le manuel d'Eric Holmboe intitulé « Evaluation of clinical competence », sur la dernière version de la « boite à outils » de l'ACGME,

ainsi que des articles de référence sur chacun des outils, que nous citerons au fur et à mesure (Holmboe et al., 2018; Holmboe & Iobst, 2020; Yudkowsky et al., 2019).

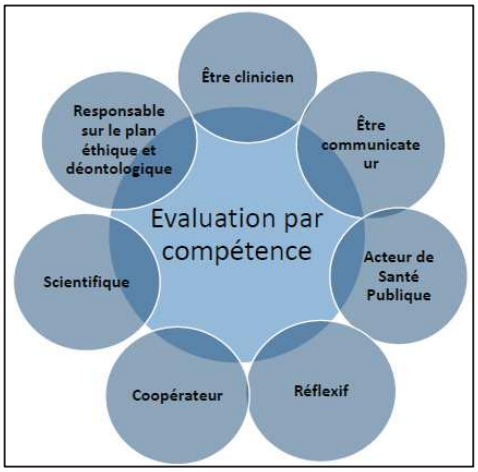


Figure 4 – 7 domaines de compétence définis dans la formation médicale française

Parmi les 7 domaines de compétences retenues par la conférence des doyens français et qui seront les compétences fondatrices des programmes de formation des futurs médecins, les outils développés suite à notre questionnement pour évaluer les apprenants dans le domaine de la médecine d'urgence, évalueront la compétence de clinicien et de communicateur (Figure 4). Utilisés et mesurant à cette étape de leur développement l'observation seule, ils ne mesurent pas réellement la compétence de réflexivité, même si cela fait partie de l'un des objectifs de l'utilisation de la simulation. Nous décrirons donc dans ce chapitre les outils qui permettent d'évaluer ces compétences spécifiques, grâce aux 4 méthodes décrites par Holmboe.

Les évaluations écrites, encore représentées en majorité par les QCM ou les QROC, permettent d'évaluer principalement les connaissances médicales, dans les registres des connaissances déclaratives. Ayant des qualités docimologiques incontestables, elles continuent de jouer un rôle important dans l'évaluation, et notamment pour les étudiants de deuxième cycle. En effet, elles sont complémentaires aux autres méthodes d'évaluation et, même si elles doivent être rédigées avec rigueur, réflexion pour répondre aux critères de qualité d'une évaluation, elles peuvent être

valides (si elles représentent bien l'ensemble du contenu à évaluer et qu'elles discriminent les apprenants), fiables (dans la mesure où elles sont reproductibles et cohérentes). Elles sont facilement utilisables et faisables, à moindre coût et sont encore largement acceptées par les enseignants et les étudiants (Holmboe, 2018 p 9 ; 2020, p 12). De plus, lorsque les résultats et données de ces évaluations sont utilisées pour réaliser une rétroaction auprès de l'apprenant, alors elles permettent à l'apprenant de se placer dans une démarche d'amélioration et d'évolution. Aux Etats-Unis, il existe les In-training exam, qui sont des évaluations écrites composées de QCM, mais réalisées en lien avec la thématique du stage hospitalier, dans un esprit de continuité pour l'apprentissage et l'évaluation des connaissances (Holmboe & Iobst, 2020).

On peut ajouter, dans ce registre, les outils écrits qui permettent d'évaluer plus spécifiquement le raisonnement clinique des apprenants. Récemment recensés dans une revue de la littérature, ils sont représentés par de nombreux exercices, qui rapportent une situation, comme une sorte de simulation écrite. En voici quelques outils, parmi les plus utilisés actuellement :

- les QCM (à condition que ces dernières ne demandent pas seulement une restitution, mais également une mobilisation et une combinaison des connaissances pour trouver la réponse correcte),
- les questions pour lesquelles il faut choisir la meilleure réponse parmi des distracteurs qui sont tous potentiellement justes (single best answer ou extending matching questions)
- les dossiers progressifs rédactionnels qui demandent aux apprenants de rédiger un court texte argumentant leur prise de décision dans une situation clinique (modified essay questions)
- les « patient management problems », constitués de scénario cliniques complexes, au sein desquels les apprenants doivent exercer un raisonnement pour prendre des décisions qui leur permettent d'avancer dans la prise en charge
- les tests de concordance de scripts (Daniel, 2019).

Ainsi, les évaluations écrites se situent aux niveaux 1 et 2 de la pyramide de Miller, évaluant les connaissances des apprenants, mais également, grâce à certains outils, leurs capacités à les mobiliser en situation complexe, en utilisant par exemple un cas clinique issu de la réalité de la pratique.

En face à face ou face à plusieurs évaluateurs, l'apprenant mis en situation d'évaluation orale aura pour objectif de mobiliser ses connaissances pour investiguer et tenter de résoudre une situation clinique. C'est le cas avec l'apprentissage par résolution de problème, initialement créé pour un travail de groupe où chaque apprenant apporte une part de construction à la résolution d'une problématique clinique. Cependant, le même déroulement peut être réalisé avec un seul apprenant, lors d'une évaluation sommative.

Les évaluations orales permettent d'évaluer la compétence de clinicien, mais également celle de communicateur. Au sein de la compétence de clinicien, elles s'intéressent au raisonnement clinique. Il existe un score d'évaluation du raisonnement clinique, qui a pour objectif d'être utilisé en observation directe, mais qui peut également l'être en situation simulée ou lors d'un oral. Elles permettent alors de se situer au troisième et quatrième niveau de la pyramide de Miller. En effet, l'apprenant peut montrer comment il mobilise ses connaissances pour résoudre des problèmes complexes et prendre des décisions, mais il peut également montrer sa capacité à formuler des idées, à les synthétiser et à les adapter à un cas clinique contextualisé. De plus, il montre son habileté à communiquer et à mobiliser ses capacités et montre ses compétences personnelles (interpersonal skills en anglais) (Yudkowsky et al., p 127-140, 2019). Les évaluations orales permettent également d'appréhender les questions éthiques et le professionnalisme des apprenants (niveau 3 de Miller). Celles qui sont le plus utilisées outre-Atlantique sont : le mini multiple interview (MMI) et le Chart-Stimulated Recall, mais qui est en lien avec le lieu de travail puisqu'il s'agit de poser des questions quant à la situation d'un patient puis de délivrer une rétroaction sur les réponses de l'apprenant (Philibert, 2018).

Organisées avec de multiples oraux, en utilisant un tableau de type blueprint pour identifier et multiplier les différents contenus évalués, associés à l'utilisation de scores validés et de formateurs formés au sein d'une organisation rigoureuse et anticipée, les évaluations orales ont tout à fait leur place dans une évaluation de compétence, au sein d'un programme d'évaluation.

Les évaluations de performance regroupent de nombreuses méthodes d'évaluation car elles s'intéressent à ce qu'un apprenant peut faire plutôt qu'à ce qu'il sait ou connait. Ainsi, les oraux font également partie de cette catégorie, dès lors qu'ils permettent à l'apprenant de montrer ce qu'il peut faire : raisonner, diagnostiquer etc. L'évaluation basée sur la simulation (simulation-based assessment, SBA), définie par le fait d'évaluer des compétences grâce à la mise en situation simulée fait partie de cette catégorie d'évaluations.

Le terme simulation fait référence à la représentation donnée d'une tâche ou d'une situation réelle. Ainsi, un oral avec l'évaluateur qui « joue » le patient peut être considérée comme de la simulation, de même qu'une évaluation procédurale sur un dispositif simulé (une tête d'intubation par exemple) ou encore via une simulation en ligne, grâce à des patients et à des situations virtualisées. Les examens cliniques objectifs structurés (ECOS) entrent dans cette catégorie d'évaluation des performances grâce à la simulation car ils sont mis en œuvre avec des patients simulés et standardisés ou avec des mannequins ou dispositifs simulés. Nous détaillerons les différentes évaluations par simulation que les institutions ont à disposition dans le chapitre suivant.

Les évaluations sur le lieu de travail (workplace-based assessment) concernent l'évaluation d'entretiens avec un patients, des familles, mais également les relations de l'étudiant avec l'équipe hospitalière, la secrétaire, ou encore une évaluation de compte-rendu ou d'observations médicales (Holmboe E., During S., Hawkins, R., p 61-90, 2018). Ces évaluations fondées sur des sources et des regards variés sont nommées « évaluation 360° » ou Multi-source feedback (MSF). Les évaluations sur le lieu de travail vont des observations informelles des apprenants à

des systèmes formels et parfois complexes d'évaluation, reposant sur le recueil de données multiples illustrant la performance de l'apprenant au sein du contexte clinique (Yudkowsky, 2020, p 7).

De nombreux outils d'évaluation par observation directe existent, mais le plus répandu, qui a démontré sa validité et qui est utilisé dans de nombreuses disciplines, à la fois pour des internes et des externes est le Mini-CEX. Il s'intéresse à l'évaluation de la pratique clinique et notamment aux connaissances, habiletés et attitudes des apprenants. Il peut être complété par un outil d'évaluation des procédures techniques tels que l'OSATS ou le DOPS (Norcini & Burch, 2007; Reznick et al., 1997; Wragg et al., 2003). Une récente revue de la littérature lui retrouve une fiabilité qualifiée de modérée à élevée (et représentée par un alpha de Cronbach variant de 0.59 à 0.97), mais pour un nombre de stations variant de 5 à 60. Les auteurs soulignent aussi que la fiabilité dépend également du niveau l'apprenant (Kogan et al., 2009; Hejri et al., 2020). Le mini-CEX est également corrélé aux résultats des autres évaluations de performance clinique. Outre ses qualités psychométriques, il peut être utilisé à la fois en évaluation formative puisqu'il fournit une rétroaction enrichissante pour l'apprenant, et facilite la réflexivité, mais également en évaluation sommative car il permet de discriminer les apprenants, et qu'il procure une supervision de qualité.

L'évaluation des comptes rendus d'hospitalisation ou opératoires est également pratiquée outre-Atlantique et permet d'ajouter, à l'évaluation de performance et de compétence clinique, une évaluation de la qualité globale des soins délivrés aux patients, et ainsi, une évaluation de la sécurité des soins prodigués par l'apprenant (Holmboe & Iobst, 2020). Ainsi, au sein d'une approche par compétences, l'évaluation par observation directe « should be an essential component of the outcome-based education and certification » (Kogan, 2009).

Tableau 3 – Evaluer des compétences médicales : différents outils à disposition des institutions, place de la simulation, d'après l'ACGME, 2020

Domaine de compétence	**Miller**	**Méthode**	**Outil**
Clinicien *(Patient care and Procedural Skills) (medical knowledge and clinical reasoning)*	"Knows" "Knows how" "Shows how" "Does"	Tests écrits Tests écrits Oraux Evaluation de performance Evaluation sur le lieu de travail	QCM de connaissance, ITE, QCM Key features, Test de concordance de script PBL, Vignette clinique Simulation Observation directe Evaluation par les pairs, les encadrants
Ethique et déontologie *(Professionalism)*	"Does"	WBA	MSF : patients, familles, équipes Auto-évaluation Ressenti et expériences des patients
Communicateur, coopération *(interpersonal and communication skills)*	"Shows how" "Does"	Performance Test WBA	Simulation Ressenti et experiences des patients Evaluation 360° (coopération ++)
Réflexif *(practice-based Learning and Improvement)*	"Does"	WBA	Audit et rétroaction de dossiers RMM Autoévaluation
Scientifique *(practice-based Learning and Improvement)*	"Knows how"	Tests écrits	Revue de cas cliniques fondée sur la littérature
	"Does"	WBA	Audit et rétroaction de dossiers RMM
Acteur de santé publique *(system-based practice)*	« Does »	WBA	- Feedback des enseignants/tuteurs/encadrants : capacités à évoluer au sein d'un système de soins complexes - MSF, particulièrement regard interprofessionnel - conscience des coûts des soins délivrés

PBL: problem-based learning; WBA: work-based assessment; RMM: revue de morbi-mortalité; MSF: multiple sources feedback

Pour chacune des six compétences décrites par l'ACGME, les auteurs de la « boite à outils » ont recommandé des méthodes d'évaluation à utiliser. Nous les reprenons le tableau 3, mais en y ajoutant les compétences utilisées dans le référentiel français. L'objectif de ce tableau est de les synthétiser pour essayer d'avoir une vision d'ensemble des différentes compétences, de leurs méthodes d'évaluation afin d'y situer la méthode d'évaluation par la simulation.

2.3 Evaluer avec la simulation en médecine d'urgence

Il convient à présent de décrire le contexte de formation des étudiants et internes en médecine, afin de placer notre travail et l'évaluation par la simulation dans la perspective de la formation médicale en France, et plus particulièrement dans le domaine de la médecine d'urgence. Ainsi, le deuxième chapitre présente le déroulement de la formation médicale en France, puis il s'intéresse ensuite aux modalités d'enseignement par simulation, et enfin à son utilisation pour évaluer les apprenants, plus spécifiquement dans le champ d'exercice concerné par la recherche : la médecine d'urgence.

2.3.1 Un point commun entre la spécialité de Médecine d'Urgence et la formation des médecins en France : les réformes récentes

Spécialité nouvellement créée parmi les spécialités médicales, la médecine d'urgence émerge très progressivement depuis les années 80 et a été officiellement reconnue comme spécialité médicale le 13 novembre 2015 (Riou, 2016). L'arrêté qui crée le diplôme d'études spécialisées de médecine d'urgence (DESMU) est publié l'année suivante et la première promotion d'interne en médecine d'urgence voit le jour à la rentrée universitaire 2017 (République française, 2015; Riou, 2017).

C'est à travers l'évolution de la formation des médecins urgentistes que nous pouvons voir se dessiner l'apparition de la spécialité. D'abord capacité d'aide médicale urgente (1986) puis capacité de médecine d'urgence ou CMU (1998), la formation des médecins urgentistes obtient pour la première fois une reconnaissance universitaire avec la création du diplôme d'études spécialisées complémentaires de médecine d'urgence (DESCMU), en 2004 (Nemitz, 2005). La création de ce diplôme est survenue un an après la vague meurtrière de canicule qui avait mobilisé les services d'urgence. En devenant en 2015, diplôme d'étude spécialisé (DES), la médecine d'urgence est reconnue au même titre qu'une autre spécialité issue de DES et est recensée, catégorisée comme telle.

Discipline ainsi devenue spécialité médicale, la médecine d'urgence, par son universitarisation a pour ambition d'améliorer son enseignement, son rayonnement, sa place et une certaine reconnaissance au sein de l'hôpital et des universités (Allain et al., 2018; Riou, 2017). La spécialité connait le même devenir en Europe, ce qui permet les échanges entre ses différents acteurs, ainsi que leur mobilité. L'enjeu actuel pour la spécialité est de continuer la construction de la formation des médecins urgentistes, en améliorant les outils d'enseignement, mais également d'avoir une certaine attractivité vis-à-vis des étudiants en médecine, qui doivent faire le choix de leur spécialité à la fin de leur sixième année d'étude, ce qui n'était pas le cas dans l'ancien système, au sein duquel le étudiants pouvaient choisir la médecine d'urgence en plus d'un autre diplôme d'études spécialisées (Riou et al., 2014).

La réforme du second cycle, qui doit être déployée à la rentrée universitaire 2022, doit également être appréhendée par les universitaires de MU, afin d'y rendre la spécialité visible, notamment par les outils d'enseignement utilisés et qui doivent être le reflet de l'exercice multiple et varié de la médecine d'urgence (Riou, 2016).

La formation des futurs médecins en France : un parcours en trois cycles

Dans le cadre de la réforme LMD, et pour s'aligner avec l'enseignement de la médecine en Europe, les études de médecine suivent trois cycles, chacune étant close par la remise d'un diplôme. Le premier cycle est constitué des trois premières années et s'intitule diplôme de formation générale en sciences médicales (DFGSM). La première année, tronc commun entre différentes sciences de la santé (Médecine, Maïeutique, Odontologie, Pharmacie, Kiné) est appelée Parcours d'Accès Spécifique Santé (PASS) et permet l'accès dans un des différents parcours de santé. Elle fait suite à la première année commune d'études en santé (PACES), dont la réforme date de la rentrée 2020/2021 et qui a principalement supprimé le numérus clausus (Arrêté du 4 novembre 2019). Les deux années suivantes sont appelées DFGSM1 et 2 et

permettent l'obtention de la Licence Accès Santé (LAS). Ces deux années complètement la formation du premier cycle et ont pour objectif l'acquisition des connaissances scientifiques de base, dans des disciplines variées, principalement médicales mais également en sciences humaines et sociales. L'acquisition de la sémiologie est également un des objectifs principaux de ces deux années.

Le deuxième cycle des études médicales, appelé diplôme de formation approfondie en sciences médicales (DFASM), constitue les trois années suivantes et correspond à ce qui était autrefois qualifié « d'externat », en opposition à l'internat. Le terme externat est toujours couramment employé, de même que celui d'externe. Pendant l'externat, les étudiants répartissent leur temps de travail entre l'université et l'hôpital, au sein duquel ils réalisent des stages hospitaliers, dont l'organisation varie considérablement selon les facultés de médecine et les centres hospitaliers. À l'université, les étudiants répartissent leur temps entre travail personnel, enseignements magistraux, dirigés et séances de simulation, activités qui sont également très variables selon les facultés.

Au sein de Sorbonne Université, les externes sont formés au sein de Certificats Couplés à la Pratique Clinique (CCPC) qui permettent de créer une certaine continuité entre les enseignements facultaires et les disciplines des stages hospitaliers. Chaque stage dure entre trois et quatre mois. Notre travail s'intéresse, entre autres, à des étudiants qui réalisent leur CCPC de médecine d'urgence et de réanimation lors de leur 4ème année de médecine, soit leur première année de deuxième cycle, appelée DFASM1. Les étudiants entrés en 4ème année jusqu'à la rentrée 2021/2022, subiront, à l'issue de la sixième année de médecine (DFASM3), un examen national, appelé Examen National Classant (ECN), qui a pour objectif de les classer en fonction de leurs résultats afin de leur permettre de choisir leur future spécialité et l'université dans laquelle ils réaliseront leur troisième cycle. Les étudiants suivants verront leur formation transformée par la réforme en cours de mise en place, qui doit s'appliquer à la rentrée 2022 et qui a pour objectif

le développement de l'approche par compétence au sein des études de médecine, en réduisant la part de connaissances à acquérir, en utilisant des modalités d'évaluation nouvelles, telle que les ECOS mais également en permettant aux étudiants d'avoir un parcours de formation plus personnalisée (Décret n° 2021-1156, 2021).

Le troisième cycle des études médicales a également été l'objet d'une réforme récente, appliquée pour la première fois lors de la rentrée universitaire 2017-2018, qui a été celle de la naissance des premiers internes de médecine d'urgence. D'une durée de 4 ans, la formation hospitalière des futurs urgentistes est constituée de 8 stages de six mois et suit une maquette précise (Figure 5).

Les deux stages de la première année, appelée phase socle, se déroulent dans un service d'accueil des urgences (SAU, médecine d'urgence intra hospitalière) et dans un service de médecine interne. Les 4 stages suivants constituent la phase de consolidation et ont lieu dans des services de médecine intensive (médecine intensive et réanimation), au SAMU (Service d'Aide Médicale Urgente) et au SMUR (Service Mobile d'Urgence et de Réanimation), aux urgences pédiatriques et dans une structure du choix de l'interne. Enfin, la dernière année appelée approfondissement, se déroule dans un service d'accueil des urgences et dans un SAMU-SMUR, l'interne ayant alors une fonction de « docteur junior », qui lui permet de prendre en charge des patients avec puis sans supervision directe.

La formation universitaire, pilotée au niveau régional par un coordonnateur local, s'appuie sur le référentiel métier-compétence publié en 2012 dans les Annales Françaises de Médecine d'Urgence et fait l'objet d'un travail permanent de la part du Collège National des Universitaires de Médecine d'Urgence, qui a pour objectif d'améliorer l'enseignement de la MU. Les modalités d'enseignement sont confiées aux coordonnateurs locaux, avec le référentiel-métier comme base de contenu (Nemitz et al., 2012). Il existe également deux référentiels décrivant les compétences

nécessaires pour la pratique de l'échographie clinique en médecine d'urgence (Duchenne et al., 2016; Martinez et al., 2018).

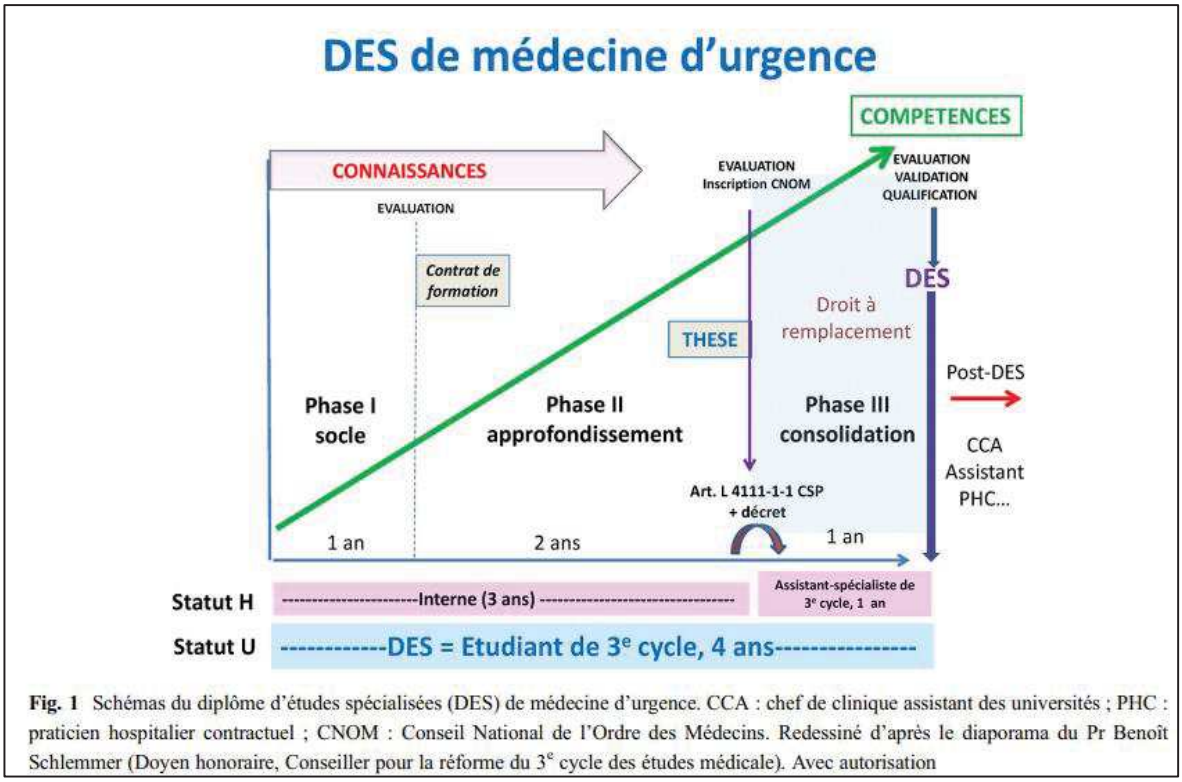


Fig. 1 Schémas du diplôme d'études spécialisées (DES) de médecine d'urgence. CCA : chef de clinique assistant des universités ; PHC : praticien hospitalier contractuel ; CNOM : Conseil National de l'Ordre des Médecins. Redessiné d'après le diaporama du Pr Benoît Schlemmer (Doyen honoraire, Conseiller pour la réforme du 3e cycle des études médicale). Avec autorisation

Figure 5 – Description du parcours hospitalo-universitaire d'un interne de médecine d'urgence, (Riou, 2016)

Pour le moment, il n'existe pas encore de référentiel pour l'évaluation des compétences, ni pour l'enseignement par la simulation, qui reste donc variable selon les universités. L'équipe de Toulouse a cependant publié une proposition de programme d'enseignement par la simulation, adaptée au contexte local, en suivant la méthode du modèle logique, qui constitue une base de travail solide pour une réflexion nationale, qui reste à l'heure actuelle un défi, compte tenu de la variabilité des moyens disponibles dans les différentes universités Une enquête est en cours afin d'améliorer et de déterminer l'offre minimale à laquelle les étudiants devraient avoir accès (Houzé-Cerfon et al., 2020).

Champ d'exercice de la médecine d'urgence

Le travail du médecin urgentiste est réparti en trois activités différentes : l'accueil hospitalier des urgences dans les services d'accueil des urgences (SAU), la médecine d'urgence extra hospitalière, organisée à partir des hôpitaux mais qui a pour but de constituer des équipes médicales mobiles dans les services de SMUR, et la régulation médicale qui consiste à réguler les appels d'urgence depuis les centres 15 (un centre SAMU par département).

Les éléments fondamentaux du champ d'exercice de la médecine d'urgence ont été rappelés lors de la création du référentiel métier-compétences pour la spécialité de médecine d'urgence, en 2012. Les deux premiers éléments décrivent la médecine d'urgence comme une spécialité sans patientèle déterminée, et dont le lieu d'exercice est hospitalier, au sein de structures privées ou publiques. En effet, puisque la discipline ne peut pas s'exercer sans l'apport de moyens humains et techniques autres que le médecin seul dans un cabinet, elle est par nature une discipline hospitalière. Le troisième élément qui fait la spécificité de la médecine d'urgence est son rapport avec la temporalité des prises en charges. L'urgentiste et son équipe doivent être en capacité de délivrer le « juste soin », à tous les patients. En situation d'urgence, le juste soin repose, en un temps réduit, sur trois éléments essentiels : la qualification de la demande du patient (ou de l'appelant, dans le cadre de la régulation médicale), l'action qui en découle et qui a pour objectif de préserver la vie ou les fonctions vitales, et enfin l'orientation du patient dans la bonne filière de soin (Nemitz et al., 2012). De cette constatation, découlent trois activités essentielles de la médecine d'urgence : trier, traiter et orienter les patients. Enfin, pour délivrer un juste soin aux patients, l'urgentiste doit également répondre à des exigences de formation professionnelle tout au long de la vie, et agir selon les principes éthiques de non-malfaisance et d'autonomie du patient. Le juste soin repose alors sur l'acquisition continue de savoirs associée à l'analyse des pratiques, au sein d'un système de soin plus ou moins en capacité de délivrer ce juste soin, d'une manière collective (Philippon, 2015). Enfin, deux autres éléments importants situent la médecine

d'urgence au cœur du système de soins : sa transversalité et sa complémentarité avec d'autres spécialités, car elle accueille tous les patients, elle partage des référentiels communs avec de nombreuses spécialités, mais également participe à de nombreuses filières de soins (Nemitz et al., 2012). Les professionnels des urgences ont également une mission sociale puisque travaillant un des seuls services à accueillir tous les patients sans limite de revenu, de temps ou d'âge. Ainsi, la dimension socio-économique s'ajoute souvent aux dimensions biomédicales et la médecine d'urgence s'inscrit donc dans un environnement complexe de travail qui la place aux frontières de nombreux champs disciplinaires (Holmboe, 2015b).

Une façon intéressante de s'intéresser aux spécificités de la médecine d'urgence est de se poser la question de ce qu'on doit enseigner aux futurs professionnels, ainsi que l'ont fait récemment les britanniques, lors de la décision d'aborder l'enseignement de la spécialité avec une approche par compétence. S'inspirant des référentiels australiens et canadiens pour définir leurs différents champs d'actions, ils complètent les capacités « génériques » professionnelles (Generic Professional Capabilities) par des activités spécifiques à la médecine d'urgence, en répondant à la question suivante : « que doivent être capable de faire les futurs urgentistes ? ». Selon eux, sept capacités spécifiques différencient les médecins urgentistes des autres spécialistes : la capacité de faire des diagnostics, celle de soigner les patients âgés, avec toutes les spécificités que cela comporte, celle de réanimer les patients lors de situations rares et complexes, celle de gérer toute situation inattendue, imprévisible qu'ils appellent « to deal with curve balls »[19], mais également d'exercer leur leadership au sein d'une équipe pluriprofessionnelle ou pluridisciplinaire et enfin, de se connaître eux-mêmes, dans le sens de connaître leurs limites, la profession de médecin urgentiste étant particulièrement sujette au stress et au burn-out (Townend et al., 2018). Quant aux Canadiens, ils ont tenu un raisonnement quelque peu différent, en partant des rôles du CanMeds (au nombre de 7) et en les déclinant spécifiquement pour la médecine d'urgence,

[19] « gérer les coups tordus », traduction par le traducteur DeepL, le 28/07/2021

créant des « tremplins », sortes d'activités pédagogiques, sur le lieu de travail, permettant de concrétiser ces différents rôles, à partir de situations authentiques (Sherbino & Frank, 2011).

D'autres auteurs soulignent également deux spécificités importantes de la médecine d'urgence : son activité « risquée », comparable à des activités industrielles à haut risque, mais dans un environnement facilement pourvoyeur d'erreurs (Walker et al., 2011). Ainsi, des auteurs ont pu mettre en évidence un taux d'erreurs médicales important dans les services de réanimation, soins intensifs ou de médecine d'urgence, milieux médicaux dont le point commun réside dans la gestion de situations d'urgence (Freund et al., 2018; Rothschild et al., 2005; Stahl et al., 2009). Les éléments pourvoyeurs d'erreurs sont également ceux qui sont caractéristiques de la pratique de la médecine d'urgence : des conditions de travail déterminées par un flux imprévisible mais souvent important de patients, un temps de prise en charge limité, le besoin de prendre des décisions dans l'urgence avec parfois des informations limitées voire nulles, la possibilité de devoir travailler avec des équipes formées instantanément, « ad hoc », et n'ayant jamais travaillé ensemble. Ces éléments sont d'autant plus présents dans les situations d'urgence vitale, auxquelles nous nous sommes intéressés dans notre travail. Une urgence vitale est définie ainsi par la HAS :

- situation où la vie du patient est en danger imminent et où il risque de décéder faute de soins ;
- le terme peut correspondre à celui d'urgence absolue (malheureusement mieux connu du grand public) ;
- toute pathologie mettant en jeu le pronostic vital immédiat rentre dans ce cadre (Haute Autorité de Santé, 2020).

A ce stade de notre exposé, il est important de rappeler que tout futur interne, puis dans une moindre mesure médecin ou chirurgien, doit pouvoir gérer à des degrés différents les situations

d'urgence vitale, et que la formation à leur prise concerne tout étudiant en médecine, qui, au premier jour de l'internat doit être en capacité de gérer les premières minutes une telle situation.

Une des façons de réduire les erreurs médicales est de s'intéresser aux habiletés non-techniques, car elles sont une partie intégrale de la prise en charge des patients, complémentaires des habiletés techniques, et impactant ainsi son devenir au même titre qu'un geste technique ou que l'application d'une connaissance (Sevdalis et al., 2008). Les habiletés non-techniques, également appelées facteurs humains comprennent les capacités de communication, de prise de décision, de leadership, de surveillance des patients et d'anticipation (Andersen et al., 2010; Fletcher et al., 2003; Guise et al., 2008; Reader et al., 2007).

Ainsi, la médecine d'urgence est une spécialité jeune, transversale, qui a pour particularité de devoir faire face à des situations de médecine et de traumatologie variées, parfois rares et graves, en s'appuyant sur des ressources techniques et non techniques, dans un milieu changeant et à risque. Ces différentes caractéristiques en font une spécialité pour laquelle l'enseignement par simulation a toute sa place, qui constitue même le « noyau incontournable » des différentes réformes de l'enseignement médical, permettant aux apprenants, et ce dès le début de leur formation, d'appréhender des notions d'urgence vitale, de travail en équipe et de gestes techniques dans un environnement protégé, à la fois pour l'apprenant, mais également pour le patient (Riou, 2017 ; HAS, 2012).

2.3.2 L'enseignement par simulation : illustration dans le champ de la médecine d'urgence

Simulation en santé : développement, principes pédagogiques et outils

Le développement de la simulation en santé possède un point commun avec celui de l'approche par compétence : apparue dans les années 60-80, son utilisation progresse considérablement à la fin des années 90, lorsque la société, les patients et les facultés de médecine prennent conscience que la sécurité des soins n'est pas optimale dans les établissements de santé

(Institute of Medicine (US), 2000; McLaughlin et al., 2008). Les spécialités pionnières dans le domaine sont l'anesthésie et la réanimation, guidés par les Pr Gaba et De Anda qui ont imaginé les tous premiers mannequins à l'université de Stanford (Gaba & DeAnda, 1988). Ce type de mannequin, actuellement développé et très utilisé existe finalement depuis le 16ème siècle en France. Créé pour enseigner les gestes de l'accouchement, il a ensuite été développé par une sage-femme, Mme du Coudray, sous le nom de « Machine ». Elle la destinait à la sensibilisation des femmes dans la France entière, afin qu'elles puissent connaître les gestes utiles de l'accouchement (Figure 6). Elle avait bien compris un des enjeux d'une formation aux gestes qui est l'apprentissage par la manipulation, l'expérience : « je pris le parti de leur rendre mes leçons palpables » (Petitcolas, 2006, p 227).

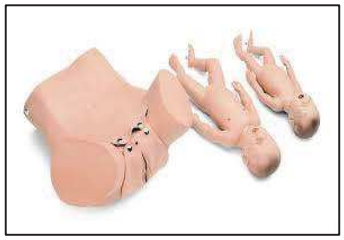
 (1)  (2)  (3)

Figure 6 – Mannequins d'accouchement : inventé par Mme du Coudray (1), basse fidélité (2) et haute-fidélité actuels (3)

Au cours des années 90, parallèlement à l'émergence de l'approche par compétence, l'utilisation de la simulation outre-Atlantique s'est rapidement disséminée à tous les champs de la médecine et de la chirurgie : médecine d'urgence, obstétrique, pédiatrie, chirurgie urologique, ORL etc. En médecine d'urgence, aux Etats-Unis, son utilisation a doublé entre les années 2003 et 2008, avec 40% puis 80% des programmes de formation des internes qui se mettent à l'utiliser (Okuda et al., 2008). En ce qui concerne les étudiants en médecine, pour lesquels la mise en place de l'approche par compétence a été décidée plus tardivement, la simulation dans les programmes

de formation aux gestes d'urgence est également bien implantée, mais dans une moindre mesure.

En France, lors d'une enquête réalisée en 2017 auprès des 25 universités sur les 29 qui ont des programmes de formation des internes de médecine d'urgence, 23 (92%) l'utilisaient de manière inégale dont 22 avec de la simulation haute-fidélité et 21 avec de la simulation procédurale. Les internes avaient une médiane de 2 à 3 journées de formation sur simulateur par an. Cependant, seulement 12 centres (52%) estimaient « plutôt » respecter un entrainement sur simulateur avant la mise en pratique sur le patient, selon l'adage « jamais la première fois sur le patient » (Allain et al., 2018). Le développement et l'intensification de l'utilisation de la simulation était principalement limités par le manque de moyens humains et matériels.

Ainsi, le leitmotiv « Jamais la première fois sur le patient » est un des piliers et objectifs de l'enseignement par simulation : le patient ne devrait pas être un outil d'apprentissage initial, mais le destinataire des compétences acquises à la faculté ou en stage, une fois que l'autonomie partielle ou complète a été validée à la faculté (Ten Cate, 2005; Vincent & Amalberti, 2015). En médecine d'urgence, par exemple, la simulation permet en effet de s'entraîner à réaliser puis maitriser des gestes techniques (tels que les sutures), mais également à la prise en charge de situations complexes (telles que les urgences vitales) qui nécessitent l'apprentissage du travail d'équipe pluriprofessionnelles et disciplinaires, ainsi que la gestion du temps de prise en charge. La simulation permet également de travailler les capacités relationnelles (annonce d'un pronostic vital engagé) et est destinée tant aux formations initiales que professionnelles (Ghazali & Casalino, 2018; Kessler et al., 2011; Lorello et al., 2014; Parent et al., 2010).

Le deuxième objectif de la simulation est de limiter les erreurs liées aux soins, grâce à une formation continue, interprofessionnelle, mais également intégrée dans une approche curriculaire, afin de permettre le transfert des apprentissages. En effet, la simulation en santé s'est inspirée des pratiques de formation des industries dites à haut risque, telles que le nucléaire

ou l'aviation qui se doivent d'avoir des exigences élevées en termes de sécurité et certaines pratiques de gestion de crise ont été transposées au champ de la médecine, comme pour le Crisis Ressource Mangement, issu du Crew Ressource Management et qui définit les habiletés non techniques à mobiliser dans des situations exceptionnelles ou à risque, comme les situations d'urgence vitale (Amalberti et al., 2005; Gaba, 2010).

Après avoir illustré ses différents objectifs, la définition de la simulation se dessine. Il s'agit, d'après Gaba, d'une « technique, et non d'une technologie, qui remplace les expériences de la « vraie » vie par des expériences encadrées, évoquant ou reproduisant les aspects fondamentaux du monde réel, d'une manière interactive » (Gaba, 2004). Cette définition peut être complétée par celle qui a été choisie lors de l'émission des recommandations de bonnes pratiques en matière de simulation en santé, pilotées par le Pr Granry et le Dr Moll en 2012 : « la simulation en santé correspond à l'utilisation d'un matériel (comme un mannequin ou un simulateur procédural), de la réalité virtuelle ou d'un patient standardisé, pour reproduire des situations ou des environnements de soins, pour enseigner des procédures diagnostiques et thérapeutiques et permettre de répéter des processus, des situations cliniques ou des prises de décision par un professionnel de santé ou une équipe de professionnels » (HAS, 2012).

Dans ces deux définitions, on entrevoit les principales pédagogies, issues des différents courants théoriques décrits ci-dessus, et qui régissent l'apprentissage au moyen de la simulation. En effet, la simulation en santé, puisqu'elle peut utiliser diverses techniques et différents moyens d'appréhender la performance et la compétence des apprenants, s'appuie également sur différents courants théoriques. Nous ne citerons ici que les principaux et en premier, le behaviorisme, lorsqu'il s'agit par exemple de mettre les apprenants en situation, puis de les soumettre à un renforcement positif ou négatif en fonction de la performance observée. L'apprentissage est centré sur l'action de l'enseignant, qui détermine l'environnement, les stimuli qui déclenchent l'apprentissage puis il délivre le renforcement qui doit modifier le comportement

de l'étudiant. Cependant, son utilisation, au sein d'un programme de simulation devrait être complété par une réflexion sur l'action (ce qui n'est pas décrit dans le behaviorisme), qui constitue, nous allons le voir un des principes fondamentaux de la pédagogie par simulation (Kay & Kibble, 2016; Pottier, 2013). Cette méthode a été décrite comme ayant un intérêt pour l'apprentissage de procédures ou d'algorithmes simples avec la simulation, même si cela peut se discuter et entre en contradiction avec les recommandations d'utiliser systématiquement des méthodes de réflexion sur l'action quand on emploie la pédagogie par simulation (Fann et al., 2013; Issenberg et al., 2005; Motola et al., 2013).

L'arrivée du constructivisme qui permet, entre autres, de prendre en considération les connaissances antérieures pour en construire de nouvelles, est une théorie qui sous-tend la simulation qui, par nature, est une méthode demandant à l'apprenant de mobiliser ses connaissances antérieures pour les confronter à une situation-problème et enfin analyser cette mobilisation et utilisation de connaissances lors du débriefing (Pottier, 2013; Shrivastava et al., 2013).

L'apport des théories socio-culturelles ou socio-cognitives, qui ont une approche plus analytique de l'apprentissage, et qui placent l'apprenant au cœur d'un système d'apprentissage collaboratif est également important pour la pédagogie de la simulation. Là où behavioristes et constructivistes ont repéré des systèmes d'apprentissage basés sur l'individu, les théories socio-cognitivistes y ajoutent l'environnement et décrivent son importance dans le processus d'apprentissage. À la lumière de ces théories, l'apprenant est vu comme un sujet, inclus dans un système socio-culturel. De plus, la connaissance n'est pas détenue par les institutions ou les enseignants seuls, mais par une communauté qui la partage et que Lave et Wenger ont dénommée « communauté de pratiques » (Bleakley, 2006). Plusieurs courants existent, et parmi eux la théorie de l'apprentissage social, développée par Bandura dans les années 70. Pour lui, l'apprentissage résulte de l'observation des comportements et de l'interaction de l'apprenant avec

son environnement. (Bandura, 2001). La simulation s'appuie également sur ce concept puisque les apprenants sont tour à tour participants et observants, ce qui ne les empêche pas d'apprendre dans chacune des deux positions. De même le rôle de l'environnement et des pairs y sont majeurs puisqu'ils participent, lors du débriefing, à l'élaboration d'une nouvelle connaissance, commune au groupe. Thomas Mann, dans un article de 2011, insiste sur l'importance de s'appuyer sur de telles théories car elles sont adaptées aux changements de la pédagogie médicale : les apprenants sont actifs, construisent leurs connaissances, au sein d'une communauté, puis ils les mettent en pratique au sein de plusieurs communautés de professionnels de santé. Ils doivent également pouvoir se former tout au long de leur carrière et plutôt que de savoir « ce qu'il faut apprendre », il doit pouvoir apprendre « comment apprendre ». Pour lui, les théories socio-cognitives ont contribué de manière significative à l'évolution de la pédagogie médicale et parmi elles, l'apprentissage périphérique légitime, théorisé par Lave et Wenger, et l'apprentissage expérientiel dont nous allons décrire plus précisément le concept (Mann, 2011). Pour conclure ce chapitre sur les apports des différentes théories de l'éducation à l'enseignement basé sur la simulation, nous empruntons aux professeurs McGaghie et Harris, qui les ont illustrées en résumant bien les résultats des apports de chacune d'elles (McGaghie & Harris, 2018), (Figure 7).

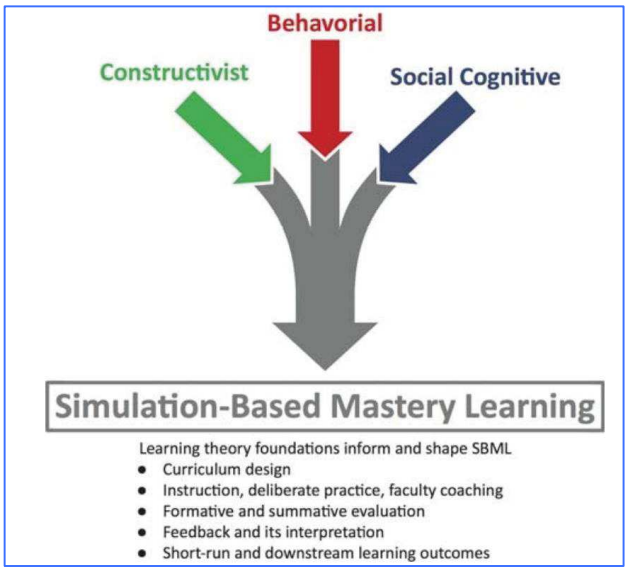


Figure 7 – Fondations théoriques de l'enseignement par simulation, in McGaghie et Harris, 2018

Le dernier modèle majeur qui a également permis de décrire et développer le modèle de formation, est celui de l'apprentissage expérientiel, défini par Kolb en 1984, suite aux travaux de Dewey, Lewin et Piaget. L'apprentissage est défini par Kolb comme « le processus par lequel le savoir est créé par la transformation de l'expérience ». Il place ainsi l'expérience au cœur du processus d'apprentissage, qui serait un mécanisme permettant de passer du concret à l'abstraction des concepts. Pour Lewin, l'apprentissage est fait par l'expérience concrète, « ici et maintenant » mais également grâce à la rétroaction que l'apprenant peut avoir de son expérience et des enseignements qu'il va en tirer. Dewey y ajoute la notion de motivation et d'impulsion de l'exercice qui sont des forces motrices pour l'apprentissage. Piaget, après avoir défini plusieurs phases de l'apprentissage, place son développement dans l'expérience et dans les concepts de réflexion et d'action. Pour lui, l'apprentissage est l'« adaptation intelligente » et il progresse à la fois grâce au rôle de l'expérience qui créée de nouveaux concepts mais également par l'intégration de l'expérience à des concepts existants. Kolb, qui a résumé les principes de l'apprentissage expérientiel dans son cycle de l'apprentissage identifie des caractéristiques communes aux quatre modèles (Figure 8). L'apprentissage, plus qu'un amas de connaissances

à délivrer et à évaluer selon des résultats, est un processus dynamique, qui permet la stimulation de l'investigation, des compétences dont l'objectif n'est pas tant le résultat mais les actions nouvelles qu'il va produire dans la quête ininterrompue des connaissances nécessaires à l'homme. Ce processus est continu et ancré dans l'expérience, et cela signifie que chaque expérience est influencée par les expériences et les connaissances antérieures mais également que chaque expérience modifie la suivante. Ainsi, tout apprentissage est « réapprentissage ». Pour que l'apprentissage fonctionne, il faut posséder quatre capacités, qui sont illustrées dans la figure 8 : celle de pouvoir réaliser des expériences concrètes, celle de l'observation réflexive, qui permet l'abstraction conceptuelle et enfin l'expérimentation active. Ainsi l'apprentissage par l'expérience, et non par les contenus ou les résultats permet, grâce à la capacité d'adaptation des apprenants de transformer ses connaissances (sans qu'elles ne soient acquises ou transmises).

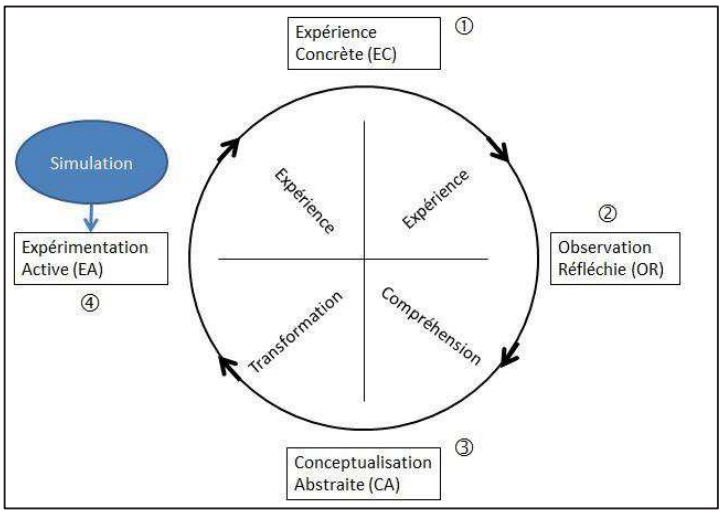


Figure 8 – Le cycle de l'apprentissage expérientiel, d'après Kolb, 1984

La simulation, en permettant à l'apprenant l'expérience d'une situation simulée, puis une réflexion sur l'action au cours du débriefing, lui permet de transformer ses connaissances antérieures, dans un cadre similaire à celui de l'apprentissage expérientiel. Cependant, comme le souligne

Blealkey, cette vision est fondée sur une vision individuelle de l'apprentissage et l'éducation médicale a besoin de théories prenant en compte la collaboration entre individus, le travail en groupe, « l'apprentissage situé » dans une communauté de pratiques, tel que l'ont décrit Lave et Wenger (Bleakley, 2006).

Les principes pédagogiques fondamentaux de la pédagogie basée sur la simulation empruntent donc à ces différentes théories ou modèle. Motola les a énoncés en 2013 dans un guide pratique pour l'enseignant en simulation. Ainsi, tout enseignement par simulation doit figurer au sein d'un programme d'enseignement, en cohérence avec les autres outils d'enseignement, ce qui lui permet d'être plus efficace. Enfin, toute action simulée devrait faire l'objet d'une rétroaction, qui peut prendre plusieurs formes, mais l'apprenant doit pouvoir avoir les moyens de réfléchir sur l'action réalisée. Ce principe fondamental de la simulation fait du débriefing un moment crucial de la session de simulation, au sein duquel se jouent les transferts d'apprentissage (Rall et al., 2000). La deliberate practice, qui permet à l'apprenant de pouvoir répéter des situations simulées, avec des objectifs connus, tout comme les seuils de validation de la performance qui permettront à l'apprenant de progresser à une étape ultérieure est également une pratique centrale de l'enseignement par simulation. De même l'apprentissage au sein d'une approche par compétence qui aura pour objectif de former des apprenants qui maîtrisent les habiletés requises et qui sont autonomes doit être un souci permanent de l'enseignement par simulation. Cet aspect de la simulation est désigné par le terme anglo-saxon suivant : Simulation-Based Mastery Learning. Les activités de simulation doivent être adaptées au niveau des apprenants, au risque d'être contre-productives et la simulation doit fournir des occasions de formation interprofessionnelles et ce, dès le début des cursus de formations en santé. Enfin, et il s'agit encore de deux arguments pour utiliser la simulation comme un des outils de l'approche par compétence, il faut pouvoir placer les apprenants dans différents contextes et environnements cliniques, afin de les entrainer aux situations cliniques complexes qu'ils auront à prendre en charge et il faut pouvoir, réaliser un

enseignement individualisé, adapté aux besoins spécifiques de chaque apprenant (Motola et al., 2013).

De nombreux outils ou techniques de simulation peuvent être utilisés pour reproduire une situation de pratique médicale. Plusieurs typologies des outils de simulation existent, elles sont basées soit sur les outils, soit en rapport avec les habiletés qu'elles permettent de travailler ou selon la technique utilisée (Alinier, 2007; Haute Autorité de Santé, 2012; McLaughlin et al., 2008). Pour présenter les différents outils à disposition, nous avons choisi la classification retenue par la haute autorité de santé, plus complète et basée sur les techniques utilisées. Ainsi la simulation peut utiliser :

- un support animal, par exemple pour apprendre les points de suture,

- un support humain avec simulation de gestes sur des cadavres (abord des voies aériennes), patient standardisé (patient volontaire ou acteur qui est formé pour jouer un scénario préétabli avec un rôle précis, par exemple une consultation dans un service d'accueil des urgences), jeu de rôle (simulation d'une situation vraisemblable, souvent pour travailler les habiletés relationnelles, mais dans laquelle le dialogue est improvisé),

- une technique synthétique avec des simulateurs patients (appelés mannequins, et qui ont un degré de réalisme varié, qui permet de qualifier la simulation de haute ou basse fidélité), des simulateurs procéduraux (qui permettent l'apprentissage de gestes techniques, de procédures complexes, comme par exemple une ponction lombaire ou l'intubation orotrachéale),

- une technique électronique représentée par la réalité virtuelle et par l'environnement 3D et les jeux sérieux (qui permettent de reproduire un environnement réaliste, à la manière d'un jeu vidéo utilisant leurs meilleures technologies, dans un but pédagogique. On peut l'utiliser en médecine d'urgence pour simuler la prise en charge de multiples victimes par exemple),

- enfin une technique « mixte » qui consiste en l'utilisation de plusieurs outils de simulation.

Il existe un outil qui n'apparait pas dans la typologie réalisée par l'HAS mais qui est décrit dans le travail d'Alinier et al. : les simulations écrites, telles que la résolution de problème ou l'étude de vignettes cliniques. Il nous parait important de le citer afin de souligner que la simulation peut revêtir de nombreuses formes, même si elle est souvent reliée à des supports techniques. Cependant, il ne faudrait pas que ces techniques soient un facteur limitant dans l'utilisation des simulations qui ont fait leurs preuves d'un point de vue pédagogique et peuvent être employées avec des outils simples et adaptés à certains objectifs pédagogiques (Alinier et al., 2007).

Dans le travail que nous allons décrire, nous avons utilisé une technique soit synthétique seule (avec des mannequins haute-fidélité lorsque les sessions se déroulaient en centre de simulation) mais également une technique hybride avec patient simulé et mannequin basse fidélité ou simulateur procédural lors des sessions de simulation in situ, au sein des structures d'urgence.

Cependant quel que soit la technique choisie, le déroulement d'un enseignement qui utilise la simulation (simulation-based education) doit suivre différentes étapes que nous décrivons brièvement et qui permettent de respecter son cadre théorique. Il faut tout d'abord tenir compte, tant pour les enseignants que pour les apprenants de l'étape qui précède la session et de celle qui la suit. La session de simulation se prépare en amont du cours lui-même. L'enseignant doit en effet définir quels vont être les objectifs pédagogiques de la formation et à partir de ces objectifs définir les outils qu'il va utiliser afin de pouvoir y répondre. Il peut, à cet effet, recommander aux étudiants d'avoir des connaissances spécifiques avant de se rendre à la séance de simulation afin qu'ils puissent les mobiliser au cours de celle-ci (Figure 9, Haute Autorité de Santé, 2012). Vient enfin le moment de la session de la simulation, structurée par trois phases indispensables. Le« pré-briefing » consiste en une présentation de la session, du matériel à disposition et a surtout pour objectif de rappeler les principes pédagogiques de la simulation afin que les apprenants l'abordent dans un climat de confiance. Vient ensuite la phase

du « scénario », initiée par un « briefing » qui est une description rapide de la situation que les apprenants vont devoir gérer. La situation se déroule et vient ensuite le « débriefing ». Il a pour objectif de revenir sur le déroulement du scénario, avec les apprenants participants (actifs), avec les tuteurs mais également avec les apprenants qui ont observé le scénario (observateurs). L'objectif est de mettre en lumière les habiletés maîtrisées par les apprenants et celles qui nécessitent un axe de progression. Ici, la mobilisation de connaissances antérieures des apprenants est fondamentale ; elle permet de savoir quelles représentations ils en ont, et comment leurs représentations entrent en conflit avec les habiletés nouvelles abordées ou utilisées pendant la simulation. Le débriefing est également constitué de plusieurs étapes, qui varient selon la méthode utilisée, mais qui ont toujours pour objectif de revenir sur les émotions des apprenants, d'explorer leur manière de résoudre le cas en allant plus loin qu'une simple observation et que les apprenants trouvent leurs axes de progression (Dieckmann et al., 2009; Motola et al., 2013). Pour la plupart des chercheurs ou utilisateurs de la simulation, le débriefing constitue la partie majeure de la simulation. Ainsi, pour Rall, "debriefing can 'make or break' a simulation session and can be attributed as the 'heart and soul' of simulator training"[20] (Rall et al., 2000).

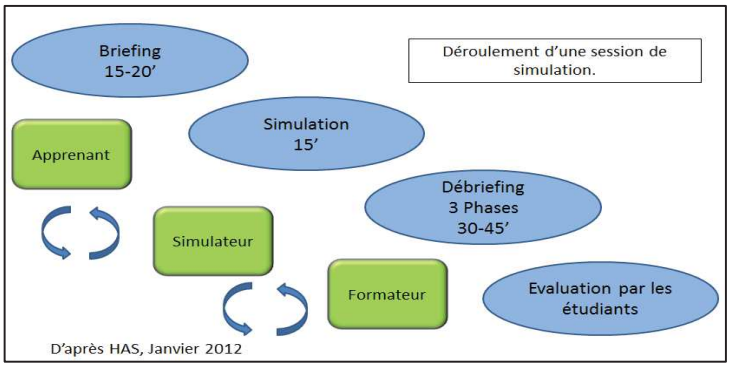


Figure 9 – Déroulement d'une session d'enseignement par simulation, HAS, 2012

[20] " le debriefing peut faire ou défaire une session de simulation, et il peut être décrit comme le « cœur et l'âme » de l'apprentissage par simulation » (traduction libre)

Ainsi, compte-tenu des spécificités de la médecine d'urgence, et de la nécessité de former des futurs professionnels compétents, la simulation apparaît comme un outil d'enseignement fondamental et à utiliser tout au long du parcours d'un futur médecin afin d'assurer une continuité d'enseignement et d'apprentissage.

Dans le champ de la formation des sciences de la santé, l'efficacité d'un outil pédagogique se mesure depuis longtemps, selon le modèle des quatre niveaux Kirkpatrick qui est, initialement, un modèle créé en 1959 pour l'évaluation de formations dans le monde industriel (Kirkpatrick, 2006; 1996). Le premier niveau (reaction) correspond à la satisfaction et à la perception des apprenants quant au dispositif de formation afin de s'assurer de leur motivation et de leur intérêt pour la formation. Le deuxième niveau (learning) s'intéresse aux nouveaux apprentissages rendus possibles grâce à la formation. Ils sont mesurés par l'apprenant lui-même (auto-évaluation) ou par une hétéroévaluation réalisée grâce à un test final, ou pendant la formation, grâce à l'observation des formateurs. Le troisième niveau (behavior) analyse les changements de comportement ou de pratiques, dans le milieu professionnel lui-même et est représenté par le transfert des apprentissages. En théorie, il doit reposer sur une observation du changement et pas uniquement sur une déclaration des apprenants, ce qui est souvent le cas. Enfin, le quatrième niveau (results) est la mesure de l'effet de la formation sur l'objectif final de celle-ci, à savoir la pertinence et l'exactitude des soins délivrés dans le domaine de la santé. Il devrait donc s'évaluer par le devenir du patient et par des critères de jugements « durs », tels que l'amélioration de la morbi-mortalité ou la diminution d'effets indésirables par exemple.

Une approche plus biomédicale de l'évaluation d'une intervention est celle de la médecine translationnelle, qui évalue l'efficacité d'une intervention au sein du système de santé en suivant toujours les trois mêmes étapes, que ce soit dans le développement d'un médicament ou d'un dispositif médical. McGaghie s'en est inspiré pour décrire les objectifs de l'évaluation d'une la

formation en santé, prenant en compte l'aspect complexe d'une formation, des compétences et de leur utilisation au cœur du système de soin, afin qu'ils soient toujours liés aux conséquences de la formation : la sécurité et la qualité des soins. Le premier niveau d'évaluation se situe au laboratoire (T1) et correspond à l'évaluation d'une intervention avec des situations simulées, le deuxième niveau s'intéresse aux conséquences sur la pratique dans des situations professionnelles (niveau 3 de Kirkpatrick), le niveau 3 s'intéresse aux bénéfices pour le patient et enfin le niveau 4, ajouté dans un second temps, analyse les coûts engendrés par l'intervention, au regard de son apport sur la santé (McGaghie, 2010; McGaghie et al., 2012).

Ainsi, en suivant ces différents niveaux d'impact, l'enseignement par simulation a pu démontrer son efficacité dans le domaine de la médecine d'urgence tout en permettant un enseignement sans risque pour l'apprenant et pour le patient. La satisfaction des apprenants a été démontrée à de nombreuses reprises, tant pour des novices que des apprenants professionnels et dans de nombreux domaines, dont celui de la médecine d'urgence (Bullard et al., 2020; Ingrassia et al., 2014; Lorello et al., 2014).

Dans de nombreux domaines et en particulier en médecine d'urgence, l'impact de la simulation est largement démontré sur l'apprentissage de gestes techniques ou de procédures nécessitant l'utilisation d'algorithmes techniques que ce soit dans des revues de la littérature ou méta-analyses ou grâce à des essais randomisés contrôlés et ce, à différents niveaux : celui de l'amélioration de la connaissance des apprenants (qui reste le niveau le moins impacté par la simulation), puis de l'amélioration et de la maîtrise des gestes ou procédures mais également ai niveau le plus important : celui du patient. En effet, les analyses de l'impact de l'enseignement par simulation (le plus souvent synthétique), ont permis de mettre en évidence des taux de succès du geste plus élevés, des complications moindres et un meilleur confort ressenti par les patients (Beal et al., 2017; Griswold-Theodorson et al., 2015; McGaghie et al., 2011a, 2014; O'Donnell et al., 2011). Cela a été vérifié en médecine d'urgence pour des gestes telles que la pose de

cathéters périphériques ou centraux, la pose de drains thoraciques, la réalisation de ponctions lombaires, ou encore la prise en charge du traumatisme crânien chez l'enfant selon des algorithmes bien décrits (Ansquer et al., 2020; Barsuk et al., 2018; Harwayne-Gidansky et al., 2021; Kessler et al., 2011).

La plupart des travaux se sont intéressés aux connaissances, procédures, habiletés techniques et à des résultats plus facilement mesurables pour les patients (tels que le confort selon une auto-évaluation du patient, le taux de complication, la douleur mesurée avec des échelles validées). Le défi reste élevé en ce qui concerne les habiletés telles que la communication, le travail en équipe qui, même si elles peuvent être mesurées par des échelles validées sont souvent moins étudiées à des niveaux élevés d'impact, alors qu'il est décrit qu'elles participent à la sécurité et à la qualité des soins. En effet, de nombreux biais existent en médecine d'urgence et dans les situations de travail en équipe et il faut pouvoir les contourner pour décrire un lien de corrélation entre la formation et son effet final pour le patient. De plus, l'extraction des données en milieu professionnelles ainsi que l'utilisation de critères de jugements pertinents pour les patients sont des éléments qui représentent également un défi pour les enseignants et chercheurs dans le domaine de la simulation (Griswold et al., 2018). Une autre inconnue en ce qui concerne l'impact de la simulation est la durée de son efficacité, plus rarement étudiée. Deux revues récentes de la littérature, dans le domaine de l'anesthésie et de la médecine d'urgence pédiatrique soulignent ces différents manques dans la recherche sur l'impact de l'enseignement par simulation (Huang et al., 2019; Young et al., 2019).

Cependant, ces nombreux marqueurs positifs de l'efficacité de la simulation pour l'apprentissage et son transfert dans les situations professionnelles, de même qu'une corrélation positive entre certains outils d'évaluation par la simulation et l'impact sur les patients, ainsi que le démontre l'équipe de Brydges en 2015 sont des arguments forts pour proposer la simulation comme outil d'évaluation de la performance et des habiletés des apprenants (Brydges et al., 2015). Ainsi,

dans le domaine de la médecine d'urgence en France, le Pr Riou explique que la simulation constitue le « moyen incontournable de cette réforme » et qu'il est indispensable pour les urgentistes de construire une formation cohérente et de qualité, reflet de la spécialité, mais également pour la rendre attirante pour les internes (Riou, 2016, p 3).

2.3.3 Evaluation par la simulation en médecine d'urgence

La simulation offre donc un outil de formation et d'évaluation qui place l'apprenant dans un environnement semi-authentique, avec la possibilité de varier les situations cliniques, les contextes, tout en fournissant une rétroaction sur l'action engagée, dans un cadre de pratique réflexive. Les situations et les contextes peuvent également être reproductibles, et relativement maîtrisées, ce qui en fait un outil d'évaluation intéressant au sein d'une approche par compétence (Wiel, E. et al., 2013). Dans ce chapitre, nous aborderons uniquement la question de l'évaluation certificative, l'approche par simulation étant formative par nature, puisque pourvoyeuse de rétroaction et d'axes de progression aux apprenants. Ainsi, nous présentons l'intérêt de la simulation pour des évaluations à enjeu dit élevé, tels que la validation d'un diplôme ou la re-certification, et dans ce cadre, ce que permet d'évaluer la simulation et avec quels critères. (Boulet, 2008). Un des éléments qui autorise les responsables de programme à utiliser la simulation comme un des outils de validation des acquis des apprenants, est son aptitude à les discriminer selon leur niveau de performance de même que la corrélation, plusieurs fois décrites entre la performance en simulation et la performance clinique. Afin de décrire ce qu'est une évaluation par la simulation et quelles en sont les aspects, nous nous sommes appuyés sur le guide de la Société Francophone de Simulation en Santé, dont le texte court est paru en 2021 et pour lequel la doctorante a également travaillé, aux recommandations canadiennes récentes appliquées à la médecine d'urgence et qui suivent le cadre de Norcini, ainsi qu'au chapitre consacré à l'évaluation par la simulation dans le livre « Practical Guide to the Evaluation of Clinical Compétence » (Hall et al., 2020; Norcini et al., 2018; SoFraSimS, 2021). Enfin, parler

d'évaluation par la simulation sous-entend que les évalués auront pu bénéficier de formation basée sur la simulation au préalable, en guise d'entrainement mais également car ils doivent s'approprier l'outil.

Les quatre thématiques générales qui définissent et problématisent une approche d'évaluation avec la simulation sont les suivantes :

- définir et choisir les habiletés et performances pertinentes à évaluer avec la simulation
- développer des outils d'évaluation avec des propriétés psychométriques valides
- évaluer la fiabilité des outils
- fournir des preuves de validité des outils, au sein de processus de validation continus (Boulet, 2008).

En premier lieu, il convient de définir quelles habiletés et capacités peuvent être évaluées par la simulation tout en rappelant qu'au sein d'une approche par compétence, il est recommandé d'utiliser plusieurs modalités d'évaluation, au sein d'un programme évaluatif et que la simulation ne peut constituer à elle seule une évaluation valide des différentes composantes menant à la compétence d'un apprenant (van der Vleuten et al., 2012). Dans le modèle de Miller, nous avons vu que la simulation se situe à un haut niveau (le troisième) et qu'elle permet donc à l'apprenant de montrer ce qu'il est capable de réaliser seul ou sous supervision (Miller, 1990). L'évaluation par simulation est donc l'évaluation d'un comportement observable à un instant précis, et elle constitue une évaluation de performance qui permet d'approcher un niveau de compétence de l'apprenant.

Les Examens Cliniques Objectifs Structurés (ECOS) représentent une des façons de pratiquer l'évaluation par la simulation qui est connue depuis les années 80, et dont les critères de qualité sont décrits depuis longtemps. Certains sont extrapolables à la simulation synthétique, qui emploie des mannequins haute-fidélité et à laquelle nous nous intéressons spécifiquement (Khan

et al., 2013). Cependant, les deux outils ne permettent pas de mesurer les mêmes habiletés, la simulation « synthétique » autorisant l'évaluation de gestes techniques (intubation orotrachéale, sutures), mais également des situations d'urgence vitale nécessitant la prise en charge des patients par une équipe pluriprofessionnelle voire multidisplinaire. Ainsi, les mannequins permettent d'évaluer des performances individuelles au sein d'une équipe, augmentant le réalisme de la situation, d'évaluer des gestes techniques en situation, d'évaluer des performances d'équipe et des situations complexes (Boet, et al., 2018a; Boet, et al., 2018b). D'un point de vue des compétences, la simulation permet d'évaluer la compétence clinique, le raisonnement clinique sous certaines conditions, ainsi que le travail d'équipe (Holmboe & Iobst, 2020).

Utiliser la simulation pour évaluer doit également répondre à des objectifs de sens : pour l'apprenant, qui doit être au courant des objectifs pédagogiques et d'évaluation (ce qui a un effet motivationnel), qui doit comprendre le sens des situations choisies (situations professionnelles importantes) ce qui permet à l'évaluation de jouer son rôle au sein de l'alignement pédagogique des activités d'un cursus. Il faut également que l'évaluation aie une validité de contenu ou écologique, avec des situations qui reflètent les situations professionnelles mais également qui permettent de prédire la compétence qu'aura l'apprenant dans son milieu professionnel, avec le patient (Griswold et al., 2018; Pelaccia, 2016).

Comme toute évaluation, celle qui utilise la simulation doit pouvoir s'appuyer sur des outils valides, fiables et reproductibles, tout en étant réalisables, acceptables et avec un effet éducationnel démontré (Norcini et al., 2011; 2018). Nous consacrons un chapitre spécifique aux qualités des outils et à leur développement dans la première partie de la méthodologie. Cependant, il est important de rappeler que pour l'évaluation par la simulation, de plus en plus d'outils sont développés, dont certains ayant fait preuve de leur validité, même si la plupart des études et recherches restent encore incomplètes ainsi que le souligne la dernière revue sur le sujet : souvent de faibles effectifs (30 apprenants en moyenne), une étude sur deux rapportant

un seul critère de validité de l'outil utilisé, et parmi ces études, la moitié ne s'intéressaient qu'à un seul groupe de niveau des apprenants, 30% s'intéressaient à la qualité du contenu des outils ou à leur fiabilité ou à leur lien avec un autre test, mais 73% qui recherchaient un lien entre le niveau de l'apprenant et le résultat au test (Cook et al., 2013). Cependant les outils seuls ne sont pas garants d'une évaluation de qualité, et les conditions de réalisation du test sont également importantes. Ainsi, en simulation, les biais qui peuvent modifier les qualités d'une évaluation sont les suivants : la qualité des scénarios et les objectifs d'évaluation qui en découlent, le contexte de l'évaluation avec le nombre et la durée des situations évaluatives, les items, les facilitateurs qui interviennent dans la simulation et enfin les évaluateurs qui doivent être formés au préalable (Hart et al., 2018; SoFraSimS, 2021; Wiel, E. et al., 2013). Une donnée validée pour les ECOS et qui semble être extrapolable pour l'évaluation par la simulation : plus le nombre de stations est important, plus l'évaluation est fiable (Charlin et al., 2000; Epstein, 2007). En effet, le nombre de situations élevé permet de faire varier les contextes, de faire varier les compétences évaluées et ainsi d'avoir un meilleur reflet du niveau de l'apprenant (Khan et al., 2013). Dans le même ordre d'idée, il est préférable d'utiliser également un grand nombre d'évaluateurs, non pas pour chaque station, mais par exemple, 14 évaluateurs seuls évaluant 14 stations permettent d'obtenir un examen plus fiable que 7 évaluateurs en binôme évaluant 7 stations. En matière de faisabilité, tout centre de simulation est dans la capacité de mettre en place une évaluation par simulation, à conditions d'avoir une certaine expertise dans le domaine et dans la limite des moyens disponibles (humains, matériels et configuration du centre) (Holmboe & Iobst,, 2020; Khan et al., 2013).

Puisqu'elle est quasiment inhérente à l'enseignement par simulation, l'évaluation formative est un format d'évaluation facilement accepté par les enseignants. En ce qui concerne l'évaluation sommative ou certificative elle se majore, mais elle est confrontée à de nombreux défis, tels que les coûts, infrastructures, les modifications qu'elle entraine auprès des enseignants et notamment

la crainte de modifier le cœur de la pédagogie et des critères de qualité de la simulation et avant d'émettre un programme intégrant l'évaluation tous ses aspects doivent être pris en considération (Hall et al., 2020).

L'impact pédagogique de l'évaluation par la simulation est élevé, car elle permet l'alignement avec les situations professionnelles, et l'évaluation de situations rares mais importantes dans la pratique clinique (telles que les situations d'urgences vitales, difficiles à évaluer en situation professionnelle). Cependant le garant de l'impact pédagogique est la manière dont est conservée la rétroaction fournie à l'apprenant et cela questionne la place du débriefing dans l'évaluation. Or, en matière de simulation évaluative, trois situations sont possibles : pas de débriefing car il s'agit d'une situation évaluative et que l'évaluation est factuelle, centrée sur la performance réalisée lors de la simulation, un débriefing intégré dans l'évaluation et qui a un impact sur la note finale, ou encore un retour d'expérience de la part de l'évaluateur, qui peut-être individuel ou en groupe, après la simulation ou différé, mais qui n'intègre pas le point de vue de l'apprenant dans la décision finale (SoFraSimS, 2021).

Ainsi, l'évaluation par la simulation semble donc adaptée à une approche par compétence si elle ne constitue pas le seul moyen d'évaluation des apprenants et si elle répond à des critères de qualité rigoureux au sein d'un processus de validation constant. Nous allons à présent nous intéresser à l'état des lieux de l'évaluation en médecine d'urgence.

Une revue de littérature, publiée en 2017 dans le *Canadian Journal of Emergency Medicine* s'est intéressée aux modalités d'évaluation des internes de la discipline (Colmers-Gray et al., 2017). Parmi les 879 articles recensés lors de la première recherche, 73 correspondaient aux critères d'inclusion de l'étude. La moitié étaient des études pilotes, et le quart concernait la description de programmes complets de formation et d'évaluation. Il est intéressant de noter que 94% des articles étaient issus d'université nord-américaines et que seulement 3% provenaient d'Europe. Les principaux résultats de l'étude mettaient en évidence des modalités d'évaluation réparties

selon trois catégories : les évaluation écrites (28.8%), les évaluations par la simulation (28.8%, principalement en lien avec une évaluation avec des ECOS) et les évaluations sur le lieu de travail (26%). La fréquence des évaluations variait de plusieurs fois par semaine à une fois par an. Il n'était pas précisé le caractère formatif ou sommatif des évaluations. Nous ne disposons pas de telles données en ce qui concerne l'Europe ou la France (si ce n'est une absence de publication de tels programmes, qui n'est pas nécessairement liée à une absence de ces programmes). En France, le Collège National des Universitaires de Médecine d'Urgence (CNUMU) a lancé un groupe de travail sur le sujet, dont la doctorante fait partie. Une enquête auprès des différents centres d'enseignement par simulation est en cours, mais d'après les différents échanges qui ont lieu au sein du CNUMU, il semble que la simulation ne constitue pas une modalité d'évaluation fréquente des internes de médecine d'urgence français.

Cependant, quelques outils valides existent dans le champ de la médecine d'urgence et nous allons brièvement décrire leurs caractéristiques, ce qui nous permettra de justifier le choix de notre travail qui a résidé dans le développement d'un nouvel outil d'évaluation. Nous avons retenu les principales échelles validées et qui s'intéressaient à des performances ou à des compétences intéressant la médecine d'urgence. Il ne s'agit pas d'une liste exhaustive, mais elle reflète les différentes possibilités apportées par ces échelles. Nous les avons classifiées selon ce qu'elles évaluent : le domaine de compétence, la mesure d'une performance individuelle ou d'une performance d'équipe et le type de professionnel évalué, le type d'échelle (check-list, score de notation globale, score de performance globale), les compétences évaluées et enfin leur possibilité d'évaluer des scénario précis et donc très contextualisés ou bien de s'appliquer à de nombreuses situations cliniques.

Les pionniers en matière de simulation et d'évaluation par la simulation sont les anesthésistes, qui ont, dès la fin des années 90, commencé à développer des outils de validation de la performance. Ainsi, de nombreux outils s'intéressent à la performance de l'anesthésiste seul (qui

exerce son leadership), au sein d'une équipe, ou bien au travail d'équipe. La première échelle validée est l' Anaesthetists' non-technical skills, ANTS (Fletcher et al., 2003). Certaines de ces échelles ont été dérivées pour la médecine d'urgence puisque les situations d'urgence vitale de ces deux disciplines sont parfois les mêmes ou ont des points communs majeurs (qui pourraient s'apparenter à des invariants). Une méta-analyse récente dans le domaine de l'anesthésie a identifié qu'en matière d'évaluation des habiletés non techniques d'un anesthésiste, l'échelle ANTS était celle dont le processus de validation été le plus développé (Boet et al., 2018). Le même travail, mené par la même équipe, mais s'intéressant cette fois à l'évaluation des habiletés non techniques d'une équipe prenant en charge une urgence vitale a mis en évidence des qualités identiques pour l'échelle Team Emergency Assessment Measure, appelée TEAM (Cooper et al., 2010) . Il s'agit de l'échelle que nous avons utilisé dans la troisième partie de notre travail. La plupart des échelles validées se sont intéressés soit à des habiletés non-techniques (telle que les ressources de management d'une situation de crise (Crisis Ressource Management ou CRM), la communication d'équipe, le sentiment d'efficacité personnelle, le travail au sein d'une équipe…), soit à des habiletés techniques telles que l'application d'un algorithme (échelle ABCDE) ou à la réalisation d'un geste technique en dehors d'un scénario complexe (Guly, 2003; Oriot et al., 2012; Peran et al., 2020; Thim et al., 2012).

Certaines équipes ont développé des échelles qui s'intéressent à la fois aux habiletés techniques et non techniques, dans le domaine des urgences pédiatriques principalement (échelles STAT, LCS-KCS, PALS) ou dans le domaine de la médecine d'urgence mais pour des infirmiers ou pour des situations cliniques très précises, telles que la pneumopathie, le syndrome coronarien aigu ou le choc septique (Hall et al., 2012; Hart et al., 2018).

Ainsi, la plupart des échelles validées est destinée à l'évaluation de professionnels de santé déjà experts ou compétents, et ne s'intéressent qu'à des habiletés techniques ou à des habiletés non techniques. Dans le cadre d'une évaluation de performance par la simulation, à destination de

novices, d'étudiants peu performants ou de niveau de performance intermédiaire il n'existait pas d'outils qui permettait d'évaluer à la fois des habiletés techniques (délivrer de l'oxygène, mobiliser des connaissances pertinentes, examiner le patient) et des habiletés non-techniques (formuler un diagnostic à voix haute pour communiquer avec l'équipe, appeler à l'aide, présenter un patient etc.), sauf dans des situations cliniques très précises et pour lesquels un score était développé pour une situation très précise. Or, afin de certifier de futurs internes qui devraient prendre en charge des situations d'urgence vitale et des futurs urgentistes, il nous est apparu intéressant de développer de tels scores qui permettrait à la fois de valider des habiletés « basiques » essentielles à la pratique de tout futur médecin, mais également qui constitue le socle de base de la pratique de la médecine d'urgence.

Par exemple, l'échelle OSCAR évalue les habiletés non-techniques d'une équipe composée d'un anesthésiste, d'un médecin urgentiste et d'un infirmier (Walker et al., 2011) . Elle est issue de trois échelles : l'OTAS (Observational Teamwork Assessment tool for Surgery), l'ANTS et la NOTECH révisée (Non-TECHnical skill) pour les blocs opératoires. Chacune de ces échelles mesure des compétences non-techniques soit pour chaque membre de l'équipe, soit pour l'équipe dans sa globalité. OSCAR analyses six catégories de compétence non-technique : la communication, la coopération, la coordination, la sensibilisation à la surveillance du patient (monitoring/situation awareness), le leadership et la prise de décision. Pour chaque groupe de professionnel, des compétences non-techniques sont mesurées, et notées de 0 à 6. La note représente l'impact du comportement sur l'équipe et oscille entre une mise en danger grave de l'équipe (0) et un haut niveau d'aide pour l'équipe (5)[21]. Une compétence évaluée est, par exemple, pour le médecin urgentiste : « interroge le patient, recueille des informations sur son histoire et communique les informations pertinentes à l'équipe ». Dans cet exemple, le critère d'évaluation ne porte pas sur la nature des informations recueillies, mais bien sur la façon de les

[21] Team Severely Compromised (0) or High level of enhancement to team (5)

recueillir et de les communiquer à son équipe. Cela suggère que l'apprenant est un apprenant expérimenté qui maîtrise déjà les informations à recueillir. Or dans notre situation, nous souhaitons pouvoir analyser la capacité de jeunes apprenants à recueillir les éléments pertinents auprès du patient, puis celle de les communiquer à ses collègues. Mais avant tout, il faut pouvoir notifier quels éléments sont indispensables à connaître dans une situation clinique.

3. Problematique

Ainsi, l'apprentissage des sciences de la santé, et notamment dans le cursus des études médicales, peut être amélioré grâce à l'utilisation de la simulation, qui a prouvé son efficacité, à la fois sur la satisfaction des apprenants, mais également sur l'amélioration de leurs compétences, et, dans de rares cas, pour l'amélioration de la prise en charge des patients (Cook et al., 2011; Curtis et al., 2013; McLaughlin et al., 2008).

De plus, bien qu'encore utilisée de manière hétérogène dans les parcours initiaux de formation, l'utilisation de la simulation tend à se développer, en tant qu'outil de formation. Elle permet d'immerger les apprenants dans des situations dites semi-authentiques, tout en respectant leur sécurité et celle des patients, et semble être un des outils appropriés pour l'apprentissage et l'évaluation des compétences des étudiants en médecine, dans le domaine des urgences, et notamment pour évaluer leur capacité à réagir devant une situation d'urgence vitale, à laquelle ils sont parfois peu exposés pendant leurs stages hospitaliers, et devant lesquelles ils sont rarement en autonomie.

Cependant, son ancrage dans les cursus des études médicales est varié, que cela soit lors de la formation commune des étudiants ou lors du parcours de spécialisation, et dans un but de formation (Allain et al., 2018; Granry, 2012). En effet, l'arrêté de 2013 qui définit les modalités des deux premiers cycles des études médicales mentionne une formation par la simulation, mais sans la définir ou la cadrer, ni même la formaliser. Dans le cadre de l'évaluation, là encore, la

simulation est peu utilisée pour les étudiants en médecine, de même que pour les internes ou les médecins certifiés en France (Arrêté du 8 avril 2013 relatif au régime des études en vue du premier et du deuxième cycle des études médicales, 2013). Outre-Atlantique, c'est une activité qui se développe et qui permet de valider des cursus de spécialité ou de certifier à nouveau les médecins (Ahmed et al., 2010; Blew et al., 2010; Boet et al., 2014), mais son utilisation reste un défi pour les universités, notamment dans le cadre de l'approche par compétence, au sein de laquelle, l'évaluation elle-même reste un défi (Epstein, 2007).

Un des défi consiste notamment à appréhender les différentes caractéristiques et critères de qualité de l'évaluation au sein d'une approche par compétences, et de développer des outils adaptés à cette approche, mais qui seront également acceptables au sein des systèmes de formation actuels, qui, pour la plupart n'ont pas adopté tous les critères de l'approche par compétence et notamment, la continuité entre les activité d'enseignement et d'évaluation, un enseignement centré sur le rythme de progression des apprenants, un jugement évaluatif qui repose sur des mesures multiples, réalisées à l'aide d'outils variés et des critères de jugement en rapport avec la progression de l'apprenant, qui fournissent un rétroaction qui fait sens pour l'apprenant plus qu'avec des mesures quantifiées, sous forme de notes, mais qui fournissent une rétroaction de faible qualité.

Ainsi, le travail de thèse se situera entre approche par compétences (APC), et cadre de travail actuel, qui est issu d'une approche par objectifs, mais tend à développer l'APC. Actuellement il n'existe aucune faculté de médecine qui a mis en place une approche par compétence dans son intégralité, mais la future réforme va donner une place de plus en plus importante aux compétences, grâce à la décision de les évaluer à la fin de la sixième année de médecine principalement avec la mise en place d'Examens Cliniques Objectifs Structurés (ECOS) nationaux. Notre travail se situe entre deux approches évaluatives : l'approche behavioriste avec une évaluation centrée sur la mesure et l'observation, actuellement en vigueur dans les facultés,

et une approche par compétence qui fait son entrée dans les études médicales et qui donne une place importante à l'évaluation formative, en situation (Fontaine & Loye, 2017). Pour autant, la disparition des notes au profit d'une validation des acquis grâce à système « réussite/échec », préconisé dans l'APC, n'est pas encore à l'ordre du jour.

Plusieurs questions se posent alors et notamment celle du paradoxe entre un outil de formation qui est de plus en plus répandu dans les facultés (la simulation), mais qui ne semble pas encore trouver une place « officielle » dans les programmes de formation médicale, alors même qu'il s'agit d'une formation qui nécessite des investissements importants de la part des facultés. De nombreux obstacles sont connus et expliquent la variabilité d'implantation de la simulation : obstacles financiers, mais également de temps et de formation des enseignants (Hosny et al., 2017; Sawaya et al., 2021). Une autre hypothèse, si l'on raisonne dans le champ de l'activité, pourrait être celle d'un manque de place reconnue à l'université, pour une méthode d'enseignement qui se trouve à la frontière entre l'université et l'hôpital et qui pourrait constituer en elle-même un nouveau système d'activité avec ses propres règles, ses propres enseignants (Berragan, 2013). Enfin, la simulation ne fait pas encore partie des outils d'évaluation des compétences des étudiants, même si la récente réforme des études médicales introduit une évaluation des compétences grâce à des situations simulées dans le cadre des ECOS. Ceci pourrait également expliquer sa variabilité d'utilisation dans les facultés de médecine. En effet, si elle devenait un outil d'évaluation intégré dans les programmes de formation et d'évaluation, alors la simulation serait ainsi reconnue par les étudiants mais également par les enseignants, puisque l'évaluation constitue une étape à part entière du processus de formation et d'apprentissage (Vial, 2012). La simulation comme outil d'évaluation permettrait alors un meilleur alignement des activités d'apprentissage et d'évaluation, ce qui permettrait de lui donner une place plus importante dans la formation.

Ainsi, il serait intéressant de comprendre ce que l'évaluation par la simulation peut apporter aux enseignants, apprenants, institutions au sein d'une approche par compétences, et plus spécifiquement dans le cadre du cursus de la médecine d'urgence. Dans le cadre des compétences requises en médecine d'urgence, il est établi que les étudiants en médecine ne sont pas prêts à prendre en charge les patients au premier jour de leur internat, puis plus tard, au cours de leur cursus (Drummond et al., 2016; McEvoy et al., 2014; Tofil et al., 2014; Xi et al., 2015). Nous avons imaginé développer un score d'évaluation qui permettrait de mettre en place une évaluation formative puis sommative des apprenants, lors des moments identifiés comme étapes « clés » de leur formation initiale, les préparant au premier jour de l'internat, puis au premier jour de leur mise en autonomie, au début de la phase de consolidation qui correspond à la 4ème année de l'internat pour le cursus de médecine d'urgence.

Objectifs de la recherche

Puisqu'un score associant l'évaluation d'habiletés techniques et non-techniques, dont nous avons vu qu'elles sont essentielles à une prise en charge adaptée des urgences vitales, n'existait pas dans le cadre de l'évaluation d'étudiants en médecine ou d'internes en MU, ou alors uniquement pour des situations très spécifiques définies par le diagnostic final relié à la situation plus qu'à la situation initiale elle-même (appelée encore situation de départ), le premier objectif de la recherche était de développer un score, en identifiant trois situations de départ pour lesquelles évaluer les apprenants grâce à la simulation aurait du sens. Les scores ont été nommés « ACAT » pour Acute Care Assessment Tool, afin d'avoir une portée anglosaxonne pour la publication de la recherche. A chacun des scores a été adjoint un numéro, en fonction de la situation clinique (1, 2, 3).

Une fois que les situations de départ seraient déterminées, l'objectif était de développer, pour chacune d'elle, le contenu des scores, puis d'en tester les qualités psychométriques qui

permettraient de déterminer leur utilisation future, avec pour objectif principal d'obtenir des scores qui pallieraient au manque d'équité perçu par les apprenants (Philippon et al, 2021).

Une fois que ces critères analysés, l'objectif était d'interroger l'utilisabilité des scores créés, d'analyser leur potentielle place dans le cursus de médecine d'urgence, et également d'étudier plus largement la place de l'évaluation au sein d'un enseignement par simulation, afin de dégager éventuellement les avantages et obstacles de l'outil.

CHAPITRE 2 : METHODOLOGIES DE LA RECHERCHE

Ce chapitre est destiné à exposer le cadre dans lequel nous avons développé nos différentes méthodes de recherche. Il fixe dans un premier temps le cadre global de développement d'un outil d'évaluation, puis décrit ensuite les différents outils méthodologiques utilisés.

1. COMMENT CREER UN OUTIL D'EVALUATION VALIDE ?

La difficulté de l'évaluation des compétences tient au fait qu'une compétence ne se résume pas à un comportement observable. Nous l'avons vu, elle est le résultat de plusieurs composantes, qui ne sont pas toutes analysables par l'observation seule. A cette difficulté s'ajoute celle de la fabrication des outils d'évaluation, qui doit être guidée par la nécessité d'obtenir des outils puis un programme d'évaluation qui mesurent, évaluent, analysent ce pour quoi ils ont été créés. Ainsi l'outil d'évaluation doit être porteur de sens, et pas uniquement producteur de mesures. Le point majeur dans l'utilisation d'un outil n'est pas tant la mesure qu'il produit mais l'interprétation qu'on en fait. Dans ce chapitre, nous reprendrons les évolutions de la notion de validité d'un outil, puis nous nous arrêterons sur deux auteurs qui ont particulièrement défini la manière de développer un outil d'évaluation.

1.1 Validité d'un test : différents cadres théoriques et leur évolution

Initialement, la validité d'un test fait référence à la pertinence de sa mesure, c'est-à-dire le degré avec lequel il mesure réellement ce qu'il est censé mesurer. Au sens plus large de son utilisation, la validité ne représente pas uniquement une propriété intrinsèque d'un test, mais elle doit être appréciée et analysée dans le contexte de réalisation de ce dernier, en tenant compte de son mode d'administration et de l'usage qui est ensuite fait de l'interprétation des résultats (André et al., p 132, 2015).

Ainsi, pour la majeure partie des auteurs, la validité d'un test s'appuie sur l'interprétation que l'on fait des scores et au sens qui leur est donné plutôt qu'aux scores eux-mêmes (Boulet & Swanson, 2004; Downing, 2003, p 832). Pour Downing, elle est le « sine qua non » de l'évaluation car sans preuve de validité, l'évaluation médicale n'a pas de signification propre : "The assessment itself is never said to be « valid » or « invalid » rather one speaks of the scientifically sound evidence presented to either support or refute the proposed interpretation of assessment scores, at a particular time period in which the validity evidence was collected" [22] (Downing, 2003, p 830). Dans le contexte de l'évaluation des compétences, un test valide pour évaluer la compétence de clinicien d'un étudiant, devra donc prouver que l'étudiant est capable d'interroger un patient, de l'examiner, puis d'émettre des hypothèses diagnostiques. Le test doit donc pouvoir évaluer une composante non observable de la compétence « clinicien » : le raisonnement clinique. On imagine alors qu'une évaluation composée de QCM aura des difficultés à être valide. Mais si l'on veut évaluer la reconnaissance simple de certains signes cliniques par un étudiant, alors un QCM bien construit pourrait être considéré comme valide.

1.2 Différentes formes de validité

Le concept de validité est donc central dans la création et l'utilisation d'un test et c'est pourquoi il nous a paru intéressant de décrire l'évolution de ses différentes conceptions. Plusieurs formes de validité ont été décrites par les auteurs, pour aboutir à deux concepts majeurs : celui de validité « unifiée » et celui de validité « argument », respectivement attribués à Messick et Downing, puis à Kane.

Les différentes formes de validité sont décrites dans les *Standards for educational and psychological tests* encore appelés *« The standards »*. En 1954, La première version des « Standards » décrit deux types de validité : validité de contenu, validité reliée à un critère et

[22] « On ne parle jamais d'évaluation elle-même valide ou invalide, mais on parle plutôt des preuves rassemblées scientifiquement pour appuyer ou réfuter l'interprétation proposée des scores d'évaluation, à une période donnée, pendant laquelle les preuves de validité ont été recueillies.» (traduction libre)

validité conceptuelle ou « de construit ». Ces trois types de validité ont été utilisés jusque dans les années 90.

La validité de contenu, ou encore validité manifeste est le travail qui s'inscrit autour du contenu d'un test et qui a pour objectif de déterminer que son contenu qualitatif est fidèle au concept évalué. Par exemple, si l'on souhaite évaluer la compétence d'un étudiant à gérer une situation d'urgence, alors il faut s'assurer que les différentes facettes de cette compétence sont présentes dans l'évaluation : savoir mettre un patient en condition en urgence, réaliser un interrogatoire pertinent, avec un mode de communication adapté, mobiliser ses ressources etc. (André et al., 2015, 133; Cook et al., 2014, p 234).

La validité reliée à un critère, encore appelée concomitante ou prédictive, est la forme de validité qui permet de relier ou non le résultat d'un test à celui d'un autre test, considéré comme « gold standard », même si en éducation cela est rarement le cas. En général le test qui sert de référence est celui qui est couramment utilisé, et il est comparé à un nouveau score, à une nouvelle façon d'évaluer. Dans l'hypothèse de l'absence de test standard validé (qui pourrait être en médecine une appréciation en situation réelle, avec pourquoi pas un retour du patient), les chercheurs ont ajouté la troisième forme de validité, dite « de construit » (André et al., 2015, p 132; Cook et al., 2014, p 234).

La notion de construit est majeure dans la recherche de la compréhension de la validité d'un test. Il s'agit, selon Downing, de concepts ou de principes abstraits, qui sont déduit de comportements observés et expliqués par une théorie. Par exemple, la réussite universitaire est un construit qui est déduit des résultats obtenus lors de différents formats d'examens portant sur un domaine de connaissances bien défini (Downing, 2003, p 831). Ainsi la validité « de construit », permet de reconnaître comme valide un test qui lierait entre eux des objets ou des attributs intangibles, non observables, avec des objets ou attributs observés lors d'examens. L'analyse de la relation entre

construit et comportement ou mesure observée est fondée sur une relation supposée entre observation et théorie.

Par exemple, on estime qu'un étudiant a réussi une prise en charge d'un patient qui présente une détresse vitale si dans les cinq minutes il a réalisé un certain nombre d'actions (observables) et que l'état de son patient s'améliore (observable). Mais le lien entre ce que l'évaluateur observe et la réussite de l'étudiant est sous-tendu par un concept qui définit des critères de réussite, et qui ne sont pas tous observables. La mesure du concept est possible car il a été décrit, circonscrit et que ces étapes ont permis d'en extraire des éléments observables (André et al., 2015, p 132). Cette forme de validité sera donc testée en mesurant les attributs observables et en évaluant leur relation théorique au construit (Cook et al., 2015, 561).

Finalement, dans les années 90, les chercheurs et notamment Messick, constatent que les différentes formes de validité, associées aux différentes mesures psychométriques de la reproductibilité d'un test ont les éléments pour constituer un cadre commun à la validité « de construit » et permettent de supporter ou de réfuter un lien entre mesure et construit. Les différentes formes de validité sont alors abandonnées au profit d'un concept unique : la validité « unifiée » de Messick, en 1989 (Cook et al., 2015; Downing, 2003).

La validité unifiée s'intéresse, en plus des caractéristiques de fidélité et de reproductibilité d'un test, aux conséquences de ce dernier. En effet, tout l'enjeu d'une évaluation, si elle est certificative, est de permettre à des professionnels compétents d'exercer tandis que ceux qui ne le sont pas doivent à nouveau s'entraîner avant d'avoir des responsabilités. L'usage et l'interprétation des données devient donc un élément de la validité d'un test, qui doit pouvoir discriminer les étudiants, non pas en fonction d'un niveau de réussite par rapport à un niveau d'apprentissage ou un standard non valide, mais par rapport à leur capacité réelle à être des professionnels compétents (Cook et al., 2015; Downing, 2003).

Ce concept a suscité des débats au sein de la communauté scientifique au sein de laquelle, quelques auteurs se demandent si Messick n'introduit pas une composante de la validité qui dépasse ce pour quoi un test doit être valide, à savoir qu'il doit réellement mesurer ce qu'il prétend mesurer. D'après Scriven, les conséquences issues des mesures faites par un test n'entrent pas dans la validité de ce test, et ne font donc pas partie du processus de sa validation mais elles relèvent plutôt de son utilité, de l'analyse de l'utilisation qui en est faite (Scriven, 2002). C'est dans ce contexte que Kane apporte une nouvelle réflexion sur la validité. Il travaille alors autour d'un cadre pour le processus de validation, en quatre étapes, elles-mêmes réparties en deux entités qu'il nomme « arguments ». Le premier type d'argument est l'argument de validité et le deuxième est l'argument d'interprétation ou usage du test (M. Kane et al., 2005; 2013).

La validité est un concept mouvant, avec des définitions qui peuvent changer selon les cadres de réflexion et l'usage que l'on veut en faire (Figure 10). C'est un concept qui continue à faire débat et qui est toujours l'objet d'interrogations et de recherches. Entre validité rapportée à des propriétés psychométriques, validité plus élargie ou encore validité reliée à des critères, les possibilités de cadre sont larges et variées (André et al., 2015, p 137). Pour notre travail dont une des questions centrales est savoir comment créer un score de qualité et donc un score valide, il nous a semblé intéressant d'approfondir certains de ces concepts afin de pouvoir choisir un cadre. Nous allons maintenant exposer le cadre qui est le plus utilisé dans les études de pédagogie médicale et de simulation en santé et qui a nourrit notre réflexion quant à la démarche de validation des scores d'évaluation.

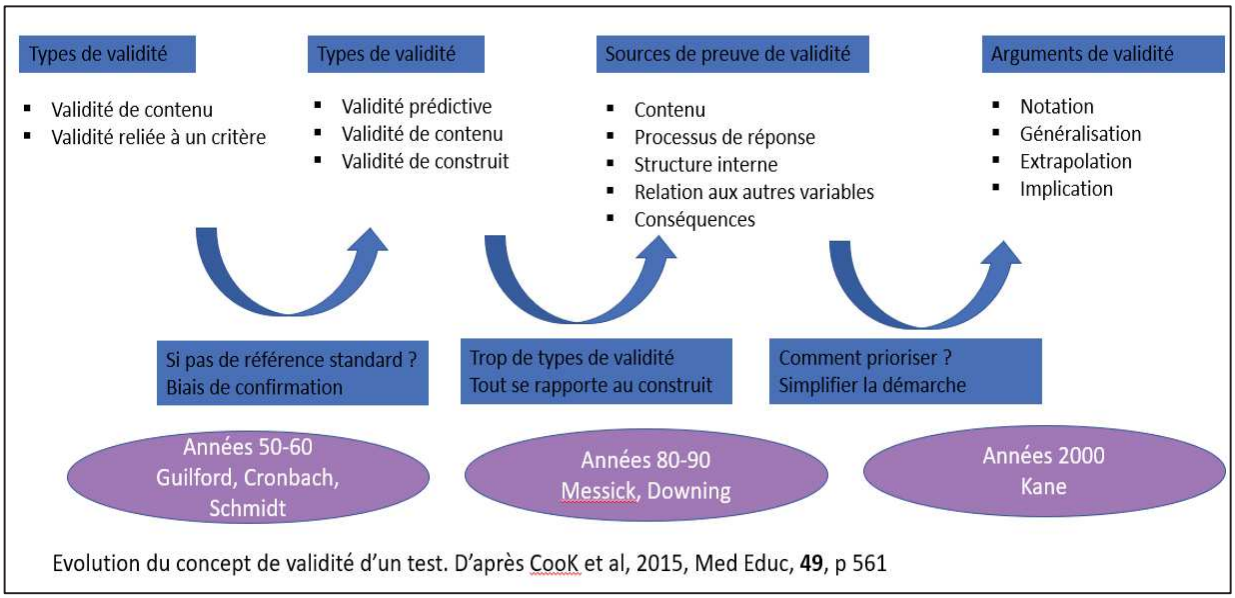


Figure 10 – Evoluation du concept de validité d'un test, d'après Cook et al, 2015

1.3 La validité unifiée et les cinq sources de validité de Messick et Downing

Le cadre théorique de Downing s'appuie sur un modèle de validité unifiée, dont les preuves sont collectées à partir de cinq sources distinctes. En effet, il compare la recherche de preuve de validité à une démarche scientifique en partant du principe que toutes les évaluations doivent « faire avec » des construits. Les construits sont un ensemble d'inférences, de comportements observables qui doivent, au mieux, refléter le construit. Pour Downing et Messick, la validité de contenu ou la validité reliée à un critère sont des exemples d'inférences qu'il faut analyser pour rassembler des preuves de validité. Ainsi, pour lui la démarche de validation d'un test est semblable à une hypothèse scientifique que le chercheur doit tester. Il doit donc rassembler des données et tester, c'est-à-dire réfuter ou approuver la validité de leur apport au test. Un examen n'est donc pas valide ou invalide mais il a plus ou moins prouvé que l'interprétation qui en découle est valide (Downing, 2003, p 830).

Le plus important est ainsi plus le processus de validation d'un test que le « cachet » valide ou non. Ce processus repose sur la formulation d'une hypothèse, la collection et l'évaluation critique

de données issues du test et enfin la recherche d'un lien entre les inférences et le construit. Si on compare à l'évaluation d'un nouveau traitement médical, il faut qu'un nouveau test fournisse des données valides pour un objectif précis (évaluer telle compétence dans tel domaine, comparable avec par exemple le critère de jugement principal dans les essais thérapeutiques). Ce critère doit avoir du sens (j'évalue la capacité à communiquer, car elle est centrale dans la consultation, ou bien j'évalue la mortalité car c'est ce qui fait sens pour une pathologie létale). L'évaluation doit être valide à un instant donné (à tel moment du cursus par exemple, à telle temps de prise en charge du patient grave) et pour une population donnée (qui peut être une profession, un niveau d'expertise et pour terminer le parallèle avec la recherche clinique, un groupe de malade spécifique tel que les patients en détresse respiratoire aigüe âgés de 18 ans au moins et qui ont une suspicion d'infection).

L'intérêt de cette approche consiste donc en la collecte de différentes données, pour construire une « preuve » de validité. Pour Downing, il existe cinq principales « sources of validity evidence », que nous avons traduit par « sources de preuves de validité » (Downing, 2003). Toutes les évaluations ne requièrent pas que toutes les sources soient analysées. En effet, dans le cadre d'une évaluation formative par exemple, les conséquences de l'évaluation ne sont pas aussi importantes (si l'on s'assure que le but de l'évaluation est atteint, à savoir, donner des axes de progression) que lors d'une évaluation certificative qui non seulement doit valider les apprenants compétents, ne pas valider les apprenant qui ne le sont pas mais également doit parfois pouvoir classer les étudiants.

Les cinq sources de preuve de validité sont les suivantes :

- l'étude du contenu : il s'agit de la description rigoureuse des différentes étapes ayant permis de créer le contenu de l'évaluation (différents scénario utilisés, items retenus, options de réponse, et instructions aux apprenants et aux évaluateurs). La question de la preuve du contenu s'intéresse au reflet que le contenu de l'examen offre du domaine

évalué : est-ce que grâce au(x) scénario(s) choisi, et aux items d'évaluation retenus, je vais réellement évaluer la réflexivité de l'apprenant ? Les outils à disposition peuvent être d'anciennes évaluations, une consultation d'expert ou bien l'élaboration d'un « blueprint » (Downing, 2003, p 832)

- l'étude du processus de réponse : il est défini par la preuve de l'intégrité des données, en lien avec ce que l'on veut évaluer. Il s'agit d'une étape qui vise à objectiver des potentielles sources d'erreur dans l'administration du test, dans son contenu. Il s'agit d'un « contrôle qualité » avant l'analyse plus en profondeur de l'outil d'évaluation (Downing, 2003, p 834).

- l'étude de la structure interne : s'intéresse aux caractéristiques psychométriques d'un outil, à ses caractéristiques propres telles que sa fiabilité (analysée par la reproductibilité des différents items, des stations ou encore selon les observateurs), l'analyse des différents items et de leurs capacités discriminatoires, leur difficulté, mais également sa généralisabilité (Downing, 2003, p 834).

- l'étude de la relation aux autres variables : il s'agit d'étudier la relation entre les scores obtenus grâce à un outil d'évaluation et un autre score obtenu avec une autre mesure du même domaine théorique, du même construit ou de la même compétence. L'intérêt est d'étudier la relation entre ces deux mesures qui peut être soit positive (car elle mesure le même construit), soit négative ou nulle (car les deux mesures devraient être indépendantes) (Downing, 2003, p 835).

- l'étude des conséquences d'un outil : s'intéresse à l'impact de l'évaluation sur les évalués : scores obtenus, décisions à prendre à partir des scores, devenir des évalués et à une plus large échelle, impact de l'évaluation sur l'apprentissage et l'enseignement. Cette étape s'intéresse donc à la définition du seuil de validation d'un étudiant, à son

utilisation en fonction des niveaux de compétence ou d'expérience et aux conséquences en termes de certification (Downing, 2003, p 836).

L'exposé théorique de ces concepts est accessible dans un article de Downing, publié en 2003, et l'illustration pratique, associée à l'analyse de son utilisation pour des études de simulation sont bien documentées dans la revue systématique de Cook, publiée en 2014.

Pour conclure, en reprenant les critères de qualité d'une évaluation, les critères de développent d'un outil d'évaluation, voici quelques définitions des différents critères, qui sont utilisés dans ce sens dans la suite de la thèse (Norcini 2010, 2018, Pelaccia, 2015, p 120, 357-69).

- ⇨ Validité : fait référence au degré de conformité avec lequel l'instrument mesure ce qu'il a pour objet de mesurer tant dans le contenu, que dans la construction d'une évaluation et dans sa capacité à identifier le niveau des apprenants.
- ⇨ Fiabilité : concerne la constance d'une évaluation dans sa capacité à mesurer de manière fiable une performance, quel que soit l'évaluateur (reproductibilité), les conditions d'évaluation (début ou fin de journée) et le choix des questions. Une évaluation fiable assure l'équité envers les étudiants et permet de prendre des décisions à enjeu élevé.
- ⇨ Objectivité : le critère d'objectivité est proche de celui de fiabilité car il porte sur la concordance des jugements émis par des évaluateurs différents mais il s'intéresse plus à ce qui constitue la bonne note, ou le bon jugement plutôt que sur la reproductibilité psychométrique. Norcini le définit comme une équivalence entre deux jugements portés par deux personnes ou institutions différents, quant à la décision finale qui sera prise. La formation des enseignants à l'évaluation (pour l'utilisation d'une grille, pour la relecture d'un devoir…) permet d'améliorer l'équivalence des jugements.
- ⇨ Faisabilité : illustre la facilité de réalisation d'une épreuve, (institution, apprenants, évaluateurs, enseignants). Il est donc constitué des moyens humains (nombre

d'évaluateurs), matériels (espace nécessaire, ordinateurs, mannequins …), temporels (temps engagé pour la réalisation de l'évaluation, pour sa correction) et financiers.

- ⇨ Acceptabilité : ce critère est en rapport avec les représentations que se font les apprenants et les enseignants d'une évaluation et à la manière dont elle sera accueillie par toutes les parties-prenantes. Les évaluations innovantes font parfois l'objet de résistances et sont donc de faible acceptabilité, qu'il faut explorer pour aider au changement.
- ⇨ Effets sur l'apprentissage : il s'agit des effets de l'évaluation sur l'apprentissage, détaillés plus haut. Ils peuvent être moteurs ou inhibiteurs et sont également liés avec l'alignement entre activité d'apprentissages et activités évaluative. Sargeant et al. soulignent qu'ils sont également en lien avec la nature et la qualité de la rétroaction fournie aux étudiants, qui peut aller d'une simple note (rétroaction pauvre) à des commentaires plus qualitatifs, précisant les directions à prendre pour les apprentissages futurs (rétroaction élevée) (Sargeant et al., 2011).

Ainsi, pour développer nos trois outils d'évaluation, nous avons choisi de suivre le cadre de Downing, qui permet d'identifier clairement les différentes étapes de développement d'un outil et qui est le plus fréquemment utilisé dans les sciences de la santé, ce qui nous permettra de comparer notre travail à celui d'autres équipes. Le cadre de Kane, plus récent et qui présente une vue plus globale des outils d'évaluation aurait également pu être utilisé, mais puisque nous en sommes au stade de développement de l'outil et pas encore de son utilisation et de sa généralisation, le cadre de Downing nous a semblé plus pertinent (Cook et al., 2015; Downing, 2003; Kane, 2013).

2. DEVELOPPEMENT DU CONTENU DU SCORE : CHOIX DE LA METHODE DELPHI (ARTICLE 1)

Nous rapportons actuellement la façon dont nous avons établi le score par la méthode Delphi. L'objectif premier de notre travail a été de construire un score qui permettrait d'évaluer les compétences des étudiants en médecine d'urgence, dans trois situations cliniques. En l'absence de score préexistant, de données empiriques, et en l'absence de recommandations, nous nous sommes tournés vers l'utilisation des « consensus group methods » ou « groupes de consensus ». Plusieurs types de méthodes existent pour mesurer ou développer un consensus d'experts : la méthode Delphi, la méthode Delphi modifiée, la Technique du Groupe Nominal et la méthode RAND (Humphrey-Murto, et al., 2017c).

Toutes ces méthodes reposent sur une même modalité. Il s'agit de la consultation collective d'experts qui a pour objectif de trouver un consensus, qui aurait alors plus de crédit qu'une prise de décision individuelle ou d'un comité restreint de personnes. Leur objectif déclaré est d'offrir « a structure that supports the democratic representation of a wide range of opinion »[23] (Humphrey-Murto et al., 2017c, p 994).

2.1 La méthode Delphi

2.1.1 Historique

Développée dans les années 50, par des chercheurs de la RAND Corporation (Research and Development Corporation), la méthode Delphi a dû servir pour la première fois dans le domaine de la défense. Son nom fait référence à la pythie de Delphes qui rendait l'oracle lorsqu'elle était interrogée, en général sur des préoccupations qui concernaient le futur. Les oracles rendus n'étaient cependant pas dépourvus d'ambiguïté. La première étude, sponsorisée par l'US Air

[23] « Une structure qui permet une représentation démocratique issue d'une gamme étendue d'opinions » (traduction libre).

Force et menée par la RAND avait pour objectif d'interroger des experts à prédire des probabilités d'attaques ainsi que leur intensité pendant la guerre froide (Turoff & Linstone, 2002, p 10).

Dans le domaine de la santé, et plus précisément de la pédagogie médicale, la méthode Delphi est souvent utilisée pour déterminer la composition d'un programme national d'enseignement, ou encore la composition d'un programme de simulation (Nayahangan et al., 2018). L'objectif est ainsi d'harmoniser les pratiques d'enseignement, à l'aide d'avis d'experts. Il en est de même en clinique ou cette méthode est parfois utilisée pour créer des recommandations ou encore des scores cliniques qui seront ensuite validés par des recherches avec les patients (Garrouste-Orgeas et al., 2009).

2.1.2 Critères de qualité d'une consultation par Delphi

Afin d'être valide, le chercheur qui utilise une consultation d'experts par Delphi se doit de respecter des critères qui permettent d'assurer une recherche de qualité (Jones & Hunter, 1995; Murphy et al., 1998) . Cinq critères sont primordiaux à respecter avec l'utilisation d'un Delphi (Tableau 4).

Tableau 4 – Cinq critères de qualité et six étapes de la méthode Delphi

Critères de qualité	**Etapes**
Anonymat des experts	1. Problématique de recherche
Itération des consultations	2. Revue de la littérature
Feedback	3. Création du questionnaire
Quantification des réponses	4. Tours anonymes, gérés par mail
Structuration des réponses	5. Feedback individuels et collectifs
	6. Résultats

Le caractère anonyme des experts revêt une grande importance car il évite que certains experts qui auraient un statut hiérarchique ou une reconnaissance importante de leurs pairs influencent

la réponse des autres. Chaque réponse est ainsi faite selon une réflexion individuelle, apportée au collectif.

La répétition des consultations s'exprime à travers les différents « tours » du Delphi et permet aux experts de modifier leur avis, suite à l'avis des autres experts. Elle leur permet également de faire de nouvelles propositions au fur et à mesure de l'enquête.

Pour chaque expert, la répétition des consultations permet un rétrocontrôle (feedback), sur ses propres réponses et sur celles des autres experts. Il a alors la possibilité de comparer ses réponses à celles du panel, d'y ajouter des éléments de discussion qui seront décisionnels pour atteindre le consensus.

L'analyse statistique fournit aux experts un résumé quantifiable des réponses du panel ainsi que des siennes propres. Les méthodes de quantification varient, de même que leur analyse et les définitions des seuils d'accord entre les experts sont très différentes d'une étude à l'autre. De nombreuses possibilités existent et ont été décrites dans un livre dédié à la méthode Delphi, par Turoff et Linstone, en 2002 .

Enfin, le fait que les réponses soient structurées par l'équipe de recherche, qu'elles entrent dans un cadre d'analyse et que le retour aux experts soit structuré, permet aux experts d'échanger dans un cadre d'analyse rigoureux.

Les critères de qualité respectés, le déroulement d'une consultation d'expert en suivant la méthode Delphi a plusieurs avantages. Elle peut être réalisée à distance car les experts du fait de leur anonymat ne doivent pas se rencontrer et ils sont contactés par courrier ou maintenant par email. Elle permet de regrouper un grand nombre d'experts d'horizons différents tout en restant flexible dans son utilisation, avec un coût moindre (Ekionea et al., 2011; Humphrey-Murto et al., 2017a). Pour ces différentes raisons nous avons choisi cette méthode pour construire les différentes parties de nos scores.

2.1.3 Différentes étapes d'un Delphi

Après avoir respecté les critères de qualité, la recherche basée sur la méthode Delphi se doit de respecter plusieurs étapes que nous allons ici énumérer rapidement et qui sont également reprises dans la méthodologie de l'article ci-dessous (Humphrey-Murto, et al., 2017b).

Le processus de recherche faisant partie intégrante de l'utilisation d'un Delphi, la première étape d'une telle consultation est la problématique de recherche, qui aboutit à une question spécifique. Dans notre situation, il s'agit de vouloir mettre en place des évaluations de compétences grâce à la simulation, dans le domaine spécifique des urgences, et en particulier de la prise en charge de trois situations d'urgences vitales. La question que nous nous posons alors et que nous allons poser aux experts est de définir des items, qui permettraient d'évaluer des étudiants dans une situation donnée. L'objectif du Delphi est d'établir pour chaque situation clinique, des items qui ne varient pas en fonction du contexte (par exemple étiologie de la situation, prise en charge hospitalière ou extra-hospitalière...) afin de pouvoir utiliser les scores pour une même famille de situation, mais dans plusieurs contextes.

La deuxième étape consiste à effectuer une revue de la littérature. À ce propos, nous avons effectué un examen de la littérature sur les scores déjà existants, grâce à PubMed. Il en ressort principalement ce que nous avons exposé dans la problématique : la plupart des scores existants analysent les compétences techniques et non techniques séparément et ils sont en général destinés à des professionnels de santé en formation continue, ou bien à des équipes pluriprofessionnelles.

Il faut ensuite établir les questionnaires qui seront envoyés aux experts (cf. article) puis les différents « tours » de consultation débutent. Après chaque tour, les réponses des experts sont analysées, individuellement et collectivement, puis un mail avec ces différentes réponses à

nouveau envoyées avec le tour suivant (étape de feedback). La dernière étape consiste en un report résumé des résultats (article et dans notre situation création de trois scores d'évaluation).

2.1.4 Limites de la méthode Delphi

Certains auteurs nous le rappellent avec vigueur, estampiller Delphi une consultation d'experts ne consiste pas en un label de qualité et les étapes sus-décrites doivent avoir été bien respectées (Humphrey-Murto, et al., 2017c). L'utilisation de groupe de consensus fait ainsi face à de nombreuses critiques qu'il ne faut pas ignorer alors que nous allons en utiliser une. Certains surnomment le consensus rien de plus qu'une « ignorance collective » alors que d'autres plaident pour un contrôle par les pairs (Jones & Hunter, 1995).

Les pièges du Delphi existent et il faut les connaitre. Il faut ainsi veiller à ne pas exposer ses opinions avant la consultation, en veillant à formuler les questions et recommandations aux experts sans les influencer et il faut bien veiller à explorer les différents désaccords. Le choix des experts est également primordial (Turoff & Linstone, 2002, p 6). Afin d'évaluer la validité d'un Delphi, certains auteurs préconisent plusieurs points de repère tels que la durée entre les tours, qui devrait être inférieure à trois mois, le choix des experts et leur stabilité (ils poursuivent l'étude à tous les tours) et enfin l'existence d'une modification des premières idées ou questions soumises aux experts (Landeta, 2006).

2.2 Méthodologie de l'étude

Nous avons mené une étude en deux temps, de novembre 2017 à janvier 2018 puis de mars à octobre 218. Les deux périodes consistaient en une consultation d'un groupe d'experts, d'abord pour déterminer les situations cliniques à évaluer, puis à la réalisation d'une méthode de consensus de groupe Delphi pour déterminer un score pour chaque famille de situation. L'avis d'un comité d'éthique n'était pas requis pour ce type de recherche (Arrêté du 12 avril 2018).

2.2.1 Méthodes Delphi : critères utilisés

Lors de la deuxième période, nous avons utilisé une méthode Delphi, en respectant les 5 principes fondamentaux d'une consultation d'expert qui sont : le caractère anonyme des participants, la répétition des consultations, faite de feedback sur les réponses des experts, des interactions entre les experts contrôlées par les investigateurs et une analyse statistique des réponses (Jones & Hunter, 1995; Murphy et al., 1998). Nous avons également respecté les différentes étapes recommandées: la réflexion autour de la question de recherche puis l'analyse de la littérature, le choix des experts, la rédaction des questions et du questionnaire, les différents tours de consultation d'experts, dont chacun est suivi d'un retour sur les réponses de chacun, la définition du consensus et enfin, l'analyse des résultats (Humphrey-Murto et al., 2017b).

2.2.2 Composition du panel d'experts

Les experts de la première consultation étaient les 29 coordonnateurs régionaux du diplôme d'études spécialisées en médecine d'urgence (DESMU). Pour la méthode Delphi, nous avons défini des critères de sélection des experts : ils devaient être francophones, avoir des responsabilités dans l'enseignement de la médecine d'urgence et/ou avoir une expertise acquise par une longue expérience d'enseignement (> 5 ans). Par ailleurs, il fallait également avoir des notions de l'enseignement par simulation ainsi que de ses spécificités. Ils ont été recrutés par mail, et les critères leur ont été soumis, afin de vérifier que leur profil correspondait bien. Ils pouvaient également suggérer des experts qui selon eux, correspondaient aux critères. Ils étaient anonymes les uns des autres. Notre objectif était d'avoir 20 experts par situation clinique.

2.2.3 Envois des questionnaires

Le premier questionnaire a été envoyé aux 29 coordonnateurs de DESMU régionaux, à qui nous avons demandé de choisir puis de classer par ordre d'importance 10 items correspondant à des situations cliniques de départ qui, selon eux, devraient faire l'objet d'une évaluation des étudiants

par la simulation. Les 10 items étaient à choisir parmi 36 items figurant dans le référentiel de formation à destination des étudiants en médecine « Urgences et défaillances viscérales aiguës » (Collégiale des Universitaires de Médecine d'urgence (CNUMU) & Collège National des Enseignants de Thérapeutiques (APNET), 2017). Suite à la consultation, nous avions prévu de sélectionner les trois situations cliniques ayant obtenu le plus de points.

Le deuxième questionnaire a été envoyé aux experts sélectionnés qui ont accepté de participer à l'étude. Il avait pour but d'établir un score à partir d'un grand nombre d'items choisis par les investigateurs selon le référentiel de formation, mais également selon les différentes recommandations en vigueur pour chaque situation clinique. Ces items pouvaient être d'ordre cognitif (par exemple, rechercher les facteurs de risque cardio-vasculaire), technique (réaliser un massage cardiaque externe de qualité) ou encore non technique (appeler la famille, se présenter au patient, communiquer avec l'équipe etc.).

Lors du premier tour, nous avons demandé aux experts de noter les items qui apparaissaient être pertinents à la constitution d'un score qui permettrait d'évaluer un étudiant avec la simulation. Nous avions précisé que le score devrait pouvoir évaluer des étudiants en formation initiale, mais ayant différents niveaux de compétence, de la fin du 2ème cycle jusqu'à la phase de consolidation du 3ème cycle. Les experts devaient noter chaque item grâce à une échelle de Likert, allant de 1 (pas du tout pertinent) à 6 (tout à fait pertinent). Nous avons également demandé aux experts d'ajouter des items qui leur paraissaient avoir été oubliés. Le deuxième tour s'est déroulé de manière identique, après retrait ou ajout de certains items, et renvoi aux experts des items retenus et des commentaires de chacun d'entre eux, de manière anonyme. Certains items pouvaient également être reformulés et fusionnés suite à l'analyse des commentaires. Le troisième tour était différent : tous les items retenus au tour 1 et 2 étaient soumis à évaluation, afin que les experts aient un aperçu global du score. La notation des items devenait binaire : 1 pour item indispensable, 0 pour item à ne pas retenir dans le score final.

2.2.4 Définition des consensus, arrêt du Delphi et analyse statistique

Pour le Delphi, nous avons déterminé à l'avance qu'il serait composé de trois tours, et que l'objectif était d'obtenir des scores composés de 20 catégories. Chaque catégorie pouvait être composée d'un seul ou de plusieurs items. Pour le premier et le deuxième tour, le consensus positif était atteint lorsque plus de 75% des experts estimaient l'item pertinent, soit un premier percentile (Q1) supérieur à 5 : l'item était conservé pour le futur score. Le consensus négatif était défini par 75% des experts qui estimaient l'item non pertinent soit un troisième percentile (Q3) inférieur à 5. Tous les items ne correspondant pas à ces définitions étaient remis au jugement des experts. Pour le troisième tour, comme le score était binaire, le consensus était défini par un accord d'expert > 85%. A chaque tour, les remarques des experts ont été également prises en compte et analysées en les codant manuellement pour modifier la formulation de certains items, les adapter à la simulation et pour en regrouper certains entre eux.

Les scores d'évaluation créés se nomment ACAT 1, 2 ou 3 pour Acute Care Assessment Tool. Le premier score s'intéresse à l'évaluation de la performance d'un apprenant qui prend en charge une situation d'arrêt cardiaque, le deuxième une situation de coma et le troisième une situation de détresse respiratoire aigüe.

3. Analyse du processus de reponse et de la structure interne des scores

Les deuxièmes et troisièmes étapes de développement d'un score s'intéressent au processus de réponse au test ainsi qu'à sa structure interne, dont la reproductibilité, la cohérence entre les items et leur utilisabilité comptent parmi les principaux composants. La deuxième étape de notre recherche a consisté en une évaluation succincte du processus de réponse au score, et en une analyse de la reproductibilité intra et inter-observateur entre deux évaluateurs indépendants dans deux situations différentes : l'utilisation du score grâce à des vidéos de situations de simulation,

puis l'utilisation du score dans des conditions de formation réelles. Nous nous sommes également intéressés à son utilisabilité.

3.1 Analyse du processus de réponse (Article 2)

Dans le cadre de Messick et Downing, la deuxième phase de construction d'un test consiste en une analyse du processus de réponse qui vise à objectiver des sources potentielles d'erreur dans l'administration du test, dans son contenu. Il s'agit d'un « contrôle qualité » avant l'analyse plus en profondeur de l'outil d'évaluation (Downing, 2003, p 834). Il peut être réalisé sous forme quantitative avec des tests psychométriques (étude de la consistance interne, étude de la reproductibilité) et/ou sous forme qualitative, grâce à l'analyse de son utilisabilité avec des situations évaluatives. Cette étape permet également de définir et d'analyser la manière dont sera administrée le test aux apprenants, ainsi que la façon de noter les différents items. Dans notre recherche, nous avons à la fois mené une étude qualitative en utilisant les scores obtenus par la méthode Delphi et en les confrontant à des situations simulées filmées, mais également déterminé la manière dont nous allions pondérer les différents items. A ce stade, nous avons également décidé d'ajouter aux scores ACAT une échelle d'analyse de la performance globale des apprenants.

3.2 Analyse de la reproductibilité d'un score (Article 2 et 3)

La reproductibilité (reproducibility) d'un score est une de ses caractéristiques psychométriques, qui participe à prouver sa validité et sa fiabilité (reliability). La reproductibilité intra ou inter-observateur permet de déterminer si deux observateurs (ou plus) examinant la même situation vont produire le même résultat. La reproductibilité est une des caractéristiques de la fiabilité d'un test et bien souvent la fiabilité est « réduite » à la reproductibilité des mesures faites par deux observateurs, ou à la reproductibilité des mesures dans le temps (Downing, 2004). Dans cette perspective, elle s'intéresse donc à la mesure plus qu'à l'instrument lui-même. Pour compléter cette caractéristique on peut également s'intéresser à la consistance de la mesure pour un même

apprenant. La fiabilité renvoie également à une notion plus qualitative de l'analyse d'un test. Dans le cadre de Downing, la fiabilité d'une mesure, représentée par sa consistance et sa reproductibilité est une des sources majeures de validité d'un test.

Les résultats d'évaluation de performance, telles que les évaluations au lit du patient, sur simulateur, ou certaines évaluations orales dépendent le plus souvent de la consistance de l'évaluateur et donc de la reproductibilité du test utilisé. Ainsi, pour les évaluations qui reposent sur un jugement humain, la première source d'étude de fiabilité est la reproductibilité des jugements portés au moyen de l'outil utilisé. La plus grande menace pesant sur le test étant la faible consistance d'un même évaluateur, ou la faible reproductibilité de la mesure entre deux évaluateurs (Downing, 2004, p 1008) : "the internal consistency reliability of the rating scale (all items rated for each student) may be of some marginal interest to establish some communality for the construct assessed by the rating scale, but interrater reliability is surely the most important type of reliability to estimate for rater-type assessments"[24]. Le cadre méthodologique utilisé pour les études de reproductibilité des scores est fondé sur les recommandations de Downing, ainsi que sur celles de Kottner, publiées en 2011 et accompagnées de recommandations de publications pour de telles études (Downing, 2004, Kottner, 2011).

Dans la recherche présentée ici, l'analyse de la reproductibilité permet l'analyse de la mesure de concordance entre deux mesures d'une seule variable (ici la performance de l'étudiant en situation d'évaluation par la simulation). La reproductibilité est représentée classiquement par le ratio entre la variance du score réel (true score) et la variance du score total. Plusieurs méthodes existent pour évaluer cette variance :

[24] « la fiabilité de la cohérence interne de l'échelle d'évaluation (tous les éléments évalués pour chaque étudiant) peut présenter un intérêt marginal pour établir une certaine cohérence pour le construit évalué par l'échelle d'évaluation, mais la fiabilité inter-évaluateurs est certainement le type de fiabilité le plus important à estimer pour les évaluations reposant sur un jugement émis par un évaluateur » (traduction libre).

- le calcul du pourcentage d'agrément entre les observateurs mais il ne prend pas en compte le rôle du hasard dans le pourcentage obtenu ;
- le calcul du coefficient de kappa (qui est une sorte de coefficient de corrélation, et qui tient compte de la proportion de hasard dans les résultats obtenus) ;
- la théorie de la généralisabilité, la méthode la plus élégante selon Downing, mais également la plus complexe puisqu'elle prend en compte l'erreur attribuable à chaque « élément » de l'évaluation : l'évaluateur, l'évalué, les items du test. Il faut donc que l'évalué puisse être dans plusieurs situations d'évaluation pour apprécier la variance de sa performance au sein de chaque situation
- le coefficient de corrélation intra-classe (CCI), (un peu moins élégant !) : plus accessible, et qui utilise des analyses de variance pour estimer la variance associée aux différents facteurs en jeu dans la fiabilité des mesures au test évalués. Il permet d'étudier aussi bien la variabilité d'un évaluateur que celles entre plusieurs évaluateurs.

Les différents coefficients qui illustrent la reproductibilité varient entre 0 et 1. Pour considérer que la reproductibilité est atteinte, il faut qu'il soit supérieur à 0,6. Lorsqu'il est supérieur à 0,8, la reproductibilité est considérée excellente (Ciccetti, 1994). Cependant, afin de traduire ces chiffres pour qu'ils aient un sens éducatif, il a été souvent considéré que pour des évaluations à enjeu élevé, la reproductibilité des mesures du test devrait être supérieure à 0,9, de même qu'elle devrait être entre 0,8 et 0,9 pour des évaluations sommatives, alors qu'elle peut se situer entre 0,7 ou 0,8 pour des évaluations formatives ou réalisées à un plan local (Downing, 2004, p 1010).

La méthode employée pour mesurer la reproductibilité dépend de la nature des variables analysées. Ainsi, pour étudier la reproductibilité du score ACAT, qui est une variable continue, il est recommandé d'utiliser le coefficient de corrélation intra-classe (ICC) qui permet l'estimation de la concordance globale des mesures. L'ICC se situe entre 0 et 1 et l'absence d'agrément à l'agrément parfait.

Pour étudier la reproductibilité du score de performance globale, qui est une variable catégorielle, comportant plus de deux catégories, il est recommandé d'utiliser le Kappa de Cohen pondéré (Kottner et al., 2011). Il est utile pour savoir si les scores entre les observateurs ont varié dans une faible ou dans une large mesure. Par exemple, un enseignant doit évaluer si un étudiant a été tout le temps, souvent, parfois ou jamais présent en cours. Alors si les enseignants estiment que l'étudiant était souvent ou parfois là versus tout le temps et jamais, le niveau de discordance passe de faible à élevé. Ce score de Kappa pondéré permet d'obtenir une valeur élevée de concordance quand les réponses correspondent quasiment et à l'inverse si la différence entre les réponses est large, alors le coefficient sera bas. Le coefficient de Kappa varie théoriquement entre -1 et 1 et son interprétation est la suivante :

0 = agrément dû à la chance;
0.10–0.20 = agrément léger;
0.21–0.40 = agreement faible;
0.41–0.60 = agrément modéré;
0.61–0.80 = agrément substantiel;
0.81–0.99 = agrément fort;
1 = agrément parfait.

3.3 Analyse de la cohérence interne (ou consistance) du score (Article 3)

Il s'agit ici d'analyser la corrélation entre les items d'un score et donc de s'intéresser à la cohérence avec laquelle le score représente le concept évalué. Il est également possible de s'intéresser aux sous-catégories d'un score, qui représentent parfois des concepts différents les uns des autres, notamment dans une évaluation de situation clinique qui peut s'intéresser à la fois au raisonnement clinique et à la façon de réaliser un geste au cours de la situation (Yudkowsky et al., 2019, p 39). Il est également possible de mesurer la cohérence d'un résultat donné par un test, pour un même apprenant, à deux moments différents. Son niveau de

performance ne devrait alors pas varier de manière significative et le test devrait pouvoir donner les mêmes résultats (en ce qui concerne la conséquence du test par exemple).

Pour analyser la cohérence interne du test, on utilise le plus souvent le coefficient alpha de Cronbach. En mesurant la proportion de variance au sein d'un score, il représente la mesure de l'homogénéité de ce score. Un score est considéré comme cohérent dès lors que tous ses items convergent vers la même intensité de réponse ou encore que chaque item d'un test évalue de la même manière un domaine de compétence. Ainsi, plus les items sont corrélés entre eux et au score total du score, plus le score a une cohérence élevée. Il est ainsi possible d'évaluer la cohérence entre les items (inter-items corrélation) et la cohérence entre chacun des items et le résultat total obtenu au score (item-total corrélation). Le coefficient alpha de Cronbach s'interprète ainsi (Downing, 2004) :

0.60–0.69 : cohérence questionnable
0.70–0.79 : cohérence acceptable, pour des examens à faible enjeu (local)
0.80–0.90 : bonne cohérence, acceptable pour des examens à enjeu modéré (fin d'année)
> 0.90 : excellente cohérence, nécessaire pour des examens à enjeu élevé (diplôme fin d'études).

3.4 Méthodologie de l'étude du processus de réponse et de la reproductibilité (Article 2)

Ainsi, pour analyser le processus de réponse de réponse, nous avons mis en place une étude qui reposait sur l'analyse de vidéo de situations simulées.

3.4.1 Design de l'étude

Nous avons mené une étude prospective, s'appuyant sur l'analyse de vidéo de situations simulées pour chacune des trois situations évaluées. Les séances de simulation ont eu lieu entre septembre 2019 et octobre 2020 et elles concernaient soit des étudiants en 4ème année de médecine soit des internes en fin de première année d'internat de médecine d'urgence.

Deux évaluateurs indépendants ont analysé le processus de réponse en déterminant la cotation des items et en réalisant une analyse qualitative de chaque item. Cette analyse a été réalisée en utilisant le score et le processus de notation sur des vidéos « test », à raison de trois pour chaque score. Les évaluateurs devaient faire une analyse qualitative de chaque item afin de voir s'il était utilisable dans trois situations variées, avec des apprenants de niveau varié. A l'issue de ce processus l'énoncé définitif des items et leur cotation a été déterminée.

Chaque score a été utilisé par deux évaluateurs indépendants, pour noter les situations simulées filmées. Les évaluateurs étaient médecins urgentistes et formateurs en simulation. Ils notaient les performances observées de 0 à 20, selon les consignes et les définitions utilisées pour chaque item de chaque score. Chaque évaluateur devait noter les performances à deux reprises, avec au moins deux mois d'intervalle entre chaque visualisation des vidéos. Au cours de la recherche un des deux évaluateurs a dû arrêter la lecture des vidéos et il a été remplacé par un troisième évaluateur. Le premier évaluateur (la doctorante) a pu lire et analyser toutes les vidéos, et ses analyses ont été comparées à celles des deux autres.

Les évaluateurs ont reçu une formation courte d'une heure, qui portait sur le contenu des scores, et sur la manière de noter les comportements observés. De même les différents niveaux du score de performance globale étaient décrits (novice, débutant avancé, compétent, efficient, expert). La notation devait se faire au plus des conditions réelles d'observation en simulation, à savoir sans interrompre la visualisation de la vidéo, en notant « au fil de l'eau », mais avec la possibilité de prendre des notes.

3.4.2 Objectifs de l'étude

L'objectif principal de l'étude était de valider les scores ACAT en analysant la reproductibilité de chaque score, entre deux évaluateurs et pour un même évaluateur.

Les objectifs secondaires étaient également d'analyser le processus de réponse des scores et l'analyse de la capacité du score à s'adapter à plusieurs situations cliniques.

Le critère de jugement principal était la reproductibilité inter et intra-observateur. Les critères de jugement secondaires étaient représentés par l'analyse du pourcentage d'items non remplis lors de l'observation, soit parce que l'évaluateur avait oublié (case vide sans justification), soit parce que l'item n'était pas pertinent pour la situation donnée (case avec mention « ne s'applique pas »).

3.4.3 Participants et descriptions des sessions de simulation

Les vidéos enregistrées étaient issues de sessions de simulation de deux centres de simulations différents, avec deux niveaux d'apprenants différents. Le premier centre délivrait des sessions de simulations aux étudiants de 4ème année de médecine, qui réalisaient leur module de formation intitulé « Certificat Couplé de Pratique Clinique – Urgences et Réanimation ». Au cours du module, ils participaient à trois sessions de simulation, dont certaines portaient sur les situations évaluées par les scores et pour lesquelles nous avons pu obtenir des enregistrements vidéo. Le deuxième centre avait mis en place des journées de formation pour les internes de fin de première année d'internat en médecine d'urgence au sein desquelles, les trois situations évaluées par les scores étaient mises en place et ont pu être filmées et enregistrées.

Les scénarios des sessions de simulation ont été créés par les enseignants responsables des différents ateliers de simulation, et donc indépendamment de l'étude réalisée. Ils devaient durer au moins 10 minutes, et étaient destinées à être réalisés par une équipe de deux à quatre étudiants, sans notion d'interprofessionnalité, puisque les sessions de simulation se déroulaient dans le cadre de la formation initiale, qui n'est pas réalisée avec les autres professions dans les deux facultés où s'est déroulée l'étude. Ainsi, les apprenants devaient parfois jouer un autre rôle que le leur (infirmière, interne si externe et médecin junior si interne).

Tous les apprenants ont donné leur consentement pour être filmés et enregistrés, soit au début de l'année universitaire dans le cadre de l'enseignement par simulation, soit au début de la journée de formation, pour les internes de médecine d'urgence. Les vidéos ont été collectées de manière prospective et la collecte des données a été approuvée par le conseil de faculté dans lequel siège les représentants des étudiants et des enseignants.

3.4.4 Analyse des données

Les données étaient analysées grâce au logiciel NCSS. La reproductibilité inter et intra-observateur a été mesurée en utilisant le coefficient de corrélation intra-classe (CCI) ou intra-class corrélation coefficient (ICC). L'utilisabilité et la pertinence des scores ont été analysées à l'aide de pourcentage de remplissage des items et du pourcentage d'items non remplis car non applicables à la situation.

4. Analyse de la relation du score aux autres variables (article 3)

Une des autres sources de validité d'un test est sa relation aux autres variables qui caractérisent la performance de l'apprenant (dans notre situation). Cela peut-être un autre test validé ou bien un autre score (tel que le score de performance globale dans notre étude) ou encore le niveau de l'apprenant. Cette dernière variable est intéressante à étudier, dans la mesure où un test doit également pouvoir différencier les apprenants en fonction de leur niveau d'expertise, surtout s'il est certificatif.

Pour étudier ces différentes relations, on utilise la corrélation entre les deux variables analysées, qui se réfère à une potentielle relation (ou non) entre deux variables, alors que l'agrément, utilisée pour la reproductibilité entre les variables, s'intéresse à leur concordance. Par exemple, deux séries d'observations fortement corrélées peuvent présenter une faible concordance alors que deux séries de valeurs qui concordent sont fortement corrélées. Ainsi, le score total étudié peut être concordant entre deux observateurs, il ne sera pas nécessairement corrélé au niveau de

l'apprenant. Des mesures peuvent être corrélées entre elle, mais pas concordantes. L'analyse de la corrélation s'attarde donc à donner des informations sur la force de la relation entre deux valeurs/deux groupes de valeurs, mais ne capture pas la cohérence/l'agrément entre ces valeurs entre les évaluateurs et finalement la manière dont elles ont été obtenues (Kottner et al. 2011; Stolarova et al. 2014). Cependant, des hypothèses concernant la reproductibilité inter-observateur sont souvent traitées avec des analyses de corrélation. La faille de ces conclusions réside dans le fait qu'une corrélation linéaire peut être obtenue entre deux évaluateurs, même s'ils diffèrent systématiquement l'un de l'autre et même sans un seul accord. En revanche un accord n'est atteint que lorsque les points se trouvent sur la ligne d'égalité des deux évaluations.

Pour étudier la corrélation du score ACAT à d'autres variables, nous avons pu nous intéresser à la corrélation entre :

- le score ACAT et le score de performance globale
- le score ACAT et le score TEAM qui est un score d'analyse de la performance des habiletés non-techniques d'une équipe et qui est le seul score validé en français (Maignan et al., 2016, Annexe 2)
- le score ACAT et le niveau des apprenants, car nous avons pu recruter, lors de la troisième étude, des internes travaillant aux urgences de niveau d'étude différent.

Pour chacune de ces analyses, nous avons utilisé le score de corrélation de Pearson.

5. Analyse des consequences du test (article 3)

Une des conséquences d'un test est déterminée par le fait, pour un évalué, de le réussir ou pas. Le seuil de réussite d'un test est le score qui détermine une frontière conceptuelle entre une performance acceptable et une autre qui ne l'est pas (Cusimano, 1996). Dans le cadre de développement d'un test selon le cadre de Downing, l'établissement du seuil correspond ainsi à ce qu'il appelle « les conséquences du test », qui peuvent être d'ordre institutionnel (une

certification) ou pédagogiques (prenant la forme d'une rétroaction ou de la réalisation d'une nouvelle activité d'enseignement par exemple).

En 2007, les *Standard setting* ont présenté et commenté une quinzaine de méthodes ou variantes envisageables pour fixer des standards parmi lesquelles, la méthode d'Angoff, qui est la plus fréquemment utilisée et s'appuie sur une quarantaine d'années de recherche et de mise en pratique (Cizek & Bunch, 2007; Hurtz & Auerbach, 2003). De plus, elle est assez simple d'utilisation et est facile à appliquer même pour des novices, ce qui était notre cas puisque nous l'utilisions pour la première fois (Wheaton & Parry, 2012). Ainsi la méthode a démontré à plusieurs reprises qu'elle fournissait, par rapport aux autres méthodes, un meilleur équilibre entre adéquation technique et praticabilité (Cizek & Bunch, 2007, p 82).

Par ailleurs, au-delà du fait que la méthode d'Angoff est facilement utilisable, son intérêt principal réside dans le fait qu'elle est fondée sur le contenu du score et pas sur la performance de l'individu par rapport à la performance du groupe, ce qui est le cas de la méthode des groupes limites ou de la méthode des groupes contrastés (Cusimano, 1996). En effet, ces méthodes fixent le seuil de réussite par rapport à l'ensemble de ce qu'a produit le groupe et qui utilisent la note « seuil » par rapport à la note médiane d'un groupe proche du niveau minimum acceptable pour la première ou bien selon les distributions des étudiants d'un groupe compétent ou non pour la deuxième (Homer et al., 2017).

A l'inverse, l'utilisation de la méthode d'Angoff s'appuie sur la création d'un panel d'experts qui estiment, item par item, la probabilité de réussite d'apprenants « minimalement compétents ». La définition de minimalement compétent est la suivante : ce sont des apprenants dont la performance est suffisante pour le niveau requis selon les avis préalables des experts. Le terme de « juste suffisant » peut également correspondre à ce qu'on essaie de définir avec la méthode d'Angoff. La Société Spécialiste des Examens au Canada décrit assez bien le concept de minimalement compétent : il faut s'imaginer sur le lieu de travail (et donc finalement revenir à

l'activité) et penser à ses collègues. Certains sont des « vedettes », identifiés comme compétents voire experts grâce à la qualité et au rendement de leur travail. D'autres ont une qualité de travail plus mauvaise, voire de devraient peut-être pas avoir le droit d'exercice de la profession. Entre ces deux extrêmes, il existe un groupe qui constitue le niveau de compétence minimale. Pour un test, le candidat limite appartient à ce groupe. L'intérêt d'identifier des apprenants qui ne sont ni clairement incompétents ou compétents est d'éviter d'éliminer des apprenants qui sont en fait compétents (que l'on pourrait voir comme des faux négatifs, si on fait référence à un test diagnostic) car le seuil est trop élevé et, à l'inverse, de valider des apprenants incompétents (ou faux positifs) à cause d'un seuil trop bas (Canada's Testing Company, 2014).

Plusieurs déclinaisons de la méthode d'Angoff existent. La plus répandue est celle qui cherche à déterminer le pourcentage de réussite d'un étudiant minimalement compétent pour un item donné. La question ainsi posée aux experts est la suivante : « parmi 100 étudiants de ce niveau, combien auront la compétence minimum requise pour réussir cet item ? ». Ou encore, « combien de candidats minimalement compétents répondront à cette question correctement ? ». Les experts déterminent alors un pourcentage pour chaque question et le seuil est obtenu en calculant la moyenne des probabilités de réussite pour chaque item, puis la moyenne des moyennes de chaque item. La méthode d'Angoff modifiée, qui a pour objectif d'améliorer le processus, introduit une discussion entre les experts, après les premières appréciations après une étude des variations des pourcentages donnés par les experts, jusqu'à atteindre une variation inférieure ou égale à 30%, pour chaque item. La procédure se déroule en deux ou trois étapes (*rounds*) entre lesquelles les experts reçoivent différentes informations et s'efforcent, lors de la discussion, de diminuer leurs divergences d'estimation. Une fois les experts accordés, on calcule la somme des probabilités de chaque expert, pour chaque item, et en faisant la moyenne des scores obtenus, on obtient le score qui identifie le seuil de réussite (Figure 11) (Hurtz & Auerbach, 2003; Ricker, 2006).

Plusieurs possibilités et modifications de la méthode existent. Certains chercheurs recommandent ainsi de former les experts au niveau attendu pour définit un apprenant minimalement compétent afin d'avoir des estimations plus fines de la réponse aux items, d'autres estiment que cela n'est pas nécessaire. La question n'est pas tranchée (Hurtz & Auerbach, 2003; Ricker, K.L., 2006). Une autre possibilité est également de définir si oui ou non un apprenant répondra ou agira correctement. Elle est appelée la méthode Angoff oui/non, mais reste peu utilisée car pourvoyeuse de biais importants. Elle a été modérée par la méthode Angoff à trois niveaux qui introduit, en plus du oui/non une option de « peut-être » (Plake & Impara, 1997; Yudkowsky et al., 2008).

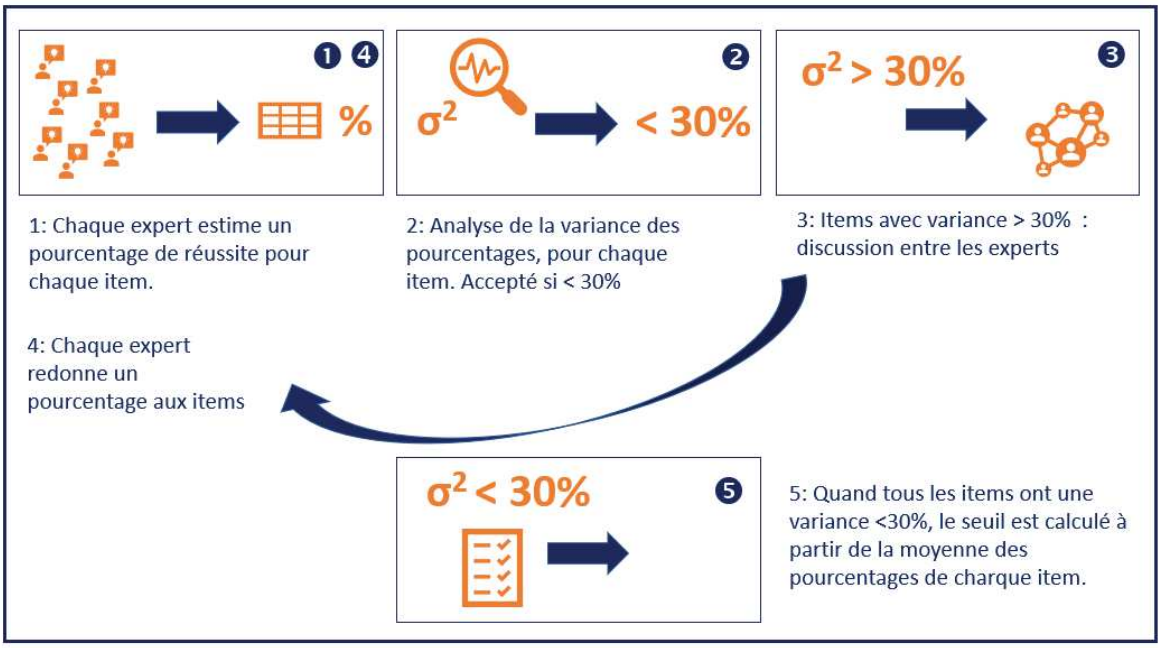


Figure 11 – Différentes étapes de la Méthode d'Angoff, réalisée pour l'étude et envoyée aux experts avec la fiche explicative

6. Analyse de l'acceptabilite et de la perception du score par les enseignants (article 3)

6.1 Etude qualitative de la perception des scores par les enseignants (Article 3)

Afin de recueillir la perception des enseignants-évaluateurs quant à l'utilisation et au contenu des scores ACAT, nous avons mené une étude qualitative au moyen de deux outils : un questionnaire et une enquête par focus groupes, proposés à l'ensemble des enseignants ayant participé à l'évaluation des apprenants.

Menant une enquête « qualitative », plusieurs choix se sont posés quant aux méthodes à employer : l'observation des sessions de simulation, l'entretien individuel, l'entretien collectif ou encore la réalisation de questionnaires. Nous avons retenu la méthode de l'entretien collectif, complété par des questionnaires distribués aux enseignants après les sessions de simulation.

6.1.1 Cadre théorique : la théorie ancrée

Le cadre de notre recherche s'inscrit dans celui de la théorie ancrée qui permet de s'appuyer sur les données relevées sur le terrain pour émettre des hypothèses de travail futur. Dans le cadre de notre recherche, il s'agit d'émettre des hypothèses ou thématiques de réflexion quant à l'utilisation des scores ACAT, selon les perspectives des enseignants. L'idée de ce cadre de recherche est de ne pas avoir « une théorie préconçue en tête » avant le début de la recherche (Corbin & Strauss, 2008; Kennedy & Lingard, 2006; Lingard et al., 2008). Avec cette approche, les théories sont un ensemble de concepts développés et reliés entre eux par l'analyse des déclarations ou des observations réalisées sur le terrain. Les concepts émergent des données recueillies et sont ensuite analysés au regard de la littérature ou d'autres théories existantes.
Utiliser ce cadre de travail impose aux chercheurs de travailler leur relation à l'objet de recherche, et leurs représentations de cet objet. Dans le cas présent, il a fallu à la doctorante prendre du recul par rapport aux scores et réfléchir à sa place dans la recherche puisqu'elle allait mener les entretiens après avoir également participé à la mise en place de la simulation in situ dans un

centre. Ainsi, la doctorante a pu prendre le temps d'analyser ses représentations propres de l'évaluation par la simulation, grâce à l'analyse de la littérature exposée dans la première partie de ce travail, mais également grâce aux différents entretiens et discussions avec le directeur du travail et avec ses collègues du laboratoire de simulation, étrangers à la recherche. L'objectif de ces réflexions était de pouvoir identifier les représentations présentes au moment de la recherche, liées à l'institution et à ses changements actuels (notamment une place plus importante donnée à l'approche par compétences dans les études de médecine), liées à sa formation propre et au travail de terrain en centre de simulation ou aux côtés des étudiants dans les services d'urgence. De plus, lors des entretiens, la doctorante connaissait certains participants et il a donc fallu travailler la distance nécessaire à adopter, ainsi que le rappel régulier de l'anonymisation des entretiens.

Nous avons choisi de réaliser des entretiens afin de laisser la parole aux enseignants, en partant de leur vécu des sessions de simulation, chose qu'ils n'auraient sans doute pas pu faire avec des questionnaires. L'espace de parole créé a donc pour objectif de recueillir des données objectives (leur expérience personnelle par rapport à la simulation, à l'évaluation), mais également subjectives (leur jugement de telle situation). Par l'utilisation de l'entretien nous avons voulu avoir la capacité d'obtenir de « nombreuses anecdotes » et à partir d'elles, objectiver des données (Beaud, 1996, p 241). Beaud rappelle en effet l'importance des anecdotes pour « placer l'entretien du côté des pratiques sociales » et pour, à travers les détails importants des récits de situations, dégager des situations sociales. Comme le suggère Giraud dans son texte intitulé « Les mots pour faire dire et écrire », le concret permet de dégager des idées théoriques et évite de poser des questions de ce genre « que pensez-vous de ? » qui peuvent être bloquantes pour l'enquêté (Giraud, 2010, p 54-55). Utiliser le langage oral est un atout qui va également dans ce sens, de même que l'analyse du langage non verbal qui est possible lors d'un entretien. Il faut néanmoins avoir à l'esprit que le discours formulé par l'enseignant n'est pas la réalité, mais une réalité construite, en fonction de ses représentations, du contexte de l'entretien et de la place qu'il

donne à l'enquêteur. Il devra donc ensuite être traduit en étant le plus « vrai » possible et enfin interprété en fonction de ce contexte (Beaud, 1996, p 236-237; Olivier de Sardan, 1989, p 132-33).

6.1.2 L'entretien collectif

Nous avons ensuite retenu l'entretien collectif pour recueillir nos données. L'un des arguments qui justifie le choix de l'entretien collectif est la possibilité qu'offre cette méthode d'obtenir plusieurs expériences des enseignants, pour, comme le dit Giraud, avoir une « représentation multiple de cas » (Giraud, 2010, p 41). Les anecdotes ayant toutes leur importance dans l'entretien, permettre leur multiplication nous est apparu intéressant. Par l'utilisation du focus group nous multiplions également le nombre d'enquêtés mais cela n'est pas le point majeur, puisque comme le souligne Beaud, la représentativité ne se fait pas grâce au nombre (contrairement aux études quantitatives), mais par la qualité des entretiens obtenus (Beaud, 1996, pp 233-235). Ainsi les différentes expériences permettront de dégager plusieurs significations à la perception des enseignants quant à l'utilisation des scores et peut-être plusieurs angles d'éclairages d'une même situation.

L'entretien collectif peut également permettre d'engager la discussion entre les enseignants pour « faciliter le recueil de la parole individuelle » qui est au centre de nos préoccupations puisque l'objectif n'est pas d'avoir un récit de groupe mais bien plusieurs récits au sein d'un groupe (Duchesne, 2004, p 11). L'intérêt de cette méthode est donc de susciter la discussion entre les enseignants, avec un effet de groupe qui limite les inhibitions individuelles.

Un troisième intérêt de ce type d'entretien, et non des moindres, est la mise en relief de « significations partagées », ainsi que les appelle Duchesne dans son ouvrage sur les entretiens collectifs (Duchesne & Haegel, 2015, p 35). Puisque le groupe des enseignants partage la même profession, les mêmes mises en situation et une formation à l'enseignement par simulation

commune, certaines opinions seront partagées par l'ensemble du groupe, d'autres discutées selon leurs expériences et cela permettra, au travers des discours personnels et des points qui font consensus ou au contraire sont conflictuels, de dégager des axes de réflexion et d'analyse. Même si les entretiens collectifs ont tendance à mettre à jour les normes du groupe, certains chercheurs considèrent qu'ils peuvent également dégager des désaccords et enrichir les discussions (Duchesne & Haegel, 2015, p 36-37). Ainsi l'entretien collectif peut mettre à jour des interactions spécifiques au sein d'un groupe, ce qui peut éclairer l'objet de recherche (Duchesne & Haegel, 2015), p 40). Dans notre situation, les interactions des enseignants et leur rapport à l'évaluation par simulation pourraient nous aider à répondre à notre question de départ.

Mais il existe des limites à son utilisation et il faut les connaître avant de l'utiliser. Ce sont des limites inhérentes à l'interaction de groupe : un participant qui domine et retient la parole alors que d'autres participent moins, et des réticences de la part des participants à exprimer leurs idées, de peur d'être jugés ou d'avoir à subir une confrontation de la part d'un autre membre du groupe. L'animateur doit donc pouvoir faire attention à distribuer la parole s'il le faut et à motiver la parole personnelle des participants.

6.1.3 Le questionnaire

Le choix de faire un questionnaire a été motivé par le fait qu'il nous permettait de recueillir des avis plus nombreux, bien que moins détaillés. En étant conscient que le questionnaire ne participe pas à la « description exhaustive d'une pratique » (De Singly, 2106, p 20), nous voulions savoir comment était perçue l'utilisation des scores par les enseignants.

A la lecture du manuel de François de Singly sur le questionnaire, nous avons pu utiliser quelques-unes de ses recommandations. Pour des approches de pratiques, réaliser des questions avec une notion de temporalité, au passé composé, et pas au présent, car l'emploi de ce dernier engage plus souvent l'opinion que la pratique (De Singly, 2106, p 61). Grâce à notre lecture, nous avions en tête les biais d'un questionnaire qui, même s'il est anonyme, reste une

déclaration de pratiques et d'opinions, que les interrogés peuvent modifier, ou sous-estimer pour « sauver la face » (De Singly, 2106, p 72-73).

6.2 Méthodologie de l'étude

6.2.1 Design de l'étude

Il s'agissait d'une étude prospective, multicentrique, s'appuyant sur des situations simulées in situ, réalisées au sein de sept services de médecine d'urgence intra et extrahospitaliers. Les séances de simulation se sont déroulées du 15 novembre 2020 au 15 avril 2021. Au total, 7 services de médecine d'urgence ont accepté de participer à l'étude, dont cinq à Paris, un à Rouen et un à Strasbourg (réparti sur trois sites différents). Pour chacune de ces séances, deux évaluateurs devaient utiliser les scores en aveugle et noter la performance réalisée par l'interne qui gérait la situation, accompagné d'une équipe complète des urgences : infirmiers et aide-soignant pour les situations d'urgence intra-hospitalières et infirmier et ambulancier pour les situations d'urgence extrahospitalières. L'objectif était de réaliser 30 simulations pour chacun des trois scores, soit 90 simulations in situ au total. Pour chaque simulation, les équipes seraient composées de 3 à 6 participants (interne, infirmiers, aides-soignants ou ambulanciers).

La recherche effectuée est une recherche de type MR-004, ce qui implique qu'elle n'avait pas besoin de soumission à un comité de protection des personnes (CPP). Cependant, le protocole de l'étude a été soumis au comité d'éthique de l'université de Strasbourg et a reçu le numéro d'accréditation suivant Unistra/CER/2020-17. Les participants devaient signer un consentement de participation à l'étude (Annexe 3 et 4).

6.2.2 Objectifs de l'étude et critères de jugement

L'étude s'intéressait à la validation des scores d'évaluation dans différents contextes (avec des scénarios et des situations cliniques variées pour un même score d'évaluation), et avait pour

objectif d'évaluer en premier lieu leur reproductibilité inter-observateurs, lors d'une situation réelle d'évaluation et pas seulement sur des vidéos, comme cela était le cas de l'étude précédente.

L'objectif principal de l'étude était d'analyser la reproductibilité des scores d'évaluation des étudiants et internes en médecine, dans un contexte de travail en équipe. Le critère de jugement principal était la reproductibilité inter-observateur des scores ACAT 1, 2, 3.

Les objectifs secondaires étaient les suivants :

1. Analyser la reproductibilité du score de compétence globale
2. Étudier la nature de la relation entre le résultat numérique des scores ACAT (sur 20) et le score de compétence globale (variant de 1 à 5)
3. Étudier la nature de la relation entre le résultat des scores ACAT et le résultat obtenu au score TEAM. Le score TEAM est un score d'analyse d'habiletés non techniques pour une équipe prenant en charge un patient en situation d'urgence vitale. Parmi les différents scores d'évaluation du travail d'équipe validés, nous avons choisi ce dernier car il a été validé pour la médecine d'urgence et en français (Maignan et al., 2016)
4. Étudier la cohérence interne du score
5. Analyser la faisabilité d'utilisation du score en situation réelle et en fonction des différents scénario utilisés
6. Étudier la relation entre le résultat obtenu au score et le niveau de l'apprenant, déterminé par son niveau d'avancement dans l'internat et par sa spécialité (médecine d'urgence ou non) afin d'évaluer si les scores ACAT avaient une validité de construit de qualité
7. Fixer et analyser, grâce à la consultation d'expert, les différents seuils de réussite pour chaque score et pour chaque niveau d'apprenant (externe en fin de cursus, interne en première année de médecine d'urgence et interne en troisième année de médecine d'urgence, avant de début l'année de docteur junior)
8. Analyser les perceptions des enseignants sur les scores : faisabilité et facilité d'utilisation

des scores, intérêt de leur utilisation.

Les critères de jugement secondaires étaient les suivants :

1. Analyser la reproductibilité inter-observateur du score de compétence globale
2. Analyser la corrélation entre score ACAT et score de compétence globale
3. Analyser la corrélation entre les scores ACAT et le score TEAM
4. Analyser la cohérence inter-items grâce au calcul de l'alpha de Cronbach
5. Analyser la faisabilité de l'utilisation des scores « en situation réelle de formation » avec le pourcentage du nombre d'items non remplis
6. Analyser la différence des moyennes obtenues au score, en fonction du niveau de l'apprenant.
7. Fixer un seuil de réussite pour différents niveaux d'apprenants, grâce à la méthode d'Angoff
8. Analyse de l'acceptabilité des scores par les enseignants évaluateurs, au moyen de questionnaires et d'entretiens par focus groupes

6.2.3 Participants et description des sessions de simulation

Depuis quelques années, la simulation in situ se développe et permet de réaliser l'enseignement par simulation dans les unités de soins, avec les membres de l'équipe clinique dans leur propre environnement de travail. Aux urgences, la simulation in situ apporterait ainsi plusieurs atouts, en comparaison avec l'entrainement en dehors des services de soins. Grâce à la multiplication des entrainements pluriprofessionnels, elle permet d'améliorer les compétences des équipes des urgences, mais également d'identifier les dysfonctionnements présents dans les services et enfin grâce à ces deux atouts, de modifier le système à risque dans lequel évolue le patient (Petrosoniak et al., 2017). Des travaux réalisés aux urgences ont déjà pu mettre en évidence l'impact positif de la simulation in situ sur les pratiques des professionnels de santé et sur la prise

en charge des patients (Knobel et al., 2018; Lighthall et al., 2010; Wang et al., 2019). De même, la simulation in situ a déjà illustré sa capacité à identifier les conditions pouvant entraîner des erreurs dans l'environnement de travail, de les appréhender afin d'en diminuer l'incidence (Patterson et al., 2013; Wheeler et al., 2013). Il a ainsi été décidé d'utiliser cette méthode d'enseignement avec ses conditions spécifiques de mise en place pour poursuivre le processus de validation des scores et notamment pour pouvoir le faire en situation proche des conditions de travail tant par l'environnement matériel (les urgences ou l'extrahospitalier) que par l'environnement humain (le travail en équipe).

Les sessions de simulation se déroulaient selon les recommandations de l'HAS (Haute Autorité de Santé, 2012), à savoir :

- préparation en amont des scénarios, par les auteurs de l'étude (Tableau 5) ;
- préparation de l'environnement de simulation 30 à 45 minutes avant le début des sessions : par les formateurs. Un des rôles de ces formateurs sera d'identifier les conditions ne permettant pas de réaliser la session : apprenant absent, refusant la formation, service des urgences trop surchargé (définition à réaliser dans chacun des services participant à l'étude) ;
- environnements possibles pour réaliser la formation : unité mobile d'hospitalisation (UMH), environnement extrahospitalier (pour les équipes de SMUR), salle d'accueil des urgences vitales (SAUV), box d'examen aux urgences ou encore chambre d'hospitalisation de l'Unité d'Hospitalisation de Courte Durée (UHCD) ;
- briefing des apprenants : selon une check-list adaptée à l'environnement et à la simulation in situ, fourni aux formateurs avec le matériel pour l'étude ;
- scénario simulé : durée de 15 à 20 minutes. Le scénario peut utiliser un patient simulé, formé au préalable ou un mannequin basse fidélité qui permet de réaliser un massage cardiaque externe. Le patient simulé peut-être perfusé grâce à un système spécialement

mis en place pour l'étude et s'il doit être intubé, une tête d'intubation est fournie aux apprenants et mise à côté du patient simulé ;

- pendant le scénario : utilisation des scores ACAT 1, 2 ou 3 en fonction du scénario et du score TEAM par deux évaluateurs qui observent la simulation, avec pour impératif qu'ils ne puissent pas communiquer (l'un est près du patient, l'autre en retrait, voir caché pour ne pas gêner la simulation) ;
- débriefing : durée de 10 à 15 minutes, mené par les formateurs qui pouvaient être (ou pas) les évaluateurs, selon les recommandations de bonne pratique (Kolbe & Grande, 2015; Rall et al., 2000).

Les séances de simulation étaient annoncées à l'avance aux apprenants. Selon les services, elles se déroulaient soit pendant les heures de travail des apprenants, si les conditions du service le permettaient ou bien juste ou juste après la prise de poste des soignants. Chaque service avait la possibilité de décider de son organisation. Afin de pouvoir décider si la session de simulation pourrait avoir lieu, des critères de go/ no go avaient été déterminés et pouvait être appliqués par les services, en fonction de leurs contraintes locales. Les critères étaient les suivants : disponibilité de l'environnement et du matériel (box des urgences, chambre d'hospitalisation dans l'unité d'hospitalisation de urgences, salle d'accueil des urgences vitales ou UMH), sous-effectif du personnel médical, infirmier ou aide-soignant, flux patient ne permettant pas de réaliser la simulation, arrivée d'un patient critique motivant l'arrêt de la séance. Les investigateurs de l'étude devaient relever, pour chaque centre, le nombre de simulations non réalisées ou stoppées, ainsi que le critère ayant mené à la décision.

Les participants étaient recrutés au sein des services d'urgence ou de SMUR, et il leur était proposé de participer volontairement à des sessions de formation in situ, proposées par les formateurs des différents services et en lien avec l'encadrement des soignants. Tous les services ont réalisé des scénarios de médecine d'urgence intra-hospitalière et deux d'entre eux ont

également mis en place des sessions de simulation au sein de leur Services Mobiles d'Urgence et de Réanimation permettant de réaliser des scénarios de médecine d'urgence extrahospitalière. Avec un objectif de 90 sessions de simulation au total, chacun des six services devait réaliser entre 12 et 13 sessions de simulation. Chaque session de simulation était composée d'un scénario qui correspondait à l'un des scénarios écrits pour l'étude et dont les environnements, présentation cliniques et diagnostics étiologiques variaient (Tableau 5). Chaque scénario a également été écrit pour une situation d'urgence intra hospitalière et pour une situation extrahospitalière.

Tableau 5 – Différents scénarios utilisés pour chaque situation clinique évaluée

Situation clinique	Arrêt cardiaque	Coma non traumatique	Détresse respiratoire aigu
Scénario 1	Intoxication aux tricycliques	Etat de mal épileptique	Covid
Scénario 2	Syndrome coronarien aigu	Hypertension intracrânienne	Œdème aigu du poumon
Scénario 3	Embolie pulmonaire et RSP	Intoxication éthylène glycol	Décompensation de BPCO
Scénario 4	BAV et hyperkaliémie	Intoxication opiacés	Pneumothorax compressif
Scénario 5	Choc anaphylactique	Intoxication benzodiazépines	Asthme aigu grave

RSP : rythme sans pouls, BAV : bloc auriculo-ventriculaire, BPCO : bronchopneumopathie chronique obstructive.

Les critères d'inclusion des apprenants et des formateurs étaient les suivants :

Apprenants : professionnels des urgences : aide-soignant, infirmiers, internes, étudiants infirmiers ou étudiants en médecine. Postés lors de la simulation in situ, ou inclus lors de périodes programmées de formation par la simulation in situ. Ils devaient avoir au préalable signé le consentement éclairé de participation à l'étude et donné leur accord pour que nous utilisions les données issues de leur participation.

Formateurs : professionnels de médecine d'urgence, médecins infirmiers ou aides-soignants, formés à l'enseignement sur simulateur et qui ont reçu une formation à l'utilisation des scores. Les formateurs avaient eux aussi donné leur accord pour participer à l'étude, pour observer les simulations in situ et utiliser le score en situation.

La présente recherche a remporté un appel à projet financé par l'INSERM, via la Fédération Hospitalo-Universitaire (FHU) IMPEC (IMProving Emergency Care). Ainsi, nous avons disposé d'un budget de 15 000 euros pour acheter le matériel nécessaire aux séances de simulation : mannequins basse-fidélité, têtes d'intubations, défibrillateur semi-automatique d'entrainement, scopes d'entrainement. Ainsi, tous les services ont pu compléter leur équipement et avoir le même matériel à disposition.

6.2.4 Conduite des entretiens et description du questionnaire

Le cadre retenu pour l'enquête qualitative était celui d'une recherche inductive, décrite ci-dessus. L'entretien choisi était semi-directif, au moyen de questions préparées dans un guide d'entretien que la doctorante avait réalisé et préalablement testé avec une collègue de l'unité de simulation de Strasbourg (Annexe 5). Du fait des conditions épidémiques et des localisations géographiques plurielles des centres, il a été décidé de mener les entretiens par visioconférences. L'outil utilisé ne permettait pas d'enregistrer la vidéo, mais nous avons pu enregistrer les échanges entre les participants, via l'ordinateur et au moyen d'un dictaphone afin de ne pas perdre des données si un des outils dysfonctionnait.

Il a été proposé à tous les enseignants ayant utilisé le score de participer aux entretiens, via des emails envoyés aux investigateurs principaux de chaque centre. Des relances étaient prévues, afin de pouvoir organiser des focus groupes de 5 à 6 participants. Lors de l'entretien, les enseignants commençaient par se présenter, afin de savoir de quel centre chacun venait, puis la consigne suivante était énoncée : « Merci beaucoup de vous joindre à ce groupe d'entretien collectif. Après vous être présentés, vous échangerez sur la simulation in situ et sur l'utilisation des scores d'évaluation à partir de vos expériences respectives ». La doctorante a modéré les focus groupes en endossant le rôle de facilitateur, et en posant des questions aux enquêtés si les thématiques présentes dans le guide d'entretien n'avaient pas été abordées lors des discussions. Le guide d'entretien s'intéressait à la manière dont les enseignants avaient utilisé

les scores et à leur expérience quant à leur utilisabilité, puis à la perception qu'avaient les enseignants des performances évaluées et enfin à une utilité potentielle pour les enseignants, dans leur pratique future. Les questions étaient principalement centrées sur l'expérience vécue par les enseignants, autour de ces scores.

Le questionnaire destiné aux enseignants a été réalisé par la doctorante. Il faisait partie des documents remis aux investigateurs principaux de chaque centre, qui devaient le remettre aux enseignants soit en format papier soit via un questionnaire électronique Google Form®.
Le questionnaire était composé de 23 questions ouvertes et de questions fermées, réparties en 4 parties : expériences des formateurs avec la simulation, mise en place des sessions de la formation in situ, utilisabilité des scores et données démographiques (métier, centre). Les questions fermées suivaient le modèle de Likert avec des cotations allant de 1 à 4, ou une case intitulée « ne se prononce pas ». Ce modèle de notation a été choisi pour qu'il ne puisse pas y avoir de neutralité et que les enseignants puissent donner un avis soit en accord soit en désaccord avec la proposition (Annexe 6).

6.2.5 Analyse des données

Toutes les données ont été anonymisées dès le recueil avec un codage pour le centre investigateur, le numéro de la session de simulation, le numéro du scénario et un numéro pour chaque participant. Les données descriptives des participants, et des sessions de simulation sont exprimées en moyenne et écart type. Le recueil du nombre d'items remplis ainsi que celui du nombre d'item ayant pu être côtés sera exprimé en pourcentage. L'analyse du critère de jugement principal et le calcul de l'accord entre les observateurs a été effectuée avec le calcul du coefficient d'IntraClass Correlation (ICC) pour les scores ACAT et avec le Kappa de Cohen pondéré pour le score de performance globale. La corrélation entre les résultats obtenus avec le score ACAT et les scores de performance globale, le score TEAM et le niveau de l'apprenant est analysée avec

le coefficient de corrélation de Pearson. Enfin, le calcul du coefficient alpha de Cronbach pour chacun des scores ACAT permettait d'évaluer la cohérence du score.

L'évaluation de la perception des enseignants est réalisée au moyen de questionnaires et par une analyse qualitative des focus groupes. La retranscription des entretiens a été faite manuellement par l'investigateur principal, à l'aide d'un logiciel en ligne « otranscribe », gratuit et qui est uniquement un soutien, une aide, mais pas une retranscription automatique. Ils ont été anonymisés par l'enquêtrice. Le codage des entretiens a été réalisé séparément par la doctorante et par une étudiante du Master 2 de pédagogie médicale de Strasbourg, totalement étrangère à l'étude, permettant ainsi une double analyse des données recueillies. L'analyse des données était concomitante au recueil de celles-ci afin de pouvoir adapter les questions abordées dans les entretiens suivants. Il était prévu de réaliser des focus groupes jusqu'à l'absence d'obtention de nouvelles données pour ainsi obtenir un recueil « saturé » des perceptions des enseignants.

CHAPITRE 3 : RESULTATS

Nous présentons ici les résultats des différentes étapes du processus de validation des trois outils d'évaluation des apprenants par la simulation, en commençant par la description de la création de leur contenu, puis à l'analyse de leur reproductibilité et enfin à leur utilisation lors de sessions de simulation in situ.

1. DEVELOPPEMENT D'OUTILS D'EVALUATION DES ETUDIANTS EN MEDECINE D'URGENCE : UNE ETUDE NATIONALE PAR LA METHODE DELPHI (ARTICLE 1)

La consultation des experts en deux temps a permis de choisir trois situations cliniques pour lesquelles il apparaissait pertinent de développer des scores d'évaluation, puis de sélectionner leur contenu.

1.1 Choix des trois situations cliniques

Parmi les 29 coordonnateurs régionaux français sollicités, 21 (72%) ont participé et permis d'identifier les 10 situations cliniques de départ à évaluer, les trois premières étant : l'arrêt cardio-respiratoire, la détresse respiratoire aigüe et le coma non traumatique (Tableau 6)

Tableau 6 – Classification des 10 premières situations cliniques qui devraient être évaluées par la simulation.

Items du programme de formation	**Classement**
Arrêt cardiaque	1
Détresse respiratoire aiguë de l'adulte / Insuffisance respiratoire aigue	2
État de choc	3
Accidents vasculaires cérébraux	4
Syndromes coronaires aigus	5
Coma non traumatique de l'adulte	6
Œdème de Quincke et anaphylaxie	7
Polytraumatisé, traumatisé abdominal, traumatisé thoracique, traumatisé oculaire	8
État confusionnel et trouble de la conscience chez l'adulte	9
Hémorragies digestives	10

1.2 Sélection du contenu des scores

1.2.1 Participation et description des experts

Nous avons identifié 58 experts francophones qui répondaient aux critères d'inclusion et leur avons également demandé de nous proposer des experts qui seraient en mesure de participer à l'étude, selon les critères définis. Ils ont été sollicités par mail, et ont indiqué en retour leur volonté de participer ou non à l'étude. Les 58 experts contactés ont indiqué connaitre 28 experts qui répondaient aux critères d'inclusion. Ainsi sur les 86 experts sollicités, 51 (59%) ont accepté de participer au Delphi. Parmi eux, cinq exerçaient en Suisse ou en Belgique, et la majeure partie (49) travaillait dans un centre hospitalo-universitaire (Tableau 7).

Tableau 7 – Description des 51 experts qui ont participé au Delphi

	Nombre	(%)		Nombre	(%)
Total participants	**51**				
Fonction - Spécialité			**Lieu d'exercice**		
Universitaire MU	17	(33)	**FRANCE**	46	(90)
Universitaire AR	3	(6)	Auvergne-Rhône-Alpes	6	(23)
Médecin Urgentiste	26	(51)	Bourgogne-Franche-Comté	3	(6)
Médecin AR	3	(6)	Centre-Val de Loire	1	(2)
Autre	2	(4)	Grand Est	2	(4)
Responsable DES-MU/DESC	9	(18)	Hauts-de-France	2	(4)
Chef de service	8	(16)	Ile de France	23	(45)
			Nouvelle Aquitaine	4	(8)
Type d'hôpital			Occitanie	4	(8)
CHU	49	(96)	Pays de la Loire	1	(2)
CH	2	(4)	**SUISSE**	2	(4)
			BELGIQUE	3	(6)

MU : médecine d'urgence ; AR : Anesthésie-Réanimation ; DES-MU : diplôme d'études spécialisées en médecine d'urgence ; DESC : diplôme d'études spécialisées complémentaires
CHU : centre hospitalo-universitaire ; CH : centre hospitalier

1.2.2 Sélection du contenu des scores

Afin d'avoir 20 experts pour la constitution de chaque score, nous avons demandé à 5 experts de participer à l'établissement de plusieurs scores. Le taux de réponse a varié entre 80 et 95%, malgré les différentes relances entre chaque tour (Tableau 8).

Tableau 8 – Taux de réponse des experts pour chaque famille de situation et à chaque tour de Delphi

Famille de situation	**Tour 1** n (%)	**Tour 2** n (%)	**Tour 3** n (%)
ACR	19 (95)	18 (90)	18 (90)
DRA	17 (85)	17 (85)	16 (80)
Coma non traumatique	17 (85)	17 (85)	17 (85)

ACR : arrêt cardio-respiratoire ; DRA : détresse respiratoire aigüe.

Pour chaque situation clinique, à partir d'un nombre d'items variable, les différentes étapes de consultation des experts ont permis de retenir après le troisième tour de consultation, respectivement 24,26 ou 30 items pour les situations d'ACR, de DRA et de coma. Au final, nous avons pu regrouper ces items en 20 catégories pour chaque score, afin d'établir un score sur 20 (Figure 12). Cela a pu être possible grâce à l'analyse des remarques des experts et à la nature de la notation de chaque item qui, n'étant pas binaire, permet de regrouper plusieurs items dans une catégorie (par exemple, des items d'examen clinique ont pu être ainsi regroupés). Une première version de chaque score a ainsi pu être obtenue après ce travail.

1.2.3 Analyse des remarques et des discussions des experts

L'analyse textuelle des commentaires des experts a permis d'affiner chaque score avant son utilisation et avant une deuxième analyse qui fera partie de l'étude du processus de réponse dans le cadre du processus de validation des scores. De plus, les experts ont mis en évidence des questions relatives à la mise en pratique de l'évaluation et qui seront pertinentes pour l'utilisation

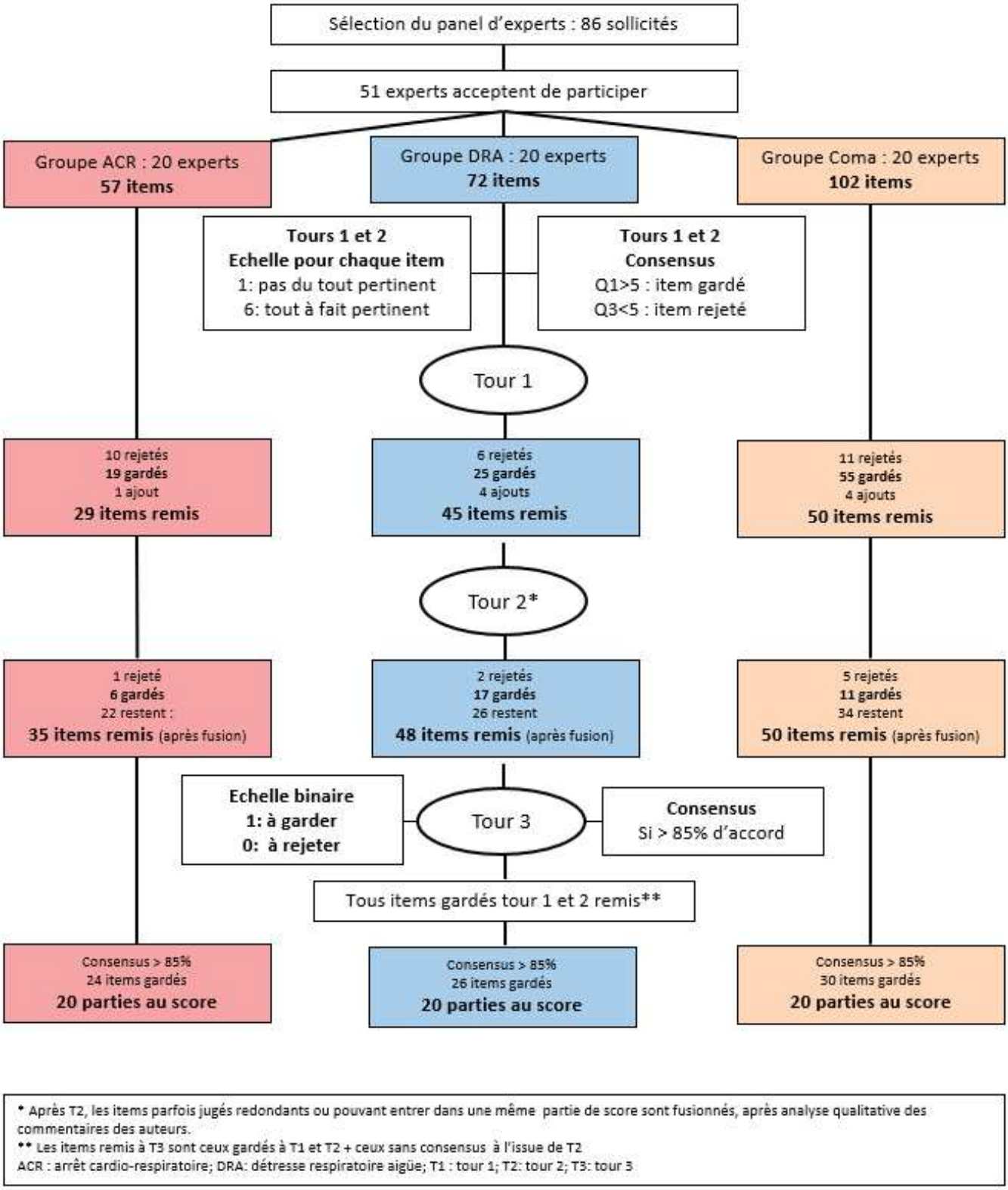
ultérieure des scores. Nous rapportons ici les principaux débats retirés de l'analyse, réalisée grâce à un codage manuel de la doctorante.

La principale difficulté exprimée par les experts était l'absence de contextualisation, et ce pour deux raisons. La première est que le public évalué sera de niveau variable : des étudiants de deuxième cycle et des internes de phase d'approfondissement qui ne doivent pas avoir les mêmes compétences ainsi que le résume cet expert : *« comment fait-on la différence entre externe, DESMU1/2/3 (E7) ? »*. L'objectif du score étant de pouvoir les discriminer, chaque item a été discuté et remis ou retiré dans ce sens. Les items qui ont été le plus discutés étaient ceux en rapport avec l'intubation orotrachéale (IOT) et les compétences requises variables pour chaque catégorie d'étudiant. La deuxième difficulté de contextualisation résidait dans le concept même de famille de situation et dans le fait que les situations sont assez variées, d'étiologie variable ce qui a perturbé les experts, habitués à raisonner à partir d'un cas clinique : contexte précis, problématique précise, étiologie précise.

Le deuxième point qui a fait débat parmi les experts est en lien avec les limites et pratiques de la simulation et notamment avec son niveau de réalisme, plus particulièrement des interrogations sur la fidélité du mannequin. Les experts ont exprimé des doutes quant à certains éléments demandés dans les scores, principalement ceux relevant de l'examen clinique, tels que le calcul du score de Glasgow, la recherche de lésions cutanées (marbrures, purpura) ou encore la recherche de signe de gravité respiratoires : *« la mesure de la fréquence respiratoire n'est pas intuitive avec le mannequin(E2), comment peut-on calculer le score de Glasgow avec un mannequin ? (E3)»*. L'absence de fidélité totale était pour quelques experts une impossibilité de se servir de cet outil pour évaluer les compétences cliniques des étudiants. Après débat, il s'est avéré que la majeure partie des experts a considéré qu'il s'agissait d'un problème inhérent à la simulation, que les étudiants le rencontreraient pendant les phases d'apprentissage et qu'une pratique adaptée de la simulation (constituée d'entrainements réguliers mais également de

briefings rappelant ces facteurs limitants) viendrait limiter les difficultés : *« on apprend pas à soigner des simulateurs mais des patients, il faut donc un facilitateur ou un briefing ad hoc (E1) ; ces éléments sont indispensables pour le raisonnement clinique et le facilitateur pourra aider à y répondre (E12) »* . Les experts ont donc insisté sur l'absolue nécessité de préciser ces écueils lors du briefing, de recourir à un facilitateur qui répond aux questions des étudiants à ce propos, mais également de ne pas couper la phase de débriefing qui pourrait permettre de comprendre certaines réflexions que les étudiants n'auraient pas eu à voix haute.

Enfin, les experts ont discuté des méthodes d'évaluation qui pourraient être employées. Choisir les items qui feront partie d'un score a permis aux experts de se poser des questions sur les conditions de l'évaluation, qui n'étaient pas détaillées, volontairement, dans la présentation de l'étude. Cela leur a permis de faire des propositions sur le déroulement possible de l'évaluation et sur la composition du score (binaire ou pas, avec des items pondérés ou pas) : *« Afin de complexifier un peu la problématique : le fait de savoir si tous les critères sont remplis (réponse binaire) n'est pas la même chose que comparer des scores (réponses quantitatives). Dans le deuxième cas, difficile d'envisager un score ou chaque item apporte le même nombre de points. Un étudiant autiste qui limite au maximum les interruptions de massage aura mieux acquis la gestion de l'ACR qu'un bon communicateur (équipe, famille etc.…) qui tarde à masser. Attribuer plus de points aux actions qui ont un impact pronostic validé serait intéressant (il y en a peu) (E25) » ; « Difficile de grader les items de façon binaire oui/non (E43) ».*



* Après T2, les items parfois jugés redondants ou pouvant entrer dans une même partie de score sont fusionnés, après analyse qualitative des commentaires des auteurs.
** Les items remis à T3 sont ceux gardés à T1 et T2 + ceux sans consensus à l'issue de T2
ACR : arrêt cardio-respiratoire; DRA: détresse respiratoire aigüe; T1 : tour 1; T2: tour 2; T3: tour 3

Figure 12 – Résultats du Delphi pour chaque situation clinique

2. Etude du processus de reponse et de la reproductibilite des scores (Article 2)

A partir des scores obtenus lors de la consultation d'experts, il a été possible de réaliser une analyse du processus de réponse, selon les étapes préconisées par Downing (Downing, 2003), puis d'étudier leur reproductibilité intra et inter-observateur. Nous décrivons ici l'aboutissement de la récupération de séances de simulation filmées, et les résultats de leurs analyses.

2.1. Description des séances de simulation

Soixante vidéos ont pu être collectées pendant la période d'inclusion, décalée jusqu'en octobre 2020 suite à l'arrêt des sessions de simulation à l'université de mars à juin 2020 pour les internes et de mars à novembre 2020 pour les étudiants de 4ème année de médecine. Il a été ainsi décidé d'enregistrer plus de sessions dédiées aux internes de médecine d'urgence, quitte à ne pas pouvoir comparer les deux populations.

Parmi les soixante vidéo, dix ont dû être exclues car le fichier contenait des défauts les rendant inutilisables et sept ont dû être exclues car l'enregistrement avait été démarré trop tard et les premières minutes du scénario manquaient. Avec une moyenne de 3 participants par situation simulée, 54 participants (36 internes, 18 étudiants en médecine) ont participé à une situation simulée d'arrêt cardiaque (score 1), 41 participants (21 internes, 14 étudiants en médecine) ont prise en charge une situation simulée de détresse respiratoire aigüe et 49 participants (39 internes, 10 étudiants en médecine), ont participé à une prise en charge d'une situation simulée de coma. Le temps médian de chaque scénario était respectivement de 8 minutes [95% IC 7,2; 8], 12,2 minutes [95% IC 11,3; 12,7] and 12,3 min [95% IC 11,6; 13] pour les situations d'arrêt cardiaque, de détresse respiratoire et de coma (Tableau 9).

Tableau 9 – Caractéristiques des participants et des différents scenarios de simulation

Score	ACAT 1		ACAT 2			ACAT 3		
	N vidéo	%		N vidéo	%		N vidéo	%
Scenario			**Scenario**			**Scenario**		
Total	18	(100)	Total	15	(100)	Total	15	(100)
Asystolie	8	(45)	Pneumonie	10	(67)	EME	11	(73)
SCA, FV	8	(45)	AAG	3	(20)	Intoxication	4	(27)
CA	2	(10)	OAP	2	(13)			
Median time (min) [95% CI]	8 [7,2-9,5]			12,2 [11,4-12,7]			12,3 [11,5-13]	
Participants								
Total	54	(100)		41	(100)	49	(100)	
DESMU1	36	(66)		27	(66)	39	(80)	
EM	18	(34)		14	(34)	10	(20)	

ACAT : Acute Care Assessment Tool ; SCA : Syndrome coronarien Aigu ; FV : fibrillation ventriculaire ; DESMU1: interne de 1ère année de médecine d'urgence; EM: étudiant en médecine; AAG : asthme aigu grave ; OAP : œdème aigu du poumon ; EME : état de mal épileptique

2.2 Etude du processus de réponse

Chaque item de chacun des scores ACAT a été analysé par deux urgentistes, également expérimentés en enseignement par simulation, grâce à l'analyse de trois vidéos pour chaque score ACAT. Le but de l'étude du processus de réponse était de rechercher les potentielles sources d'erreur telles que des items imprécis, mal décrits ou des marqueurs de temps non adaptés à la réalisation des actions requises. Les items n'ont pas été changés car ils avaient été sélectionnés par lors du processus Delphi, mais ils ont été clarifiés. Par exemple, pour le score ACAT1, évaluant les situations d'arrêt cardiaque : 3 items ont été précisés et 3 autres ont vu leur processus de notation modifié. Ainsi l'item « appeler à l'aide » est devenu « reconnaissance de l'arrêt et appeler à l'aide ou dire « c'est un arrêt cardiaque ». Pour les deux autres scores ACAT, 5 et 10 ont été précisés, aucun n'a vu son processus de notification modifié.

Après cette première étape, il a été décidé d'adopter une échelle de notation à trois niveaux : 0 – ½ - 1, correspondant aux descriptions suivantes : non fait, incomplet, fait complètement. Une autre option est de préciser que l'item n'est pas adapté au scénario évalué. Ces trois niveaux ont été choisis car ils permettent de pondérer le score tout en ayant une échelle de notation simple

d'utilisation, adaptée à des évaluations qui doivent se dérouler pendant l'observation de la performance, qui est de courte durée.

2.3 Reproductibilité et utilisabilité

Trois évaluateurs indépendants ont donc relu les vidéos à deux reprises. Un évaluateur a relu les 48 vidéos, ce qui représente une durée de lecture d'environ 17 heures et les deux autres évaluateurs ont relu respectivement 27 et 21 vidéos (soit environ 10h et 7h chacun).

La reproductibilité inter et intra-observateur varie entre 0.88 et 0.97 pour chaque paramètre analysé et est rapportée dans le tableau 10.

Tableau 10 – Reproductibilité inter et intra-observateur pour chaque score ACAT, mesuré par deux évaluateurs indépendants

	ACAT 1		**ACAT 2**		**ACAT 3**	
	CCI	95% IC	CCI	95% IC	CCI	95% IC
Reproductibilité Intra-Observateur						
O1: M0 – M2	0.88	[0.69 – 0.95]	0.89	[0.70-0.96]	0.98	[0.94 – 0.99]
O2: M0 – M2	0.97	[0.93 – 0.99]	0.87	[0.66 – 0.95]	0.95	[0.86 – 0.99]
Reproductibilité Inter-Observateur						
M0: O1 – O2	0.95	[0.70 – 0.95]	0.96	[0.89 – 0.99]	0.91	[0.76 – 0.97]
M2: O1 – O2	0.9	[0.72 – 0.96]	0.9	[0.7 – 0.97]	0.97	[0.91 – 0.99]

ACAT: Acute Care Assessment Tool; O1: Observateur 1 ; Observateur 2 : O2 ; M0: observation initiale, M2: observation à 2 mois (au moins), CCI: Coefficient de Corrélation Intra-classe; IC: intervalle de confiance

Afin d'évaluer l'utilisabilité du score, l'analyse s'est également intéressée au pourcentage de remplissage des items et à leur pertinence. Le pourcentage de remplissage des items variait entre 96 et 100% pour tous les items avec, parmi les soixante items (20 par score), cinquante qui étaient remplis à 100%. En ce qui concerne la pertinence des items, un total de cinq items n'étaient pas toujours remplis car pas adaptés à la situation. Pour quatre d'entre eux, ils étaient remplis dans 73% des cas, alors que pour un item, qui concerne la gestion de la famille lors d'une prise en charge d'un arrêt cardiaque, seuls 27,7% des scénarios étaient concernés par le contenu

de l'item. Le tableau 11 relève les items incomplets (*remplissage)* pour chaque score et ceux qui n'ont pas été utilisés dans toutes les situations (*pertinence)*.

Tableau 11 – Pourcentage de remplissage des items et pertinence, tous scenarios confondus.

Score	Item	% remplissage	pertinence
ACAT 1 (AC)	Ensemble des items (20)	99,7%	-
	Utilisation de l'adrénaline adaptée aux recommandations	97,2%	100%
	Intubation: anticiper et préparer le matériel adapté	98,6%	100%
	Initier et organiser le traitement étiologique	98,6%	100%
	Gérer de la famille/des témoins	**100%**	**27,7%**
ACAT 2(Coma)	Ensemble des items (20)	99,6%	-
	Recueillir les éléments pertinents de l'histoire de la maladie	98,3%	100%
	Recueillir les traitements	98,3%	100%
	Appeler un correspondant adapté pour de l'aide/demander une orientation	96,6%	100%
	Protéger les voies aériennes supérieures : choix de la technique, argumentation, explications à l'équipe	**100%**	**73,3%**
ACAT 3 (DRA)	Ensemble des items (20)	99,6%	-
	Installer le patient en position semi-assise	98,3%	100%
	Rechercher les signes de défaillance respiratoire	98,3%	100%
	Rechercher les signes de défaillance hémodynamique	98,3%	100%
	Appeler un correspondant adapté pour de l'aide/demander une orientation	98,3%	100%
	Argumenter et expliquer le choix de la ventilation au sein de l'équipe	**100%**	**73,3%**
	Intubation: anticiper et préparer le matériel adapté	**100%**	**73,3%**
	Préparer le ventilateur : anticiper et régler	**100%**	**73,3%**

ACAT: Acute Care Assessment Tool; AC: arrêt cardiaque; DRA: détresse respiratoire aigüe
En gras : les 5 items moins pertinents indifféremment des différents scénario

3. Etude de la validite externe du score : analyse de sa reproductibilite (Article 3)

L'étude rapportée ici avait pour objectif d'étudier la validation externe des scores ACAT, en les utilisant dans des conditions réelles de simulation et dans sept services d'urgences intra et extrahospitaliers différents. L'objectif principal était également d'étudier la reproductibilité inter-observateur en situation réelle, l'utilisabilité des scores, mais également de comparer les résultats obtenus avec d'autres marqueurs de performance, tels que le score de performance globale, un autre score d'évaluation (le score TEAM) qui s'intéresse à l'évaluation du travail en équipe, et au niveau de l'apprenant. De plus, nous nous sommes également intéressés à la fixation d'un seuil de réussite et au point de vue des formateurs à propos des scores ACAT, mais également à propos de l'évaluation par la simulation.

3.1 Démographie : apprenants, formateurs, scénarios

Dans les sept centres, 314 apprenants ont participé à 104 sessions de simulation in situ, dont 13 situations en extrahospitalier et 91 en intrahospitalier. Parmi les apprenants, 85 internes, 161 infirmiers, 60 aides-soignants et 8 ambulanciers ont participé (Tableau 12). Pour chaque scénario, l'équipe était donc constituée au minimum d'un interne, d'un ou deux infirmiers. Il n'y avait pas d'aide-soignant pour chaque session de simulation en intrahospitalier, mais en revanche chaque situation en extrahospitalier a pu être jouée, par des équipes complètes (un interne, un infirmier, un ambulancier).

37 formateurs ont mis en place les sessions de simulation in-situ, dont 26 médecins, 10 IDE et une aide-soignante. Comme cela était prévu par le protocole, chaque service était libre d'adapter les sessions in situ selon son organisation : soit pendant les horaires de travail des apprenants, soit dans des moments précédant ou suivants leur temps de travail, soit pendant des temps de formation identifiés en dehors des heures de travail. Deux centres ont réalisé les séances de simulation pendant les heures de travail, trois autres centres ont choisi de les mettre en place

juste avant les heures de travail des équipes infirmières et aides-soignantes mais avec des internes en poste, deux centres avaient libéré des temps identifiés pour la formation et enfin un centre a opté pour les deux solutions en fonction du contexte choisi pour la simulation (horaires programmées en dehors des heures de travail pour l'extrahospitalier et horaires consécutives ou précédents la simulation pour l'intrahospitalier).

Tableau 12 – Participants de l'étude e-Simsit

	SERVICES							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
PARTICIPANTS								
INTERNES								
MEDECINE D'URGENCE								**67**
MU1	3	9	6	4	3	1	4	30
MU3	7	1	1	1	-	14	13	37
SPECIALITE AUTRE								**18**
1ère année	-	-	1	4	4	-	3	12
2ème année	2	-	-	-	-	-	-	2
3ème année	4	-	-	-	-	-	-	4
TOTAL INTERNES	***16***	***10***	***8***	***9***	***7***	***15***	***20***	***85***
IDE	***22***	***30***	***12***	***21***	***22***	***33***	***21***	***161***
AS	***16***	***4***	***6***	***5***	***9***	***14***	***6***	***60***
AMBULANCIERS							***8***	***8***
TOTAL	**54**	**44**	**26**	**35**	**38**	**62**	**55**	**314**
FORMATEURS								
MEDECINS	1	2	1	5	5	5	7	26
IDE	1	-	3	-	4	2	-	10
AS	-	-	-	-	1	-	-	1
TOTAL	**2**	**2**	**4**	**5**	**10**	**7**	**7**	**37**

MU : médecine d'urgence ; IDE : infirmier diplômé d'état ; AS : aide-soignant

Le nombre total de scénario était de 104, soit 43 scénarios d'arrêt cardiaque, 30 scénarios de coma et 31 scénarios de détresse respiratoire aigüe, nous permettant ainsi d'atteindre le minimum de 30 scénarios par situation clinique (Tableau 13). Pour chaque situation, chacun des

cas cliniques a pu être mis en place au moins 5 fois, sauf pour le scénario 4 du score de détresse respiratoire aigüe, et qui constituait en la prise en charge d'un pneumothorax compressif.

Tableau 13 – Répartition des scénarios et annulation des sessions

	SERVICES							
	1	2	3	4	5	6	7	TOTAL
SCENARIO								**104**
ARRET CARDIAQUE	**7**	**6**	**3**	**5**	**8**	**5**	**9**	**43**
1. Intoxication tricycliques	1	1	-	1	1	1	1	6
2. Syndrome coronarien aigu	1	2	1	1	1	1	2	9
3. Embolie pulmonaire (RSP)	2	1	1	1	3	1	4	13
4. Hyperkaliémie (BAV3)	2	1	-	1	3	1	-	8
5. Choc anaphylactique	1	1	1	1	-	1	2	7
COMA	**3**	**3**	**3**	**5**	**3**	**5**	**8**	**30**
1. Etat de mal épileptique	1	1	1	1	-	1	2	6
2. Hypertension intracrânienne	-	1	2	1	-	1	1	6
3. Intoxication éthylène glycol	1	-	-	1	1	1	1	5
4. Intoxication opiacés	1	-	-	1	1	1	2	6
5. Intoxication benzodiazépines	-	1	-	1	1	1	2	5
DETRESSE RESPIRATOIRE AIGUE	**5**	**6**	**2**	**5**	**2**	**5**	**6**	**31**
1. Covid	1	2	-	1	-	1	2	6
2. Œdème aigu du poumon	1	2	1	1	-	1	1	5
3. décompensation BPCO	1	1	1	1	1	1	-	6
4. Pneumothorax sous VNI	1	-	-	1	-	1	1	4
5. Asthme aigu grave	1	1	-	1	1	1	2	7
TOTAL	**15**	**15**	**8**	**15**	**13**	**15**	**23**	**104**
SESSIONS ANNULEES								
Sous-effectif personnel	-	-	1	1	-	-	2	4
Flux de patients	-	-	1	-	-	-	-	1
Locaux non disponibles	-	-	-	1	-	-	-	1
TOTAL	**0**	**0**	**2**	**2**	**0**	**0**	**2**	**6**

RSP : rythme sans pouls ; BAV3 : bloc auriculo-ventriculaire de 3ème degré ; BPCO : broncho-pneumopathie chronique obstructive ; VNI : ventilation non invasive

3.2 Fiabilité de scores : consistance interne et reproductibilité inter-observateurs

A nouveau, comme lors de l'étude sur vidéo, l'analyse de la mesure des scores par deux opérateurs indépendants retrouvait une bonne reproductibilité inter-observateurs à la fois pour les trois scores ACAT, mais également pour les scores de performance globale (SPG), dans

chacune des trois situations étudiées. Pour les scores ACAT, l'ICC varie entre 0.89 et 0.95, ce qui traduit une excellente reproductibilité inter-observateur. Quant à la reproductibilité du score de performance globale, elle est également bonne, avec un coefficient kappa de Cohen pondéré qui varie entre 0.76 et 0.84 (Tableau 14).

Tableau 14 – Reproductibilité des scores ACAT et des scores de performance globale

Inter-rater reliability	**ACAT 1**		**ACAT 2**		**ACAT 3**	
	CCI	95% IC	CCI	95% IC	CCI	95% IC
O1 – O2	0.95	[0.93 – 0.98]	0.89	[0.77 – 0.95]	0.92	[0.83 – 0.96]
	SPG ACAT 1		**SPG ACAT 2**		**SPG ACAT 3**	
	Kappa[a]	95% IC	Kappa[a]	95% IC	Kappa[a]	95% IC
O1 – O2	0.80	[0.67 – 0.93]	0.76	[0.60 – 0.92]	0.84	[0.70 – 0.98]

ACAT : Acute Care Assessment Tool ; O1 : Observateur 1 ; Observateur 2 : O2 ; CCI : Coefficient de Corrélation Intra-classe ; IC : intervalle de confiance ; SPG : score de performance globale
[a] Kappa : Kappa pondéré de Cohen

La cohérence interne des scores, analysée avec le coefficient alpha de Cronbach, par la corrélation entre les résultats aux différents items variait entre 0,73 et 0,8, ce qui en fait une cohérence acceptable (Tableau 15).

Tableau 15 – Cohérence interne des scores : corrélation entre les items

	ACAT 1	**ACAT 2**	**ACAT 3**
Alpha de Cronbach	0.79	0.8	0.73
Alpha de Cronbach standardisé	0.80	0 .82	0 .72

ACAT : Acute Care Assessment Tool

De plus, comme pour la deuxième étude, nous avons étudié les taux de remplissage des items, ainsi que la pertinence des items. Les résultats sont présentés dans le tableau 16.

Tableau 16 – Pourcentage de remplissage et pertinence des items. Etude e-simsit

Score	Item	Completeness	Relevance
ACAT 1 (CA)	Overall item (20)	**99,2%**	-
	Time of the cardiac arrest	97.7%	100%
	Give adrenaline/amiodarone according to the guidelines	96,5%	100%
	Assess rhythm	98.9%	100%
	CPR 30/2, monitor for fatigue	97.7%	100%
	Tracheal intubation: anticipate and prepare the material	96,5%	**97,9%**
	Rapidly outlines a plan/strategy and ask for equipment	96,5%	100%
	To manage the family/bystanders	98.9%	**58,4%**
ACAT 2 (Coma)	Overall item (20)	**99.1%**	-
	Realize and verbalize Glasgow score	96,6%	100%
	Review patient's past history	96,6%	100%
	To look for scalp bruising or hematoma	98,3%	100%
	Communicate clearly to team the patient's history and clinical examination's details	96,6%	100%
	Organize the etiological treatment	98,3%	100%
	Anticipate the patient's outcome: outlines a plan/strategy and ask for equipment	98,3%	100%
	Organize the patient hospital's outcome: call for intensivist/out-of-hospital regulation	98,3%	100%
	Ask and check for a vascular access	**100%**	**96,6%**
	Blood examination	**100%**	**86,2%**
	Neuroimaging	**100%**	**81%**
ACAT 3 (ARF)	Overall items (20)	**99%**	-
	Interview the patient/bystander	98,4%	100%
	Dyspnea: presence and characteristics	98,4%	100%
	Gather respiratory symptoms	98,4%	100%
	Complementary investigations	**98,4%**	**97,8%**
	Organize the complete treatment	87%	100%
	Provide team with information and the diagnosis plan	98,4%	100%
	Installation in a half-seated position	**100%**	**94,5%**

ACAT: Acute Care Assessment Tool; CA: Cardiac Arrest; ARF: Acute Respiratory Failure
In bold type : the five non relevant to every scenario items

3.3 Lien avec les autres marqueurs de performance

Trois différents marqueurs ont été étudiés, afin d'analyser les niveaux de corrélation des scores ACAT avec d'autres marqueurs de performance des apprenants et afin de savoir s'ils étaient discriminants. Le score ACAT est bien corrélé avec le score de performance globale et avec le score de performance globale, établis par le même observateur (Tableau 17).

Tableau 17 – Corrélation entre les scores ACAT et les scores TEAM et SPG			
Score	Coefficient rho de Pearson	95% IC	R2
ACAT et score de performance globale			
ACAT 1	0.6	[0.37-0.76]	0.55
ACAT 2	0.72	[0.48-0.85]	0.52
ACAT 3	0.83	[0.66-0.91]	0.74
ACAT et score TEAM			
ACAT 1	0.77	[0.61-0.88]	0.6
ACAT 2	0.67	[0.41-0.86]	0.45
ACAT 3	0.74	[0.52-0.86]	0.55

ACAT : Acute Care Assessment Tool; SPG: score de performance globale, IC: intervalle de confiance à 95% R2 : coefficient de dispersion

Tableau 18 – Comparaison des scores selon le niveau de l'apprenant							
Score	Autre spécialité		MU1		MU3		p
	Moyenne	DS	Moyenne	DS	Moyenne	DS	
ACAT 1	9.3	3.0	11.0	2.6	13.5	2.9	<0.001
ACAT 2	9.6	3.1	12.2	3.1	12.3	2.6	0.114
ACAT 3	10.9	2.2	11.4	3.2	13.9	2.0	0.019

En ce qui concerne le lien entre le résultat obtenu au score et le niveau de l'apprenant, les analyses statistiques retrouvent une association statistiquement significative pour les scores ACAT 1 et 3, mais pas pour le score ACAT 2 (Tableau18).

3.4 Fixation du seuil de réussite

L'utilisation de la méthode Angoff a permis de fixer des seuils de réussite en fonction des niveaux attendus des apprenants. Six experts ont participé à la détermination de trois seuils pour trois niveaux d'apprenants différents : externes en fin de sixième année (et ainsi au premier jour de l'internat), interne en fin de phase socle de médecine d'urgence et enfin interne en fin de phase d'approfondissement, avant de devenir « docteur junior » et d'être en autonomisation progressive.

Pour les scores ACAT 2 et 3, une seule consultation a permis de déterminer les seuils, les variations entre les notes des experts étant inférieures à 30%. Dans le cas du score ACAT 1, il a fallu un deuxième tour de consultation, les experts n'ayant pas pu se mettre d'accord sur 5 items, et dans le cas du niveau attendu d'un externe. Les seuils de validation obtenus sont présentés dans le tableau 19.

Tableau 19 – Seuils de réussite selon la méthode d'Angoff, taux de succès parmi les internes de l'étude

Score	Externe		MU1		MU3	
	Seuil	%	Seuil	%	Seuil	%
ACAT 1	7	NSAP	12	33	16	5
ACAT 2	8	NSAP	13	30	17	0
ACAT 3	8	NSAP	12	50	16	25

3.5 Expérience des formateurs avec les scores ACAT

Le point de vue des formateurs a donc été recueilli en utilisant deux méthodes : un questionnaire remis par l'investigateur principal à tous les formateurs à la fin de l'étude et des entretiens au moyen de focus groupes.

3.5.1 Résultats des questionnaires

18 formateurs qui ont participé aux évaluations ont rempli le questionnaire. Ils représentaient 49% des formateurs de l'étude, sachant que tous n'avaient pas utilisé le questionnaire car ils n'étaient pas évaluateurs. Nous n'avons pas pu récupérer cette information pour tous les centres. Parmi eux, 15 médecins et 3 infirmières, qui étaient âgés en moyenne de 35 ans (DS +/-3,7) et travaillaient aux urgences depuis 7 ans en moyenne (DS +/- 3). Cinq centres étaient représentés, les formateurs des hôpitaux de Lariboisière et de l'hôpital européen Georges Pompidou n'ayant pas rempli le questionnaire. Les formateurs ne trouvaient pas le score pertinent pour des externes (0/18), mais en revanche, ils le trouvaient pertinents pour des internes de médecine d'urgence, toute année confondue (6 pour des phases socles, 8 pour des phases d'approfondissement, 8

également pour les phases de consolidation et seulement 4 pour des médecins urgentistes formés). Onze formateurs sur 18 estimaient que le score de performance globale était un outil indispensable à la note et pour deux d'entre eux, il était même intéressant seul.

Parmi les 18 questions qui concernaient en détail les scores, 14 ont reçu un taux de réponse de 100%, des réponses à 4 questions manquaient pour 2 participants. Les 4 questions qui n'ont reçu que 16 réponses au lieu de 18 étaient les suivantes : les scénarios proposés étaient réalistes ; le score est utile pour guider le feedback formatif délivré à l'apprenant ; le score aide le formateur pour le débriefing et le score est un reflet des compétences requises en médecine d'urgence.

En ce qui concerne l'évaluation de l'utilisabilité des scores, la plupart des évaluateurs estime que le score est facile à utiliser (respectivement 83%, 67% et 83% pour les scores 1, 2 et 3). Les items sont rapportés comme étant compréhensibles et pertinents (avec une variation allant de 67% à 83% selon les scores et la question posée) (Figures 13-15, items 1-2-3).

Pour les experts, les scores seraient utilisables pour évaluer des internes de médecine d'urgence (78% pour le score ACAT 1, et 61% pour les scores ACAT 2 et 3). Mais cela est moins évident en ce qui concerne les médecins urgentistes (78% pour le score ACAT 1, 56% pour les deux autres scores). A l'inverse, les évaluateurs ne pensent pas utiliser ces scores pour évaluer des externes (39% y sont favorables pour le score ACAT 1 et 33% pour les scores ACAT 2 et 3). Ces chiffres se confirment puisqu'à la question de leur utilisation dans le futur, une majorité souhaite se servir des scores pour évaluer les internes (78%, tout scores confondus), une moitié souhaite les utiliser pour évaluer les médecins urgentistes (78% pour le score ACAT 1 et 56% pour les scores ACAT 1 et 2) et 44% les utiliseraient pour évaluer des externes, tout scores confondus (Figures 13-15, items 4-6 et 16-18).

Les évaluateurs ont un avis partagé sur la possibilité d'évaluer la performance d'équipe, puisque pour chacun des trois scores, 50% d'entre eux estiment qu'elle n'est pas évaluée avec cet outil.

De même pour le raisonnement clinique (61% estiment que le score ACAT 1 ne permet pas de l'évaluer, et 50% pour les scores ACAT 2 et 3). Enfin, les évaluateurs perçoivent un intérêt plus important à utiliser ce score pour évaluer la communication du médecin (72% pour le score ACAT 1 et 56% pour les deux autres scores). Cependant, pour la majorité des évaluateurs, les scores sont un outil d'évaluation des compétences requise en médecine d'urgence (94% pour le score ACAT 1 et 81% pour ACAT 2 et 3) (Figure 13-15, items 7-10).

Les évaluateurs estiment que les scores constituent une plus-value parmi les scores déjà existants (entre 61 et 72% d'entre eux) et pour 72%, ils rapportent pouvoir utiliser ces scores en complément des scores d'évaluation des habiletés non-techniques. Enfin, une majorité des évaluateurs (entre 72 et 83%) estime que le score de performance globale est une aide supplémentaire pour apprécier la performance de l'apprenant (Figures 13-15, items 11-13).

Enfin, 75% des évaluateurs rapportent une utilité à se servir des scores pour guider l'évaluation formative des apprenants et leur procurer une rétroaction fondée sur les items de chaque score, de même que 75 à 81% d'entre eux considère que les scores sont une aide pour le débriefing (Figures 13-15, items 14,15).

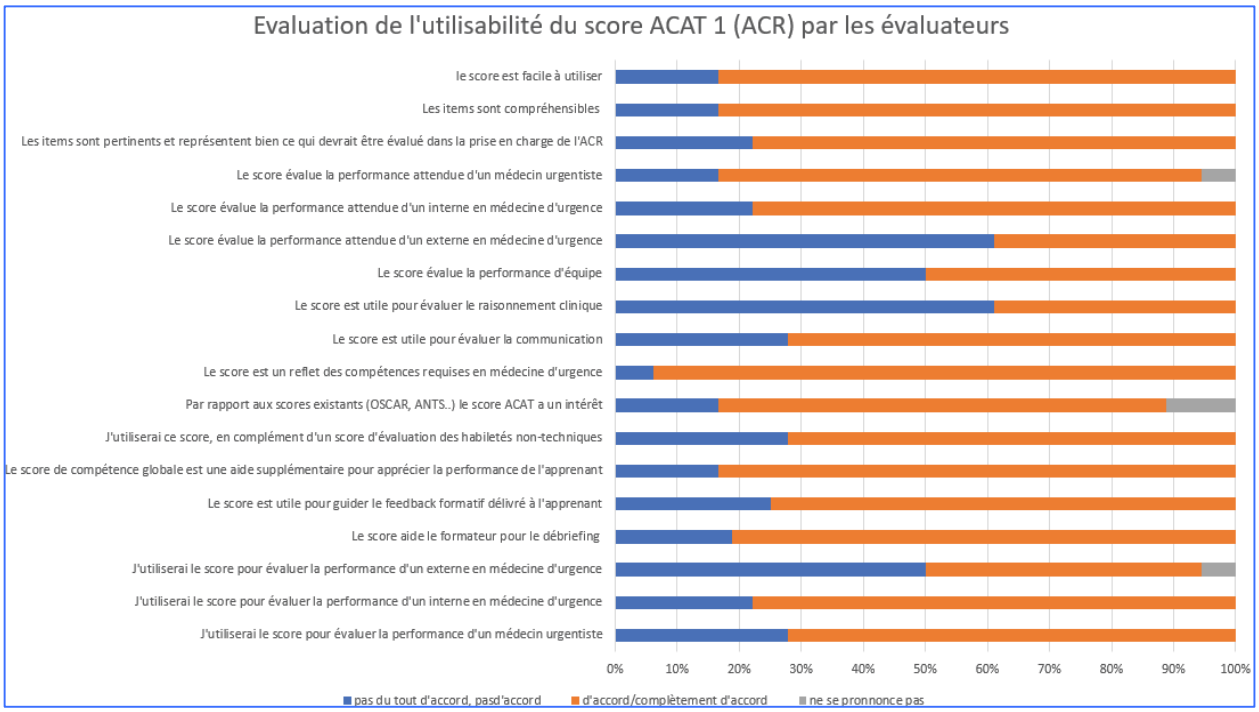


Figure 13 : utilisabilité du score ACAT 1 (arrêt cardiaque)

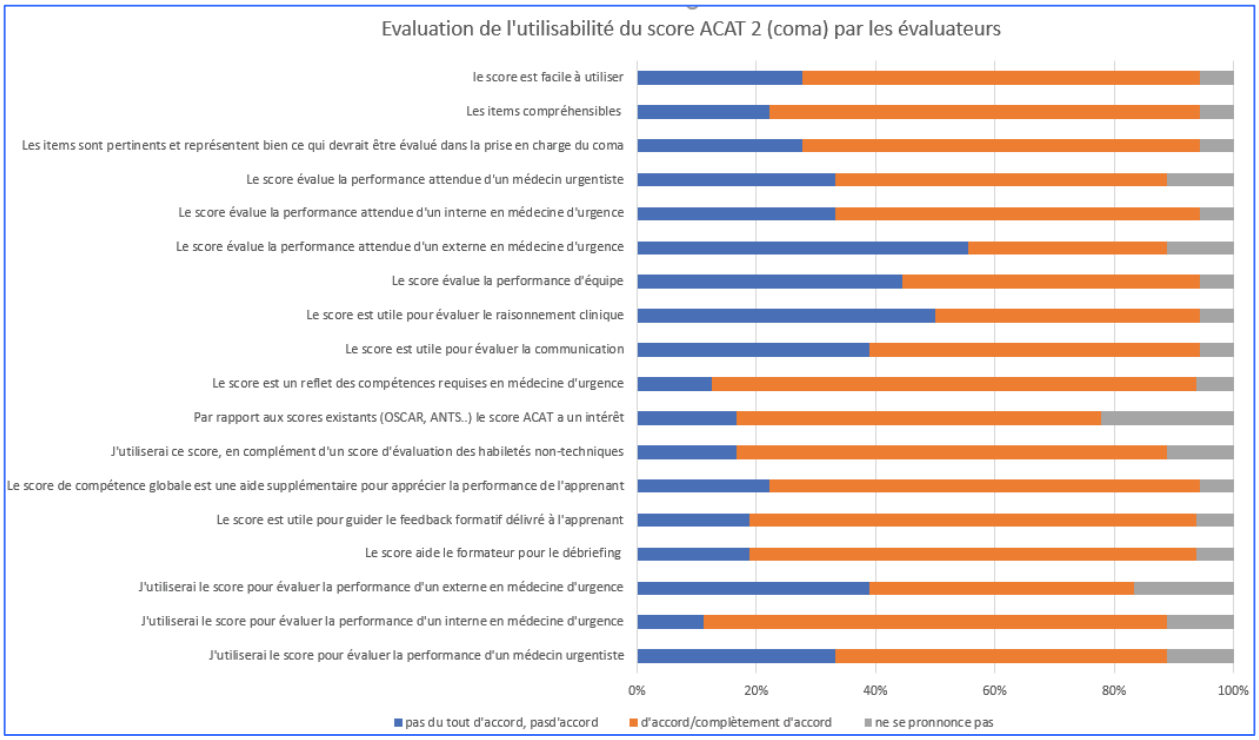


Figure 14 : utilisabilité du score ACAT 2 (coma)

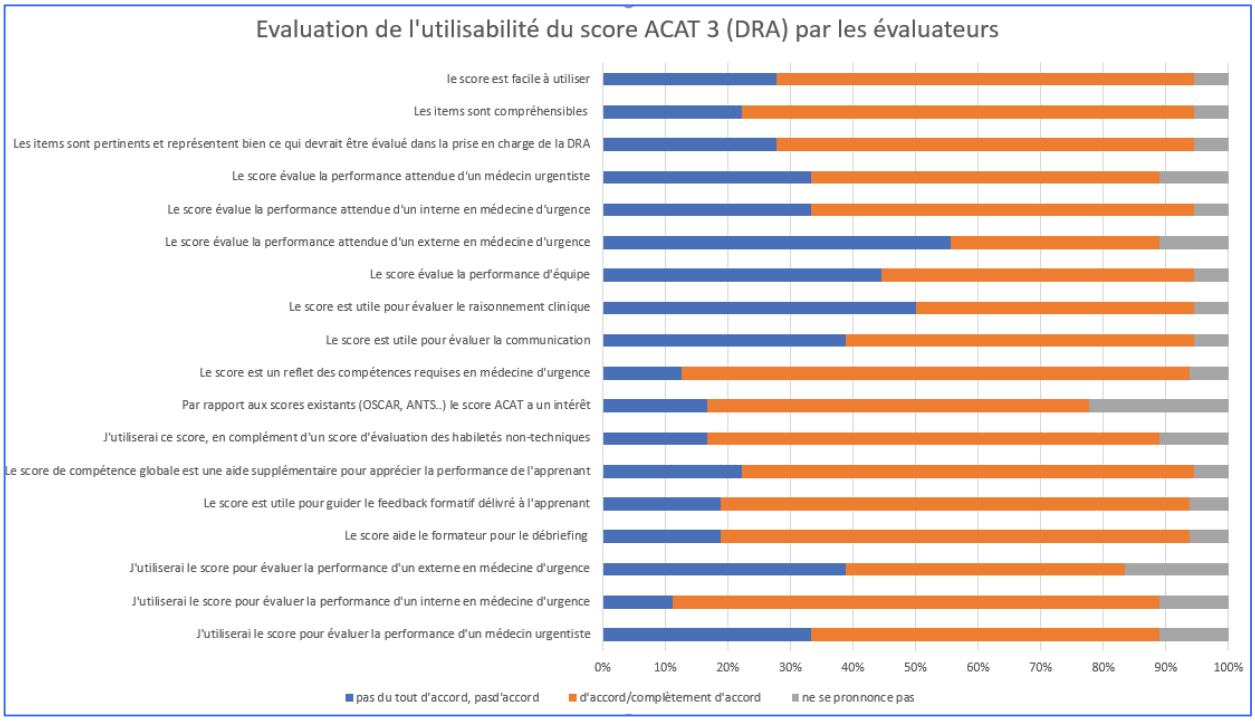


Figure 15 : utilisabilité du score ACAT 3 (détresse respiratoire aigüe)

3.5.2 Résultats des focus groupes

Deux focus groupes ont pu être réalisés entre mai et juin 2021, à distance, compte-tenu de la localisation nationale des enquêtés. Ils étaient composés chacun de 5 formateurs. Tous les centres investigateurs ont été représentés dans les focus groupes. Le premier focus groupe a duré 1h05 et était composé d'évaluateurs-formateurs venant des centres de Strasbourg, de Rouen, et de deux hôpitaux parisiens, la Pitié-Salpêtrière l'Hôpital Européen Georges Pompidou. Le deuxième focus groupe a duré 1h23 et était composé d'évaluateurs-formateurs originaires des hôpitaux parisiens de Bichat, Laribosière, Cochin, Pitié-Salpêtrière et de l'hôpital de Strasbourg.

Après relecture par deux évaluateurs indépendants, la catégorisation des remarques soulevées par les évaluateurs concernait les scores (contenu, utilisabilité, public visé), leur rôle dans le débriefing, la place de la simulation pour évaluer les apprenants, la légitimité qu'ils pouvaient

avoir pour évaluer les apprenants, la profession que devait avoir l'évaluateur et enfin le cadre formatif ou certificatif que devait avoir l'évaluation.

Après analyse thématique des différentes catégories et mise en relation des verbatim, cinq thématiques principales reflétaient l'expérience et la perception qu'avaient les évaluateurs quant à l'utilisation des scores ACAT, en lien, pour trois d'entre eux avec les critères de qualité d'un outil d'évaluation (Norcini et al., 2011) : leur utilisabilité, leur validité, l'effet qu'ils peuvent avoir sur l'apprentissage et sur les pratiques des enseignants, leur acceptabilité et la représentation qu'ils se font de l'utilisation de la simulation pour évaluer des apprenants, et enfin leur utilisation dans un contexte de simulation interprofessionnelle.

Utilisabilité des scores

En rapportant leur expérience d'utilisation des scores, les évaluateurs avaient des retours différents en ce qui concerne la facilité d'utilisation des scores, les uns relevant un nombre d'items trop important et parfois trop denses ou peu clairs alors que les autres ont pu facilement utiliser le score et s'approprier des items très, voire trop précis : *« je les trouve pas si facile que ça à appliquer, y'a beaucoup de trucs, beaucoup d'items, qui sont plutôt simples on est d'accord, mais il y en a beaucoup. Et parfois pas toujours clairs, je sais pas comment mais ils gagneraient à être un peu moins denses (E9) » ; « sur le ressenti j'ai trouvé que ce n'était pas forcément facile, déjà vu le nombre d'items, d'utiliser ce score et euh… je les trouvais parfois un peu trop précis fin avec beaucoup de euh…. beaucoup d'items à évaluer (E10) ; « j'ai trouvé que c'était suffisamment clair, concis, je pense pas qu'il y ait d'erreur d'interprétation sur euh sur quand tu coches une case (E4) » ; « C'était assez simple à utiliser c'est vrai, c'est facile à utiliser, j'ai connu des scores sui sont beaucoup plus compliqués à utiliser (E1) ».*

Quant au score de performance globale, sa simplicité est appréciée des évaluateurs, même si certains le trouvent trop simple, ce qui entraine des difficultés avec son utilisation qui ne repose

pas sur des critères leur apparaissant comme concrets : *« j'ai préféré le score ACAT, avec tous les items, perso celui-là il aide pas trop je trouve (E1) » ; « oui l'ACAT il est plus précis , sinon tu as une image plus globale mais on sait pas sur quoi ça repose en tout cas euh ouais je préfère qu'il y ait les deux plutôt que juste le score de perf globale et quitte à en choisir j'préfère celui où il y a le plus d'items (E4) ; moi je trouve que le SPG c'était bien, je le trouvais intéressant, il collait bien avec l'autre score j'ai l'impression (E5) ».*

Validité des scores

En abordant les questions de public visé par l'évaluation, les performances évaluées par les scores, les évaluateurs abordent la question de leur validité et du degré de conformité avec lequel les scores ACAT permettent l'évaluation de leur objet premier.

Pour les évaluateurs, le contenu du score, même s'il semble complexe pour certains, reflète les performances requises pour la gestion des trois situations cliniques évaluées et peut être utilisé à des fins pédagogiques : *« je ne pense pas qu'il y ait d'erreur d'interprétation sur euh quand tu coches une case ou quoi que ce soit, au-delà même de l'interne ça donnait une idée globale de la prise en charge attendue. Moi ça me semblait pertinent pour chaque thème des scenarios (E4) ».* De même, il évalue les performances de leadership et de communication de l'interne avec l'équipe, ce qui était l'un de ses objectifs : *« C'est dirigé sur le team leader et sa communication (E2) ».* De plus, son utilisation, dans certains services, a permis d'implémenter un programme de formation, dont le contenu pouvait être évalué par les scores ACAT, ce qui souligne sa validité pour les formateurs, en accord avec les objectifs pédagogiques fixés : *« Moi globalement mon ressenti c'est que ça tombait très bien parce que ça colle avec euh un programme pédagogique (E1) ; ce score permet de faire la formation à chaque fois que les nouveaux internes vont arriver (E6) ».*

Le nombre d'items est relativisé par rapport à certains autres scores utilisés en médecine d'urgence, par rapport à l'expérience de l'évaluateur, par rapport au lien entre nombre d'items et validité d'un score, mais également par rapport à toutes les informations qu'ils permettent d'obtenir et de recenser : *« en fait, ceux qui ont l'habitude d'utiliser les scores diront qu'il y a quand même très peu d'échelles ou tu as que euh trois items parce que pour la validité c'est difficilement calculable donc la plupart des échelles ont de toute façon au minimum une dizaine d'items, c'est le principe même des échelles (E1) ; après il faut aussi savoir se dire qu'on a un score d'évaluation, on reste assez simple et ça nous apporte déjà plein d'informations pour eux pour améliorer les compétences (E10) ».* Cependant, afin d'affiner un peu la pertinence des items, certains évaluateurs posaient la question de la valeur accordée à chacun des items, qui, lors de l'étude était identique pour chaque des 20 items de chaque score : *« moi je me pose beaucoup de questions sur la checklist sur euh, est ce qu'on accorde le même niveau d'importance à chaque tâche ou bien est-ce qu'il y en a qui sont plus importantes pour la performance globale et ça s'est hyper compliqué je pense mais voilà (E3) »*

Le SPG, en association avec le score ACAT est perçu comme un outil qui serait plus en lien avec la compétence des apprenants, permettant aux évaluateurs de dire si l'étudiant est prêt à prendre en charge des urgences vitales, et avec quel degré d'autonomie : *« il est peut-être trop simple mais ça force à se poser la question de la compétence (E6) »*. Il est également un guide pour aborder le débriefing : *« je, j'trouve que c'est difficile de remplir le score de performance globale si on n'a pas quelque chose justement pour se baser, c'est là que ça peut être intéressant, en ayant une vision euh une vision globale et euh, ouais j'ai trouvé, c'est vrai que c très simple mais ça me permettait de voir un petit peu euh ouais comment on allait débriefer derrière aussi c'était intéressant (E10) ».*

L'autre préoccupation des évaluateurs était le public visé par l'utilisation de ces scores. En l'ayant utilisé avec des internes, mais en se posant la question de l'utiliser avec d'autres publics, ils se

posaient la question de leur applicabilité pour des externes ou pour des médecins urgentistes. Le consensus entre les deux focus groupes étaient qu'ils leur paraissaient plutôt destiné à des internes, toute année confondue : *« Je pense que pour les externes c'est trop, les internes pourquoi pas et les séniors pas assez (E9) ; Parce que l'idée est quand même à la fin d'évaluer les compétences d'interne sur euh, par la simulation (E2) ; ».* La question de l'utilisation des scores selon les différentes années de l'internat, avec une réflexion autour de la possibilité d'y faire croitre la difficulté, les scores étant jugés difficiles pour des internes de première année : *« j'me disais quand même entre un interne de premier semestre et une fin de cursus il y a un gros gap et je trouvais ça un peu sévère(E2) ; « euh je sais pas si y'a moyen de moduler un peu euh sur des euh niveler un peu en fonction de leur avancement dans le cursus (E6) » ; « c'est très intéressant ce que tu dis car au début quand on a accepté l'étude je pensais que ça s'adressait, que c'était plutôt de l'évaluation d'équipe euh expérimentée entre guillemets et ce qu'on s'est tous dit quand on a fait le truc c'est ohlala les scores pour un médecin expérimenté j'vois pas bien ou on va et euh là pour le coups j'te rejoins complètement je trouve que c'est plutôt adapté pour des jeunes internes ou des vieux externes, quelque part c'est plutôt positif entre guillemets on arrive à voir à qui se destine ce genre de score (E10) ».*

Effets de l'utilisation des scores sur l'apprentissage via la modification de la rétroaction

Les évaluateurs ont tous intégré l'utilisation de la grille d'évaluation à leur débriefing. Ils font cependant la différence entre l'utilisation des éléments recensés par le score et celle de la note, qui pour eux, n'a pas de valeur ajoutée au débriefing.

Ainsi, ils perçoivent que l'utilisation du score leur permet de fournir un cadre validé des éléments à débriefer, mais ils gardent la possibilité de s'en affranchir, pour éviter de réaliser des débriefing trop « scolaires » : *« j'm'en suis servi côté débrief, pour faire un retour cadré » (E1) ; c'est vrai que si tu fais ligne par ligne c'est peut-être un peu scolaire mais du coup on a vraiment une idée de ce qu'ils ont fait et puis de ce qu'ils doivent améliorer (E3) ; ça pouvait être aidant pour le*

débrief parce que ça cadrait un peu mais finalement euh à trop vouloir suivre euh les items on pouvait passer à côté de euh.. Non pas qu'ils étaient pas pertinents mais on pouvait passer à côté de questionnements et de choses sur lesquels les participants voulaient réagir, donc c'était utile, mais en évitant le risque de trop parler, et de toujours partir de ce que rapportaient les apprenants, comme on a appris pour faire un bon débrief (E4) ; je l'utilisais comme support mais c'est pas du euh je fais item par item pour débriefer (E8). Ainsi, même comme guide, les évaluateurs perçoivent le risque d'utiliser uniquement le score pour débriefer, mais peuvent s'en servir comme outil de dépistage des difficultés des apprenants : *« Mais le score c'est un apport objectif, c'est à dire que si les items ils sont ratés tu dis, là y'a un truc euh qui a été raté tu t'attendais à ce que le score sur cette partie soit meilleur parce qu'il y avait des choses qui marchaient et tu peux comprendre pourquoi cette partie du score a été ratée et tu peux reprendre pour comprendre à partir de l'observation, à partir des éléments de difficultés dire on attendait certains gestes ou certaines compétences non techniques et que je n'ai pas vu et j'aimerai savoir pourquoi (E7) ».* On voit ici une utilisation formative des scores, détachée de la note qu'ils ont pu mesurer.

De plus, les évaluateurs ont perçu que l'utilisation du score modifiait la rétroaction qu'un étudiant peut avoir après une évaluation, et ainsi, perçoivent un effet sur l'apprentissage : *« C'est un peu sévère mais y'a un axe de progression, un axe pédagogique, moi j'ai retrouvé quelque chose de pédagogique dans ces scores tu vois, qui peux servir à une base de progression pour l'interne, c'est un apport pour son processus d'apprentissage, car il permet de décortiquer la note (E6) ».* La note n'a pas sa place dans la rétroaction donnée aux étudiants, et pour les évaluateurs, elle n'apporte rien au débriefing qui, pour les enseignants constitue un avantage dans une évaluation certificative : *« la valeur du score je sais pas si ça va m'apporter grand-chose, mais ce que je donnerai comme axe à l'étudiant, oui (E10) ».* Ainsi, contrairement à une note brute, les

enseignants perçoivent l'intérêt de la note expliquée par le débriefing, qui va permettre d'améliorer le transfert d'apprentissage.

Acceptabilité des scores par les enseignants

Questionner les enseignants sur l'utilisation des scores ACAT dans un programme d'évaluation des internes a permis d'explorer leur représentation non pas des scores, mais plutôt de l'évaluation par la simulation, dans une vision plus globale. Après avoir regretté une perte de la bienveillance de la simulation, ils dégagent les deux cadres possibles d'utilisation de l'évaluation : le cadre formatif, acceptable et le cadre certificatif, beaucoup plus discutable.

Les enseignants ont tous rappelé qu'une des valeurs fondamentales de la simulation était le cadre bienveillant dans lequel elle se situait : *« j'trouve qu'il y a quand même une bienveillance dans la « sim » sinon personne n'en ferait ou ne participerait (E2) » ; c'est pas trop l'esprit de la simulation pour moi (E5) ; la simulation c'est un côté un peu plus bienveillant en discussion en débriefing (E7) »*. Pour eux, utiliser une note engendre une perte de bienveillance car cela entraine une façon de travailler différente : les conditions changent, de même que la rétroaction qui risque d'être focalisée plus sur ce que les apprenants n'ont pas ou mal fait : *« avant, on restait dans une optique de bienveillance et donc de pas noter, de pas obligatoirement ressortir que les trucs négatifs mais plutôt de ressortir que ce est positif et c'est là que ça m'embêtait un petit peu c'est que des fois, bah c'était un peu la catastrophe mais on essaie quand même de ressortir le positif alors que finalement d'avoir une note, ca va un peu à l'encontre de ça (E9) ; « je pense que dans le cadre de la formation et de la simulation on peut pas être dans le cadre de l'évaluation et de la note (E8) »*.

Dans un contexte formatif, nous l'avons vu, les scores seraient utiles pour cadrer l'apprentissage, pour fournir un cadre théorique valide au débriefing, mais les enseignants, dans ce cas n'ont pas besoin de fournir la rétroaction sous forme de note car elle n'a pas de sens pour eux. Les scores

sont également utiles pour suivre la progression des apprenants, et donc en adéquation avec un cadre formatif : *«pour faire un retour cadré et marquer quelques notions, et de les faire progresser donc ça peut être intéressant pour faire du suivi, de la répétition de simulation et de voir comment ils évoluent (E4) ; sinon plutôt pour faire des courbes de progression (E2) ; et l'intérêt des scores c'est que tu amènes de l'hétéroévaluation objective qui te permet de faire de calculs d'évolution et je pense que les scores détaillés ont un intérêt (E1) ».* Ainsi, certains évaluateurs pensent les utiliser à nouveau, dans ce cadre formatif dans un premier temps « *je compte bien l'utiliser, je trouve qu'ils ont l'air plutôt bons quand même et euh et je j'attends la confirmation mais l'objectif est de les utiliser bien évidement, tous les semestres à l'arrivée de chaque nouvel interne, qui auront un axe de progrès et avec les équipes du porte parce que ça répond à une problématique du service. Ça tombe donc bien, donc oui je les réutiliserai. En novembre, je peux même dire quand (E1)».*

Le cadre certificatif leur pose plus de problème, à la fois car certains trouvent les scores difficiles pour des jeunes internes, mais également car cela modifie trop le cadre de la simulation. Cependant, ils s'accordent à dire que s'il faut certifier avec l'outil simulation, il faudrait une note : *« sur de la certification ça peut paraitre intéressant autant sur de la formation je sais pas (E7) ; la note pour certifier oui, pas pour former ; pour faire du formatif on peut utiliser les scores, mais pas pour le côté sanctionnant genre j'évalue ton stage avec ça (E10) ; effectivement si tu faire de l'évaluation sanctionnante je trouve que ça serait trop sévère et euh pas utilisable dans le mode "je t'évalue pour dire si tu as réussi ou pas", Pas au stade de euh, des novices (E1) ; soit on fait de la formation et on n'est pas là pour noter soit on est là pour faire passer un examen et on note et on évalue les capacités. (8) ».* Deux évaluateurs pensent tout de même utiliser la simulation comme outil de certification des compétences des internes, via les scores ACAT entre autres et avec une note : *« moi je pense qu'il y a de la place pour les deux (E1) ; dans ma tête c'est pas*

forcément incompatible mais euh à ce moment on frôle un peu plus avec en effet une vraie note, peut être en fin de stage ou de DES ».

L'utilisation des scores et de l'évaluation reste cependant fortement liée à la faisabilité de l'évaluation, les enseignants percevant des difficultés à les utiliser selon le même schéma (c'est-à-dire avec un double regard évaluatif), dans les conditions actuelles de formation et de travail, marquées par un manque de moyens humains : « *je les réutiliserai si on a assez de formateur, euh, alors l'idéal ça serait de faire à l'arrivée des internes et à la fin, quand ils s'en vont mais euh, matériellement 'fin en humain », je sais pas trop comment faire (E2) ».*

Ainsi, en parlant des scores ACAT, les évaluateurs ont finalement abordé la question de l'acceptabilité de l'évaluation par la simulation, qu'ils ne pratiquaient pas jusqu'à présent et qui n'est pas une pratique répandue en France. Ils ont souligné deux points : bien que les scores représentent une aide pour le débriefing, pour l'appréciation de la performance globale, voire de la compétence des apprenants et donc une utilisation dans un cadre formatif acceptable, leur utilisation dans un cadre certificatif leur parait beaucoup moins envisageable, ou à certaines conditions. Ils font un lien presque indissociable entre évaluation certificative et note ; puisque dans ce cadre, l'utilisation de la note leur parait possible et justifiée.

Utilisation des scores dans un contexte de simulation interprofessionnelle

Utilisés dans le contexte de simulation interprofessionnelle, les scores posent des questions aux évaluateurs, selon trois axes : la question de quelle profession peut évaluer quel professionnel et donc celle de la légitimité en tant qu'évaluateur, l'importance du regard porté par les autres professions sur la performance de l'interne et enfin l'influence que peut avoir l'évaluation de l'interne sur celle de l'équipe.

Les évaluateurs infirmiers posaient la question de la légitimité de leur regard sur le travail de l'interne, alors que les évaluateurs médecins s'interrogeaient et semblaient trouvaient pertinent

d'avoir des retours et évaluations de la part d'évaluateurs infirmiers. Cela se traduisait à la fois dans les évaluations mais également dans la façon de mener le débriefing et de délivrer le feedback : *« même avec le score il nous était parfois difficile d'évaluer le travail de l'interne, on ne se trouve pas légitime, même si on a eu la formation et même si bien souvent on avait les mêmes remarques que les médecins lors des débriefings (E5) ; c'est vrai que nous on a eu un peu parfois des infirmiers qui trouvaient qu'on ne les avaient pas assez débriefés sur leurs compétences techniques, parce que nous il n'y avait que des médecins dans les formateurs, mais euh... est-ce que c'est du fait du score je crois pas, je pense que ça vient aussi de la compétence du débriefeur, de sa manière de faire je sais pas (E3) ; nous on n'avait pas d'infirmier dans les formateurs et euh et peut être que ça a manqué, j'ai l'impression que ça a pu manquer (E6) ».* Par ailleurs, le décalage entre formation initiale et formation continue s'ajoutait à la difficulté de l'évaluation en situation interprofessionnelle : *« ce qui m'a euh ce que j'ai trouvé un peu difficile c'est le côté interdisciplinaire et le fait de débriefer en même temps des internes en formation pour le coup que tu peux évaluer ou pas via la simulation et des infirmier déjà en poste dont c'est le métier, qui sont là, euh qu'on n'est pas là pour juger, qui doivent progresser dans ce qu'ils doivent déjà faire au quotidien mais on peut leur apprendre des choses mais en même temps on va pas leur apprendre leur métier, fin tu vois, donc sur la communication d'équipe c'est hyper facile, sur les compétences de chacun notamment IDE[25] moi j'ai trouvé ça, un peu plus délicat et les scores à double tranchant (E7)».* Cela pose la question de la difficulté et la pertinence de l'évaluation en interprofessionnel, mais sans qu'aucun des enquêtés ne l'ai remise en question d'autant que les scores n'étaient pas destinés à évaluer l'équipe mais l'interne en situation interprofessionnelle.

Pour certains évaluateurs, le fait d'évaluer l'interne pouvait être un biais qui influençait leur appréciation globale de l'équipe : *« je me demandais toujours si l'évaluation de l'interne et le fait*

[25] Infirmier Diplômé D'Etat

de se concentrer sur sa performance influençait l'évaluation de l'équipe, ça je savais pas (E10) ». Les évaluateurs soulignaient la nécessité de scores complémentaires pour évaluer l'équipe et notamment pour les habiletés non-techniques. Ainsi, le fait de devoir remplir le score TEAM dans le cadre de l'étude semble les avoir aidés pour leur débriefing d'équipe : *« Apres euh, il y a le score ACAT mais c'est vrai que moi j'étais beaucoup plus à l'aise sur l'évaluation d'équipe sur les scores génériques d'évaluation d'équipe au final. J'trouvais ça plus adapté pour le débriefing, pour l'aide au débriefing (E10) ».*

A la difficulté d'évaluer d'autres professions que la sienne, s'ajoutat le souci d'évaluer et de faire de la simulation avec ses collègues, par peur d'introduire une relation hiérarchique avec eux, ainsi que le décrit un médecin : *« y'a un souci aussi j'pense hiérarchique c'est que nous on est un peu gênés d'évaluer les infirmières avec qui on travaille au quotidien et avec qui on veut que ça se passe bien alors est-ce qu'il faut pas qu'il y ait une évaluation par des pairs, bienveillants ? Euh les inf' évaluent les inf', les AS*[26] *les AS (E8).*

D'autres évaluateurs auraient apprécié avoir une évaluation de la part des infirmiers et aides-soignants qui avaient réalisé la simulation avec l'interne. Soulignant l'apport différent qu'ils feraient en comparaison avec les évaluateurs médecins, moins centré sur les habiletés techniques, mais plus au cœur des habiletés relationnelles ou de communication. Cependant, cela a été beaucoup discuté par d'autres formateurs rappelant que c'est précisément le rôle du débriefing car il permet ce genre d'interactions : *« le PU ou le médecin et une infirmière expérimentée devraient noter les internes, et en fait la note que donne l'infirmière serait aussi importante que celle que donne le médecin, parce que finalement ils sont vachement plus sur la gestion d'équipe, alors que les PU ou médecins ils restent très médical, très respect des reco etc. [....] je trouve que le ressenti de l'équipe sur place est très intéressant (E8)» ;*

[26] Aide-soignant.e.s

CHAPITRE 4 : DISCUSSION

Les différentes études menées pour décrire le développement des scores ACAT, leur utilisabilité, et leur potentielle place dans l'enseignement et l'évaluation des étudiants et internes en médecine d'urgence ont permis de mettre en évidence plusieurs sources de validité, et de faire émerger une discussion autour de leur utilisation future.

S'inscrivant dans le cadre d'un processus de validation d'un outil d'évaluation précis et validé, le développement des scores a permis de créer des outils ayant un contenu valide, puisque reposant à la fois sur les recommandations récentes, sur une consultation de 58 experts suivant la méthode Delphi, puis sur le retour des évaluateurs nombreux, qui ont pu utiliser les scores ACAT lors de leur phase d'évaluation en situation réelle de simulation in-situ. De plus, l'analyse qualitative des items des scores ACAT a permis de préciser leur contenu.

La structure interne des scores, illustrée par une bonne consistance interne et une excellente reproductibilité intra et inter-observateur, permet également d'établir la fiabilité des scores ACAT, qui est également une des sources de validité d'un outil de mesure de la performance.

De plus, les scores ACAT ont une bonne corrélation avec les scores de performance globale et avec le score TEAM ce qui renforce leur validité, en les liant soit avec la performance globale de l'apprenant, soit avec un autre score validé, même s'il s'intéresse plus spécifiquement aux habiletés non techniques d'un leader et de son équipe lors des situations de prise en charge des urgences vitales. Si l'on fait le lien entre les mesures effectuées avec les scores ACAT et le niveau d'étude des apprenants, il apparait que les scores ACAT 1 et 3 permettent de les discriminer selon ce niveau, ce qui leur fournit une source supplémentaire de validité, dans leur relation à la variable « niveau », autrement appelée validité de construit.

Enfin, l'expérience qu'ont eue les enseignants avec l'utilisation de ces scores est riche d'enseignements : en considérant que leur contenu représente précisément les situations

évaluées, ils leur attribuent une bonne validité de contenu avec une tendance à les trouver facilement utilisables (même si cela a été sujet à débat lors des entretiens). Les enseignants soulignent que le score de performance globale est complémentaire des scores ACAT et qu'il est plus en lien avec la compétence des apprenants. Pour les enseignants, les scores ACAT pourraient être utilisés principalement avec des internes en médecine d'urgence, plus qu'avec des externes, ou alors en modifiant les exigences requises. Cela est souligné également par l'utilisation de la méthode Angoff, qui a permis de fixer des seuils à partir de 8/20 pour les externes jusqu'à 16/20 pour des internes de fin de troisième année de DESMU. Les enseignants ont également perçu l'effet positif que pourrait avoir l'utilisation des scores ACAT dans leur pratique, car ils leur fournissent un cadre validé pour le débriefing et pour donner des axes d'amélioration aux apprenants, basés sur leurs observations. L'utilisation des scores a aussi modifié leurs pratiques puisque les enseignants s'en sont tous servi pour réaliser leurs débriefings. Ainsi, les scores ACAT leur semblent être un outil intéressant, mais plutôt un cadre formatif. En effet, les enseignants émettent des réserves à la fois sur l'utilisation d'une note avec l'outil de simulation, de même que pour l'utilisation de la simulation dans un cadre d'évaluation certificative ou normative, craignant perdre un des éléments essentiels de cet outil d'enseignement : la bienveillance.

Enfin, les enseignants se sont posé la question de l'évaluation en interprofessionnelle et soulèvent des difficultés rencontrées dans ce cadre : le regard évaluatif entre les professions et la difficulté d'évaluer à la fois des apprenants en formation initiale, avec des professionnels qui viennent se former dans un cadre de formation continue.

A partir de ces différents résultats, nous allons discuter les différentes propriétés des scores qui permettront de discuter leur possible place dans le cursus actuel de formation. Puis, nous questionnerons la place des scores ACAT au sein d'une approche par compétence et la place de

l'évaluation au sein d'un enseignement par simulation. Enfin, nous exposerons les limites des différentes études puis les perspectives de recherche qui émanent de ce travail.

1. QUALITE D'UNE EVALUATION ET SCORES ACAT

La première question de ce travail de recherche portait sur la possibilité de développer des scores d'évaluation valides, selon un processus de validation rigoureux, afin de pouvoir se reposer sur des scores qui seraient utilisables lors d'une évaluation par la simulation. Nous allons donc analyser dans un premier temps les différents résultats obtenus grâce au processus de validation utilisé, via les différents critères de qualité d'une évaluation, décrits par Norcini (Norcini et al., 2011, 2018)

1.1 Validité des scores ACAT

L'utilisation d'un processus de validation rigoureux, a permis d'analyser les différents sources validité des scores ACAT, qui permettent d'établir que les résultats du score seraient adaptés à son objectif spécifique et reposent sur différentes preuves. Nous en faisons ici l'analyse.

1.1.1 Le contenu des scores

Devant le manque d'outil permettant d'évaluer à la fois les habiletés techniques et non-techniques dans des situations d'urgence vitale en formation initiale, l'objectif de notre travail était de développer ces outils, réunissant ces deux éléments fondamentaux à une prise en charge optimale des urgences vitales, et dans une optique d'enseignement basé sur la continuité de son contenu (Englander & Carraccio, 2018; Greif et al., 2021; Sevdalis & Brett, 2009).

Le choix des trois situations, réalisé grâce à la consultation des coordonnateurs régionaux du DES de médecine d'urgence, parmi 36 items issus de la formation théorique et du programme de formation en médecine d'urgence en vigueur en 2017-2018, s'est révélé être pertinent dans une vision d'approche par compétence. En effet, ce sont des situations complexes, auxquelles les étudiants sont peu exposés pendant le deuxième cycle et pour lesquelles ils ne sont pas en

autonomie lors des trois premières années de leur internat. Cependant, ils doivent être en mesure d'agir et d'initier la prise en charge initiale lorsqu'ils y sont confrontés. Ainsi, la simulation et son utilisation pour certifier les étudiants avant une autonomie complète apparaissent être des outils adaptés pour l'enseignement puis l'évaluation des étudiants. L'arrêt cardio-respiratoire correspond à une situation déjà majoritairement enseignée pour les internes de médecine d'urgence en France (Allain et al., 2018). En revanche, les autres situations retrouvées parmi les plus abordées en simulation étaient l'état de choc, la douleur thoracique et tachycardie ainsi que le traumatisé grave. Les trois premières thématiques pourraient néanmoins s'intégrer dans les situations évaluées par les scores ACAT, du fait de leur utilisabilité avec des contextes variables. L'intérêt de raisonner avec les familles de situation qui ont des points communs dans leur contenu est ainsi d'avoir un cadre assez large qui repère les invariants de chaque situation et permet ensuite de varier le contexte (Pastré, 2011; Roegiers, 2000). Il est important de préciser que, dans notre travail, même si le raisonnement est le même (trouver des invariants dans chaque situation rencontrée en médecine d'urgence), il s'agit plus de situations cliniques de départ, que de familles de situation.

Le processus de création des items a utilisé une méthode Delphi, et la consultation de 58 experts (20 par score), avec un excellent niveau de participation des experts tout au long de la durée du processus, ce qui nous permet d'obtenir des scores avec des contenus qui ont été discutés, choisis au regard de littérature dans un premier temps puis des experts dans un second temps.

L'idée de départ était de pouvoir repérer des éléments invariants de chacune des trois situations d'urgence vitale afin de pouvoir utiliser les trois scores dans des contextes cliniques variés et avec des populations d'étudiants variées. L'objectif était en effet de pouvoir créer des scores qui seraient utilisables à des niveaux très variables de l'apprentissage, ce qui a probablement engendré une difficulté dans le choix de leurs contenus. En effet, s'il est apparu important aux experts de pouvoir vérifier que des étudiants en médecine, au premier jour de leur internat, étaient

capables de réaliser les gestes de bases en situation d'urgence vitale (comme par exemple asseoir un patient qui présente une détresse respiratoire aigüe), ces éléments apparaissaient plus comme des détails pour des internes en fin de parcours, à l'aube de leur prise de fonction en autonomie. Cependant, il a été décidé de les garder, puisqu'ils étaient considérés comme essentiel à la prise en charge de la situation et qu'ils se devaient d'être maîtrisés.

Par ailleurs, les enseignants ont estimé que les scores reflétaient bien les différents éléments à prendre en compte dans chacune des situations évaluées, ce qui permet une analyse plus qualitative du contenu des scores, et renforce une des preuves de sa validité. Une partie de la validité d'un score est liée à sa capacité à discriminer les étudiants, nous aborderons cet aspect dans la partie discutant le public visé par le score.

1.1.2 Le processus de réponse

L'étude du processus de réponse, ou encore de la façon que l'on peut avoir d'utiliser les différents items retenus par les experts permet d'aborder la question de l'analyse qualitative des items, réalisées par deux évaluateurs (Article 2), mais également de parler du choix de la typologie de la notation. En effet, en matière d'évaluation sommative, deux outils d'évaluation existent : les listes de contrôle, encore appelées checklists, ou les échelles de notation globale. Pour les scores ACAT, il a été fait le choix d'utiliser une notation en trois niveaux, proche d'une liste de contrôle car en rapport avec des comportements observables (avec une nuance dans la réalisation complète, incomplète ou partielle de la tâche requise).

Alors que l'utilisation de la liste de contrôle demande à l'évaluateur de reporter s'il a observé un comportement ou une tâche lors de l'évaluation de l'étudiant, le score global lui demande de juger une performance globale. La liste de contrôle, permet de décrire si oui ou non l'action requise a été effectuée (0 ou 1) alors que dans le score global il existe le plus souvent des niveaux d'appréciation de l'action (0, 1, 2, 3…), pondérés par une appréciation globale de la performance de l'étudiant, en général à l'aide d'une échelle de Likert.

La liste de contrôle permet de faire des évaluations analytiques, basées sur une analyse détaillée des différents comportements requis, qui ont été décrits et/ou illustrés par un exemple et pas seulement par un mot (échec/satisfaisant/honorable etc.), afin d'augmenter la reproductibilité de l'échelle et d'améliorer son utilisation, plus intuitive, par des formateurs ayant reçu une formation courte. Les échelles de notation globales, au contraire, permettent d'apprécier une performance dans sa globalité et avec un point de vue plus systémique. Elles sont structurées comme des échelles de Likert et ont, en général entre trois à six niveaux de jugement. Les critères requis sont moins détaillés, mais l'intérêt de l'utilisation de ces échelles réside dans le fait qu'elles représentent mieux la performance globale de l'apprenant. Le tableau 20 résume les différences communément décrites entre les deux outils de notation.

Ainsi, la validité des listes de contrôle, serait moindre que celle des échelles d'évaluation globale qui sont plus subtiles, et plus fidèles à la réalité. Une récente revue de la littérature modère cette différence puisque parmi les 45 études comparant checklist et score global, la reproductibilité inter-observateur était similaire entre les deux outils, mais surtout lorsque le score concernait l'évaluation d'une seule tâche, de nature procédurale ou pour évaluer des habiletés techniques. En revanche, dans le cadre de l'évaluation des habiletés non techniques et notamment la communication, les échelles de notation globale semblent plus appropriées (Newble, 2004). Cependant l'échelle de notation globale garde des avantages certains en comparaison aux listes de contrôle. Elle offre en effet une meilleure fiabilité (notamment pour la fidélité inter-item et l'inter-station), la possibilité d'évaluer plusieurs tâches avec le même outil évitant ainsi de développer un outil pour chaque tâche et simplifiant les évaluations. De plus, leur variabilité dans les possibilités de notation permet de discriminer mieux les étudiants, introduisant plus de subtilité et la notion de perspectives différentes entre les évaluateurs (Hodges et al., 1999; Regehr et al., 1998; Schwartz et al., 2017). Ilgen souligne ainsi que l'appréciation l'évaluateur quant à la performance réalisée, étant moins liée à des critères limités à l'observation, ce dernier peut

émettre un jugement plus constant, mais également plus discriminant entre les performances observées (Crossley & Jolly, 2012; Ilgen et al., 2015). De ce fait, l'utilisation de l'échelle de notation globale nécessite une bonne formation des évaluateurs, qui leur permet d'avoir en tête une idée des attendus d'apprentissage pour chaque niveau de la formation (Ilgen et al., 2015). Ces résultats soulignent le fait qu'un test ne peut pas se passer de l'expertise et du jugement humain, ce qui nous développerons dans le chapitre suivant.

Tableau 20 – Caractéristiques des listes de contrôle et des échelles d'évaluation globale, d'après Ilgen et al, 2016

	Liste de contrôle	**Echelle d'évaluation « globale »**
Critères de jugement	Comportement observable	Comportement + performance globale
Métrique	Binaire	Pondérée (au moins 3 niveaux)
Utilisation	Plus intuitive Formation courte	Plus technique Nécessite formation
Objectivité	Semble plus objectif	Semble plus subjectif
Validité	Moindre	
Fiabilité	Moindre	
Discrimination de niveau	Moindre	Oui

Le choix d'une échelle à trois niveaux, permettant plus de variations qu'une liste de contrôle, mais avec des items très bien décrits et mais basés sur des comportements observables, a donc été fait pour l'évaluation de ces trois situations d'urgence vitale. L'objectif était d'essayer de combiner à la fois les critères de qualité d'une liste de contrôle, qui permet une rétroaction intéressante auprès des apprenants (ainsi que l'ont relevé les enseignants), mais à la fois une probable meilleure discrimination du niveau de l'apprenant. Associés à une échelle de notation globale pouvant aider les évaluateurs à prendre une décision sur la performance observée, nous émettons l'hypothèse que cela améliore la prise de décision, même si cela n'a pas été testé dans notre recherche. En effet, même si les scores ACAT semblent corrélés aux scores de performance globale, il reste encore à déterminer si leur utilisation synchrone permet d'améliorer la décision prise grâce à leurs mesures.

Enfin, les résultats qui concernent la pertinence des différents items en fonction des scénarios sont plutôt bons. En fonction des études (2 et 3), et sur les soixante items des trois scores ACAT, 49 items étaient pertinents dans 100% des cas, 4 étaient pertinents dans plus de 90 à 99% des cas, 6 dans 70 à 89% des cas. Cela souligne une validité de leurs contenus puisqu'ils peuvent être utilisés dans des contextes variables sans être modifiés ou adaptés par les enseignants. Un seul item faisait exception : celui qui s'intéressait à la gestion de la famille dans les arrêts cardio-respiratoires et qui n'était pertinent que dans 27 à 58% des cas selon les études. L'item a été retenu par les experts qui ont donc considéré qu'il s'agissait d'un élément majeur à prendre en compte dans les situations d'arrêt cardiaque. Or, les différents scénarios d'arrêt cardiaque utilisés pour les formations dans les deux études sur la reproductibilité n'avaient pas prévu cet élément dans les situations simulées. Nous ne pouvons qu'émettre des hypothèses pour expliquer la constatation d'un décalage entre la situation supposée « idéale » pour la formation (avec la famille à gérer) et la situation finalement mise en place par les formateurs (sans la famille). Bien souvent, le membre de la famille ou témoin est géré par l'un des formateurs, appelé facilitateur. Or, dans une situation de simulation, les formateurs doivent pouvoir gérer la technique, l'aide aux apprenants, et, dans le cas de la troisième étude, l'évaluation. Ainsi, il est probable que le manque de ressources humaines conduise à « faire l'impasse » sur un des éléments considérés important dans la situation. Cela permet d'observer que, comme dans le travail, il existe des situations de formation « idéales », voire prescrites (ici par les experts) et que les formateurs s'en accommodent et réalisent un compromis avec les ressources à disposition dans leur réalité de formateur (Olry, 1995). Une autre hypothèse est que la proposition et la sélection de cet item a été biaisée par la recherche, non pas en cours, mais par la recherche en médecine d'urgence, une équipe française ayant publié un article d'importance mettant en évidence que la présence de la famille lors de la prise en charge d'un arrêt cardiaque ne compromettait pas sa sécurité psychologique, voire l'améliorait tandis que le stress des soignants n'augmentait pas (Jabre et al., 2013).

1.2 Fiabilité

La fiabilité d'un test permet d'étudier les biais possibles à l'interprétation d'un test et ainsi que les résultats qu'il produit seront les mêmes quelles que soient les circonstances auxquelles sont soumis le test, les apprenants et les enseignants. Lorsque la fiabilité s'intéresse uniquement à la constance à mesurer que possède un test on parle également de fidélité. Il s'agit donc de la constance ou de la cohérence avec lesquelles un test évalue les performances. Nous discutons ici les éléments de fiabilité des scores ACAT, notamment en lien avec l'analyse de la structure interne des scores ACAT et du score de performance globale.

1.2.1 Reproductibilité

Les deux études qui s'intéressaient la reproductibilité ont permis d'établir une bonne voire excellente reproductibilité entre évaluateur et pour un même évaluateur à deux moments différents, que ce soit uniquement lorsque deux évaluateurs participaient à l'étude (Article 2), ou bien quand de nombreux évaluateurs ont participé (Article 3). Cela valide une partie de la fiabilité des scores ACAT. Le résultat peut être expliqué par les étapes précédentes du développement du test et notamment par l'étape d'analyse qualitative des items qui a permis de préciser certains items, en les confrontant à des situations simulées variées (Article 2).

En effet, une grande partie de la reproductibilité des items repose sur la façon qu'a un évaluateur de les utiliser. En précisant leur contenu et la façon de les noter, il a été possible de minimiser les erreurs potentiellement attribuables aux évaluateurs. Ce résultat pose la question de la formation des formateurs, indispensable avant toute évaluation sommative par la simulation car cette activité ne requiert pas les mêmes compétences que lors d'une situation d'entrainement avec la simulation (Koster & Soffler, 2021). Les évaluateurs doivent en effet avoir connaissances des bonnes pratiques en matière d'évaluation, des biais potentiellement créées par l'outil d'évaluation, de sa calibration (en fonction des attendus d'apprentissage par exemple), mais ils

doivent également maitriser l'outil qui leur permettra d'émettre un jugement sur la performance observée.

De nombreux facteurs sont confondants dans une évaluation et la connaissance de leur existence peut aider les évaluateurs à les minimiser. Parmi ces facteurs on retrouve l'effet halo, qui est l'impression suscitée chez l'évaluateur par la performance d'un étudiant pour un élément de l'évaluation, qui influencera le jugement des autres éléments. L'effet de stéréotypie est le mécanisme de classement des élèves chez l'évaluateur : une fois qu'un élève se retrouve dans une catégorie, il n'en changera plus (de Landsheere, 1976, p37). Enfin, une durée d'évaluation prolongée peut entrainer chez l'évaluateur l'évaluation non pas de la performance de l'apprenant, mais de son caractère, qui ne doit pas entrer en compte dans une évaluation de performance (Charlin et al., 2003). L'évaluateur peut également donc constituer une menace la fiabilité des évaluations s'il existe un manque de consensus sur ce qui représente une performance satisfaisante entre les évaluateurs. Cela entraine des décisions variables et inéquitables et menace ainsi la qualité des évaluations, amenant les étudiants à se sentir comme des joueurs de « loterie » qui se présentent à un examen en se posant les questions suivantes « sur qui vais-je « tomber » aujourd'hui et que va-t-il dire ? » (Holmboe et al., 2011 ; Lurie et al., 2009). Ainsi l'évaluateur fait partie des biais inhérents à l'évaluation, mais le former et utiliser des outils fiables permet de réduire ces biais.

Une réflexion intéressante de Kogan élargit la problématique de l'évaluateur et de la reproductibilité des évaluations à l'analyse plus large du système de santé. En partant du constat que la variabilité des évaluations est due à des facteurs connus tels que les caractéristiques psychométriques d'un test, mais également aux évaluateurs eux-mêmes et au contexte du déroulement de l'évaluation, elle souligne le fait que le problème de notation ou de jugement évaluatif n'est pas seulement dû à des caractéristiques pédagogiques d'un test mais qu'il est très en lien avec des problématiques cliniques. Evaluer les étudiants de manière juste et valide

reviennent à former un système de santé dans lequel les patients seront en sécurité tout en recevant des soins de qualité. L'objectif d'une évaluation est donc la qualité future des soins et les évaluateurs doivent l'avoir en tête, de même que les concepteurs des évaluations. Ainsi, elle insiste sur la formation des évaluateurs qui ne devraient pas avoir pour objectif d'être satisfaits ou pas d'une performance, mais qui devraient pouvoir dire si la performance réalisée illustre la capacité à délivrer des soins efficaces, centrés sur le patient et qui ne seront pas dangereux (Kogan et al., 2014). Le fait de réaliser des évaluations sur le lieu de travail permet également d'impliquer les cliniciens dans ce défi que représente l'évaluation des compétences.

Dans notre travail, le fait d'utiliser des listes de contrôle a permis de faire des formations plus courtes, ce qui a permis également d'obtenir des outils reproductibles et d'envisager leur généralisabilité. En effet, si la formation nécessaire à l'utilisation de l'outil s'avère trop longue, elle peut être partiellement faite, voire pas du tout, ce qui qui peut fausser la fiabilité de l'outil et son utilisation à grande échelle. Cependant, les listes de contrôle, même si elles apportent une certaine fiabilité, présentent des désavantages, et notamment pour l'évaluation des compétences, que nous aborderons dans la partie suivante.

1.2.2 Cohérence interne

Dans le dernier volet de la recherche présentée dans ce travail, le calcul de la cohérence interne des scores, c'est-à-dire de la bonne corrélation entre les items du score pour mesurer la performance de l'apprenant, était plutôt bonne. Le résultat obtenu signifie que les items des scores ACAT sont liés entre eux et que la situation est bien représentée par les items. Si un apprenant maîtrise la situation alors les items auront globalement le même niveau d'appréciation de la performance. On aurait pu s'attendre à une cohérence interne moins importante, liée au choix d'évaluer à la fois des performances techniques et non-techniques. Cependant une des hypothèses expliquant ce résultat est la bonne corrélation entre performance technique et non technique, ainsi que l'ont mis en évidence des enseignants de chirurgie, établissant un lien, voire

une prédiction entre performance non technique et performance clinique lors d'une situation d'urgence vitale (Cha et al., 2019).

1.3 Impact pédagogique et effet catalytique

Les études mises en place ne sont pas directement intéressées à l'impact pédagogique ou à l'effet catalytique des scores ACAT, mais les observations des enseignants apportent quelques éléments de réponse pour discuter ces deux aspects importants dans une évaluation de performance.

L'évaluation par la simulation, en permettant un alignement avec des performances requises ultérieurement dans la pratique future, donne du sens à l'apprentissage et donne potentiellement un effet motivationnel à l'enseignement. La motivation mobilisée serait la motivation intrinsèque, mue par la valeur de la tâche à effectuer et dont on sait qu'elle est liée à un engagement plus en profondeur dans l'apprentissage (Pelaccia & Viau, 2017). Seulement quelques d'études se sont intéressées à l'impact de l'enseignement par simulation sur la motivation des apprenants, les chercheurs utilisant plus souvent comme critère d'impact de la simulation des critères de satisfaction, de confiance en soi ou d'efficacité sur les performances des apprenants. Les quelques auteurs qui se sont posé la question autour de la motivation en lien avec la simulation ont regardé si le fait d'avoir une motivation intrinsèque élevée était corrélée à de meilleures performances en simulation, ou bien si la simulation dans une discipline (la neurochirurgie) modifiait la motivation des étudiants en médecine à choisir cette spécialité pour leur pratique future. Les premiers ne retrouvaient pas de lien entre motivation et simulation et les seconds en revanche avaient pu établir que l'enseignement par simulation avait un impact sur leur motivation à choisir une carrière en neurochirurgie (Hanrahan et al., 2018; Schulte-Uentrop et al., 2020). Une des rares études qui s'intéressait à l'effet de la simulation sur la motivation des étudiants ne retrouvait pas de modification de la motivation des apprenants apres plusieurs sessions de

simulation (Moll-Khosrawi et al., 2021). Nous n'avons pas pu analyser cet effet spécifique de l'évaluation par la simulation, mais cela pourrait constituer une piste de recherche intéressante.

Quant à l'effet catalytique, qui constitue la rétroaction fournie par l'évaluation et guide le futur apprentissage, il semble être présent grâce à l'utilisation des scores par les évaluateurs au moment des débriefings. Parmi les évaluateurs interrogés, plus de 80% d'entre eux trouvaient dans les scores ACAT une aide pour le débriefing et il en était de même pour tous les participants aux focus groupes. Ces derniers ont pu détailler leur utilisation et l'inclure dans une démarche formative, permettant de souligner les axes d'amélioration grâce au score qui leur fournissait un cadre validé et perçu comme objectif. Ainsi, même si l'effet catalytique n'a pas été étudié directement chez les apprenants, il apparait pouvoir être mis en place par les évaluateurs, ce qui renforce la qualité des scores ACAT.

1.4 Faisabilité

La faisabilité s'intéressant au côté pratique, réaliste d'une évaluation, nous nous y intéresserons par le prisme de l'utilisabilité des scores, mesurée à la fois par le pourcentage d'items complétés dans les différentes études mais également grâce aux appréciations des enseignants.

Tous les items des trois scores ont été rempli à plus de 95%, un seul l'ayant été dans seulement 87% des cas. Il s'agissait de l'item « organiser le traitement étiologique complet » dans le score s'intéressant à la détresse respiratoire aigüe (Tableau 16). Ces résultats soulignent que les scores sont utilisables à la fois devant une vidéo de situation simulation, mais également dans une situation d'évaluation en direct, alors même que les évaluateurs doivent pouvoir gérer l'interaction avec les apprenants, la gestion technique du mannequin et l'observation en vue du débriefing. En effet, dans les études réalisées, un nombre minime d'éléments manquant vient mettre en jeu la validité des scores ACAT.

Du point de vue des enseignants, 67 et 83% des enseignants ayant répondu au questionnaire estiment que les scores sont faciles à utiliser. Ce chiffre est confirmé par les participants aux focus groupes qui, dans l'ensemble n'ont pas eu trop de difficultés. Certains enseignants ont souligné le fait que les scores contenaient trop d'items et que les descripteurs des items étaient parfois compliqués à maitriser, mais ces discussions n'ont pas fait l'objet d'un consensus et ne sont pas confirmées par les pourcentages de remplissage des items. Il est possible que le fait même de participer à une étude ait biaisé ces taux de remplissage, il faudrait pouvoir réanalyser les chiffres lors d'une évaluation « de routine » si les scores ACAT sont utilisés. En ce qui concerne le score de performance globale, les enseignants estiment qu'il est simple à utiliser, ne nécessitant pas de pénibilité particulière.

La mise en place des évaluations, non testée dans les études pourraient être sujet à difficultés car nécessitant de nombreuses ressources tant techniques qu'humaines (Hosny et al., 2017; Sawaya et al., 2021). Il s'agit d'un obstacle inhérent à l'enseignement par la simulation, qui se retrouve logiquement lorsqu'il est question d'une évaluation avec cet outil. Les enseignants des focus groupes ont naturellement abordé ce sujet, y voyant également un obstacle à l'acceptabilité non pas des scores eux-mêmes mais de l'évaluation par la simulation qui nécessiterait, ainsi que le soulignent les récentes recommandations de la Société Francophone de Simulation en Santé, trois types de formateurs pour mener à bien l'évaluation. Un formateur concepteur du programme de formation contenant un volet évaluatif, un formateur opérateur permettant la mise en place des sessions de simulation, avec l'aide du premier et enfin, un formateur évaluateur, possédant des compétences et une formation spécifique, dédiées à l'évaluation proprement dite. Les recommandations de bonne pratique appuient ainsi les remarques des évaluateurs quant à la nécessité de recourir à des ressources humaines supplémentaires, qui diminuent alors la faisabilité de l'évaluation par la simulation (SoFraSimS, 2021).

Ainsi, les scores ACAT semblent avoir certaines des qualités requises pour évaluation à enjeu élevé, avec une validité et une fiabilité toutes deux démontrées par les trois études entreprises, mais également avec des possibles enjeux pédagogiques qui restent à démontrer, et une certaine faisabilité, du point de vue de leur utilisabilité. L'impact pédagogique des évaluations n'a pas encore pu être démontré avec les méthodologies de recherche employées, de même que son objectivité. Enfin son acceptabilité, très discutée par les enseignants lors des focus groupe, au regard de la place de l'évaluation au sein de l'enseignement par simulation, sera discutée dans la quatrième partie de la discussion.

2. Place des scores ACAT dans le cursus de medecine d'urgence

L'objectif initial de la recherche était de développer des scores d'évaluation de trois situations cliniques en médecine d'urgence, afin d'essayer de valider les futurs internes au premier jour de l'internat, mais également les internes de médecine d'urgence au fur et à mesure de leur évolution dans le cursus. Nous allons discuter ici, de la possible utilisation des scores en analysant leur place et leur rôle à la lumière de leur finalité éventuelle, mais également en analysant les résultats en fonction du niveau des apprenants, en fonction de la place des scores par rapport aux autres échelles d'évaluation en médecine d'urgences, et plus globalement aux autres évaluations en place.

2.1 Finalité des évaluations avec les scores ACAT

2.1.1 Des scores ayant toute leur place au sein d'une évaluation formative

Plusieurs éléments, issus de la construction des scores (comme le choix d'une notation proche d'une liste de contrôle) et du retour des enseignants, permettent d'envisager l'utilisation des scores ACAT pour une finalité formative. En effet, les listes de contrôle permettaient à au moins 75% des évaluateurs de réaliser un feedback aux apprenants, dans un cadre validé, reproductible et qui pourrait être à nouveau évalué lors d'une prochaine mise en situation. Cette utilisation des

scores conviendrait donc à un cadre formatif, qui a pour objectif de faire progresser les apprenants avec un objectif de professionnalisation, grâce à des mises en situations régulières permettant la pratique réflexive sur les expériences réalisées dans le cadre de la simulation (Epstein, 2007; Fontaine & Loye, 2017; Schartel & Metro, 2010).

Dans ce cadre formatif, il pourrait être également intéressant de donner les scores d'évaluation aux apprenants, pour qu'ils puissent réaliser une auto-évaluation de leur performance et en dégager des axes de progression. Le processus de notation des scores le permet, par sa précision et sa représentation valide des trois situations évaluées (Ilgen et al., 2015).

Utilisés dans un cadre formatif, et ainsi avec la possibilité, pour les étudiants, de réaliser plusieurs entrainements dans chacune des trois situations d'urgence vitale, les scores pourraient permettre aux enseignants d'établir des courbes d'apprentissage (soit par domaine de compétence, soit en utilisant les résultats des scores dans leur totalité). L'intérêt d'utiliser ces courbes serait double : illustrer la progression de l'apprenant (et si les scores permettent de le faire, alors cela renforce l'idée qu'ils arrivent à identifier la progression d'un même apprenant, et qu'ils sont discriminant dans le temps), et permettre de déterminer un niveau à partir duquel l'apprenant peut agir en autonomie et ainsi se soumettre à une évaluation sommative, certifiante, lui permettant de valider un certains de compétences (Bok et al., 2018; Fahim et al., 2018). Cette dernière caractéristique permettrait de s'approcher du fonctionnement d'un système centré sur l'apprenant, au sein duquel il peut adapter son apprentissage et l'évaluation de ses compétences à son propre rythme de progression et de professionnalisation (Lucey et al., 2018).

2.1.2 Les éléments manquant pour une utilisation en évaluation sommative

L'objectif initial des scores ACAT était de pouvoir proposer une évaluation normative, qui ferait partie d'un ensemble d'examens à enjeux élevés. Certaines propriétés des scores permettraient de rendre cette utilisation possible, mais il reste néanmoins du travail de recherche à réaliser pour

permettre l'utilisation des scores ACAT dans ce but. Outre la réticence des enseignants sur laquelle nous reviendrons à la fin de cette discussion, de nombreuses questions se posent encore.

Le premier point soulevé si l'on s'intéresse à une évaluation sommative avec les scores ACAT est la définition du seuil de réussite, en fonction du niveau de l'apprenant. Malgré la consultation d'experts, réalisée selon un processus rigoureux et validé, il reste encore une étape de validation des seuils des experts avec une utilisation dans des situations calibrées. Il serait intéressant en effet d'appliquer les seuils dans une population ayant différents niveaux validés et de regarder avec quelle spécificité le test détecte les apprenants compétents des autres.

Par ailleurs, les résultats fournis par le score concernent une performance observée et ne sont pas nécessairement liés avec la compétence de l'apprenant, c'est-à-dire avec sa capacité potentielle à effectuer seul la prise en charge globale requise. Il reste encore un travail qui consisterait à lier le résultat obtenu au score ACAT, avec celui obtenu au score de performance globale et avec la mise en autonomie de l'apprenant. Il s'agira donc de « définir la compétence », à partir de critères de jugements basés à la fois sur un seuil, mais également sur le score de performance globale et dans des contextes multiples. En effet, à l'image des OSCE, il ne serait pas valide d'évaluer les apprenants sur une performance isolée, dans un seul contexte. La fidélité d'une telle évaluation serait très faible puisque ne permettant pas d'apprécier la capacité de l'apprenant à mobiliser ses ressources dans des contextes variés. Dans le cadre des OSCE, il est établi qu'il faut entre 10 et 20 stations pour obtenir des résultats représentatifs de la compétence des apprenants (Epstein, 2007; Khan et al., 2013).

Le besoin de multiplier les situations évaluatives pour utiliser et poursuivre le processus de validation des scores ACAT permet également d'aborder la nécessité d'utiliser les scores dans le cadre d'un programme d'évaluation, avec plusieurs évaluations, de nature différente, qui permettent d'évaluer les différentes composantes des compétences. En effet, un instrument de

mesure idéale n'existant pas, une évaluation unique compromettrait les critères de qualité de l'évaluation, même s'ils sont tous réunis (Van Der Vleuten, 1996). Afin de déterminer la place des scores ACAT dans une évaluation sommative au sein du parcours de formation en médecine d'urgence, il apparait donc essentiel de mener une réflexion incluant les différentes évaluations nécessaires pour former un réel programme évaluatif, dont l'un des intérêts est de combiner différentes évaluations pour atténuer les compromis sur les évaluations seules, rendant ainsi le résultat total obtenu aux évaluations, supérieur à la somme des parties (van der Vleuten et al., 2012). Par exemple, il serait essentiel d'ajouter à l'évaluation des situations d'urgence vitale une évaluation axée sur les procédures nécessaires à leur gestion (intubation orotrachéale, pose d'une voie intra-osseuse, réalisation d'une ponction lombaire) pour lesquelles des scores validés existent déjà, qu'il soient spécifiques au geste ou à l'évaluation d'une procédure en général (Bodle et al., 2008; Gerard et al., 2013; Oriot et al., 2012). Mais il faudrait aussi pouvoir utiliser des outils d'analyse qualitative de la compétence, tels que les portfolios, les études de dossier et bien sûr les évaluations sur le lieu de travail.

Ainsi, certains éléments manquent encore pour utiliser les scores ACAT dans une évaluation dite à enjeu élevé, et le processus de validation se doit d'être poursuivi à la fois pour valider un « niveau de performance » en lien avec les niveaux de compétence nécessaires en fonction des différentes étapes à franchir, mais également pour intégrer les scores ACAT au sein d'une démarche évaluative intégrée au programme de formation des futurs internes et des futurs médecins urgentistes.

2.2 Public visé

La question ici posée est à la fois celle des apprenants à qui s'adresse le score et également celle de la capacité qu'ont les scores ACAT à discriminer les étudiants selon leur niveau de performance.

Initialement, les scores avaient pour vocation d'évaluer les externes à la fin du deuxième cycle des études médicales, afin de s'assurer de leur capacité à prendre en charge les premières minutes d'une situation d'urgence vitale, qu'ils seraient amenés à rencontrer dans leurs différents terrains de stage, toute spécialité confondue. Mais les scores avaient également l'objectif d'évaluer des internes de médecine d'urgence, tout au long de leur spécialisation, dans une idée de continuité entre les différents apprentissages (Englander & Carraccio, 2018). Après mise en pratique des scores, plusieurs observations, fondées à la fois sur les remarques des évaluateurs et sur les résultats de fixation des seuils de réussite, se dégagent.

2.2.1. ACAT et niveau des apprenants : la validité de construit des scores

N'ayant pas pu étudier la capacité des scores à discriminer les externes des internes, pour des raisons d'effectif insuffisant, il nous a cependant été possible de regarder si les scores identifiaient des internes de première année de DESMU en comparaison avec des internes de troisième année des DESMU et avec des internes d'autre spécialité (principalement médecine générale et gériatrie, qui réalisent des stages aux urgences).

En comparant les moyennes de trois groupes dont les niveaux sont supposés être différents, il apparaissait que les scores ACAT 1 et ACAT 3 permettaient de différencier les apprenants selon leur niveau d'évolution dans la formation ou selon leur spécialité. Cette différence n'était pas retrouvée pour le score ACAT 2. En effet, même si une différence semble exister entre les internes d'autre spécialité et les internes des de DESMU, avec une tendance pour le score à augmenter entre ces deux groupes, il n'y avait pas de différence entre les groupes d'internes de DESMU. Puisque nous n'avions pas créé de groupes homogènes artificiellement, mais uniquement intégré les internes en stage pendant l'étude, plusieurs paramètres peuvent expliquer ce résultat. Parmi les internes évalués sur une prise en charge d'un coma, ceux de troisième année avaient peut-être un niveau plus faible que le niveau attendu, et de manière plus importante que pour les autres situations de prise en charge.

Une autre hypothèse est en lien avec la formation préalable à la situation simulée utilisée pour l'évaluation. Nous n'avons pas recueilli cet élément parmi les apprenants, mais il est possible qu'une différence d'entrainement entre les participants à l'étude entraine une différence de performance en simulation. Or, les formations par la simulation en médecine d'urgence en France sont hétérogènes entre les différentes universités et services hospitalo-universitaires, ainsi qu'une enquête récente, menée par la SoFraSimS non encore publiée, l'a souligné. Par ailleurs, il n'existe pas encore de programme « officiel » de formation par la simulation en médecine d'urgence, que ce soir pour les apprenants de deuxième ou de troisième cycle. Cette situation a été retrouvée par des auteurs qui effectuaient le même processus de validation et avait pu mettre en évidence une différence importante de formation préalable à l'évaluation pour expliquer des différences entre les groupes et l'impossibilité de mettre en évidence la faculté discriminatoire du score évalué (Hall et al., 2012). Cette hypothèse questionne la place de l'entrainement par simulation avant l'évaluation par simulation, dans un souci d'alignement pédagogique, mais également de connaissance de l'outil d'évaluation, qui va guider la préparation de l'apprenant (Cilliers et al., 2010; SoFraSimS, 2021, p 21) . Cependant cela pose la question de la place de la simulation, puisqu'elle n'est pas un but en soi pour l'évaluation et la formation, mais un outil permettant la mise en situation semi-authentique pour évaluer l'apprenant avant sa mise en autonomie dans le système de soins. Un des écueils de l'évaluation par la simulation serait en effet de se préparer à soigner des mannequins et non plus des patients.

2.2.2. Utilisation des scores dans le cursus de médecine d'urgence

La formation en médecine d'urgence concerne à la fois les futurs internes qui doivent être préparés aux gestes d'urgences fondamentaux et les futurs urgentistes qui doivent apprendre leur futur métier. Combiner ces deux objectifs dans un seul score avait pour ambition de ne pas multiplier les outils d'évaluation, de respecter une certaine continuité dans les apprentissages, et de définir, à partir des mêmes outils, des niveaux progressifs de performance dans trois situations

auxquelles les apprenants sont peu confrontés pendant leur cursus mais qu'ils doivent savoir prendre en charge. Actuellement, plusieurs auteurs ont mis en évidence un manque de maîtrise quant aux situations d'urgence vitale, tant chez les futurs internes, que les chez les internes eux-mêmes (Holmboe et al., 2011; McEvoy et al., 2014; Tofil et al., 2014) mettant en évidence la nécessité d'une formation et d'une certification pour ces habiletés indispensables à tout futur interne puis urgentiste.

Plus de la moitié des enseignants interrogés via le questionnaire affirment vouloir utiliser les scores ACAT pour évaluer des externes. Ce résultat est minimisé par les entretiens, au cours desquels la majorité des enseignants estimait que ces scores étaient trop sévères, d'un niveau de difficulté dépassant ce qu'on peut attendre d'un externe en fin de second cycle. Ces résultats sont confortés par les seuils retrouvés lors de la consultation d'experts, qui sont inférieurs à la moyenne. Ce peut être une utilisation possible des scores, qui demanderai aux apprenants d'atteindre une performance de « 40% » pour être validé. Cela signifierait potentiellement que 60% des items évalués ne sont pas requis pour atteindre la performance requise et que donc le contenu du score n'est pas adapté à la performance souhaitée, même si le score reflète les éléments de la situation à évaluer. Une des solutions à cette observation serait d'ajouter aux items une qualification d'action essentielle ou non, qui représenterait la condition indispensable à la certification et aurait un lien avec les actions à engager pour prendre en charge les premières minutes d'une situation d'urgence vitale. Afin de déterminer quels items peuvent être qualifiés d' « action essentielle », il serait possible de se baser sur une consultation d'experts associée à une revue de la littérature qui s'intéresserait aux facteurs prédictifs d'une amélioration de la situation d'urgence vitale (pour chacun des trois scores).

La plupart des enseignants reconnaissaient par contre une utilité au score pour des internes, soit en formatif au début des stages afin de les aider à enseigner des compétences essentielles et de suivre la progression des étudiants, mais également car ils avaient pu utiliser les scores lors d'une

mise en situation interprofessionnelle qui leur permettait de travailler la performance de l'interne, mais également celle de l'équipe. Pour les enseignants l'utilisation des scores était pertinente dans cette situation, moins souvent rencontrée en formation initiale, qui est, le plus souvent, uniquement centrée sur une profession, ce qui supprime une grande part de réalisme aux situations enseignées puis évaluées. Cela pose la question de la pertinence de l'évaluation dans des conditions trop éloignées des situations professionnelles, ce qui est le plus fréquent. Les scores ACAT pourraient alors permettre, puisqu'ils ont été étudiés dans ce contexte interprofessionnel une évaluation en contexte plus authentique que dans un centre de simulation, loin d'une équipe « réelle » des urgences.

Ainsi, les scores ACAT pourraient trouver leur place pour évaluer des internes de médecine d'urgence, voire des externes si les conséquences du test sont adaptées ultérieurement au niveau de performance attendu. Par ailleurs, leur utilisabilité en interprofessionnel permettrait de mettre en place des évaluation in situ ou au centre de simulation, en équipe complète, apportant ainsi une validité écologique supplémentaire aux scores. De plus, cela permettrait d'inclure les scores ACAT dans une progression de l'évaluation, qui serait individuelle, puis individuelle en équipe et enfin à la fois individuelle et d'équipe, avec d'autres outils validés pour cette dernière étape, tels que le score TEAM (Maignan et al., 2016). Cette façon de procéder pourrait permettre d'intégrer progressivement les compétences nécessaires au travail d'équipe, en situation d'urgence vitale, la performance de l'équipe étant un élément indispensable à une prise en charge des patients optimale, et en toute sécurité (Marriage & Kinnear, 2016). Toutefois, leur utilisation ne serait pertinente qu'au sein d'un programme d'évaluation visant à intégrer l'évaluation des tous les domaines de compétences requis pour gérer les situations d'urgence vitale, grâce à des outils variés s'intéressant à toutes les facettes nécessaires à la mise en autonomie des internes.

3. SCORES ACAT ET APPROCHE PAR COMPETENCES

Une des questions de ce travail était de réfléchir à la position que pourraient avoir les scores d'évaluation développés pour une potentielle utilisation dans une approche par compétence. Après avoir discuté la place que pourrait avoir les scores ACAT dans le système actuel qui n'est pas encore ancré totalement dans une approche par compétence, nous allons aborder ici cette question puisque le système tend à adopter une partie de l'approche, actuellement pour le deuxième cycle des études de médecine, mais également au sein du troisième cycle qui a mis en place une année d'autonomisation des internes pour toutes les spécialités.

L'évaluation au sein d'une approche par compétence requiert différents éléments indispensables : une évaluation régulière, basée sur une continuité des éléments évalués, qui n'est plus standardisée mais personnalisée, axée sur le développement continu de l'apprenant, à son rythme, une évaluation de situations complexes et authentiques, et qui nécessite de la part de l'évaluateur une certaine culture qui lui permette d'émettre un jugement professionnel, au sein d'un programme évaluatif (Holmboe et al., 2010; Schuwirth & Van der Vleuten, 2011).

3.1 De la performance à la compétence, qu'évaluent les scores ACAT ?

La problématique induite par la structure des scores ACAT est celle du morcellement des compétences, pratique alors en vigueur dans les années 80-90, en lien avec une vision behaviouriste de l'apprentissage et contre laquelle s'inscrit l'approche par compétence (Hoang & Lau, 2018). Le fait de détailler les situations complexes à évaluer s'inscrit dans le cadre de la standardisation des évaluations, avec pour objectif de lutte contre une subjectivité qui serait nocive pour l'équité des examens (Hodges, 2013). Cependant, le morcellement des compétences, laissant imaginer que la somme des parties est égale au tout, entraine nécessairement une perte d'authenticité et de validité des outils d'évaluation si l'on s'inscrit dans

l'analyse de la compétence (Ginsburg et al., 2010). Par exemple, dans le cadre des ECOS, il est démontré que l'analyse du raisonnement clinique, n'étant pas réalisé avec les mêmes stratégies par un novice et par un expert, ne peut pas être authentiquement évalué pur ces derniers avec les listes de contrôle des ECOS qui divisent les éléments menant aux hypothèses diagnostiques (Charlin et al., 2003; Schuwirth & van der Vleuten, 2011). Dauphinee, en 1995, explique très bien ce phénomène en insistant sur l'importance de la confusion entre objectivité à tout prix et processus d'objectivation. L'objectivité à tout prix, représentée par des critères principalement psychométriques (tels que la reproductibilité, la cohérence) ne doit pas masquer les influences subjectives qui ont conduit à la réalisation d'un examen, tels que le choix des sujets, des contenus, des évaluateurs. Au contraire, il est préférable d'avoir en tête les différentes étapes de construction de l'évaluation, avec les influences subjectives de chacune étape plutôt que de se réfugier dans un test reproductible mais qui représente mal le domaine qu'il cherche à évaluer, car il le morcelle trop par exemple (Van der Vleuten et al., 1991). Ainsi, des tests en apparence « très » objectifs ne sont pas nécessairement plus fiables que des tests qui semblent « subjectifs » (Dauphinee, 1995).

Même si l'approche par compétence consiste en une vision holistique des habiletés, ressources, techniques, nécessaires au devenir d'un bon professionnel, elle n'échappe pas au morcellement de ces compétences et c'est ce qui a pu poser problème dans son implémentation, si elle est réalisée dans un but de standardisation et d'objectivité (Hoang & Lau, 2018). Le problème n'est donc pas l'approche, mais la manière de s'en saisir et de l'implémenter, qui entre en confrontation avec des visions de l'évaluation, tant chez les enseignants que chez les apprenants, ainsi que le souligne Demeester (sondage BVA-Presse régionale – Foncia, 2017, in Demeester, 2020, p 195).

La problématique soulevée par la structure des scores ACAT est donc celle de l'extrapolation de la performance vers la compétence, d'autant qu'ils n'intègrent pas le débriefing, qui permet d'analyser les processus réflexifs et cognitifs des apprenants, au résultat final, alors qu'il aurait

pu être une aide supplémentaire pour un accès à la compétence plus globale de l'apprenant. La performance, marqueur indirect de la compétence, peut être extrapolée à cette dernière si la validité des outils, notamment par leur capacité à évaluer les apprenants dans différents contextes et par leurs caractéristiques psychométriques, mais la compétence reste toujours morcelée (Andreatta & Gruppen, 2009; Downing, 2003). A ce stade de leur développement, il est possible de dire que les score ACAT peuvent évaluer la performance de l'apprenant, et que leur validité et fiabilité pourrait être un argument de compétence, mais qui reste à démontrer en déterminant un lien entre les décisions qu'ils permettent de prendre et la compétence de l'apprenant dans d'autres domaines, évaluées par d'autres examens. Utiliser alors le cadre de Kane, qui décrit plusieurs étapes pour inférer la compétence à partir des observations, pour en arriver à une possible généralisation, puis à une extrapolation des résultats pour aider à une prise de décision valide, pourrait aider à renforcer l'hypothèse d'évaluation de compétences à travers l'observation de la performance (Cook et al., 2015; M. T. Kane, 2013). Une des méthodes pour utiliser ce cadre, est d'utiliser la théorie de la généralisabilité, qui permet d'analyser la variance de tous les éléments d'un test : apprenants, scénarios, enseignants. Un test qui présente une variance en fonction du seul niveau des apprenants apparait alors comme valide et fiable (Tavares et al., 2013).

Le lien entre performance et compétence clinique, peut également être apprécié avec des critères de jugement plus en rapport avec la clinique, et avec l'objectif réel d'un programme de formation : l'aide à devenir un professionnel autonome (Carraccio & Englander, 2013; Touchie & Ten Cate, 2016). La difficulté réside également dans l'impossibilité de définir concrètement et avec exactitude une compétence, car des éléments échappent toujours à cette définition (Holmboe et al., 2017). Hoang rappelle en effet que toutes les dimensions du métier de médecin n'entrent pas dans les domaines de compétence, définis par les différents organismes référents en pédagogie médicale. Ainsi, il nomme les « non-competency domains » plusieurs domaines d'importance

dans l'exercice de la médecine, et qui ne sont pas intégrés aux six à huit domaines de compétences habituellement utilisés. Il s'agit de la sécurité du patient, de la compétence culturelle, ou de la compétence de gestion (stewardship) (Hawkins et al., 2015; Hoang & Lau, 2018).

Une des solutions serait d'utiliser les « milestones » et les « activités professionnelles confiables » qui s'appuient sur une analyse de la progression de l'apprenant, dans une analyse plus qualitative des actions menées et en lien avec un critère de jugement clinique pertinent qui es l'autonomie de l'apprenant (Carraccio & Englander, 2013; Touchie & Ten Cate, 2016). Ainsi, pour une utilisation des scores ACAT qui permettrait une extrapolation vers la compétence, il pourrait être intéressant de leur adjoindre, à la fois une prise en compte du débriefing et un lien avec des activités professionnelles confiables ou, tout du moins, avec la notion d'autonomisation de l'apprenant. Une des questions à laquelle les évaluateurs pourraient avoir à répondre suite à l'observation de la performance et suite au débriefing serait : « Selon ses compétences antérieures et selon les données collectées, quel est le risque que cet apprenant soit très en dessous des standards requis dans le futur, pour une même situation ?[27] » (Schuwirth & van der Vleuten, 2006).

3.2 Notes, jugement professionnel et subjectivité d'une évaluation de compétences

Dans le travail que nous avions mené en 2017, un des reproches qui a été fait à l'évaluation par la simulation, tant par les enseignants que par les étudiants, était l'absence d'équité entre les épreuves : conditions différentes, cas cliniques de difficulté variable et enseignants qui changent, ce qui a provoqué en partie la réalisation de ce travail (Philippon et al., 2017; 2021). De même, dans ce travail, cette constatation est retrouvée par les enseignants, qui ne pensent pas pouvoir s'affranchir des notes dans une évaluation certificative, tout en estimant que la note ne leur est

[27] "How big is the risk of this student performing seriously below the standard in a future case, given his or her history and the newly collected information?" (traduction libre)

d'aucune utilité, en simulation, pour qualifier la performance et encore moins la compétence. Pour les enseignants des focus groupes, la note est un outil de mesure, qui répond à des exigences institutionnelles, mais qui n'est pas adapté à l'outil d'enseignement, ce qui peut être néfaste pour l'outil « simulation » en lui-même, alors qu'on sait qu'il participe à un enseignement de qualité.

Ces remarques introduisent la notion de la nature du jugement donné après une évaluation, de même que celle de subjectivité qui sont, pour certains auteurs, indissociable de l'évaluation des compétences, en situation complexe. En effet, une des autres façons d'évaluer la compétence est de tenir compte de la part de subjectivité de la compétence, ainsi que de la part de subjectivité existant dans toute évaluation.

La part de la subjectivité inhérente à la compétence tient au fait qu'elle ne s'exprime qu'en situation et donc à travers la variation des situations. Une fois ces éléments acceptés par les évaluateurs, cela peut les aider à diminuer la connotation négative de la subjectivité, en reconnaissant qu'elle est indissociable de la compétence (Hoang & Lau, 2018; Hodges, 2013).

Scallon, en affirmant que l'évaluation ne peut pas se réduire à une « activité » mécanique » introduit la notion de jugement dans l'évaluation (Scallon, 2015, p 18). Il s'agit selon lui, d'une étape nécessaire de l'évaluation qui s'apparente à un jugement professionnel et certainement pas à un jugement de valeurs. De plus, il ne s'agit pas d'un jugement vague, mais d'une évaluation basée sur des faits, observés, qui dépendent néanmoins d'un contexte et de l'appréciation d'un regard professionnel et pas personnel. Le risque de cette manière d'évaluer est de tomber dans le jugement personnel, de valeur, ou de l'étudiant, ainsi que le souligne Gérard. Cependant, une fois ce risque identifié, il est possible et même nécessaire de former les enseignants à cette nouvelle pratique, qui s'apparente à une nouvelle culture (Gérard, 2013).

Il faut cependant être vigilant avec la notion de subjectivité qui, comme le souligne Gérard, ne doit en aucun cas être synonyme d'arbitraire. Elle intervient, au contraire, entre le « rêve inaccessible d'objectivité » et le « refus de l'arbitraire » dans les évaluations (Gérard, 2002). En

réexploitant un article de Cardinet, datant de 1992, Gérard souligne en effet que l'objectivité est nécessaire, car elle permet à l'étudiant évalué de percevoir sa « vraie valeur », elle est souhaitée afin de ne pas tomber dans le biais du jugement de la personne plutôt que de sa compétence, mais elle est impossible pour de nombreuses raisons. Les enseignants sont différents, les contextes aussi (Gérard, 2002, p 2). Elle est en fait, selon Gérard, intrinsèquement composée d'éléments subjectifs, qui ne doivent cependant pas être flous ou arbitraires.

Dans notre situation, la subjectivité résiderait dans le choix des situations évaluatives (leur nombre, leur niveau de difficulté) et dans la manière dont les enseignants vont appliquer strictement les critères descriptifs des items, malgré une liste de contrôle détaillée, qui n'est pas toujours suivie « à la lettre » par les évaluateurs, en fonction d'impressions subjectives qu'ils peuvent avoir. C'est à ce moment qu'intervient le jugement professionnel, dans la dernière étape de validation ou non des étudiants.

Gérard nomme ce processus « l'examen entre l'adéquation entre critères et indicateurs, ou la question du sens ». Il explique qu'il peut expliquer un écart entre les normes attendues (mettre l'oxygène dans les deux minutes) et le jugement fait par l'évaluateur qui, pourra estimer que les autres éléments en présence. Par exemple, les étudiants délivrent l'oxygène au bout de trois minutes (décalage entre l'indicateur souhaité et la réalité), mais par ailleurs, ils constatent une bonne communication des étudiants, qui ont accompli d'autres gestes, dans un ensemble global de bonne réalisation du scenario. Ils peuvent ainsi décider de « rattraper » cet écart à la norme. L'évaluateur constate un « écart entre la norme et la réalité, mais il a donné du sens à tous les éléments en présence » (Gérard, 2002, p 8).

Enfin, Gérard explique qu'il serait vain de vouloir éviter la subjectivité car elle est présente dans toutes les étapes d'une évaluation : dans le choix des objectifs (qui doivent être néanmoins pertinents), dans le choix des critères d'évaluation et dans celui des informations à recueillir ou pas (par exemple, se présenter au patient est un critère qui, n'était pas présent dans toutes les échelles d'évaluation, selon la présentation clinique de ce dernier), mais également dans le choix

de la stratégie de recueil (ici, deux enseignants qui observent, mais permettent également au scénario de se dérouler sans encombre) et enfin, dans la dernière étape, celle qui donne du sens et permet aux évaluateurs de confronter les informations recueillies aux critères exigés (Gérard, 2002).

Ainsi, plutôt que de vouloir l'éviter, au risque de créer des évaluations floues sous couvert d'objectivité, avoir conscience de la présence de la subjectivité dans le processus d'évaluation paraît majeur. Il ne s'agit pas de mettre en place un système injuste, peu valide ou fiable, mais de tenir compte de la complexité de la situation, des éléments subjectifs qui la composent pour pouvoir l'évaluer au mieux. Scallon insiste sur la nécessité de former et de sensibiliser les enseignants, dont ce n'est pas la culture, afin qu'ils puissent adopter une « démarche emprunte de subjectivité » mais qui reste professionnelle et basée sur des critères précis (Scallon, 2015, p 28). Il s'agit ici d'encadrer et d'objectiver la subjectivité (Romainville, 2011).

A propos de l'exercice d'évaluation étudié, la notion de subjectivité apporte deux réflexions. La première est qu'effectivement les critères et indicateurs d'évaluation devraient pouvoir être validés, afin de rendre l'évaluation plus valide et fiable et ainsi équitable et non pas arbitraire. La deuxième est qu'il apparait important, si l'évaluation de situations complexes veut progresser et se développer, de modifier la culture enseignante et étudiante. A ce sujet, la mise en place d'une évaluation formative serait peut-être une étape indispensable à ce changement de vision.

L'autre solution, déjà exposée au chapitre précédent est d'intégrer des épreuves qualitatives pour évaluer les compétences, au sein d'une approche méthodologique mixte d'évaluation dans un programme global d'évaluation des compétences (Jick, 1979).

Ainsi, les questions soulevées par l'objectivité ou la subjectivité d'une évaluation pose la question de la culture évaluative présente dans les institutions, qui s'appuient le plus souvent sur des critères uniquement objectifs pour certifier les apprenants. Or, nous l'avons vu, cela est un danger

pour le développement de l'approche par compétence, qui a déjà été compromis pour cette raison dans les années 80.

La nécessité d'objectivité dans l'évaluation peut être attachée à cette culture, partagée par les enseignants, et par certains apprenants (comme cela était le cas avec les étudiants en médecine de la première recherche). Les institutions et enseignants, s'il veulent pouvoir utiliser l'approche par compétence dans son entièreté, doivent pouvoir en accepter toutes les facettes (Nousiainen et al., 2017). Cela passe en partie par un soutien aux enseignants, par des formations sur leur place au sein d'une évaluation de compétences et notamment sur la nécessité de reconnaître que l'évaluation est un jugement, que leur jugement professionnel y a sa place. Cela passe par la reconnaissance de l'évaluateur comme une source fiable d'informations, mais également comme source d'erreurs, rectifiées par la possibilité de multiplier les regards évaluatifs : par l'évaluation des connaissances, par la supervision directe en stage et les mises en situations variées et non pas par une évaluation réalisée par deux évaluateurs pour une situation (Gauthier et al., 2018; Khan et al., 2013; Schuwirth & van der Vleuten, 2020). Il s'agit également de connaître les limites des évaluateurs plutôt que d'imaginer qu'elles sont annulées par une évaluation vraisemblablement objective (SoFraSimS, 2021).

Dans cette perspective, l'évaluation par la simulation a un rôle à jouer puisqu'elle permet une certaine reproductibilité et standardisation, tout en laissant la place à la variation des contextes, à des échelles de notation globales, et en établissant un lien avec des critères de jugement tels que l'autonomie et la situation de travail (Epstein, 2007; Khan et al., 2013).

4. Place de l'evaluation sommative au sein d'un enseignement par simulation

Une des questions de ce travail était de réfléchir à la position que pourraient avoir les scores d'évaluation développés pour une potentielle utilisation dans une approche par compétences et au sein du cursus de médecine d'urgence, mais également d'aborder la place que peut avoir

l'évaluation au sein d'un programme d'enseignement par simulation. Les réactions des enseignants interrogés seront un point de départ à cette réflexion qui nous permettra de conclure cette discussion en abordant l'évaluation par la simulation à travers des questions d'éthique pour la formation, mais également pour les patients.

4.1 Ethique de la simulation

L'appréciation mitigée de la part des enseignants quant à la simulation évaluatrice remet en question l'intérêt de la développer pour les étudiants qui sont débutants. Les enseignants ont notamment regretté la disparition de la bienveillance, qui serait une condition selon eux indispensable à l'utilisation de la simulation et qui disparaitrait avec l'utilisation des scores ACAT en cas de mauvaise note ou de jugement estimé trop « sévère » par les enseignants. Au travers des principes éthiques de la simulation nous allons tenter d'éclairer leurs réticences et voir ce que les principes d'éthiques et de pédagogie en simulation peuvent apporter à la réflexion.

La simulation est souvent considérée comme éthique, car elle est jugée bénéfique et utile du point de vue de l'apprentissage, considéré meilleur et efficient grâce à cette technique. Parmi les règles de bonnes pratiques définies, l'évaluation formative fait partie des outils à disposition des enseignants, aidés « d'outils d'aide à la progression ». Quant à l'évaluation normative ou sommative, les recommandations précisent ceci : « Indépendamment de l'évaluation formative, l'évaluation peut être sommative dans le cadre de la formation initiale, de la (re)certification des professionnels de santé, sous réserve de disposer d'outils docimologiques validés. » (HAS, 2012).

La simulation semble être actuellement un outil indispensable dans les formations de santé, mais la question qui se pose ici est « peut-on tout faire avec la simulation ? » (Collange & McKenna, 2013) et notamment, peut-on évaluer de jeunes étudiants par le biais de la simulation, tout en respectant les principes éthiques de formation ? Dans le domaine de la bioéthique, quatre principes définissent une action qui serait éthique : la bienfaisance, la non malfaisance,

l'autonomie et la justice. Comme l'ont fait Homerin et Roumanet pour une situation d'évaluation de soins infirmiers par la simulation, nous nous posons également la question de savoir si les principes de bienfaisance et de non malfaisance, remis en cause par les enseignants peuvent être respectés et appliqués malgré la situation d'évaluation (Homerin & Roumanet, 2014).

Dans la situation étudiée, d'après les enseignants, le principe de bienfaisance pourrait ne pas être respecté si la rétroaction fournie par la note à l'évaluation s'avérait décevoir l'étudiant. En revanche, les principes de la simulation (formation en petit groupe, avec briefing et débriefing) étant conservés, les « garde-fous » de la pédagogie fondée sur la simulation sont respectés et limitent cette impression de « malveillance ».

Le principe de non malfaisance n'est pas respecté si l'on considère que la rétroaction fournie par les notes ne semble pas adéquate à l'enseignement par la simulation. Par ailleurs, il faudrait analyser le stress engendré par une telle évaluation, en le comparant à la fois au stress perçu pendant une évaluation facultaire, çà laquelle les apprenants sont régulièrement soumis et en le comparant au stress engendré par une séance de simulation, dont on sait qu'il est présent. (Bong et al., 2010). Cependant, comme le précisent les recommandations de bonne pratique, l'évaluation devrait se faire avec des outils validés, ce qui est le cas de notre démarche (HAS, 2012) .

La remarque des enseignants soulève la question de la double posture qu'ils doivent prendre en simulation, qu'elle soit formative ou évaluative certifiante. Cette posture est souvent ambivalente entre bienveillance (pour créer un environnement d'apprentissage rassurant) et le besoin de rétroaction, basée sur un jugement et donc une évaluation de l'action mais qui a pour finalité la progression. Il s'agit donc pour les formateurs d'accompagner pour faire progresser (position qui est la plus souvent survalorisée, mise au premier plan, et en accord avec ce que nous retrouvons dans notre recherche), mais également de contrôler les apprentissages, ce qui est vu comme plus coercitif par les enseignants (Houzé-Cerfon et al., 2019).

Ainsi la perte de la bienveillance parait problématique aux yeux des enseignants, mais elle peut être contrôlée par le maintien des bonnes pratiques de la simulation, la prise de conscience par les enseignants d'une posture double et ce, même dans l'évaluation formative, et atténuée par l'utilisation d'outils validés, qui font sens pour les enseignants et pour les apprenants.

Un autre point majeur est celui de la place de l'évaluation par la simulation au sein du système d'enseignement et dans une éthique de sécurité des soins pour les patients. En effet, si l'on considère que le leitmotiv, devenu central dans la formation des professionnels de santé « Jamais la première fois sur le patient », s'applique à la formation, on peut également se poser la question de son application dans la certification, dans le sens où cette dernière permettrait d'éviter de réaliser des prises en charge médicales en autonomie complète avant d'avoir été certifié, en partie par la simulation. L'évaluation par la simulation, trouvant ici sa place entre évaluation de connaissances facultaire, le plus souvent utilisée et évaluation au lit du patient, une fois que l'étudiant peut agir en autonomie supervisée puis en autonomie seule. Cela permettrait ainsi d'améliorer la sécurité des patients et aurait, à ce titre, une portée sociétale (Amalberti et al., 2005; Collange & McKenna, 2013; Philippon et al., 2021).

4.2 Des atouts pour évaluer

Finalement, une fois les questions éthiques travaillées et abordées en respectant les règles de bonne pratique, il apparait que l'évaluation par la simulation a des atouts non négligeables pour participer à un programme d'évaluation sommative ou certificative. Comme nous l'avons souligné plus haut, elle permet d'évaluer les compétences, à partir de l'observation des performances, en évaluant la capacité de mobilisation des ressources d'un apprenant, dans des situations complexes, semi-authentiques, mais avec la possibilité d'une validité écologique élevée. (ref).

Par ailleurs son utilisation de la pratique réflexive grâce au débriefing doit pouvoir être conservée et mise en avant comme un avantage de cet outil d'évaluation. En effet, parmi les bonnes pratiques pédagogiques en simulation, le débriefing arrive en tête des étapes à pratiquer

systématiquement (Boet et al., 2013; Motola,et al, 2013) puisqu'il permet aux enseignants et aux apprenants de s'assurer que tous les objectifs pédagogiques planifiés ont été abordés, et qu'il complète l'apport de l'expérience, qui, sans débriefing n'est pas suffisante pour un apprentissage de qualité (Issenberg et al., 2005; Kolb, 1984)

De plus, le débriefing permet de limiter certains risque d'une mise en situation simulée, dans le but d'assurer une « sécurité psychologique » aux apprenants, déjà soumis à un stress important (Bong et al., 2010). En se basant sur un contenant cadré, délimité, le débriefing permet aux apprenants de revenir sur leurs émotions, d'aborder leurs difficultés, dans un contexte de bienveillance et de regard positif (Kolbe & Grande, 2015), ce qui limite l'aspect de perte de bienveillance.

Contrairement aux autres évaluations certificatives habituellement pratiquées, la simulation apporte donc une rétroaction élevée, qui lui donne un atout supplémentaire si l'on regarde ses critères de qualité. Ce peut être un handicap et une barrière puisque cela nécessite du temps supplémentaire, sauf si le débriefing est réalisé sous la forme d'une auto-évaluation cadrée sur la forme d'un débriefing actuel ou bien sous la forme de rétroaction plus basique mais reprenant les points essentiels abordés selon la méthode « +/delta+ » les points acquis, appelés « + » et les points à retravailler, appelés « delta+ » (Motola et al., 2013). Par ailleurs, il a été mis en évidence que lors de débriefing d'équipe, les apprenants progressaient de la même manière selon qu'ils avaient un débriefing « classique » avec évaluateur ou un débriefing cadré sans évaluateur (Houzé-Cerfon et al., 2020).

Ainsi la perte de la bienveillance parait problématique aux yeux des enseignants, mais elle peut être contrôlée par le maintien des bonnes pratiques de la simulation, la prise de conscience par les enseignants d'une posture double et ce, même dans l'évaluation formative, et atténuée par l'utilisation d'outils validés, qui font sens pour les enseignants et pour les apprenants. De plus la simulation offre la possibilité de réaliser une rétroaction sous forme de débriefing, qui assure à la

fois la sécurité psychologique des apprenants et le transfert des apprentissages. L'évaluation prend alors la forme d'une évaluation qui « juge » mais qui participe également à la « valorisation de ceux qu'elle porte » et elle respecte alors deux « exigences fondamentales pour la pratique évaluative : être méthodologiquement fiable et éthiquement défendable » (Demeester, 2020, p 5-10) Ainsi, la plupart des référentiels de formation et d'enseignement par la simulation conseille l'intégration totale de la simulation dans les curricula, et de ce fait, d'intégrer des sessions d'évaluation, tout en respectant la sécurité des apprenants (Hall et al., 2020; Motola et al., 2013; SoFraSimS, 2021).

5. Limites des etudes

Chacune des trois études présente des limites, que nous allons résumer et décrire afin de faciliter et de modérer l'interprétation des résultats.

5.1 Développement du contenu des scores

Le choix des items et des contenus des scores présente les limites d'une méthode basée sur la consultation d'experts qui, même si elle reposait sur un échantillon large et sur une démarche de qualité suivie peut avoir des biais. Par ailleurs, la consultation d'experts ne remplace pas l'étude de l'activité et dans ce cas, l'étude de l'activité professionnelle à travers l'observation de situation d'urgences vitales. Même si cette démarche de consultation d'expert est celle qui est la plus employée dans l'univers de la pédagogie médicale et des sciences de la santé, l'apport du terrain, avec la mise à jour des contradictions et tensions qui le traversent pourrait être intéressant, notamment dans le cadre de l'approche par compétence, qui est intimement lié à l'objectif final de la formation : un professionnel compétent sur le terrain (Engeström, 2000; Humphrey-Murto et al., 2017a).

La méthodologie de création du contenu a également été limitée par l'étude incomplète du processus de réponse qui a été uniquement qualitative, reposant sur une analyse de la

signification précise des items, et pas sur une analyse initiale de leur cohérence, et de leur pertinence pour un nombre varié de situations. Cela aurait peut-être permis d'améliorer la validité de contenu des scores, et leur utilisation pour de plus nombreuses situations, sans avoir besoin de les modifier.

Les enseignants ont une appréciation variable du contenu des scores et de leur utilisabilité, notamment quant aux nombres d'items, qui parait adapté pour certains et trop important pour d'autres. Cela peut être dû à la fois à leur inexpérience en termes d'évaluation par la simulation, les enseignants qui connaissaient d'autres scores estimaient que les scores ACAT correspondaient aux normes en la matière, mais également à un manque de formation à l'utilisation des scores. En effet la formation, volontairement brève afin de pouvoir rendre les scores facilement utilisables, n'était peut-être pas assez complète, notamment dans la description de ce qui peut être attendu pour des apprenants de niveau variable. Il aurait probablement fallu proposer aux évaluateurs de s'entrainer à noter sur des vidéos avant de le faire en situation réelle.

5.2 Structure interne, relation aux autres variables et conséquences des scores

Les différentes analyses de la reproductibilité inter-observateurs comportent un biais principal : celui de n'avoir pas été en aveugle du niveau des apprenants, qui peut influencer la manière de noter. Cependant, dans l'étude qui s'intéresse à la reproductibilité seule (Article 2), les conséquences ne sont pas majeures, car tous les évaluateurs étaient au courant du niveau des apprenants, et que leurs observations n'étaient pas utilisées pour savoir si les scores ACAT discriminaient de manière adaptée les apprenants. En revanche, dans l'étude utilisant la simulation in situ, cela a pu influencer les moyennes des apprenants et donc l'analyse de la discrimination des scores.

De plus, dans cette étude, les évaluateurs étaient en aveugle l'un de l'autre, mais évaluaient au même moment, dans la même pièce. Ils n'étaient pas censés échanger à propos de leur notation

précise, mais on peut imaginer que le fait d'être ensemble a pu cependant influencer les décisions prises, pour des éléments spécifiques de notation.

Enfin, une des limites pour conclure à la cohérence des scores ACAT et à leur utilisation pour de nombreux scénarios d'une même situation de départ est l'absence d'analyse de la performance du même apprenant, ou de la même équipe, au sein de plusieurs scénarios. En effet, s'il avait été possible d'évaluer les apprenants lors de plusieurs situations de simulation avec les trois scores ACAT, à plusieurs reprises (avec la même mise en place que des ECOS par exemple), il aurait été possible d'étudier la cohérence d'un même score pour mesurer le niveau de performance d'un apprenant, ainsi que d'étudier la cohérence des scores ACAT entre eux, et la cohérence d'un même score pour plusieurs scénarios. Cela aurait été possible, en utilisation le cadre de validation de Kane, qui permet, grâce à l'utilisation de la théorie de la généralisabilité, d'étudier la variance de plusieurs éléments d'une évaluation à savoir : les apprenants, les scores, les scénarios et les enseignants. Ainsi, on s'attendrait à trouver des résultats identiques pour un même étudiants en fonction de la variation des scénarios, en fonction de la variation des scores ACAT et en fonction des différents évaluateurs. Cela permettrait une extrapolation des résultats de nos recherches (Cook et al., 2015; Kane, 2013; Laveault, 2008).

De plus, avoir utilisé le score TEAM pour étudier la corrélation avec d'autres tests validés peut apparaitre comme une limite. En effet, le score TEAM, composé de 11 items, évalue principalement la performance globale de l'équipe (sauf pour deux items centrés sur le leadership), alors que les scores ACAT s'intéressent à la performance individuelle du médecin. Ce choix a été fait car le score TEAM était la seule grille d'évaluation d'habiletés non-techniques validée en français, et qu'il est recommandé d'utiliser des grilles validées dans sa propre langue afin d'éviter des erreurs de traduction et de sémantique (SoFraSimS, 2021). La corrélation moyenne entre les scores ACAT peut être expliquée par le fait que les deux échelles ne mesurent

pas exactement la même chose, à la fois dans la nature des habiletés mais également dans la performance évaluée (individuelle ou d'équipe).

La définition des seuils à partir desquels un apprenant est considéré comme ayant atteint les objectifs requis pour son niveau reste encore à travailler, possiblement grâce à l'utilisation des scores ACAT avec des vidéos ou des groupes d'étudiants répondant aux standards requis, qui restent également à définir, pour chaque étape clé de la formation, ce qui a pu manquer aux évaluateurs.

5.3 Faisabilité et impact pédagogique

La faisabilité d'une telle évaluation et son utilisation par un nombre conséquent de centres et de formateurs ont pu être démontrées dans la dernière étude de la thèse. En revanche, il reste tout un travail à effectuer sur l'implantation des scores, mais également de la simulation « évaluative » au sein des programmes de formation actuels, qui sont en cours de construction et dans lesquels, il faudra intégrer l'évaluation. Ce travail a permis d'identifier certains obstacles, certains éléments nécessaires à la mise en place d'un tel programme, ambitieux, mais qui répond aux exigences de formation actuelles des futurs médecins.

Un des éléments manquant de notre recherche est l'étude de l'impact de l'utilisation des scores, et de l'évaluation par la simulation sur les apprenants. Nous avions pu mener ce travail auprès d'étudiants en 4ème et 5ème année de médecine mais à propos d'autres scores et d'autres situations évaluatives (Philippon et al, 2021). Il aurait été pertinent de voir se confirmer les hypothèses dégagées par le premier travail. Cependant, alors que nous nous sommes intéressés au développement des scores et à leurs caractéristiques intrinsèques, nous n'avons pas mis en place de réelle situation d'évaluation sommative, ce qui ne nous permettait pas d'étudier son impact sur les apprenants.

6. Perspectives de recherche

6.1 Extrapolation de la validité des scores

Quelques limites à l'extrapolation de la validité des scores subsistent, notamment quant à la définition des seuils de détection de la performance des apprenants. Le processus de validation d'un score étant un processus toujours en mouvement et jamais terminé, les étapes suivantes qu'il nous faudrait poursuivre seraient : la validation des seuils de réussite à partir de groupes « calibrés » d'apprenants, et après avoir défini précisément ce qui est attendu d'eux en fonction de leur niveau de progression ; l'identification, pour chaque groupe d'apprenants et pour chaque items, de marqueurs de performances en lien avec la compétence et avec l'autonomisation des apprenants, afin d'être plus en accord avec une démarche d'approche par compétences, pour abandonner progressivement la mesure behaviouriste des comportements des apprenants. Il faudrait ainsi pouvoir s'appuyer sur l'utilisation de jalons prédéfinis, ainsi que sur des activités professionnelles confiables.

6.2 Place de l'évaluation formative

Au-delà des barrières « habituelles » que peut rencontrer la simulation, l'acceptabilité de l'évaluation par la simulation semble être problématique parmi les enseignants, dans ce qu'elle nuirait à la règle éthique de non-malfaisance qui prime au sein des principes pédagogiques de la simulation. Dans une vision plus globale de l'évaluation de compétences, qui s'appuie sur des marqueurs de progression des apprenants, dans une temporalité adaptée à leur évolution personnelle, et détachée de marqueurs métriques, une des façons, pour les institutions et directeurs de programme d'enseignement, de faire accepter la simulation comme outil d'évaluation à enjeu élevé (mais dépourvu des attributs métriques habituels d'une telle évaluation), pourrait être d'implanter l'évaluation formative de manière plus formelle, afin de préparer enseignants et apprenants à l'étape suivante qui serait celle de l'évaluation certificative, avec la simulation. Il pourrait être en effet intéressant de regarder si l'implantation des scores

ACAT, une fois les seuils validés et le lien avec quelques marqueurs cliniques pertinents réalisé, serait plus facilement acceptée après une utilisation dans un but initialement formatif, puis secondairement dans un but certificatif, les enseignants ayant acquis une certaine nouvelle « culture évaluative ».

Dans ce cadre, il serait également intéressant d'explorer une façon d'adjoindre le débriefing à la décision finale, qui pourrait être fondée à la fois sur les scores ACAT, mais également sur la manière dont l'apprenant a fait part de réflexivité lors du débriefing, a pu apporter des éléments de raisonnement et semble avoir intégré les axes de progression qu'il pourrait travailler suite à l'évaluation.

Par ailleurs, une étude de l'impact de l'évaluation en analysant son retentissement sur les pratiques des apprenants, sur leur motivation et sur leur sentiment de compétence permettrait d'affiner à la fois l'apport des scores, mais également l'apport de l'évaluation par la simulation. De même, il faudrait pouvoir étudier la mise en place des évaluations au sein d'un programme composé d'évaluations multiples, qui utilise de nombreux outils pour évaluer les performances des apprenants, afin de pouvoir en dégager une idée de leur compétence.

6.3 Être « bon en simulation » : un dessein suffisant ?

Une autre question se pose dans l'évaluation et la certification des compétences par la simulation. Il s'agit de la question du transfert des compétences validées vers le patient. Finalement, le but ultime de l'évaluation en simulation n'est pas d'être « bon en simulation » mais bien de devenir un futur professionnel compétent. On s'approche ici du principe de la médecine translationnelle qui voudrait voir les effets de l'enseignement sur le patient et non plus sur le mannequin ou la copie. Des recherches restent à mener dans ce sens, mais elles s'avèrent difficiles, car elles comportent de nombreux biais (McGaghie, et al, 2011).

Certains auteurs ont pu les réaliser quand il s'agissait de gestes techniques. Par exemple en mesurant le taux de succès de ponction lombaire chez les enfants après un entrainement sur

simulateur (Kessler et al, 2011). Il faudrait également évaluer l'impact d'une formation par simulation non pas sur la satisfaction de l'apprenant, ou sur ses connaissances, mais sur la satisfaction du patient et la façon de réaliser convenablement un geste en situation réelle. Pour cela, il faut inclure la simulation dans les programmes de formation, donner la possibilité aux étudiants en médecine de la pratiquer plus et surtout, créer des programmes et des équipes de recherche qui permettent, du début de la formation à la pratique avec le patient, d'évaluer les effets de la formation pratique, en ayant conscience des biais induits par l'expérience en stage, les différences de mise en condition ou encore les difficultés liées au travail en équipe (McGaghie et al, 2012).

Pour ce faire, une telle recherche doit pouvoir s'intégrer dans un système de santé dans lequel la simulation est bien implantée, ce qui n'est pas encore tout à fait le cas en France. On pourra objecter le fait que pour l'implanter mieux, il faut qu'elle puisse prouver son efficacité au plus haut niveau, à savoir en ce qui concerne les soins délivrés aux patients mais également en terme de coûts et d'organisation des soins (Kirkpatrick, 2006).

Cela constitue néanmoins une piste de recherche intéressante, qui permettrait de valider les pratiques enseignantes basées sur la simulation et de valider la place des scores ACAT au sein d'une telle démarche à la fois pour des étudiants de deuxième cycle et pour des internes de médecine d'urgence.

6.4 Recherches dans le cadre de l'activité

Une des perspectives de recherche intéressante pour l'évaluation en médecine d'urgence au sein d'une approche par compétence, serait de compléter la démarche rapportée dans ce travail, fondée sur la consultation d'experts et l'analyse de la littérature, avec une démarche plus en lien avec l'activité. Par activité, nous entendons l'analyse du travail en médecine d'urgence, avec ses règles, tensions, contradictions, mais également l'analyse du travail en simulation et en formation: de l'apprenant à l'évaluateur, en passant par le formateur en simulation, mais également par le

formateur sur les terrains de stage. L'analyse que permettrait ce travail permettrait à la fois d'affiner les critères d'évaluation à associer aux scores ACAT, mais également de leur permettre d'évaluer le travail lui-même, et les apprenants en situation de travail plutôt qu'en situation simulée. Une des façons d'aborder la question serait de réaliser des observations de terrain, à la fois dans les structures d'urgence, mais également dans les structures de formation. L'utilisation du cadre théorique de l'activité d'Engeström, ainsi que celui de la didactique professionnelle qui permet d'identifier les situations apprenantes de travail, avec leurs différentes composantes, permettrait de réaliser ce travail (Engeström, 2001; Pastré et al., 2006). Cette approche permettrait également de lier les résultats obtenus aux scores avec des critères de jugement clinique, et ainsi pallier à la problématique de la réussite unique sur simulateur. De plus, cela permettrait de questionner à nouveau la place de la simulation au sein de l'enseignement en santé : comme un système d'activité à part entière ou comme un outil supplémentaire à la disposition des enseignants. Cette perspective nous permettrait de faire le lien entre la recherche actuelle qui nous a permis d'approfondir les principes clés d'une évaluation en s'initiant à ce concept via la création d'un outil, et notre recherche précédente, dans le cadre de l'activité, mais qui était axée sur la perception des apprenants (Philippon et al., 2021). Le chainon manquant à ces recherches étant la finalité du programme d'enseignement et d'évaluation : le système de soins, et plus spécifiquement pour notre discipline de recherche, les terrains d'exercice de la médecine d'urgence que sont les services hospitaliers, extrahospitaliers et de régulation médicale.

CONCLUSION

A travers l'étude du processus de développement de trois scores d'évaluation, qui avaient pour objectif d'évaluer les habiletés techniques et non-techniques, à la fois pour des étudiants et des internes dans le champ de la médecine d'urgence, la recherche effectuée a permis de créer des scores valides, fiables et utilisables dans de nombreuses situations et contextes cliniques. Le processus de validation n'est néanmoins pas terminé, de nombreuses étapes restant à franchir pour utiliser les scores dans des conditions valides et au sein d'un programme d'évaluation.

Dans notre travail, nous avons déployé des scores à la fois adaptés au système d'enseignement actuel, au sein duquel les évaluations restent attachées à une mesure trop morcelée de la performance, et adaptés à une approche par compétence qui tend à se développer. Cela nous permet d'entrevoir un rôle potentiel pour l'évaluation par la simulation et pour les scores ACAT qui se trouvent à la frontière entre les deux systèmes. Grâce à l'utilisation des scores et à la perception qu'en ont eu les enseignants, on aperçoit un paradoxe entre des enseignants, étudiants, institutions qui restent attachés à un système métrique d'évaluation, reposant sur une supposée objectivité des résultats obtenus et une réalité mise à jour lorsque l'on utilise une évaluation de performance tel que cela est le cas avec l'évaluation par la simulation : le système d'évaluation objective présente des limites, et les résultats qu'il fournit (en l'occurrence les notes), ne sont pas en accord avec la rétroaction que nécessite une telle activité, fondée sur l'expérience et sur ce que l'apprenant peut en dégager pour sa progression future. Ainsi, la simulation pourrait être un moyen d'accompagner le changement de regard sur les évaluations et notamment sur la place qui est faite aux évaluations dites « subjectives », mais qui présentent de réels critères de validité et de fiabilité.

Ainsi la simulation, dans son aspect évaluatif, pourrait avoir un rôle à jouer, à condition que son déploiement repose sur des outils validés, et sur une utilisation de l'évaluation formative

concomitante au déploiement de l'évaluation sommative. En effet, les enseignants s'interrogent quant à la perte de la bienveillance qui suivrait l'introduction de l'évaluation, alors que celle-ci est un composant inhérent à tout enseignement par simulation, qui se doit d'établir et de maintenir une sécurité psychologique chez les apprenants. Il s'agit d'un des obstacles identifiés dans ce travail de recherche et qui est relativement nouveau en comparaison aux obstacles habituellement retrouvés lorsque l'on s'intéresse à l'enseignement par simulation. Ainsi, de même qu'il s'agirait de prendre en compte la nécessité d'un changement de culture évaluative de la part des enseignants, des apprenants et des institutions, utiliser la simulation pour évaluer les performances requerrait probablement un questionnement et des aides pour limiter les conséquences négatives de l'évaluation en simulation, à la fois sur les apprenants, mais également sur le système d'enseignement par simulation.

Enfin, les outils développés ne peuvent être utilisés sans être intégrés à la fois dans un programme de formation en médecine d'urgence duquel découlerait un programme d'évaluation, reposant sur des outils de mesure variés, au sein desquels la simulation et les scores ACAT doivent trouver leur place.

BIBLIOGRAPHIE

Ahmed, K., Jawad, M., Dasgupta, P., Darzi, A., Athanasiou, T., & Khan, M. S. (2010). Assessment and maintenance of competence in urology. *Nature Reviews. Urology*, *7*(7), 403-413

Albano, MG., & d'Ivernois, J. F. (2001). Quand les médecins se font pédagogues. *Les Cahiers pédagogiques, 399*. Consulté le 7 février 2016 à http://www.cahiers-pedagogiques.com/Quand-les-medecins-se-font-pedagogues

Alinier, G. (2007). A typology of educationally focused medical simulation tools. *Medical Teacher*, *29*(8), e243-250

Allain, M., Kuczer, V., Longo, C., Batard, E., & Le Conte, P. (2018). Place de la simulation dans la formation initiale des urgentistes : Enquête nationale observationnelle. *Annales françaises de médecine d'urgence*, *8*(2), 75-82

Allal, L. (2013). *Evaluation : Un pont entre enseignement et apprentissage à l'université.* In Romainville, M., Goasdoué R. & M. Vantourout (Eds.), Evaluation et enseignement supérieur (pp. 21-40). Bruxelles: De Boeck

Amalberti, R., Auroy, Y., Berwick, D., & Barach, P. (2005). Five System Barriers to Achieving Ultrasafe Health Care. *Annals of Internal Medicine*, *142*(9), 756-764

Andersen, P. O., Jensen, M. K., Lippert, A., & Østergaard, D. (2010). Identifying non-technical skills and barriers for improvement of teamwork in cardiac arrest teams. *Resuscitation*, *81*(6), 695-702

André, N., Loye, N., & Laurencelle, L. (2015). La validité psychométrique : Un regard global sur le concept centenaire, sa genèse, ses avatars. *Mesure et évaluation en éducation*, *37*, 125-148

Andreatta, P. B., & Gruppen, L. D. (2009). Conceptualising and classifying validity evidence for simulation. *Medical Education*, *43*(11), 1028-1035

Andrews, J. S., Bale, J. F., Soep, J. B., Long, M., Carraccio, C., Englander, R., Powell, D., & EPAC Study Group. (2018). Education in Pediatrics Across the Continuum (EPAC) : First Steps Toward Realizing the Dream of Competency-Based Education. *Academic Medicine*, *93*(3), 414-420

Ansquer, R., Oriot, D., & Ghazali, D. A. (2021). Evaluation of Learning Effectiveness After a Simulation-Based Training Pediatric Course for Emergency Physicians. *Pediatric Emergency Care*, *37*(12): e1186-e1191

Aouni, Z. (2012). Démystification d'une pédagogie émergente : L'approche par les compétences. *Entreprendre & Innover*, *11-12*, 120-126

Bandura, A. (2001). Social cognitive theory : An agentic perspective. *Annual Review of Psychology*, *52*, 1-26

Barsuk, J. H., Cohen, E. R., Williams, M. V., Scher, J., Jones, S. F., Feinglass, J., McGaghie, W. C., O'Hara, K., & Wayne, D. B. (2018). Simulation-Based Mastery Learning for Thoracentesis Skills Improves Patient Outcomes : A Randomized Trial. *Academic Medicine*, *93*(5), 729-735

Batalden, P., Leach, D., Swing, S., Dreyfus, H., & Dreyfus, S. (2002). General competencies and accreditation in graduate medical education. *Health Affairs (Project Hope)*, *21*(5), 103-111

Baudouin, J. (2002). La compétence et le thème de l'activité : vers une nouvelle conceptualisation didactique de la formation. In Dolz J. (Ed). *L'énigme de la compétence en éducation*, pp. 147-168. Louvain-la-Neuve: De Boeck Supérieur

Beal, M. D., Kinnear, J., Anderson, C. R., Martin, T. D., Wamboldt, R., & Hooper, L. (2017). The Effectiveness of Medical Simulation in Teaching Medical Students Critical Care Medicine : A Systematic Review and Meta-Analysis. *Simulation in Healthcare, 12*(2), 104-116

Beaud, S. (1996). L'usage de l'entretien en sciences sociales. Plaidoyer pour l'« entretien ethnographique ». *Politix*, *9*(35), 226-257

Bernabeo, E. C., Holtman, M. C., Ginsburg, S., Rosenbaum, J. R., & Holmboe, E. S. (2011). Lost in transition : The experience and impact of frequent changes in the inpatient learning environment. *Academic Medicine, 86*(5), 591-598

Berragan, L. (2013). Conceptualising learning through simulation: An expansive approach for professional and personal learning. *Nurse Education in Practice*, *13*(4), 250-255

Bleakley, A. (2006). Broadening conceptions of learning in medical education : The message from teamworking. *Medical Education*, *40*(2), 150-157

Blew, P., Muir, J. G., & Naik, V. N. (2010). The evolving Royal College examination in anesthesiology. *Canadian Journal of Anaesthesia, 57*(9), 804-810

Bloodgood, R. A., Short, J. G., Jackson, J. M., & Martindale, J. R. (2009). A change to pass/fail grading in the first two years at one medical school results in improved psychological well-being. *Academic Medicine*, *84*(5), 655-662

Bloom B, Engelhart M, Furst E, Hill W, Krathwohl D. (1956). *Taxonomy of Educational Objectives. The Classification of Educational Goals; Handbook I : Cognitive Domain*. Longmans, Green

Bodle, J. F., Kaufmann, S. J., Bisson, D., Nathanson, B., & Binney, D. M. (2008). Value and face validity of objective structured assessment of technical skills (OSATS) for work-based assessment of surgical skills in obstetrics and gynaecology. *Medical Teacher*, *30*(2), 212-216

Boelen, C., Heck, J. E., & Health, W. H. O. D. of D. of H. R. for. (2000). *Définir et mesurer la responsabilité sociale des facultés de médecine* (WHO/HRH/95.7).[en ligne], consulté le 2 avril 2017. https://apps.who.int/iris/handle/10665/66532

Boet, S., Etherington, N., Larrigan, S., Yin, L., Khan, H., Sullivan, K., Jung, J., & Grantcharov, T. (2019). Measuring the teamwork performance of teams in crisis situations : A systematic review of assessment tools and their measurement properties. *BMJ Quality & Safety*, *28*, 327-337

Boet, S., Jaffrelot, M., Naik, V. N., Brien, S., & Granry, J.-C. (2014). La simulation en santé en Amérique du Nord : État actuel et évolution après deux décennies. *Annales Françaises d'Anesthésie et de Réanimation, 33*(5), 353-357

Boet, S., Larrigan, S., Martin, L., Liu, H., Sullivan, K. J., & Etherington, N. (2018). Measuring non-technical skills of anaesthesiologists in the operating room : A systematic review of assessment tools and their measurement properties. *British Journal of Anaesthesia*, *121*(6), 1218-1226

Boet, S., Savoldelli, G., & Granry, J.-C. (Éds.). (2013). *La simulation en santé. De la théorie à la pratique*. Springer. Paris

Bok, H. G. J., de Jong, L. H., O'Neill, T., Maxey, C., & Hecker, K. G. (2018). Validity evidence for programmatic assessment in competency-based education. *Perspectives on Medical Education*, *7*(6), 362-372

Bong, C. L., Lightdale, J. R., Fredette, M. E., & Weinstock, P. (2010). Effects of simulation versus traditional tutorial-based training on physiologic stress levels among clinicians : A pilot study. *Simulation in Healthcare*, *5*(5), 272-278

Boulet, J. R. (2008). Summative Assessment in Medicine : The Promise of Simulation for High-stakes Evaluation. *Academic Emergency Medicine*, *15*(11), 1017-1024

Boulet, J. R., & Swanson, D. (2004). Psychometric challenges of using simulation for high-stakes assessment. In *Simulators in critical care education and beyond* (4e éd.), p. 119-130. Society of Critical Care Medicine

Brailovsky, C., Miller, F., & Grand'Maison, P. (1998). L'évaluation de la compétence dans le contexte professionnel. *Service social*, *47*(1-2), 171-189

Brydges, R., Hatala, R., Zendejas, B., Erwin, P. J., & Cook, D. A. (2015). Linking Simulation-Based Educational Assessments and Patient-Related Outcomes : A Systematic Review and Meta-Analysis. *Academic Medicine*, *90*(2), 246-256

Bullard, M. J., Fox, S. M., Heffner, A. C., Bullard, C. L., & Wares, C. M. (2020). Unifying Resident Education : 12 Interdisciplinary Critical Care Simulation Scenarios. *MedEdPORTAL: The Journal of Teaching and Learning Resources*, *16*, 11009

Burg, F. D., Brownlee, R. C., Wright, F. H., Levine, H., Daeschner, C. W., Vaughan, V. C., & Anderson, J. A. (1976). A method for defining competency in Pediatrics. *Journal of Medical Education*, *51*(10), 824-828

Canada's Testing Company. (2014). *The Angoff method of standard setting.*[en ligne], consulté le 17 juin 2020.
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwixq-T9p_j0AhUP1BoKHclMB0gQFnoECB0QAQ&url=https%3A%2F%2Fbcrsp.ca%2Fsites%2Fdefault%2Ffiles%2Fdocuments%2FAngoff%2520Method%2520Article.pdf&usg=AOvVaw3IqiQ0gUT9hCE4qts5R3YE

Cangiarella, J., Fancher, T., Jones, B., Dodson, L., Leong, S. L., Hunsaker, M., Pallay, R., Whyte, R., Holthouser, A., & Abramson, S. B. (2017). Three-Year MD Programs : Perspectives From the Consortium of Accelerated Medical Pathway Programs (CAMPP). *Academic Medicine*, *92*(4), 483-490

Carraccio, C., Englander, R., Holmboe, E. S., & Kogan, J. R. (2016). Driving Care Quality : Aligning Trainee Assessment and Supervision Through Practical Application of Entrustable Professional Activities, Competencies, and Milestones. *Academic Medicine*, *91*(2), 199-203

Carraccio, C. L., & Englander, R. (2013). From Flexner to competencies : Reflections on a decade and the journey ahead. *Academic Medicine: Journal of the Association of American Medical Colleges*, *88*(8), 1067-1073

Carraccio, C., Wolfsthal, S. D., Englander, R., Ferentz, K., & Martin, C. (2002). Shifting paradigms : From Flexner to competencies. *Academic Medicine: Journal of the Association of American Medical Colleges*, *77*(5), 361-367

Cha, J. S., Anton, N. E., Mizota, T., Hennings, J. M., Rendina, M. A., Stanton-Maxey, K., Ritter, H. E., Stefanidis, D., & Yu, D. (2019). Use of non-technical skills can predict medical student performance in acute care simulated scenarios. *The American Journal of Surgery*, *217*(2), 323-328

Charlin, B., Bordage, G., & Vleuten, C. V. D. (2003). L'évaluation du raisonnement clinique. *Pédagogie Médicale*, *4*(1), 42-52

Charlin, B., Roy, L., Brailovsky, C., Goulet, F., & van der Vleuten, C. (2000). The Script Concordance test : A tool to assess the reflective clinician. *Teaching and Learning in Medicine*, *12*(4), 189-195

Ciccetti. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology : Normative assessment. *Psychological Assessment*, *6*(4), 284-290.

Cilliers, F. J., Schuwirth, L. W., Adendorff, H. J., Herman, N., & van der Vleuten, C. P. (2010). The mechanism of impact of summative assessment on medical students' learning. *Advances in Health Sciences Education*, *15*(5), 695-715

Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting : A Guide to Establishing and Evaluating Performance Standards on Tests* (New e. édition). SAGE Publications, Inc.

Collange, O., & McKenna, J. (2013). Éthique et simulation en santé. In S. Boet, G. Savoldelli, & J.-C. Granry (Éds.), *La simulation en santé De la théorie à la pratique*. Paris: Springer. 177-183

Collégiale des Universitaires de Médecine d'urgence (CNUMU), & Collège National des Enseignants de Thérapeutioques (APNET). (2017). *Urgences. Défaillances viscérales aiguës. Situations exceptionnelles. 2ème éd.* Paris : Med-Line

Colmers-Gray, I. N., Walsh, K., & Chan, T. M. (2017). Assessment of emergency medicine residents : A systematic review. *Canadian Medical Education Journal*, *8*(1), e106-e122

Commission des Communautés Européennes. (1995). *Livre blanc sur l'éducation et la formation - Enseigner et apprendre - Vers la société cognitive.* Publications Office of the European Union.

[en ligne] consulté le 1er mars 2021. http://op.europa.eu/fr/publication-detail/-/publication/d0a8aa7a-5311-4eee-904c-98fa541108d8/language-fr

Conférence des Doyens des facultés de Médecine, Collège National des Enseignants en Médecine. (2020). *La R2C expliquée sous l'angle pédagogique—Livret enseignant. [en ligne],* consulté le 2 janvier 2021. https://cncem.fr/node/384

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments : A practical guide to Kane's framework. *Medical Education*, *49*(6), 560-575

Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. (2013). Technology-Enhanced Simulation to Assess Health Professionals : A Systematic Review of Validity Evidence, Research Methods, and Reporting Quality. *Academic Medicine*, *88*(6), 872-883

Cook, D. A., Hatala, R., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., Erwin, P. J., & Hamstra, S. J. (2011). Technology-enhanced simulation for health professions education : A systematic review and meta-analysis. *JAMA*, *306*(9), 978-988

Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R., & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in Health Sciences Education: Theory and Practice*, *19*(2), 233-250

Corbin, J., & Strauss, A. (2008). *Basics of Qualitative Research (3rd ed.) : Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Inc.

Coulet, J.-C. (2016). Les notions de compétence et de compétences clés : L'éclairage d'un modèle théorique fondé sur l'analyse de l'activité. *Activités*, *13*(13-1), Article 1 [en ligne] consulté le 21 novembre 2020. https://doi.org/10.4000/activites.2745

Crahay, M. (2006). Dangers, incertitudes et incomplétude de la logique de la compétence en éducation. *Revue française de pédagogie. Recherches en éducation*, *154*, 97-110.

Crossley, J., & Jolly, B. (2012). Making sense of work-based assessment : Ask the right questions, in the right way, about the right things, of the right people. *Medical Education*, *46*(1), 28-37

Curtis, J. R., Back, A. L., Ford, D. W., Downey, L., Shannon, S. E., Doorenbos, A. Z., Kross, E. K., Reinke, L. F., Feemster, L. C., Edlund, B., Arnold, R. W., O'Connor, K., & Engelberg, R. A. (2013). Effect of communication skills training for residents and nurse practitioners on quality of communication with patients with serious illness : A randomized trial. *JAMA*, *310*(21), 2271-2281

Cusimano, M. D. (1996). Standard setting in medical education. *Academic Medicine*, *71*(10), S112-20

Daniels, V. J., & Pugh, D. (2018). Twelve tips for developing an OSCE that measures what you want. *Medical Teacher*, *40*(12), 1208-1213

Dauphinee, W. D. (1995). Assessing Clinical Performance : Where Do We Stand and What Might We Expect? *JAMA*, *274*(9), 741-743

De Ketele J.-M., & Gérard, F.-M. (2005). La validation des épreuves d'évaluation selon l'approche par les compétences. *Mesure et évaluation en éducation*, *28*(3), 1-26

de Landsheere, G. (1976). *Evaluation continue et examens : Précis de docimologie* (4ème éd). Paris : Fernand Nathan

De Singly, F. (2106). *Le questionnaire* (4ème éd). Paris : Armand Colin

Déclaration commune des ministres européens de l'éducation. (1999). *Déclaration de Bologne. L'espace européen de l'enseignement supérieur. [en ligne], consulté le 12 novembre 2020.* http://www.mes.tn/tempus/tempus05/990719_Bologna_Declaration-Fr.pdf

Demeester, A., De Giorgi, B., Gouchan, Y. (2020). *L'évaluation à l'épreuve du contexte. Pratiques et réflexions.* Aix-Marseille : Presses Universitaires de Provence (PUP)

Dieckmann, P., Molin Friis, S., Lippert, A., & Ostergaard, D. (2009). The art and science of debriefing in simulation : Ideal and practice. *Medical Teacher*, *31*(7), e287-294

Downing, S. M. (2003). Validity : On meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830-837

Downing, S. M. (2004). Reliability : On the reproducibility of assessment data. *Medical Education*, *38*(9), 1006-1012

Drummond, D., Arnaud, C., Thouvenin, G., Guedj, R., Duguet, A., de Suremain, N., & Petit, A. (2016). [Newly formed French residents in pediatrics are not well prepared for conducting pediatric resuscitation after medical school]. *Archives De Pédiatrie*, *23*(2), 150-158

Duchenne, J., Martinez, M., Rothmann, C., Claret, P.-G., Desclefs, J.-P., Vaux, J., Miroux, P., Ganansia, O., & membres de la commission des référentiels de la SFMU. (2016). Premier niveau de compétence pour l'échographie clinique en médecine d'urgence. Recommandations de la

Société française de médecine d'urgence par consensus formalisé. *Annales françaises de médecine d'urgence*, *6*(4), 284-295

Duchesne, S., & Haegel, F. (2015). *L'entretien collectif*. Paris : Armand Colin

Ekionea, J.-P. B., Bernard, P., & Plaisent, M. (2011). Consensus par la méthode Delphi sur les concepts clés des capacités organisationnelles spécifiques de la gestion des connaissances. *Recherches Qualitatives,* 29(3), 168-192

Engestrom, Y. (2000). Activity theory as a framework for analyzing and redesigning work. *Ergonomics*, *43*(7), 960-974

Engeström, Y. (2001). Expansive learning at work : Toward an activity theoretical reconceptualization. *Journal of education and work*, *14*(1), 133-156

Englander, R., Cameron, T., Ballard, A. J., Dodge, J., Bull, J., & Aschenbrener, C. A. (2013). Toward a common taxonomy of competency domains for the health professions and competencies for physicians. *Academic Medicine, 88*(8), 1088-1094

Englander, R., & Carraccio, C. (2018). A Lack of Continuity in Education, Training, and Practice Violates the "Do No Harm" Principle. *Academic Medicine*, *93*(3S), S12-S16

Englander, R., Frank, J. R., Carraccio, C., Sherbino, J., Ross, S., Snell, L., & ICBME Collaborators. (2017). Toward a shared language for competency-based medical education. *Medical Teacher*, *39*(6), 582-587

Epstein, R. M. (2007). Assessment in medical education. *The New England Journal of Medicine*, *356*(4), 387-396

Fahim, C., Wagner, N., Nousiainen, M. T., & Sonnadara, R. (2018). Assessment of Technical Skills Competence in the Operating Room : A Systematic and Scoping Review. *Academic Medicine*, *93*(5), 794

Fann, J. I., Sullivan, M. E., Skeff, K. M., Stratos, G. A., Walker, J. D., Grossi, E. A., Verrier, E. D., Hicks, G. L., & Feins, R. H. (2013). Teaching behaviors in the cardiac surgery simulation environment. *The Journal of Thoracic and Cardiovascular Surgery*, *145*(1), 45-53

Farrell, S. E. (2005). Evaluation of Student Performance : Clinical and Professional Performance. *Academic Emergency Medicine*, *12*(4), 302.e6-e10

Fazio, S. B., Ledford, C. H., Aronowitz, P. B., Chheda, S. G., Choe, J. H., Call, S. A., Gitlin, S. D., Muntz, M., Nixon, L. J., Pereira, A. G., Ragsdale, J. W., Stewart, E. A., & Hauer, K. E. (2018). Competency-Based Medical Education in the Internal Medicine Clerkship : A Report From the Alliance for Academic Internal Medicine Undergraduate Medical Education Task Force. *Academic Medicine*, *93*(3), 421-427

Ferguson, P. C., Kraemer, W., Nousiainen, M., Safir, O., Sonnadara, R., Alman, B., & Reznick, R. (2013). Three-year experience with an innovative, modular competency-based curriculum for orthopaedic training. *The Journal of Bone and Joint Surgery. American Volume*, *95*(21), e166.

Flanagan, B. (2008). Debriefing : Theory and techniques. *Manual of simulation in healthcare.* Oxford : Oxford University Press, 155-170

Fletcher, G., Flin, R., McGeorge, P., Glavin, R., Maran, N., & Patey, R. (2003). Anaesthetists' Non-Technical Skills (ANTS) : Evaluation of a behavioural marker system†. *British Journal of Anaesthesia*, *90*(5), 580-588

Flexner, A. (1910). *Medical Education in the United States and Canada—A report to the Carnegie Foundation for the Advancement of Teaching.* (Bulletin Number Four, p. 364). [en ligne], consulté le 12 mars 2017.
http://archive.carnegiefoundation.org/pdfs/elibrary/Carnegie_Flexner_Report.pdf

Fontaine, S., & Loye, N. (2017). L'évaluation des apprentissages : Une démarche rigoureuse. *Pédagogie Médicale*, *18*(4), 189-198

France. Ministère de l'enseignement supérieur et de la recherche, Ministère des armées, Ministère des solidarités et de la santé, *Arrêté du 4 novembre 2019 relatif à l'accès aux formations de médecine, de pharmacie, d'odontologie et de maïeutique,* Journal officiel, n°0257, 5 novembre 2019, ESRS1930498A

France. Ministère de l'enseignement supérieur, de la recherche et de l'innovation, Ministère des armées, Ministère des solidarités et de la santé, *Décret n° 2021-1156 du 7 septembre 2021 relatif à l'accès au troisième cycle des études de médecine, 2021-1156,* Journal officiel, n°0209, 8 septembre 2021, ESRS2112241D

France. Ministère des solidarités et de la santé, *Arrêté du 12 avril 2018 fixant la liste des recherches mentionnées au 3o de l'article L. 1121-1 du code de la santé publique,* Journal officiel, n°0089, 17 avril 2018, SSAP1810240A

France. République française. *Décret n°2016-1597 du 25 novembre 2016 relatif à l'organisation du troisième cycle des études de médecine et modifiant le Code de l'éducation,* Journal officiel, n°0276, 27 novembre 2016, MENS1620996D

Frank, J. R., & Danoff, D. (2007). The CanMEDS initiative : Implementing an outcomes-based framework of physician competencies. *Medical Teacher*, *29*(7), 642-647

Frank, J. R., Snell, L. S., Ten Cate, O., Holmboe, E. S., Carraccio, C., Swing, S. R., Harris, P., Glasgow, N. J., Campbell, C., Dath, D., Harden, R. M., Iobst, W., Long, D. M., Mungroo, R., Richardson, D. L., Sherbino, J., Silver, I., Taber, S., Talbot, M., & Harris, K. A. (2010). Competency-based medical education : Theory to practice. *Medical Teacher*, *32*(8), 638-645

Collège royal des médecins et chirurgiens du Canada. Frank JR, Snell L, Sherbino J, Boucher A, rédacteurs.(2015). *Référentiel de compétences CanMEDS 2015 pour les médecins*, [en ligne], consulté le 2 février 2021. https://www.royalcollege.ca

Freund, Y., Goulet, H., Leblanc, J., Bokobza, J., Ray, P., Maignan, M., Guinemer, S., Truchot, J., Féral-Pierssens, A.-L., Yordanov, Y., Philippon, A.-L., Rouff, E., Bloom, B., Cachanado, M., Rousseau, A., Simon, T., & Riou, B. (2018). Effect of Systematic Physician Cross-checking on Reducing Adverse Events in the Emergency Department : The CHARMED Cluster Randomized Trial. *JAMA Internal Medicine*, *178*(6), 812-819

Gaba, D. M. (2004). The future vision of simulation in health care. *Quality and Safety in Health Care*, *13*(suppl 1), i2-i10

Gaba, D. M. (2010). Crisis resource management and teamwork training in anaesthesia. *British Journal of Anaesthesia*, *105*(1), 3-6

Gaba, D. M., & DeAnda, A. (1988). A comprehensive anesthesia simulation environment : Re-creating the operating room for research and training. *Anesthesiology*, *69*(3), 387-394

Garcia, J., Coste, A., Tavares, W., Nuño, N., & Lachapelle, K. (2015). Assessment of competency during orotracheal intubation in medical simulation. *British Journal of Anaesthesia*, *115*(2), 302-307

Garrouste-Orgeas, M., Boumendil, A., Pateron, D., Aergerter, P., Somme, D., Simon, T., Guidet, B., & ICE-CUB Group. (2009). Selection of intensive care unit admission criteria for patients aged 80 years and over and compliance of emergency and intensive care unit physicians with the

selected criteria : An observational, multicenter, prospective study. *Critical Care Medicine*, *37*(11), 2919-2928

Gauthier, G., Couture, S., & St-Onge, C. (2018). Jugement évaluatif : Confrontation d'un modèle conceptuel à des données empiriques. *Pédagogie Médicale*, *19*(1), 15-25

Gérard, F.-M. (2002). L'indispensable subjectivité de l'évaluation. *Antipodes*, *156*, 26-34

Gérard, F.-M. (2013). L'évaluation au service de la régulation des apprentissages : Enjeux, nécessités et difficultés. *Revue française de linguistique appliquée*, *XVIII*(1), 75-92

Gerard, J. M., Kessler, D. O., Braun, C., Mehta, R., Scalzo, A. J., & Auerbach, M. (2013). Validation of Global Rating Scale and Checklist Instruments for the Infant Lumbar Puncture Procedure: *Simulation in Healthcare*, *8*(3), 148-154

Ghazali, D. A., & Casalino, E. (2018). La simulation : Développement d'un outil pédagogique devenu un paradigme en médecine d'urgence. *Annales françaises de médecine d'urgence*, *8*(2), 73-74

Gilbert, P. & Yalenios, J. (2017). I. L'évaluation de la performance en perspective. Dans : Patrick Gilbert éd., *L'évaluation de la performance individuelle* (pp. 11-28). Paris: La Découverte

Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K., & Regehr, G. (2010). Toward authentic clinical evaluation : Pitfalls in the pursuit of competency. *Academic Medicine: Journal of the Association of American Medical Colleges*, *85*(5), 780-786

Giraud, C. (2010a). « Les techniques d'enquête en sociologie », In de Singly F., Martin O., Giraud C, *Nouveau manuel de sociologie*. Paris: Armand Colin, 38-51

Giraud, C. (2010b). « Les mots pour faire dire et écrire »,. In de Singly F., Martin O., Giraud C, *Nouveau manuel de sociologie*. Paris: Armand Colin, 52-67

Greif, R., Lockey, A., Breckwoldt, J., Carmona, F., Conaghan, P., Kuzovlev, A., Pflanzl-Knizacek, L., Sari, F., Shammet, S., Scapigliati, A., Turner, N., Yeung, J., & Monsieurs, K. G. (2021). European Resuscitation Council Guidelines 2021 : Education for resuscitation. *Resuscitation*, *161*, 388-407

Griswold, S., Fralliccardi, A., Boulet, J., Moadel, T., Franzen, D., Auerbach, M., Hart, D., Goswami, V., Hui, J., & Gordon, J. A. (2018). Simulation-based Education to Ensure Provider Competency Within the Health Care System. *Academic Emergency Medicine*, *25*(2), 168-176

Griswold-Theodorson, S., Ponnuru, S., Dong, C., Szyld, D., Reed, T., & McGaghie, W. C. (2015). Beyond the Simulation Laboratory : A Realist Synthesis Review of Clinical Outcomes of Simulation-Based Mastery Learning. *Academic Medicine*, *90*(11), 1553

Guise, J.-M., Deering, S. H., Kanki, B. G., Osterweil, P., Li, H., Mori, M., & Lowe, N. K. (2008). Validation of a tool to measure and promote clinical teamwork. *Simulation in Healthcare*, *3*(4), 217-223

Hall, A. K., Chaplin, T., McColl, T., Petrosoniak, A., Caners, K., Rocca, N., Gardner, C., Bhanji, F., & Woods, R. (2020). Harnessing the power of simulation for assessment : Consensus recommendations for the use of simulation-based assessment in emergency medicine. *Canadian Journal of Emergency Medicine*, *22*(2), 194-203

Hall, A. K., Pickett, W., & Dagnone, J. D. (2012). Development and evaluation of a simulation-based resuscitation scenario assessment tool for emergency medicine residents. *CJEM*, *14*(3), 139-146

Hanrahan, J., Sideris, M., Tsitsopoulos, P. P., Bimpis, A., Pasha, T., Whitfield, P. C., & Papalois, A. E. (2018). Increasing motivation and engagement in neurosurgery for medical students through practical simulation-based learning. *Annals of Medicine and Surgery*, *34*, 75-79

Hart, D., Bond, W., Siegelman, J. N., Miller, D., Cassara, M., Barker, L., Anders, S., Ahn, J., Huang, H., Strother, C., & Hui, J. (2018). Simulation for Assessment of Milestones in Emergency Medicine Residents. *Academic Emergency Medicine*, *25*(2), 205-220

Harwayne-Gidansky, I., Askin, G., Fein, D. M., McNamara, C., Duncan, E., Delaney, K., Greenberg, J., Mojica, M., Clapper, T., & Ching, K. (2021). Effectiveness of a Simulation Curriculum on Clinical Application : A Randomized Educational Trial. *Simulation in Healthcare,* [en ligne], consulté le 25 juillet 2021. https://journals.lww.com/simulationinhealthcare/Abstract/9000/Effectiveness_of_a_Simulation_Curriculum_on.99344.aspx

Haute Autorité de Santé. (2012). *Guide de bonnes pratiques en simulation en santé*. [en ligne], consulté le 12 avril 2015. https://www.has-sante.fr/upload/docs/application/pdf/2013-01/guide_bonnes_pratiques_simulation_sante_guide.pdf

Haute Autorité de Santé. (2020). *Évaluation de la prise en charge de l'urgence vitale en établissement selon le référentiel de certification.* Certification des établissements de santé -

Fiche pédagogique - Urgence Vitale. [en ligne], consulté le 2 juillet 2021. https://www.has-sante.fr/plugins/ModuleXitiKLEE/types/FileDocument/doXiti.jsp?id=p_3222201

Hawkins, R. E., Welcher, C. M., Holmboe, E. S., Kirk, L. M., Norcini, J. J., Simons, K. B., & Skochelak, S. E. (2015). Implementation of competency-based medical education : Are we addressing the concerns and challenges? *Medical Education*, *49*(11), 1086-1102

Hayward, L. (2015). Assessment is learning : The preposition vanishes. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 27-43

Hirsh, D., Gaufberg, E., Ogur, B., Cohen, P., Krupat, E., Cox, M., Pelletier, S., & Bor, D. (2012). Educational outcomes of the Harvard Medical School-Cambridge integrated clerkship : A way forward for medical education. *Academic Medicine*, *87*(5), 643-650

Hoang, N. S., & Lau, J. N. (2018). A Call for Mixed Methods in Competency-Based Medical Education : How We Can Prevent the Overfitting of Curriculum and Assessment. *Academic Medicine*, *93*(7), 996-1001

Hodges, B. (2013). Assessment in the post-psychometric era : Learning to love the subjective and collective. *Medical Teacher*, *35*(7), 564-568

Hodges, B., Regehr, G., McNaughton, N., Tiberius, R., & Hanson, M. (1999). OSCE checklists do not capture increasing levels of expertise. *Academic Medicine: Journal of the Association of American Medical Colleges*, *74*(10), 1129-1134

Holmboe E., During S., Hawkins, R. (2018). *Practical Guide to the Evaluation of Clinical Competence* (2nd Ed.). Elsevier

Holmboe, E. S. (2015a). Realizing the promise of competency-based medical education. *Academic Medicine: Journal of the Association of American Medical Colleges*, *90*(4), 411-413

Holmboe, E. S. (2015b). Emergency Medicine : On the Frontlines of Medical Education Transformation. *Western Journal of Emergency Medicine*, *16*(6), 801-803

Holmboe, E. S., Sherbino, J., Englander, R., Snell, L., Frank, J. R., & ICBME Collaborators. (2017). A call to action : The controversy of and rationale for competency-based medical education. *Medical Teacher*, *39*(6), 574-581

Holmboe, E. S., Sherbino, J., Long, D. M., Swing, S. R., & Frank, J. R. (2010). The role of assessment in competency-based medical education. *Medical Teacher*, *32*(8), 676-682

Holmboe, E. S., Ward, D. S., Reznick, R. K., Katsufrakis, P. J., Leslie, K. M., Patel, V. L., Ray, D. D., & Nelson, E. A. (2011). Faculty Development in Assessment : The Missing Link in Competency-Based Medical Education: *Academic Medicine*, *86*(4), 460-467

Holmboe E.S. & Iobst, W.F. (2020). *ACGME Assessment Guidebook.*[en ligne], consulté le 3 mars 2021, https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjny-TMuMryAhVBqxoKHQ2UDckQFnoECAQQAQ&url=https%3A%2F%2Fwww.acgme.org%2FPortals%2F0%2FPDFs%2FMilestones%2FGuidebooks%2FAssessmentGuidebook.pdf&usg=AOvVaw0p-ERP_m_XzJumCwaJtguq

Homer, M., Pell, G., & Fuller, R. (2017). Problematizing the concept of the « borderline » group in performance assessments. *Medical Teacher*, *39*(5), 469-475

Homerin, M.-P., & Roumanet, M.-C. (2014). Évaluation des étudiants infirmiers en situation simulée : En quête de sens et d'éthique. *Recherche en soins infirmiers*, *118*(3), 38-51

Hosny, S. G., Johnston, M. J., Pucher, P. H., Erridge, S., & Darzi, A. (2017). Barriers to the implementation and uptake of simulation-based training programs in general surgery : A multinational qualitative study. *The Journal of Surgical Research*, *220*, 419-426.e2

Houzé-Cerfon, C., Boet, S., Marhar, F., Saint-Jean, M., & Geeraerts, T. (2019). L'éducation interprofessionnelle des équipes de soins critiques par la simulation : Concept, mise en œuvre et évaluation. *La Presse Médicale*, *48(7-8), 780-787*

Houzé-Cerfon, C. H., Boet, S., Saint-Jean, M., Cros, J., Vardon-Bounes, F., Marhar, F., Couarraze, S., Der Sahakian, G., Mattatia, L., Nicolle, L., Balen, F., Charpentier, S., Bounes, V., & Geeraerts, T. (2020). Effect of combined individual-collective debriefing of participants in interprofessional simulation courses on crisis resource management : A randomized controlled multicenter trial. *Emergencias*, *32*(2), 111-117

Houzé-Cerfon, C.-H., Lauque, D., Wiel, E., Bounes, V., & Charpentier, S. (2020). Conception d'un programme d'enseignement par simulation dans le DES de médecine d'urgence selon la méthode du modèle logique. *Annales françaises de médecine d'urgence*, *10*(1), 14-30

Huang, J., Tang, Y., Tang, J., Shi, J., Wang, H., Xiong, T., Xia, B., Zhang, L., Qu, Y., & Mu, D. (2019). Educational efficacy of high-fidelity simulation in neonatal resuscitation training : A systematic review and meta-analysis. *BMC Medical Education*, *19*(1), 323

Humphrey-Murto, S., Varpio, L., Gonsalves, C., & Wood, T. J. (2017a). Using consensus group methods such as Delphi and Nominal Group in medical education research. *Medical Teacher*, *39*(1), 14-19

Humphrey-Murto, S., Varpio, L., Wood, T., Gonsalves, C., Ufholz, L.-A., Mascioli, K., Wang, C., & Foth, T. (2017b). The Use of the Delphi and Other Consensus Group Methods in Medical Education Research. *Academic Medicine*, *92*(10), 1491-1498

Humphrey-Murto, S., Wood, T. J., & Varpio, L. (2017c). When I say … consensus group methods. *Medical Education*, *51*(10), 994-995

Hurtz, G. M., & Auerbach, M. A. (2003). A Meta-Analysis of the Effects of Modifications to the Angoff Method on Cutoff Scores and Judgment Consensus. *Educational and Psychological Measurement*, *63*(4), 584-601

Ilgen, J. S., Ma, I. W. Y., Hatala, R., & Cook, D. A. (2015). A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Medical Education*, *49*(2), 161-173

Ingrassia, P. L., Ragazzoni, L., Tengattini, M., Carenzo, L., & Della Corte, F. (2014). Nationwide program of education for undergraduates in the field of disaster medicine : Development of a core curriculum centered on blended learning and simulation tools. *Prehospital and Disaster Medicine*, *29*(5), 508-515

Institute of Medicine (US) Committee on Quality of Health Care in America. Kohn L.T., Corrigan, J.M. & Donaldson M.S. (Éds.). (2000). *To Err is Human : Building a Safer Health System*. US: National Academies Press

Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Lee Gordon, D., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning : A BEME systematic review. *Medical Teacher*, *27*(1), 10-28

Jabre, P., Belpomme, V., Azoulay, E., Jacob, L., Bertrand, L., Lapostolle, F., Tazarourte, K., Bouilleau, G., Pinaud, V., Broche, C., Normand, D., Baubet, T., Ricard-Hibon, A., Istria, J., Beltramini, A., Alheritiere, A., Assez, N., Nace, L., Vivien, B., … Adnet, F. (2013). Family Presence during Cardiopulmonary Resuscitation. *New England Journal of Medicine*, *368*(11)

Jick, T. D. (1979). Mixing Qualitative and Quantitative Methods : Triangulation in Action. *Administrative Science Quarterly*, *24*(4), 602-611

Jones, J., & Hunter, D. (1995). Consensus methods for medical and health services research. *BMJ (Clinical Research Ed.)*, *311*(7001), 376-380

Jonnaert, P. (2017). La notion de compétence : Une réflexion toujours inachevée. *Éthique publique, (19)*1, [En ligne], consulté le 29 décembre 2020. http://journals.openedition.org/ethiquepublique/2932

Jouquan, J. (2002). L'évaluation des apprentissages des étudiants en formation médicale initiale. *Pédagogie médicale*, *3*(1), 38-52

Jouquan, J. (2007). De l'approche par objectifs à l'approche par compétences. Faut-il jeter le bébé avec l'eau du bain ? *Pédagogie médicale*, *8*(4), 197-198

Jouquan, J., & Parent, F. (2015). *Comment élaborer et analyser un référentiel de compétences en santé ?* (1ère éd.). Bruxelles: De boeck supérieur

Kane, M., Crooks, T., & Cohen, A. (2005). Validating Measures of Performance. *Educational Measurement: Issues and Practice*, *18*(2), 5-17

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1-73

Kay, D., & Kibble, J. (2016). Learning theories 101 : Application to everyday teaching and scholarship: Table 1. *Advances in Physiology Education*, *40*(1), 17-25

Kennedy, T. J. T., & Lingard, L. A. (2006). Making sense of grounded theory in medical education. *Medical Education*, *40*(2), 101-108

Kessler, D., Auerbach, M., Pusic, M., Tunik, M., & Foltin, J. (2011). A randomized trial of simulation-based deliberate practice for infant lumbar puncture skills. *Simulation in Healthcare*, *6*(4), 197-203

Khan, K. Z., Gaunt, K., Ramachandran, S., & Pushkar, P. (2013). The Objective Structured Clinical Examination (OSCE) : AMEE Guide No. 81. Part II: Organisation & Administration. *Medical Teacher*, *35*(9), e1447-e1463

Kirkpatrick, D. L. (2006). Seven keys to unlock the four levels of evaluation. *Performance Improvement*, *45*(7), 5-8

Kirkpatrick, D. (1996, January). Great ideas revisited. *Training & Development*, *50*(1), 54+

Knobel, A., Overheu, D., Gruessing, M., Juergensen, I., & Struewer, J. (2018). Regular, in-situ, team-based training in trauma resuscitation with video debriefing enhances confidence and clinical efficiency. *BMC Medical Education*, *18*(1), 127

Kogan, J. R., Conforti, L. N., Iobst, W. F., & Holmboe, E. S. (2014). Reconceptualizing Variable Rater Assessments as Both an Educational and Clinical Care Problem. *Academic Medicine*, *89*(5), 721

Kogan, J. R., Holmboe, E. S., & Hauer, K. E. (2009). Tools for Direct Observation and Assessment of Clinical Skills of Medical Trainees : A Systematic Review. *JAMA*, *302*(12), 1316-1326

Kogan, J. R., Whelan, A. J., Gruppen, L. D., Lingard, L. A., Teunissen, P. W., & Ten Cate, O. (2018). What Regulatory Requirements and Existing Structures Must Change If Competency-Based, Time-Variable Training Is Introduced Into the Continuum of Medical Education in the United States? *Academic Medicine*, *93*(3S) S27-S31

Kolb, D. A. (1983). *Experiential Learning : Experience as the Source of Learning and Development* (1re éd.). Englewood Cliffs: Prentice Hall

Kolbe, M., & Grande, B. (2015). Briefing and debriefing during simulation-based training and beyond : Content, structure, attitude and setting. *Best Practice & Research Clinical Anaesthesiology*, *29*(1), 87-96

Koster, M. A., & Soffler, M. (2021). Navigate the Challenges of Simulation for Assessment : A Faculty Development Workshop. *MedEdPORTAL: The Journal of Teaching and Learning Resources*, *17*, 11114

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, *64*(1), 96-106

Krupat, E., & Dienstag, J. L. (2009). Commentary : Assessment Is an Educational Tool. *Academic Medicine*, *84*(5), 548-550

Landeta, J. (2006). Current validity of the Delphi method in social sciences. *Technological Forecasting and Social Change*, *73*(5), 467-482

Langevin, S., & Hivon, R. (2007). En quoi l'externat ne s'acquitte-t-il pas adéquatement de son mandat pédagogique ? Une étude qualitative fondée sur une analyse systématique de la littérature. *Pédagogie médicale*, *8*(1), 7-23

Laveault, D. (2008). *Chapitre 9.* Mesure critériée et théorie de la généralisabilité : Applications aux mesures cognitives. In Grégoire, J (Ed.). *Evaluer les apprentissages. Les apports de la psychologie cognitive* (Vol. 2). Bruxelles: De Boeck Supérieur

Le Boterf, G. (1994). *De la compétence : Essai sur un attracteur étrange (3ème).* Paris: Les Editions d'Organisation

Lemaitre, D., & Hatano, M. (2007). *Usages de la notion de compétence en éducation et formation.* France : L'Harmattan

Lemenu, D., & Heinen, E. (2015). *Comment passer des compétences à l'évaluation des acquis des étudiants*. Bruxelles : De Boeck Supérieur

L'expérience : Journal de médecine et de chirurgie. (1842). (2)

Lighthall, G. K., Poon, T., & Harrison, T. K. (2010). Using in situ simulation to improve in-hospital cardiopulmonary resuscitation. *Joint Commission Journal on Quality and Patient Safety*, *36*(5), 209-216

Lingard, L., Albert, M., & Levinson, W. (2008). Grounded theory, mixed methods, and action research. *BMJ*, *337*, a567

Lorello, G. R., Cook, D. A., Johnson, R. L., & Brydges, R. (2014). Simulation-based training in anaesthesiology : A systematic review and meta-analysis. *British Journal of Anaesthesia*, *112*(2), 231-245

Lucey, C. R., Thibault, G. E., & Ten Cate, O. (2018). Competency-Based, Time-Variable Education in the Health Professions : Crossroads. *Academic Medicine*, *93*(3S) S1-S5

Lurie, S. J., Mooney, C. J., & Lyness, J. M. (2009). Measurement of the General Competencies of the Accreditation Council for Graduate Medical Education : A Systematic Review: *Academic Medicine*, *84*(3), 301-309

Maignan, M., Koch, F.-X., Chaix, J., Phellouzat, P., Binauld, G., Collomb Muret, R., Cooper, S. J., Labarère, J., Danel, V., Viglino, D., & Debaty, G. (2016). Team Emergency Assessment Measure (TEAM) for the assessment of non-technical skills during resuscitation : Validation of the French version. *Resuscitation*, *101*, 115-120

Mann, K. V. (2011). Theoretical perspectives in medical education : Past experience and future possibilities: Pedagogy: past and future. *Medical Education*, *45*(1), 60-68

Marriage, B., & Kinnear, J. (2016). Assessing team performance – Markers and methods. *Trends in Anaesthesia and Critical Care, 7-8,* 11-16

Martinez, M., Duchenne, J., Bobbia, X., Brunet, S., Fournier, P., Miroux, P., Perrier, C., Pès, P., Chauvin, A., & Claret, P.-G. (2018). Deuxième niveau de compétence pour l'échographie clinique en médecine d'urgence. Recommandations de la Société française de médecine d'urgence par consensus formalisé. *Annales françaises de médecine d'urgence*, *8*(3), 193-202

McEvoy, M. D., Dewaay, D. J., Vanderbilt, A., Alexander, L. A., Stilley, M. C., Hege, M. C., & Kern, D. H. (2014). Are fourth-year medical students as prepared to manage unstable patients as they are to manage stable patients? *Academic Medicine*, *89*(4), 618-624

McGaghie, W. C. (2010). Medical Education Research As Translational Science. *Science Translational Medicine*, *2*(19), 19cm8

McGaghie, W. C., Draycott, T. J., Dunn, W. F., Lopez, C. M., & Stefanidis, D. (2011). Evaluating the impact of simulation on translational patient outcomes. *Simulation in Healthcare*, *6* (7), S42-S47

McGaghie, W. C., & Harris, I. B. (2018). Learning Theory Foundations of Simulation-Based Mastery Learning. *Simulation in Healthcare*, *13*(3S), S15-S20

McGaghie, W. C., Issenberg, S. B., Barsuk, J. H., & Wayne, D. B. (2014). A critical review of simulation-based mastery learning with translational outcomes. *Medical Education*, *48*(4), 375-385

McGaghie, W. C., Issenberg, S. B., Cohen, E. R., Barsuk, J. H., & Wayne, D. B. (2012). Translational educational research : A necessity for effective health-care improvement. *Chest*, *142*(5), 1097-1103

McGaghie, W. C., Sajid, A. W., Miller, G. E., Telder, T. V., Lipson, L., & Organization, W. H. (1978). *Competency-based curriculum development in medical education : An introduction*. World Health Organization. [en ligne], consulté le 2 février 2021. https://apps.who.int/iris/handle/10665/39703

McLaughlin, S., Fitch, M. T., Goyal, D. G., Hayden, E., Kauh, C. Y., Laack, T. A., Nowicki, T., Okuda, Y., Palm, K., Pozner, C. N., Vozenilek, J., Wang, E., Gordon, J. A., & on behalf of the SAEM Technology in Medical Education Committee and the Simulation Interest Group. (2008).

Simulation in Graduate Medical Education 2008 : A Review for Emergency Medicine. *Academic Emergency Medicine*, *15*(11), 1117-1129

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine: Journal of the Association of American Medical Colleges*, *65*(9 Suppl), S63-67

Moll, W. (1968). History of American Medical Education. *Medical Education*, *2*(3), 173-181

Moll-Khosrawi, P., Zöllner, C., Cronje, J. S., & Schulte-Uentrop, L. (2021). The effects of simulation-based education on medical students' motivation. *International Journal of Medical Education*, *12*, 130-135

Mortaz Hejri, S., Jalili, M., Masoomi, R., Shirazi, M., Nedjat, S., & Norcini, J. (2020). The utility of mini-Clinical Evaluation Exercise in undergraduate and postgraduate medical education : A BEME review: BEME Guide No. 59. *Medical Teacher*, *42*(2), 125-142

Motola, I., Devine, L. A., Chung, H. S., Sullivan, J. E., & Issenberg, S. B. (2013). Simulation in healthcare education : A best evidence practical guide. AMEE Guide No. 82. *Medical Teacher*, *35*(10), e1511-1530

Mundell, W. C., Kennedy, C. C., Szostek, J. H., & Cook, D. A. (2013). Simulation technology for resuscitation training : A systematic review and meta-analysis. *Resuscitation*, *84*(9), 1174-1183

Murphy, M. K., Black, N. A., Lamping, D. L., McKee, C. M., Sanderson, C. F., Askham, J., & Marteau, T. (1998). Consensus development methods, and their use in clinical guideline development. *Health Technology Assessment*, *2*(3), i-iv, 1-88

Musselin, C., Froment, E., & Ottenwaelter, M.-O. (2007). Le Processus de Bologne : Quels enjeux européens ?. Un entretien avec Christine Musselin et Eric Froment. *Revue internationale d'éducation de Sèvres*, *45*, 99-110

Nemitz B. (2005). L'évolution de l'enseignement de la médecine d'urgence jusqu'à la naissance du DESC. *Annales françaises de médecine d'urgence*,*28*, 329-332

Nemitz, B., Carli, P., Carpentier, F., Ducassé, J.-L., Giroud, M., Pateron, D., Pelloux, P., Riou, B., & Schmidt, J. (2012). Référentiel métier-compétences pour la spécialité de médecine d'urgence. *Annales françaises de médecine d'urgence*, *2*(2), 125-138

Newble, D. (2004). Techniques for measuring clinical competence : Objective structured clinical examinations. *Medical Education*, *38*(2), 199-203

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

Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., Galbraith, R., Hays, R., Kent, A., Perrott, V., & Roberts, T. (2011). Criteria for good assessment : Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, *33*(3), 206-214

Norcini, J., Anderson, M. B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., Hays, R., Palacios Mackay, M. F., Roberts, T., & Swanson, D. (2018). 2018 Consensus framework for good assessment. *Medical Teacher*, *40*(11), 1102-1109

Norcini, J., & Burch, V. (2007). Workplace-based assessment as an educational tool : AMEE Guide No. 31. *Medical Teacher*, *29*(9), 855-871

Norris, T. E., Schaad, D. C., DeWitt, D., Ogur, B., Hunt, D. D., & Consortium of Longitudinal Integrated Clerkships. (2009). Longitudinal integrated clerkships for medical students : An innovation adopted by medical schools in Australia, Canada, South Africa, and the United States. *Academic Medicine*, *84*(7), 902-907

Nousiainen, M. T., Caverzagie, K. J., Ferguson, P. C., Frank, J. R., & ICBME Collaborators. (2017). Implementing competency-based medical education : What changes in curricular structure and processes are needed? *Medical Teacher*, *39*(6), 594-598

O'Donnell, J. M., Goode, J. S., Henker, R., Kelsey, S., Bircher, N. G., Peele, P., Bradle, J., Close, J., Engberg, R., & Sutton-Tyrrell, K. (2011). Effect of a simulation educational intervention on knowledge, attitude, and patient transfer skills : From the simulation laboratory to the clinical setting. *Simulation in Healthcare, 6*(2), 84-93

Okuda, Y., Bond, W., Bonfante, G., McLaughlin, S., Spillane, L., Wang, E., Vozenilek, J., & Gordon, J. A. (2008). National growth in simulation training within emergency medicine residency programs, 2003-2008. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, *15*(11), 1113-1116

Olivier de Sardan. (1989). Le réel des autres. *Cahiers d'études africaines*, *29*(113), 127-135

Olry, P. (1995). *OLRY P*, *La* formation aux prises avec le travail réel dans une industrie de process*. Éducation permanente,(122),*1993-2

O'Neill, P., Baxter, C. M., & Morris, J. (1999). Does awarding a medical degree with honours act as a motivator or demotivator to student learning? *Medical Education*, *33*(8), 566-571

Oriot, D., Darrieux, E., Boureau-Voultoury, A., Ragot, S., & Scépi, M. (2012). Validation of a performance assessment scale for simulated intraosseous access. *Simulation in Healthcare*, *7*(3), 171-175

Pangaro, L., & Ten Cate, O. (2013). Frameworks for learner assessment in medicine : AMEE Guide No. 78. *Medical Teacher*, *35*(6), e1197-1210

Parent, R. J., Plerhoples, T. A., Long, E. E., Zimmer, D. M., Teshome, M., Mohr, C. J., Ly, D. P., Hernandez-Boussard, T., Curet, M. J., & Dutta, S. (2010). Early, Intermediate, and Late Effects of a Surgical Skills "Boot Camp" on an Objective Structured Assessment of Technical Skills : A Randomized Controlled Study. *Journal of the American College of Surgeons*, *210*(6), 984-989.

Pastré, P. (2011). *La didactique professionnelle* (1re éd.). Paris: Presses Universitaires de France

Pastré, P., Mayen, P., & Vergnaud, G. (2006). La didactique professionnelle. *Revue française de pédagogie. Recherches en éducation*, *154*, 145-198

Patterson, M. D., Geis, G. L., Falcone, R. A., LeMaster, T., & Wears, R. L. (2013). In situ simulation : Detection of safety threats and teamwork training in a high risk emergency department. *BMJ Quality & Safety*, *22*(6), 468-477

Pelaccia, T. (2016). *Comment mieux former et évaluer les étudiants en médecine et en sciences de la santé* ? Bruxelles: De Boeck Supérieur

Pelaccia, T., Delplancq, H., Triby, E., Leman, C., Bartier, J.-C., & Dupeyron, J.-P. (2008). La motivation en formation : Une dimension réhabilitée dans un environnement d'apprentissage en mutation. *Pédagogie médicale*, *9*(2), 103-121

Pelaccia, T., & Triby, E. (2011). La pédagogie médicale est-elle une discipline ? *Pédagogie Médicale*, *12*(2), 121-132

Pelaccia, T., & Viau, R. (2017). Motivation in medical education. *Medical Teacher*, *39*(2), 136-140

Perrenoud, P. (2002). D'une métaphore à l'autre : Transférer ou mobiliser ses connaissances ? In Dolz, J., Ollagnier E. *L'énigme de la compétence en éducation*. Bruxelles: De Boeck Supérieur, p. 45-60

Petitcolas. (2006). Le mannequin de Mme du Coudray ou comment former les accoucheuses au XVIIIe siècle. *La Revue du Praticien*, *56,* 226-229

Petrosoniak, A., Auerbach, M., Wong, A. H., & Hicks, C. M. (2017). In situ simulation in emergency medicine : Moving beyond the simulation lab. *Emergency Medicine Australasia, 29*(1), 83-88

Philibert, I. (2018). Using Chart Review and Chart-Stimulated Recall for Resident Assessment. *Journal of Graduate Medical Education*, *10*(1), 95-96

Philippon, A.-L. (2017). Evaluation des compétences des étudiants de 4ème année de médecine dans la prise en charge des urgences vitales : Quelle place pour la simulation ? Adjectif.net, [en ligne], consulté le 3 janvier 2018. http://www.adjectif.net/spip. http://www.adjectif.net/spip/spip.php?article451&lang=fr

Philippon, A.-L., Truchot, J., De Suremain, N., Renaud, M.-C., Petit, A., Baron, G.-L., & Freund, Y. (2021). Medical students' perception of simulation-based assessment in emergency and paediatric medicine : A focus group study. *BMC Medical Education*, *21*(1), 586

Philippon, S. (2015). Le juste soin. *Perspective soignante, 54(12), 25-33*

Picard, J.-F., & Mouchet, S. (2009). *La métamorphose de la médecine*. Paris: Presses Universitaires de France

Plake, B. S., & Impara, J. C. (1997). Standard Setting : An Alternative Approach. *Journal of Educational Measurement*, *34*(4), 353-366

Postiaux, N., Bouillard, P., & Romainville, M. (2010). Référentiels de compétences à l'université. Usages, rôles et limites. *Recherche et formation*, *64*, 15-30

Pottier, P. (2013). Théories de l'apprentissage et simulation Le point de vue du professionnel de santé-enseignant. In S. Boet, G. Savoldelli, & J.-C. Granry (Éds.), *La simulation en santé De la théorie à la pratique*. Paris: Springer, p 15-14

Powell, D. E., & Carraccio, C. (2018). Toward Competency-Based Medical Education. *New England Journal of Medicine*, *378*(1), 3-5

Rall, M., Manser, T., & Howard, S. K. (2000). Key elements of debriefing for simulator training. *European Journal of Anaesthesiology*, *17*(8), 516-517

Reader, T. W., Flin, R., & Cuthbertson, B. H. (2007). Communication skills and error in the intensive care unit. *Current Opinion in Critical Care*, *13*(6), 732-736

Regehr, G., MacRae, H., Reznick, R. K., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine: Journal of the Association of American Medical Colleges*, *73*(9), 993-997

Reznick, R., Regehr, G., MacRae, H., Martin, J., & McCulloch, W. (1997). Testing technical skill via an innovative « bench station » examination. *American Journal of Surgery*, *173*(3), 226-230

Ricker, K.L. (2006). Setting Cut-Scores : A Critical Review of the Angoff and Modified Angoff Methods. *The Alberta Journal of Educational Research*, *52*(1), 53-64

Riou, B. (2016). 13 novembre 2015 : Terrorisme, résilience, et espoir. *Annales françaises de médecine d'urgence*, *6*(1), 1-2

Riou, B. (2017). 2017 : L'an 1 du diplôme d'études spécialisées de médecine d'urgence. *Annales françaises de médecine d'urgence*, *7*(1), 1-4

Riou, B., Carli, P., Carpentier, F., Kopferschmitt, J., Conte, P. L., Lauque, D., Levraut, J., & Veber, B. (2014). Combien formons-nous de médecins urgentistes en France ? *Annales Françaises de Médecine d'urgence*, *4*(1)

Roegiers, X. (2000). *Une pédagogie de l'intégration : Compétences et intégration des acquis dans l'enseignement*. Paris: De Boeck

Roegiers Xavier. (2012). *Quelles réformes pédagogiques pour l'enseignement supérieur ? Placer l'efficacité au service de l'humanisme.* Avec la collaboration de Miled Mohamed, Ratziu Ioan, de Letor Caroline, de Étienne Richard, Hubert Gaëlle, de Dali Mohamed et al. Bruxelles: De Boeck Supérieur

Rohe, D. E., Barrier, P. A., Clark, M. M., Cook, D. A., Vickers, K. S., & Decker, P. A. (2006). The benefits of pass-fail grading on stress, mood, and group cohesion in medical students. *Mayo Clinic Proceedings*, *81*(11), 1443-1448

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

Romainville, M., Goasdoué, R., & Vantourout, M. (2013). *Evaluation et enseignement supérieur.* Bruxelles: De Boeck Supérieur

Rothschild, J. M., Landrigan, C. P., Cronin, J. W., Kaushal, R., Lockley, S. W., Burdick, E., Stone, P. H., Lilly, C. M., Katz, J. T., Czeisler, C. A., & Bates, D. W. (2005). The Critical Care Safety Study : The incidence and nature of adverse events and serious medical errors in intensive care. *Critical Care Medicine*, *33*(8), 1694-1700

Sargeant, J., Eva, K. W., Armson, H., Chesluk, B., Dornan, T., Holmboe, E., Lockyer, J. M., Loney, E., Mann, K. V., & van der Vleuten, C. P. M. (2011). Features of assessment learners use to make informed self-assessments of clinical performance : Informed self-assessment in learners. *Medical Education*, *45*(6), 636-647

Sawaya, R. D., Mrad, S., Rajha, E., Saleh, R., & Rice, J. (2021). Simulation-based curriculum development : Lessons learnt in Global Health education. *BMC Medical Education*, *21*(1), 33

Scallon, G. (1999). L'évaluation sommative et ses rôles multiples. [en ligne], consulté le 2 avril 2017. https://mobile.eduq.info/xmlui/handle/11515/34522?locale-attribute=en

Scallon, G. (2007). *L'évaluation des apprentissages dans une approche par compétences.* Bruxelles: De Boeck Supérieur

Scallon, G. (2015). *Des savoirs aux compétences : Explorations en évaluation des apprentissages*. Bruxelles: De Boeck Supérieur

Schartel, S. A., & Metro, D. G. (2010). Evaluation : Measuring performance, ensuring competence, achieving long-term excellence. *Anesthesiology*, *112*(3), 519-520

Schmidt, H. G., Rotgans, J. I., & Yew, E. H. (2011). The process of problem-based learning : What works and why: What works and why in problem-based learning. *Medical Education*, *45*(8), 792-806

Schulte-Uentrop, L., Cronje, J. S., Zöllner, C., Kubitz, J. C., Sehner, S., & Moll-Khosrawi, P. (2020). Correlation of medical students' situational motivation and performance of non-technical skills during simulation-based emergency training. *BMC Medical Education*, *20*(1), 351

Schuwirth, L. W. T., & Van der Vleuten, C. P. M. (2011). Programmatic assessment : From assessment of learning to assessment for learning. *Medical Teacher*, *33*(6), 478-485

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2006). A plea for new psychometric models in educational assessment. *Medical Education*, *40*(4), 296-300

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2011). General overview of the theories used in assessment : AMEE Guide No. 57. *Medical Teacher*, *33*(10), 783-797

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2020). A history of assessment in medical education. *Advances in Health Sciences Education: Theory and Practice*, *25*(5), 1045-1056

Schwartz, C. E., Powell, V. E., & Rapkin, B. D. (2017). When global rating of change contradicts observed change : Examining appraisal processes underlying paradoxical responses over time. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *26*(4), 847-857

Sevdalis, N., & Brett, S. J. (2009). Improving care by understanding the way we work : Human factors and behavioural science in the context of intensive care. *Critical Care*, *13*(2), 139

Sevdalis, N., Davis, R., Koutantji, M., Undre, S., Darzi, A., & Vincent, C. A. (2008). Reliability of a revised NOTECHS scale for use in surgical teams. *The American Journal of Surgery*, *196*(2), 184-190

Sherbino, J. & Frank J.R. (2011). *Les tremplins d'enseignement de CanMEDS. Médecine d'urgence*.[en ligne], consulté le 23 mars 2020. https://www.royalcollege.ca

Shrivastava, S., Shrivastava, P., & Ramasamy, J. (2013). Problem-based learning : Constructivism in medical education. *Education for Health*, *26*(3), 197

SoFraSimS. (2021). *Guide SoFraSimS – Évaluation Sommative et Simulation en Santé – Texte Court*. [en ligne], consulté le 2 juin 2021. https://sofrasims.org/evaluation-sommative-et-simulation-en-sante/

Stahl, K., Palileo, A., Schulman, C. I., Wilson, K., Augenstein, J., Kiffin, C., & McKenney, M. (2009). Enhancing patient safety in the trauma/surgical intensive care unit. *The Journal of Trauma*, *67*(3), 430-433; discussion 433-435

Stolarova, M., Wolf, C., Rinker, T., & Brielmann, A. (2014). How to assess and compare inter-rater reliability, agreement and correlation of ratings : An exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. *Frontiers in Psychology*, *5* [en ligne], consulté le 12 avril 2018. https://doi.org/10.3389/fpsyg.2014.00509

Tardif, J., Fortier, G., & Préfontaine, C. (2006). *L'évaluation des compétences : Documenter le parcours de développement*. Montréal : Chenelière Éducation

Tardif, M., & Lessard, C. (1999). *Le travail enseignant au quotidien : Contribution à l'étude du travail dans les métiers et les professions d'interactions humaines*. Québec: Presses de l'Université Laval

Tavares, W., Boet, S., Theriault, R., Mallette, T., & Eva, K. W. (2013). Global Rating Scale for the Assessment of Paramedic Clinical Competence. *Prehospital Emergency Care*, *17*(1), 57-67

Ten Cate, O. (2005). Entrustability of professional activities and competency-based training. *Medical Education*, *39*(12), 1176-1177

Ten Cate, O. (2017). Competency-Based Postgraduate Medical Education : Past, Present and Future. *GMS Journal for Medical Education*, *34*(5)

Tofil, N. M., Dollar, J., Zinkan, L., Youngblood, A. Q., Peterson, D. T., White, M. L., Stooksberry, T. N., Jarrell, S. A., & King, C. (2014). Performance of anesthesia residents during a simulated prone ventricular fibrillation arrest in an anesthetized pediatric patient. *Pediatric Anesthesia*, *24*(9), 940-944

Touchie, C., & Ten Cate, O. (2016). The promise, perils, problems and progress of competency-based medical education. *Medical Education*, *50*(1), 93-100

Townend, W., Long, J., Munro-Davies, L., & Beet, E. (2018). How do we educate the next generation of emergency physicians : RCEM 50. *Emergency Medicine Journal, 35*(3), 159-161

Turoff, M., & Linstone, H. A. (Eds.). (2002). *The Delphi Method : Techniques and Applications*. [en ligne] consulté le 25 février 2019. https://web.njit.edu/~turoff/pubs/delphibook/index.html

Van Der Vleuten, C. P. M. (1996). The assessment of professional competence : Developments, research and practical implications. *Advances in Health Sciences Education*, *1*(1), 41-67

Van der Vleuten, C. P., Norman, G. R., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity : Issues of reliability. *Medical Education*, *25*(2), 110-118

Van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, *34*(3), 205-214

Vial, M. (2012). *Se repérer dans les modèles de l'évaluation. Méthodes—Dispositifs—Outils.* Bruxelles: De Boeck Supérieur

Vincent, C., & Amalberti, R. (2015). Safety in healthcare is a moving target. *BMJ Quality & Safety*, *24*(9), 539-540

Walker, S., Brett, S., McKay, A., Lambden, S., Vincent, C., & Sevdalis, N. (2011). Observational Skill-based Clinical Assessment tool for Resuscitation (OSCAR) : Development and validation. *Resuscitation*, *82*(7), 835-844

Wang, C. J., Lin, S. Y., Tsai, S. H., & Shan, Y. S. (2019). Implications of long-term low-fidelity in situ simulation in acute care and association with a reduction in unexpected cardiac arrests : A retrospective research study. *PloS One*, *14*(3), e0213789

Wheeler, D. S., Geis, G., Mack, E. H., LeMaster, T., & Patterson, M. D. (2013). High-reliability emergency response teams in the hospital : Improving quality and safety using in situ simulation training. *BMJ Quality & Safety*, *22*(6), 507-514

Wiel, E., Nunes, F., Cluis, E., & Lebuffe, G. (2013). Intérêts et limites de la simulation pour l'évaluation certificative des professionnels de santé. In *S. Boet, J.-C. Granry, G. Salvodelli. La simulation en santé : De la théorie à la pratique.* Paris: Springer

Wiggins, Grant. (1990). The case for authentic assessment. *Practical Assessment, Research & Evaluation*, *2*(2), [en ligne], consulté le 5 mai 2017. https://scholarworks.umass.edu/pare/vol2/iss1/2

Wragg, A., Wade, W., Fuller, G., Cowan, G., & Mills, P. (2003). Assessing the performance of specialist registrars. *Clinical Medicine (London, England)*, *3*(2), 131-134

Xi, H., Liu, J., & Gui, L. (2015). Knowledge of fresh-graduated medical students for advanced cardiovascular life support in China. *Resuscitation*, *92*, e13

Young, S., Dunipace, D., Pukenas, E., & Pawlowski, J. (2019). Can Simulation Improve Patient Outcomes? *International Anesthesiology Clinics*, *57*(3), 68-77

Yudkowsky, R., Downing, S. M., & Popescu, M. (2008). Setting standards for performance tests : A pilot study of a three-level Angoff method. *Academic Medicine: Journal of the Association of American Medical Colleges*, *83*(10 Suppl), S13-16

Yudkowsky, R., Park, Y. S., & Downing, S. M. (Éds.). (2019). *Assessment in Health Professions Education* (2e éd.). New York: Routledge

ARTICLES

1. PREMIER ARTICLE

Titre : **Evaluer les compétences des étudiants en médecine d'urgence grâce au développement de trois scores: une étude nationale par la méthode Delphi.**

Anne-Laure Philippon (1, 2), P. Hausfater (1, 3), E.Triby (2), Y.Freund (1, 3)

1) Service d'accueil des Urgences, Centre Hospitalier Universitaire Pitié-Salpêtrière, Assistance Publique – Hôpitaux de Paris (APHP), Paris, France.

2) Laboratoire Interuniversitaire des Sciences de l'Éducation et de la Communication LISEC – EA 2310, Université de Strasbourg, France.

3) Sorbonne Université, UPMC Paris Univ-06, Paris, France.

Publié en Novembre 2019, Annales Françaises de Médecine d'Urgence

DOI : https://doi.org/10.3166/afmu-2019-0199

2. DEUXIEME ARTICLE

Titre : **Usability and reproducibility of three tools to assess medical students and residents in emergency medicine**

AL Philippon[1,2], MD, A. Baud[3], M Dumont[2], MD, J, SA Remini[4], J. Leroy[5,] J Truchot MD[6], PhD, E Triby[1], PhD, Y Freund[2,5] MD, PhD.

1) Laboratoire Interuniversitaire de Sciences de l'Education (LISEC) – Learning Sciences Department, Université de Strasbourg, Strasbourg, France
2) Emergency department, Hôpital Pitié-Salpêtrière, Assistance Publique – Hôpitaux de Paris, Paris, France
3) Emergency department, Hôpital Tenon, Assistance Publique – Hôpitaux de Paris, Paris, France
4) Centre de Simulation ILumens Paris Diderot, Université de Paris, Paris.
5) Faculté de Médecine, Sorbonne Université, Paris, France.
6) Emergency Department, Hôpital Cochin, Assistance Publique – Hôpitaux de Paris, Paris, France

Accepté le 18/10/21 dans Academic Emergency Medicine Education and Training

DOI : https://doi.org/10.1002/aet2.10704

3. TROISIEME ARTICLE

Le résumé du manuscrit en cours de soumission est ici présenté.

Titre : Validation of three tools to Measure and Promote residents' clinical performance in Emergency Medicine: The Acute Care Assessment Tools (ACAT) validation study.

Auteurs : Philippon AL. (1, 2, 3), Lefèvre-Scelles A. (4), Eyer X. (5), Zumstein C. (6), Ghazali A. (7), Audibert S. (8), P.Leborgne (9), Y.Freund (1,3), E.Triby (2), J.Truchot (10)

And the FHU e-SimSit Study group: Guillaume Payan, Aurore Kolmer, Sarah Uge-Ginsberg, Clémence Bertrand, Pauline Canavaggio, Mélanie Roussel, Martin Behr, Alexandre Bitoun, Isabelle Borraccio, Walid, Anthony Chauvin.

7) Faculté de Médecine, Sorbonne Université, Paris, France
8) Laboratoire Interuniversitaire de Sciences de l'Education (LISEC) – Learning Sciences Department, Université de Strasbourg, Strasbourg, France
9) Emergency department, Hôpital Pitié-Salpêtrière, Assistance Publique – Hôpitaux de Paris, Paris, France
10) Rouen
11) Emergency Department, Hôpital Lariboisière, Assistance Publique – Hôpitaux de Paris, Paris, France
12) Faculté de Médecine, Université de Strasbourg, Strasbourg, France
13) Emergency Department, Hôpital Bichat, Assistance Publique – Hôpitaux de Paris, Paris, France
14) Emergency Department, Hôpital Européen Georges Pompidou, Assistance Publique – Hôpitaux de Paris, Paris, France
15) Emergency Department, Hôpitaux Universitaires de Strasbourg, Strasbourg, France
16) Emergency Department, Hôpital Cochin, Assistance Publique – Hôpitaux de Paris, Paris, France

ABSTRACT

Objectives: it is of upmost importance to assess competency of medical students and residents in the field of emergency medicine (EM). However, very few valid tools exist to assess both technical and non-technical skills in the specific context of EM. We previously developed the content and assess the reproducibility of three Acute Care Assessment Tools (ACAT1, 2 and 3) for three acute care conditions: cardiac arrest (1), coma (2) and acute respiratory failure (3). This study aimed to evaluate the validity, reliability and usability of the tools.

Methods: To ensure reliability, validity, usability, we conducted a prospective study which assessed the three ACAT's psychometrics data thanks to in-situ simulation interprofessional sessions in 7 emergency departments. We aimed to complete at least 30 simulation sessions for each ACAT, scored by two independent raters. The raters were not always the same. Intraclass correlation coefficients (ICC) assessed inter-rater reliability (IRR), and Cronbach's alpha coefficient assessed inter-item score correlation in order to analyze the internal consistency of each ACAT. We also assessed the correlation between the ACAT's results and the learner level of performance, according to their residency year. Finally, the ACAT usability were assessed through a survey questionnaire and two focus groups.

Results: With an analyze of 104 simulations sessions for 314 learners rated by 37 raters, we found a good reliability with all ICC superior to 0.89 and all Cronbach's alpha superior to 0.73. The ACAT 1 and 3 had a good predictive validity ($p< 0.001$; $p< 0.019$) although it was not found for ACAT 2. Raters found the three ACAT scores as valid and usable tool, even if they had some concerns about the use of the SBA in a summative way.

Conclusion: Three valid and reliable tools were developed to measure the learners' performance in three acute care conditions. Even if there is future step to pursue the validation process of the ACAT, their high reliability and validity advocated for good usability. The three ACAT can be utilized to assess for completeness of predefined tasks in three acute care broad scenario in a competency-based medical education framework.

ANNEXES

Annexe 1 – Article préliminaire au travail de la thèse

Philippon *et al. BMC Medical Education* (2021) 21:586
https://doi.org/10.1186/s12909-021-02957-5

BMC Medical Education

RESEARCH ARTICLE Open Access

Medical students' perception of simulation-based assessment in emergency and paediatric medicine: a focus group study

Anne-Laure Philippon[1,2,3*], Jennifer Truchot[3,4], Nathalie De Suremain[5], Marie-Christine Renaud[2], Arnaud Petit[2,6], Georges-Louis Baron[3] and Yonathan Freund[1,2]

Abstract

Background: Although simulation-based assessment (SBA) is being implemented in numerous medical education systems, it is still rarely used for undergraduate medical students in France. Objective structured clinical examinations (OSCEs) will be integrated into the national medical curriculum in 2021. In 2016 and 2017, we created a mannequin SBA to validate medical students' technical and psychometric skills during their emergency medicine and paediatric placements. The aim of our study was to determine medical students' perceptions of SBA.

Methods: We followed the grounded theory framework to conduct a qualitative study. A total of 215 students participated in either a paediatric or an emergency medicine simulation-based course with a final assessment. Among the 215 participants, we randomly selected forty students to constitute the focus groups. In the end, 30 students were interviewed. Data were coded and analysed by two independent investigators within the activity theory framework.

Results: The analyses found four consensual themes. First, the students perceived that success in the SBA provided them with self-confidence and willingness to participate in their hospital placements (1). They considered SBA to have high face validity (2), and they reported changes in their practice after its implementation (3). Nevertheless, they found that SBA did not help with their final high-stakes assessments (4). They discussed three other themes without reaching consensus: stress, equity, and the structure of SBA. After an analysis with activity theory, we found that students' perceptions of SBA underlined the contradictions between two systems of training: hospital and medical. We hypothesise that a specific role and place for SBA should be defined between these two activity systems.

Conclusion: The students perceived that SBA would increase self-confidence in their hospital placements and emphasise the general skills required in their future professional environment. However, they also reported that the assessment method might be biased and stressful. Our results concerning a preimplementation mannequin SBA and OSCE could provide valuable insight for new programme design and aid in improving existing programmes. Indeed, SBA seems to have a role and place between hospital placements and medical schools.

Keywords: Education, Undergraduate medical, Assessment, Simulation in healthcare, Activity theory

* Correspondence: annelaurephi@gmail.com
[1]Emergency Department, Hôpital Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, 83, bd de l'hôpital, 75013 Paris, France
[2]Sorbonne Université, Paris, France
Full list of author information is available at the end of the article

Background

Following the recommendations of the Accreditation Council for Graduate Medical Education, competency-based medical education (CBME) principles have been widely implemented in most medical education systems [1]. Therefore, student assessments must meet the requirements of a competency-based approach even though it remains a barrier to CBME development [2]. Competencies can feel abstract while being context dependent, resulting in difficulties in finding meaningful assessment tools [2, 3]. As underlined by Carraccio, assessment remains the "Achilles' heel" of CBME, even though significant progress has been made in this field [3–5]. Consequently, medical education systems face the challenge of evaluating competencies through validated, high-stakes assessments, such as the already existing NBME licensing board assessments [4, 6–8].

To achieve high-quality assessment, multiple modalities can be used to assess competencies: direct observation, multisource feedback, and simulation [4, 9]. CBME creates an opportunity for the simulation community to participate in a competency-based assessment system, both as formative (assessment for learning) or summative (assessment of learning) assessments [10, 11]. However, simulation-based assessment (SBA) offers a semiauthentic, complex environment and the opportunity to practice a full range of clinical skills without exposing real patients to any risks. For those reasons, it appears to be a suitable tool in the field of emergency medicine (EM), where health care providers manage rare and critical conditions [8, 12, 13].

The use of an objective structured clinical examination (OSCE) or mannequin-based simulations for summative assessments has emerged for postgraduate practice and interprofessional training, and their feasibility and acceptability have been demonstrated [14–17]. However, it is still not routinely used for medical students in EM. Moreover, a recent Canadian study underlined the need for research on the role and the optimal way to incorporate high-stakes summative SBAs in EM training [18]. In France, OSCE or SBA uses remain unusual and varied, but in 2022, a new curriculum reform will implement an OSCE for medical students. When this research began, the students had never participated in either OSCE or mannequin-based simulation.

Thus, we developed two mannequin SBAs within the emergency medicine and paediatric curriculum of one medical school. However, although existing research addresses how to develop and use mannequin SBA, there remains a gap regarding learners' perspectives [19, 20]. Learners' reactions and perceptions of assessment could impact their engagement in the learning and assessment processes; therefore, this issue should be taken into consideration [18, 21, 22]. Therefore, we aimed to collect information on medical students' perceptions of these new assessments.

Methods

We conducted prospective qualitative research after validation of the protocol by the French Society of Intensive Care Medicine (SRLF) ethical committee (n°16–55). All focus group participants received written and oral information and signed an informed consent form. The study method reporting followed the COREQ framework, which is a 32-item checklist generated from a systematic literature review to help authors report on qualitative studies [23].

Characteristics of the research team and reflexivity

The main investigator (ALP) has a master's degree in learning sciences and previous experience with focus group interviews. YF, JT and ALP are graduated simulation trainers and attending physicians in the emergency department. MCR is an internist doctor working in a medical education department. AP and NdS are paediatricians and trainers in the paediatrics simulation curriculum. To enhance the credibility of the results, we worked with an outside expert, GLB, who is a learning sciences professor. Before designing the study, the main investigator analysed some bias linked to her representations of simulation-based training, assessment and emergency medicine. The aim of this process was to identify pitfalls in the field, such as assumptions and beliefs regarding SBA, and to acknowledge their potential influence. Because JT also analysed the data, she underwent the same process. The main investigator introduced herself as a learning sciences student and an emergency physician conducting a research project in the medical education domain.

Study design and theoretical framework

To understand the medical students' perception of SBA, the grounded theory approach was used to produce emergent themes and theories, as we did not have a "preconceived theory in mind" [24–26]. With this approach, the theories emerge from the data and could be analysed in regard to another theory. The focus group method was chosen to foster discussions between participants and generate point-counterpoint discussion [27].

Setting: description of the simulation-based courses and assessments

The undergraduate medical curriculum lasts 6 years in France. At the end of the sixth year, undergraduate students undergo a national high-stakes assessment serving a classification purpose, which allows them to choose both a specialty and their residency university. During the final 3 years, medical students divide their time

between hospital placements and faculty courses (including lectures, tutorials or simulation-based training). The simulation courses followed a traditional structure (prebriefing, briefing, scenario, debriefing).

The study took place in a single medical school affiliated with 4 teaching hospitals and 18 urban hospitals within Paris Sorbonne University in Paris, France.

Simulation course in the emergency medicine and intensive care medicine curriculum (EM-ICMC)

Fourth-year medical students participated in two three-hour simulation-based courses. During each course, two or three students had to participate in one scenario lasting 8 to 10 min. The summative SBA took place during a third simulation session. The SBA started with a collective prebriefing, followed by medical students participating in pairs, in one EM scenario (Fig. 1). The debriefing took place in two stages: immediately after the scenario for the two "assessed" medical students and at the end of the "assessment session" with all the students. Assessors used specific assessment scores developed for each clinical case, as none existed to assess medical students in emergency medicine. The scores assessed medical students' technical and nontechnical skills. Medical students had to complete two requirements to succeed: a grade higher than 10/20 and the completion of all the mandatory items (4 to 6 among 20 according to the scores).

Simulation-based assessment in the paediatric curriculum (PC) (Fig. 2)

Fifth-year medical students participated in a three-hour simulation-based course that included three sessions on paediatric basic and advanced life support, followed by an individual SBA of a paediatrics basic-life support clinical case. The SBA took place immediately after the end of the simulation-based courses. A single assessor assessed each student's paediatric basic life support performance using a score derived from the ILCOR guidelines (Additional file 3: Annex 1, [28]).

For both SBA (paediatrics and EM), when the medical students failed, they had to undergo another assessment session. If they failed again, they had to take the entire simulation course again.

Participants' selection and data collection

A total of 125 and 90 fourth- and fifth-year medical students, respectively, participated in the two curricula. There weren't minor (< 18 years-old) students. From the overall cohort, we randomly selected forty medical students, who received an email invitation to discuss their simulation-based courses' perceptions before and after the SBA.

Constructed by ALP and JT, the interview guide used semistructured methods with predetermined, open-ended questions. It was pilot tested with voluntary non-participating students to ensure that the questions were appropriate and clear (Additional file 2: Appendix 2). Before and after the SBA proceedings, the focus groups aimed to explore the medical students' anticipation towards the SBA and to evaluate their SBA's perceptions. The focus groups took place at the medical school or the hospital according to the participants' preferences. ALP moderated all the focus groups as a facilitator and made field notes on relevant moments and on medical students' attitudes. The focus groups were audio

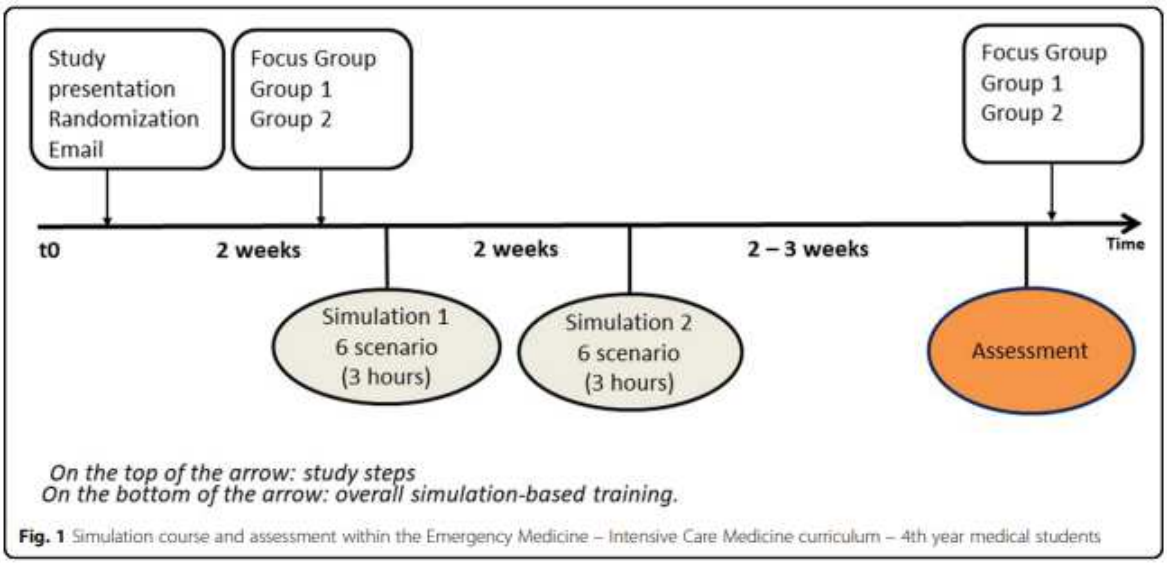


Fig. 1 Simulation course and assessment within the Emergency Medicine – Intensive Care Medicine curriculum – 4th year medical students

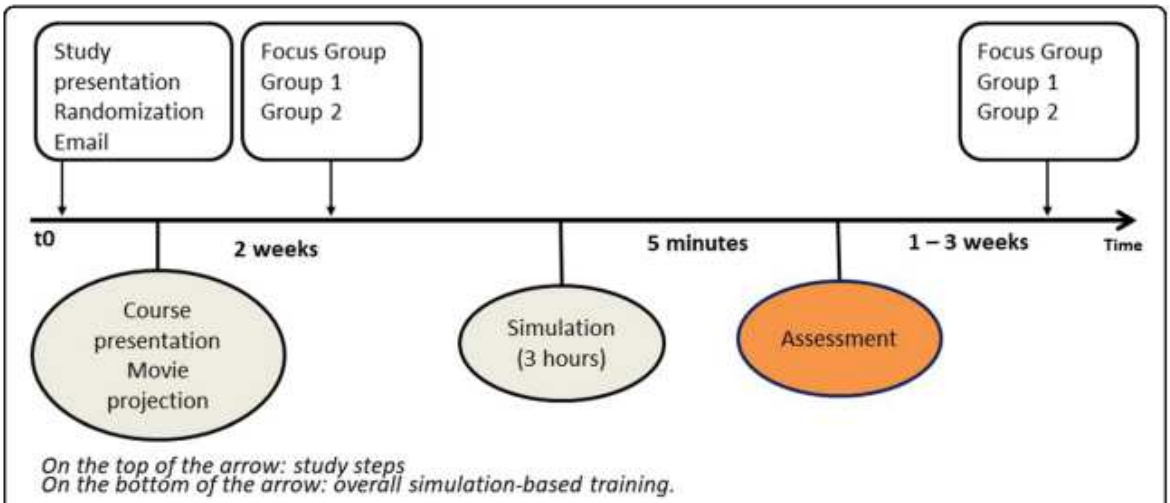


Fig. 2 Simulation course and assessment within the Pediatric Curriculum – 5th year medical students

recorded, downloaded onto a computer for storage, and then transcribed verbatim. Two native English-speaking individuals, independent from the study, translated the quotes from French to English.

Data management and analysis

Two investigators (ALP, JT) separately analysed the transcripts. With respect to the grounded theory approach, the interview transcription took place immediately after the data collection and was analysed before the next focus group using the constant comparison method [25]. After the SBA, focus groups ran on until no new themes emerged from the data. This process allowed the identification of themes and new questions or indicated theoretical saturation. The two curricula were analysed separately, as they occurred during two different periods. Focus groups were open-coded following those stages: familiarization with the transcripts with several readings identified main themes from the transcripts. We then employed open software (Iramuteq®) to perform advanced discourse analysis. The two investigators discussed the common themes before the final analysis.

Results

Among the 215 medical students who completed the two simulation-based courses, thirty medical students (14%) participated in nine focus groups: four in the PC and five in the EM-ICM (Table 1). The focus groups had a mean duration of 71 min (+/− 12 min). The ten medical students who declined to participate in the study were either unavailable ($n = 9$) or uninterested in the study ($n = 1$). After the assessment, all the medical students passed the SBA, except for one student who failed the EM-ICM curriculum.

After data analysis, seven themes emerged from the focus groups: four were consistently present across all focus groups (major themes); three others were not consistently present and were subject to debate (major inconsistent themes). We chose to designate them as major themes because they emerged from contradictory discussions and seemed to be important issues for the medical students. The quotations from participants are reported as follows: curriculum (PC/EM) + participant number.

Summary of the four major themes that reached consensus (Fig. 3)

SBA as a support to hospital placements

Most students found that SBA would prepare them for hospital placements and be additive relative to simulation-based training alone. They emphasised the lack of feedback and supervision during their placements and great variability in the training and exposure to learning objectives. Consequently, medical students saw a motivational impact of SBA that offered meaning and a willingness to face a challenging clinical environment: *"SBA makes me want to go to work to the hospital" (EM 13); "Most of the time, I feel completely disregarded, particularly during placements, during which we spend so much time, learning so little" (PC2); "We are considered the insignificant medical students. We have never been shown or trained on technical procedures. For example, in my last hospital placement, we had to beg the attending physicians to show us how to use an oxygen venti-mask" (EM17).*

Medical students perceived that SBA enhanced their self-confidence and that it would favour assertiveness within the clinical environment. The medical students

Table 1 Nine focus groups' description

Paediatrics Curriculum			Emergency and Intensive care medicine curriculum		
	Participants (N)	Duration (min)	N° FG	Participants (N)	Duration (min)
Curriculum participants (n)	90			125	
Randomized for FG	20			20	
Total FG participants	**12**			**18**	
FG Before SBA					
FG 1	6	54	*FG 5*	**6**	**49**
FG 2	6	73	*FG 6*	**6**	**82**
FG after SBA					
FG 3	3	65	*FG 7*	**6**	**75**
FG 4	6	84	*FG 8*	**6**	**82**
			FG 9	**6**	**76**

FG focus group, *SBA* simulation-based assessment

felt more confident and felt able to take more initiative. They noted as follows: "*Assessment is hard, but after I succeeded, I felt like I had seen and managed the most difficult part and realised it would never be as difficult as this. Afterwards, I felt more prepared to manage the first moments of a life-threatening situation, for example" (EM 2); "SBA success could convince attending physicians to trust me and let me perform technical procedures during my hospital placements" (PC7).*

The students highlighted that SBA could be a valid tool filling a gap in assessments during placements. They described the content, tools, and organisations of the placements' assessments as too heterogeneous, and they all pointed out the very weak validity of the final hospital placement assessments. Their main criticism was the lack of feedback on their skills. In contrast, they identified SBA as an organised assessment, probably because of the implications for medical school: *"In my final gynaecology placement assessment, I worked hard, tried to learn as much as I could, and I was asked to show the uterus on an ultrasound picture" (PC9); "As the SBA is conducted in the medical school, it's better than placement assessments: the organisation is better, and the requirements are standardised. The training and*

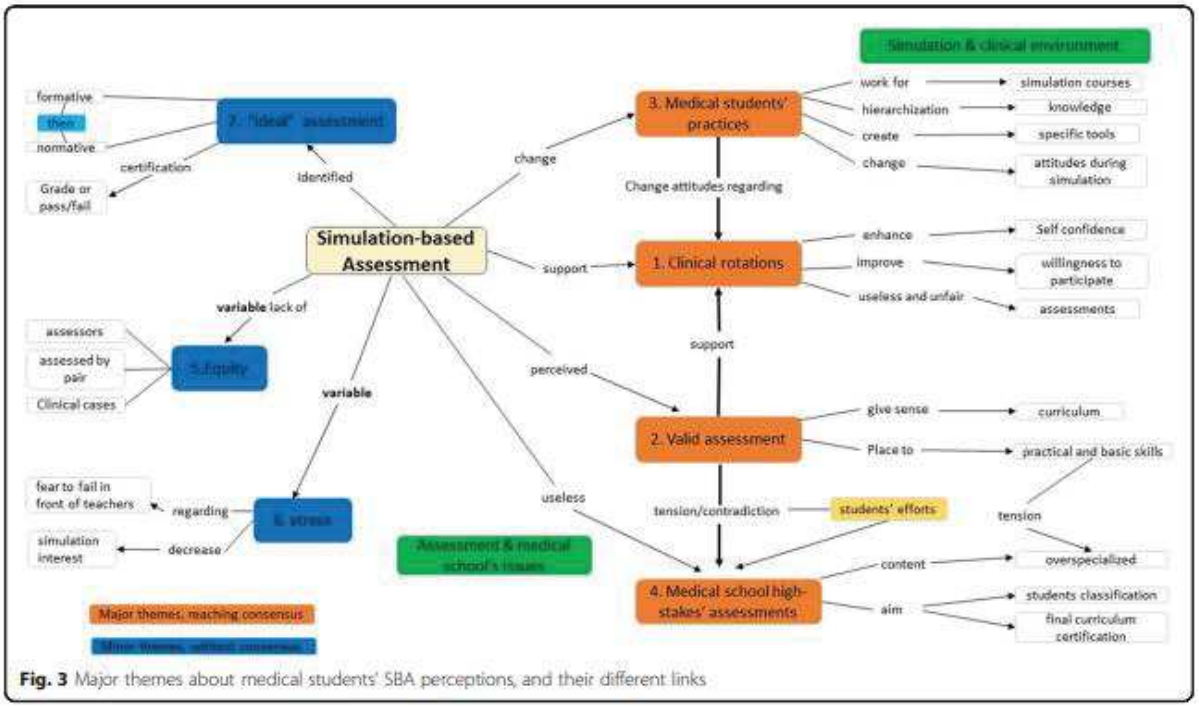


Fig. 3 Major themes about medical students' SBA perceptions, and their different links

assessment conditions can vary greatly during hospital placements [...] some of my colleagues had never had training before the final oral assessment" (PC3).

Finally, the medical students all emphasized their interest in placing practical skills at the centre of their training. After a deeper analysis of their comments, this view seemed to be associated generally with simulation-based education and not only with the existence of the assessment.

A major place for mannequin SBA within the overall assessment programmes

The medical students acknowledged that mannequin SBA is a tool with good face validity, and they felt that it centred practical skills in the curriculum. This view contrasted with their perception of medical school assessments, which were perceived as overly specialised and not always adapted to a competency mastery approach: *"It is a high-quality certification with a safe level for validation of skills" (PC 12); "SBA fixed the threshold very high" (PC 11); "The SBA allowed assessing 'the basics' although the medical school assessments often focused on the very small details" (EM5); and "it reminded us of what we must master for our future work, in contrast to knowledge that will never be useful, for example, the X mutation that leads to Y disease" (EM12).*

Some medical students also expressed doubts regarding the validity of medical school assessments and their ability to assess every competency, specifically clinical reasoning: *"It doesn't teach how to reason about clinical conditions, it is not a smart tool" (PC7); "it is impossible to assess everything! Last year, we had a test on caregiving relationship: one of the items of the test was 'to be empathic with the patient' ... I don't know who didn't select this obvious item..." (PC4).* The students found SBA to be a valid solution.

Mannequin SBA led to changes in students' practices

For the medical students, SBA was a key element leading to changes. Indeed, they reported that they did not act as they would normally do before a simulation-based course. First, they all prepared for the simulation courses, while they usually attended simulation-based courses without any previous specific work: *"Assessment was like a magic world that helped me prepare all the simulation courses differently" (PC5); "assessment compelled me to revise" (PC10); and "because I knew there would be an assessment, it forced me to deepen and organise [my] theoretical knowledge and forced me to study with a different methodology" (PC2).*

Second, they described changes in their working methods: they focused on essential knowledge and tried to organise it. This illustrated the testing effect of the SBA, because of which students had to organise and mobilise their knowledge instead of only memorising it. They accomplished this with the help of specific tools that they created themselves. All students found the lack of such tools (cognitive aids or video supports) regrettable. As the students noted, *"I had never done that before, usually when I come to a simulation course or a tutorial, I just read the corresponding chapter, and not even systematically" (PC9); "SBA helped us hierarchise and organise the knowledge before the simulation courses, with a method I had never employed before" (EM 13); "I created four to six essential points for each clinical problem I could face during the simulation" (EM2); and "because we cannot find such a practical guide in our books, I had to create my own [...] for each clinical case. This shows how much the books are unfit to the practice of medicine and only useful to train for the written assessments" (EM6).*

Medical students also perceived different attitudes during the simulation-based courses. Although they usually identified simulation-based training as a game and a pleasant course, they felt more focused and more involved than usual when the courses were not directed by a final summative assessment. For them, the changes were due to the need to identify the necessary skills and attitudes to pass the final exam: *"I felt more motivated to participate and, overall, more focused, as I wanted to understand what skills or knowledge would be useful for the final assessment but also for my practice during the hospital* placements*" (PC5); "I was involved and motivated during the training, and I pushed myself to organise my knowledge in an intelligent way" (PC4).*

Simulation-based assessment is unnecessary for written high-stakes assessments

Although they considered SBA to be a useful assessment regarding their needs for hospital placements and future internships, the medical students reported that SBA was unnecessary in preparing for their final sixth-year high-stakes assessment and for their final curriculum oral and written assessments. They did not consider SBA to be helpful for success in high-stakes assessments, as it did not assess the same knowledge or skills. Throughout their medical training, they focused their efforts on succeeding in the final assessment, which will determine their professional future. Thus, they felt that one more assessment with no relationship to this goal was not essential. This view stands in opposition to the perception of the approach being helpful for hospital placements, but again, they highlighted that their future was more important: *"It doesn't help us validate our education this year" (EM 2); "it doesn't give us bonus points for the EM-ICM module" (EM 17); "it is just another assessment" (PC 3); and "It doesn't assess the same knowledge as the written evaluation, so it doesn't help us" (PC 10).*

Summary of three other major themes inconsistent within the focus groups (Fig. 3)

A lack of equity

Most medical students who participated in the EM-ICM assessments reported an equity issue based on a great difference between the assessors, the scenario contents and the composition of the student pairs: *"I would have preferred another teammate; someone I knew would have been ideal. It's not fair because my friend did not have this issue" (PEM4); "I would have preferred cardiac arrest or pneumonia over the intoxication case" (EM 14).*

The medical students perceived a disparity between the assessors and expressed some doubts on the scales' reproducibility, grounded in their negative experience with variable content validity and reliability during the oral assessments. They also would have preferred to know the different scale contents before the final assessment session and asked for a formative assessment using the same scale rather than the summative scale. The students explained as follows: *"The scales of the oral assessments are awful, appalling, it's a real scandal" (PC8); "For SBA, it would be useful to know the scale's content before the assessment, it would be smart because, we are actually not aware what is expected from us. It's disturbing, even if it is the same for the other assessment" (EM2); and "With the simulation-based assessment, we should have the opportunity to know what exactly is expected from us, like a training session with the assessment scale and adequate feedback before the final SBA" (EM16).*

Stressful or not?

The PC medical students did not report major stress but reported stressful moments. In contrast to medical school or hospital assessments, the preparation stage was not stressful due to the light workload: *"The simple use of the word assessment is stressful, even if I knew that I would be trained and that it would be easy" (PC 4); "Our usual assessments require five months of work, whereas for this one, which concerned only a few skills, it was easier to prepare" (PC2).*

However, the few minutes before the SBA was more stressful than the time before an oral or a written test. They described the stress of being exposed to a difficult scenario in front of the assessor. However, the students also admitted that stress was not a major issue because if they passed this test, it would help to reduce any potential stress they might feel with a real patient: *"Five minutes before the assessment, I was very, very stressed. I think it was mainly due to what the other students and teachers might think of me rather than the assessment itself. But I think this is good stress, because it exists also in real life and we have to deal with it. And if we can manage the stress here, I will probably be able to manage it in real life. I prefer to make a mistake and to feel stress with a mannequin" (PC 7).*

Conversely, half of the EM-ICM students reported stress. Even if they all recognised its benefit for clinical placements, they criticised it for exposing them to another stressful event during the difficult curriculum. The students also described losing their interest in SBT because of the stress generated by SBA: *"Although we knew the scenarios, it was stressful, more than other assessments" (EM 12); "simulation must remain a fun exercise" (EM 3); and "normally simulation-based training is nice and friendly, but with the assessment, it became stressful" (EM 9).*

Practical issues and nature of the assessment

The majority of the EM-ICM students and some PC students highlighted the importance of being trained before being assessed and thus to prefer formative before summative assessments. They suggested using more formative assessment tools: *"Directors' programmes should integrate simulation earlier in the curriculum and with higher volume than currently" (PC 4); "we should have more training before the assessment, just like for any other assessment, even if it means additional working time" (EM 2).*

However, the medical students all recognised the need for a summative assessment to formally recognize their abilities and performance. Afterwards, they identified formative assessments as less valuable than summative ones, probably due to misconducts considerations because they can successfully pass only by being present on the assessment day: *"I prefer to have an assessment with a real objective and with a consequence rather than an assessment without stakes" (EM10); "Tests you can easily validate by just being present are useless and unbearable" (PC 2).*

They also compared SBA with their written assessments and discussed the value of a grade for SBA. They did not reach a consensus, but some found that a grade could be helpful to identify progress and the minimal required level, although others found that the most important marker was to pass the test and to know they could use their abilities in the clinical environment. Moreover, they pointed out the main difference from the written assessment; they all appreciated the possibility of receiving feedback just after the SBA: *"We do not need a grade, what is important is to succeed, not to be perfect" (EM 13); "the grades are important in our curriculum, and they are currently used so we know how to interpret them" (PC4); and "One of the reasons for the new interest in the SBA is personal. Because it provides specific feedback, and after the session, I would exactly know what I had to work on" (EM 12).*

The last practical issue remained in the limitation of environmental fidelity, which refers to the realism of the simulation. However, this does not seem to be an obstacle for SBA, as the medical students identified many positive aspects.

SBA highlights contradictions between two activity systems

Within the grounded theory framework, the theories emerge from the data and could be analysed in relation to another theory. For this work, the relevant theoretical framework appeared to be Engeström activity theory, which focuses on the dynamics of learning and on the learner as a participant and assists in analysing the contradictions and tensions in a given system in order to help participants change [29, 30]. When debating the issue of SBA, the students constantly mentioned the challenges they faced during hospital placement. This was one of the main emergent themes, as it was not present at first during the semidirected interviews. They also made numerous comparisons between hospital placement and university training, pointing out the contradictions with an impact on their training. Engeström activity theory allowed us to analyse and illustrate these contradictions: medical students evolve in a dual system, between the university and hospital, sharing the same subjects (medical students) but with different outcomes, different rules and division of labour (Fig. 4). The hospitals' main objectives are patient outcomes, whereas effective learning, graduation and ranking are the university outcomes. A contradiction exists between the two systems because of these different objectives. This leads to tensions for medical students, who perceive medical school as a uniform system unbiased in teaching and assessment. This is in opposition to hospital placements, where high levels of heterogeneity in the teaching and assessing methods are reported, including exposure to clinical situations.

Discussion

Our study explored students' perceptions of assessment with mannequin-based simulation. They underlined that SBA would be valuable in the clinical environment because it would enhance self-confidence and willingness to participate in patient management and would offer medical students' better integration within the sometimes hostile clinical environment. The students also perceived mannequin SBA as a tool with high face validity, with the ability to centre basic skills in the curriculum and to impact their work practices. However, they found that SBA did not prepare them for the high-stakes assessments of their curriculum.

The main reason for employing a posteriori the Engeström activity theory was the constant evocation of the medical students' placement difficulties. In the data analysis, the medical students clearly stood between two systems and their specific outcomes. The Engeström activity theory has shown that expansive learning can

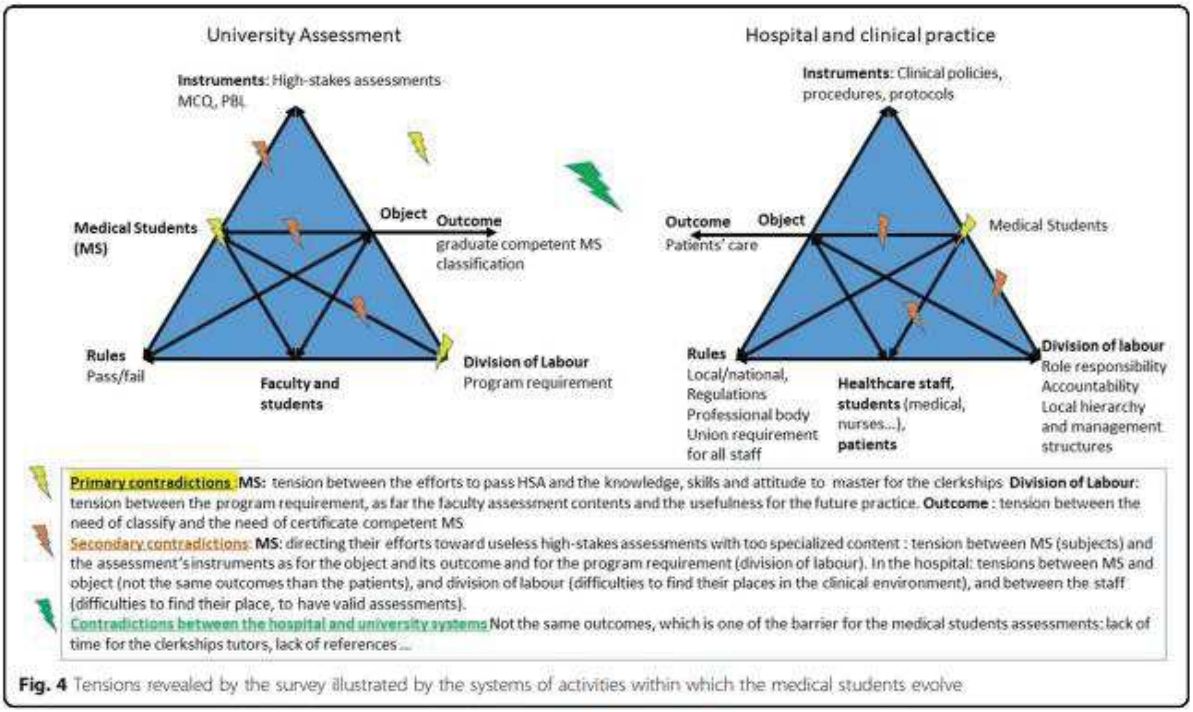


Fig. 4 Tensions revealed by the survey illustrated by the systems of activities within which the medical students evolve

occur when the two activity systems generate a new shared object and concept for their combined activity [30]. Berragan hypothesised that simulation-based training could be this shared object and the link between the two systems. Simulation generated a potential learning environment for medical students to practice and acquire clinical reasoning skills considering the context of the core formation systems [31]. Thus, we hypothesised that SBA could also have a specific role and place in the overall curriculum. In particular, SBA could be a tool to change not only the assessment systems but also the supervision and teaching methods used during clinical placements. Our data, focusing on SBA, suggested the same findings and emphasised Berragan's theory (Fig. 5). Moreover, another hypothesis is that by improving their self-confidence and assertiveness within their placements, medical students could more easily develop their professional identity. Indeed, medical students' placement is considered an experience that triggers professional identity formation, and when the students feel involved and seen as a "real doctor" by the team, it helps to construct their future professional identity [32, 33].

Another change induced by SBA was the impact on the medical students' work routines. With this assessment, students knew the aim of the courses, and they could prepare with a certain degree of autonomy. Autonomy, motivation and control in learning are factors that enable self-regulated learning and encourage students to be active in their learning process [34]. Students have intrinsic motivation to succeed, and this is associated with deepened learning and increased control of their own outcomes, which could decrease feelings of distress [35].

The students were worried about their final high-stakes assessment and reported that SBA was not an efficient tool to prepare them for this very important step in their curriculum. This observation indicates another contradiction and tension within their curriculum. As described above, the final year assessment focuses on knowledge assessment, which is more often overly specialised and unaligned with the SBA content. This issue underlines the need to align the different teaching and assessment tools. Such alignment is lacking between SBA and the final-year written assessment [36], and our results highlight the impact of a lack of alignment on students' motivation to participate in learning activities. Thus, our findings support the need for future reform to emphasise the place of SBA in the curriculum.

When the students identified SBA as an unfair assessment, they mentioned the subjectivity and lack of authenticity. However, subjectivity is one of the inherent pitfalls of a competency-based assessment [37]. The nature of competency is a multicomponent object, with exteriorised and measurable performance but also hidden components such as mobilisation of internal resources or clinical reasoning. Our hypothesis is that medical students rejected the subjectivity because it is not aligned with "students' culture" [38]. Indeed, in past decades, a valid assessment tool was defined as quantitative and objective. The challenges for faculty are to understand and deal with this subjectivity to create new assessment frameworks different from MCQ [7]. This will help educators and students to employ less objective assessments, such as multimodal ones with several tools and situations in a whole programmatic assessment [39]. Simulation has a great role to play because it employs

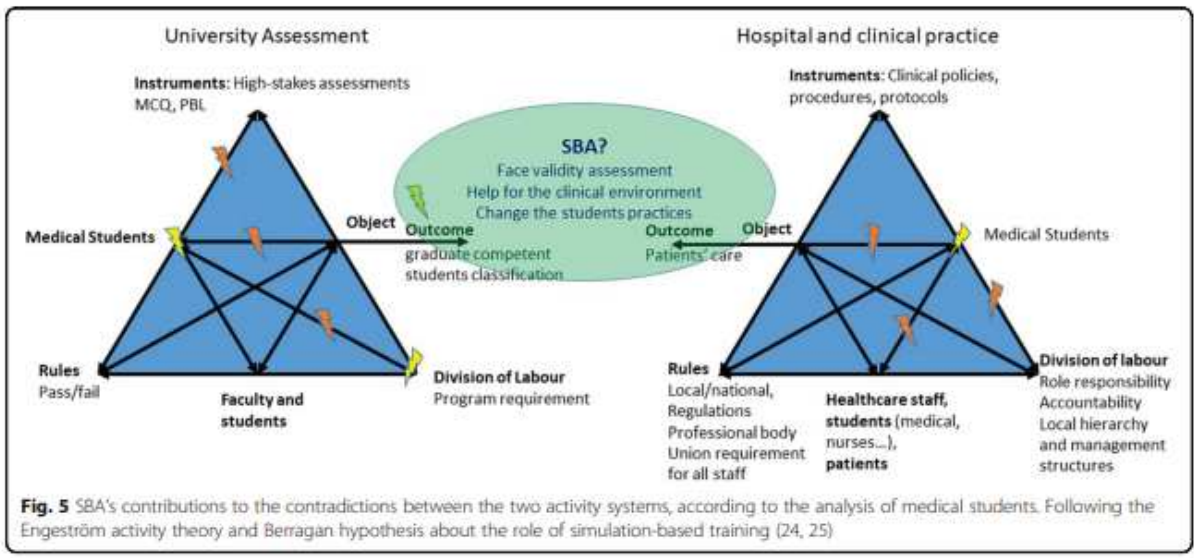


Fig. 5 SBA's contributions to the contradictions between the two activity systems, according to the analysis of medical students. Following the Engeström activity theory and Berragan hypothesis about the role of simulation-based training (24, 25)

controlled reproducible, reliable environments with controlled subjectivity. It is the main aim of OSCEs, but even if they are standardised simulation assessments, they reflect student performance and have good validity and reliability only if they use different contexts with different raters: at least 8 stations from five to 10 minutes [7, 40, 41]. Teachers and medical schools must pay attention to these issues, which seem to be important to many medical students. For this reason, the recent Canadian recommendations emphasise the need to maintain simulation courses as a "safe place". The use of formative assessments should remain a major place for explicit feedback. They also underline the current need for standardisation with reproductive, valid and reliable assessments. The use of respected professional activities would likely be seen as helpful [42].

Discussions on the practical aspects of the SBA highlighted two major characteristics of such an assessment: feedback and ratings. Normative assessment does not always provide feedback, especially in the French context, where feedback is provided through grades or classifications without qualitative feedback. With systematic debriefings, SBA could provide students with accurate feedback. However, grades were not completely approved by the students. They preferred knowing that they had the required skills without rankings [43]. Some authors also found these results, with a specific element: the more they progressed in the curriculum, the more they felt demotivation towards the ranking system [44]. Grades engage extrinsic motivation, which is linked to short-term memory and surface learning, unlike intrinsic motivation, which aids in self-development, satisfaction with an accomplished task and increased efficacy [35]. One other suggestion based on our results is the ethical issue of our assessment. As previously shown, debriefing is an essential part of simulation-based training [45, 46]. During the assessment session, the debriefing was shorter than that during the training courses. However, it was appreciated by the students because it was the first time they were provided with individual feedback immediately after an assessment. For simulation practice, it could be viewed as a short, weak debriefing and could contribute to the perception of unfairness of SBA. A possible improvement would be to give each student personal feedback with individual improvement goals [47].

Another ethical concern is the stress linked to the assessment process. Simulation-based training is supposed to be a safe environment to learn with opportunities to make errors and learn from these errors. However, this training environment has been shown to be stressful [48]. If we add stress to assessments, it could deflect SBT from one of its important aims: safe learning.

For these different reasons, caution should be applied in SBA for medical students, and we should improve our simulation tools and environments.

Limits

This study presents some limitations. First, it is a single-centre study, but two different SBAs took place, and thirty students did not have the same clinical experiences. The variety of experiences contribute to the authenticity of the study. Moreover, we obtained data saturation with the eighth focus group. Second, we missed a step in the qualitative approach, as we did not send back the findings to the participants. This would have improved the validity of the study by ensuring that the participants' ideas were accurately represented. Third, the medical students' perception highlighted the tension to which they are exposed within the two activity systems. It would have been helpful to complete the data by including observations of their activity within the simulation-based courses, assessments and placement. Moreover, even if it was not our main objective, we could have obtained insight from teachers to obtain more complete information regarding the medical students' perceptions.

Conclusion

Medical students' perceived SBA was a valid assessment tool with the capacity to enhance their self-confidence and willingness to participate in hospital placements. The assessment provided them the necessary opportunity to learn to take care of patients safely. They found that the experience had a positive impact on their practices; however, they regretted that it could be unfair, stressful and useless for their final high-stakes written assessment. The data analysis highlighted the several contradictions that medical students face within their two training systems. It was a relevant hypothesis that SBA could be an interesting link between these two systems, with a dedicated role to play in addressing the challenges faced by medical students between hospital placements and medical school requirements. These results are inspiring and should lead to the improvement and development of simulation-based assessments throughout medical school curricula. Our results describe a preimplementation mannequin SBA and OSCE and provide valuable insight for programme designers.

Abbreviations

CBME: Competency-based medical education; EM-ICMC: Emergency and intensive care medicine curriculum; EM1: Student one from EM-ICMC; OSCE: Objective structured clinical examination; PC: Paediatric curriculum; PC1: Student one from PC; SBA: Simulation-based assessment; SBT: Simulation-based training; SRLF: French Society of Intensive Care Medicine

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-021-02957-5.

Additional file 1: Appendix 1. EM-ICMC's scenarios and focus groups.

Additional file 2: Appendix 2. Example of focus groups within the paediatric curriculum.

Additional file 3: Annex 1. Pediatrics' Basic Life Support : Score used for the simulation-based assessment.

Acknowledgements

The authors wish to acknowledge Miss Jodie Battle (London, UK) and Pr Benjamin Bloom (from the Royal London Hospital, Barts Heath NHS Trust, London, UK) for their careful reading and editing of the manuscript. The authors also wish to thank the Scholarship Committee for assistance in contacting the medical students and organising the focus groups. Last, they wish to acknowledge the medical students for participating in the study.

Authors' contributions

All authors have made contributions: ALP, AP, MCR, GLB and YF designed the work. ALP conducted the interviews and constructed the semidirected interviews with the help of JT, YF, and GLB. ALP and JT performed the analysis and interpretation of the data. ALP wrote the draft, and JT, GLB, JT, NdS and AP substantively revised it. ALP, GLB and JY made the revisions. All the authors have approved the submitted version and any substantially modified version that involves the author's contribution to the study. They all have agreed both to be personally accountable for the author's own contributions and to answer questions related to the accuracy or integrity of any part of the work.

Funding

The authors declare that they received no funding for this study.

Availability of data and materials

Data supporting the results are available in a safe, secured electronic file on ALP's professional computer. All the transcripts are available if needed. Therefore, data are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

As reported in the manuscript, the SRLF ethical committee approved the study protocol, and students signed an informed consent form.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Emergency Department, Hôpital Pitié-Salpétrière, Assistance Publique-Hôpitaux de Paris, 83, bd de l'hôpital, 75013 Paris, France. [2]Sorbonne Université, Paris, France. [3]Department of Learning Sciences, EDA Laboratory, Université Sorbonne Paris Cité, Paris, France. [4]Emergency Department, SMUR, Hôpital Cochin, Assistance Publique – Hôpitaux de Paris (APHP), Paris, France. [5]Emergency Department, Trousseau Hospital, Assistance Publique – Hôpitaux de Paris (APHP), Paris, France. [6]Department of Pediatric Hematology-Oncology, Trousseau Hospital, Assistance Publique – Hôpitaux de Paris, Paris, France.

Received: 8 July 2020 Accepted: 24 September 2021
Published online: 19 November 2021

References

1. Batalden P, Leach D, Swing S, Dreyfus H, Dreyfus S. General competencies and accreditation in graduate medical education. Health Aff (Millwood). 2002;21:103–11.
2. Englander R, Carraccio C. A Lack of Continuity in Education, Training, and Practice Violates the "Do No Harm" Principle. Acad Med. 2018;93(Suppl 3):S12.
3. Carraccio CL, Englander R. From Flexner to competencies: reflections on a decade and the journey ahead. Acad Med. 2013;88(8):1067–73. https://doi.org/10.1097/ACM.0b013e318299396f.
4. Carraccio C, Englander R, Holmboe ES, Kogan JR. Driving care quality: aligning trainee assessment and supervision through practical application of Entrustable professional activities, competencies, and milestones. Acad Med. 2016;91(2):199–203. https://doi.org/10.1097/ACM.0000000000000985.
5. Ten Cate O, Chen HC, Hoff RG, Peters H, Bok H, van der Schaaf M. Curriculum development for the workplace using Entrustable professional activities (EPAs): AMEE guide no. 99. Med Teach. 2015;37(11):983–1002. https://doi.org/10.3109/0142159X.2015.1060308.
6. Lurie SJ, Mooney CJ, Lyness JM. Measurement of the General Competencies of the Accreditation Council for Graduate Medical Education: A Systematic Review. Acad Med. 2009;84(3):301–9. https://doi.org/10.1097/ACM.0b013e3181971f08.
7. Epstein RM. Assessment in medical education. N Engl J Med. 2007;356(4):387–96. https://doi.org/10.1056/NEJMra054784.
8. Holmboe ES, Ward DS, Reznick RK, Katsufrakis PJ, Leslie KM, Patel VL, et al. Faculty Development in Assessment: The Missing Link in Competency-Based Medical Education. Acad Med. 2011;86(4):460–7. https://doi.org/10.1097/ACM.0b013e31820cb2a7.
9. Govaerts MJB, van der Vleuten CPM, Schuwirth LWT, Muijtjens AMM. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. Adv Health Sci Educ Theory Pract mai. 2007;12(2):239–60. https://doi.org/10.1007/s10459-006-9043-1.
10. Beeson MS, Vozenilek JA. Specialty milestones and the next accreditation system: an opportunity for the simulation community. Simul Healthc. 2014;9(3):184–91. https://doi.org/10.1097/SIH.0000000000000006.
11. Bennett RE. Formative assessment: a critical review. Assess Educ. 2011;18(1):5–25. https://doi.org/10.1080/0969594X.2010.513678.
12. Griswold S, Fralliccardi A, Boulet J, Moadel T, Franzen D, Auerbach M, et al. Simulation-based education to ensure provider competency within the health care system. Acad Emerg Med. 2018;25(2):168–76. https://doi.org/10.1111/acem.13322.
13. Van Der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ. 1996;1(1):41–67. https://doi.org/10.1007/BF00596229.
14. Ahmed K, Jawad M, Dasgupta P, Darzi A, Athanasiou T, Khan MS. Assessment and maintenance of competence in urology. Nat Rev Urol. 2010;7(7):403–13. https://doi.org/10.1038/nrurol.2010.81.
15. Doughty CB, Kessler DO, Zuckerbraun NS, Stone KP, Reid JR, Kennedy CS, et al. Simulation in pediatric emergency medicine fellowships. Pediatrics. 2015;136(1):e152–8. https://doi.org/10.1542/peds.2014-4158.
16. McMurray L, Hall AK, Rich J, Merchant S, Chaplin T. The nightmares course: a longitudinal, multidisciplinary, simulation-based curriculum to train and assess resident competence in resuscitation. J Grad Med Educ. 2017;9(4):503–8. https://doi.org/10.4300/JGME-D-16-00462.1.
17. Langdon MG, Cunningham AJ. High-fidelity simulation in post-graduate training and assessment: an Irish perspective. Ir J Med Sci. 2007;176(4):267–71. https://doi.org/10.1007/s11845-007-0074-2.
18. Chaplin T, Thoma B, Petrosoniak A, Caners K, McColl T, Forristal C, et al. Simulation-based research in emergency medicine in Canada: Priorities and perspectives. CJEM. 2020;22:103–11.
19. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ Theory Pract. 2014;19(2):233–50. https://doi.org/10.1007/s10459-013-9458-4.
20. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ. 2015;49(6):560–75. https://doi.org/10.1111/medu.12678.
21. Watling CJ, Kenyon CF, Schulz V, Goldszmidt MA, Zibrowski E, Lingard L. An exploration of faculty perspectives on the in-training evaluation of residents. Acad Med. 2010;85(7):1157–62. https://doi.org/10.1097/ACM.0b013e3181e19722.
22. Cilliers FJ, Schuwirth LWT, Herman N, Adendorff HJ, van der Vleuten CPM. A model of the pre-assessment learning effects of summative assessment in medical education. Adv Health Sci Educ Theory Pract. 2012;17(1):39–53. https://doi.org/10.1007/s10459-011-9292-5.

23. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care. 2007;19(6):349–57. https://doi.org/10.1093/intqhc/mzm042.
24. Corbin J, Strauss A. Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory. Thousand Oaks: SAGE Publications, Inc.; 2008. [Cited may 5th 2021]. Disponible sur: http://methods.sagepub.com/book/basics-of-qualitative-research
25. Lingard L, Albert M, Levinson W. Grounded theory, mixed methods, and action research. BMJ. 2008;337(aug07 3):a567. https://doi.org/10.1136/bmj.39602.690162.47.
26. Kennedy TJT, Lingard LA. Making sense of grounded theory in medical education. Med Educ. 2006;40(2):101–8. https://doi.org/10.1111/j.1365-2929.2005.02378.x.
27. Stalmeijer RE, Mcnaughton N, Van Mook WNKA. Using focus groups in medical education research: AMEE guide no. 91. Med Teach. 2014;36(11):923–39. https://doi.org/10.3109/0142159X.2014.917165.
28. Maconochie IK, Bingham R, Eich C, López-Herce J, Rodríguez-Núñez A, Rajka T, et al. European resuscitation council guidelines for resuscitation 2015. Resuscitation. 2015;95:223–48.
29. Engestrom Y. Activity theory as a framework for analyzing and redesigning work. Ergonomics. 2000;43(7):960–74. https://doi.org/10.1080/001401300409143.
30. Engeström Y. Expansive learning at work: toward an activity theoretical reconceptualization. J Educ Work. 2001;14(1):133–56. https://doi.org/10.1080/13639080020028747.
31. Berragan L. Conceptualising learning through simulation: an expansive approach for professional and personal learning. Nurse Educ Pract. 2013;13(4):250–5. https://doi.org/10.1016/j.nepr.2013.01.004.
32. Maitra A, Lin S, Rydel TA, Schillinger E. Balancing forces: medical students' reflections on professionalism challenges and professional identity formation. Fam Med. 2021;53(3):200–6. https://doi.org/10.22454/FamMed.2021.128713.
33. Kay D, Berry A, Coles NA. What experiences in medical school trigger professional identity development? Teach Learn Med. 2019;31(1):17–25. https://doi.org/10.1080/10401334.2018.1444487.
34. White CB. Smoothing out transitions: how pedagogy influences medical students' achievement of self-regulated learning goals. Adv Health Sci Educ Theory Pract. 2007;12(3):279–97. https://doi.org/10.1007/s10459-006-9000-z.
35. Pelaccia T, Viau R. Motivation in medical education. Med Teach. 2017;39(2):136–40. https://doi.org/10.1080/0142159X.2016.1248924.
36. Schuwirth LWT, van der Vleuten CPM. General overview of the theories used in assessment: AMEE guide no. 57. Med Teach. 2011;33(10):783–97. https://doi.org/10.3109/0142159X.2011.611022.
37. Van der Vleuten CP, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. Med Educ. 1991;25(2):110–8. https://doi.org/10.1111/j.1365-2923.1991.tb00036.x.
38. Laqueur T. Boys in White: Student Culture in Medical School. BMJ. 2002;325:721.
39. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. Med Educ. 2012;46(1):38–48. https://doi.org/10.1111/j.1365-2923.2011.04098.x.
40. Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part II: Organisation & Administration. Med Teach. 2013;35(9):e1447–63. https://doi.org/10.3109/0142159X.2013.818635.
41. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part I: an historical and theoretical perspective. Med Teach. 2013;35(9):e1437–46. https://doi.org/10.3109/0142159X.2013.818634.
42. Hall AK, Chaplin T, McColl T, Petrosoniak A, Caners K, Rocca N, et al. Harnessing the power of simulation for assessment: consensus recommendations for the use of simulation-based assessment in emergency medicine. CJEM. 2020;22(2):194–203. https://doi.org/10.1017/cem.2019.488.
43. Rohe DE, Barrier PA, Clark MM, Cook DA, Vickers KS, Decker PA. The benefits of pass-fail grading on stress, mood, and group cohesion in medical students. Mayo Clin Proc. 2006;81(11):1443–8. https://doi.org/10.4065/81.11.1443.
44. O'Neill P, Baxter CM, Morris J. Does awarding a medical degree with honours act as a motivator or demotivator to student learning? Med Educ. 1999;33(8):566–71. https://doi.org/10.1046/j.1365-2923.1999.00369.x.
45. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach. 2005;27(1):10–28. https://doi.org/10.1080/01421590500046924.
46. Dieckmann P, Molin Friis S, Lippert A, Ostergaard D. The art and science of debriefing in simulation: ideal and practice. Med Teach. 2009;31(7):e287–94. https://doi.org/10.1080/01421590902866218.
47. Motola I, Devine LA, Chung HS, Sullivan JE, Issenberg SB. Simulation in healthcare education: a best evidence practical guide. AMEE guide no. 82. Med Teach. 2013;35(10):e1511–30. https://doi.org/10.3109/0142159X.2013.818632.
48. Bong CL, Lightdale JR, Fredette ME, Weinstock P. Effects of simulation versus traditional tutorial-based training on physiologic stress levels among clinicians: a pilot study. Simul Healthc. 2010;5(5):272–8. https://doi.org/10.1097/SIH.0b013e3181e98b29.

Publisher's Note

Annexe 2 – Score TEAM traduit et validé en Français (Maignan et al, 2016)

Mesure d'Evaluation d'une Equipe d'Urgence (TEAM)

Introduction

Ce questionnaire de compétences non-techniques a été conçu pour une évaluation observationnelle permettant une notation valide, fiable et réalisable des équipes d'urgence médicale (par exemple les équipes de réanimation et de traumatologie). Le questionnaire devra être complété par des cliniciens experts pour une évaluation précise de la performance et un retour d'information sur le leadership de l'équipe, sur le travail en équipe, sur la compréhension de la situation et sur la gestion des tâches. Des suggestions d'aide à l'évaluation sont proposées s'il y a lieu. L'échelle suivante devra être utilisée pour chaque item :

Jamais/Presque jamais	Rarement	A peu près la moitié du temps	Souvent	Toujours/Presque toujours
0	1	2	3	4

Identification de l'équipe

Date : Heure : Lieu :
Chef d'équipe : Equipe :

Leadership : partant du principe que le chef d'équipe est soit désigné, soit qu'il se soit dégagé par rapport au reste de l'équipe ou qu'il soit le expérimenté. Si aucun chef d'équipe n'apparait, répondez par « 0 » à la question 1 et « 0 » à la question 2.	0	1	2	3	4	
1. Le chef d'équipe a informé l'équipe de ce que l'on attendait d'elle en donnant les directives et les ordres						
2. Le chef d'équipe a maintenu une perspective globale. *Suggestions : contrôle des procédures cliniques et de l'environnement ?* *Rester « non-interventionniste » selon le cas. Délégation appropriée.*						
Travail en équipe : Les évaluations devront inclure l'équipe en totalité, c'est-à-dire le chef d'équipe et l'équipe collectivement (à plus ou moins grande échelle)	0	1	2	3	4	
3. L'équipe a communiqué de façon efficace. *Suggestions : communication verbale, non-verbale et écrite.*						
4. L'équipe a travaillé ensemble pour compléter à bien les tâches requises en temps voulu.						
5. L'équipe a agi avec sang-froid et de façon contrôlée. *Suggestions : émotions appropriées ? Problèmes de la gestion des conflits ?*						
6. Le moral de l'équipe était positif *Suggestions : soutien approprié, confiance, esprit, optimisme, détermination ?*						
7. L'équipe s'est adaptée aux changements de situation *Suggestions : Adaptation dans leur rôle professionnel ?* *Changements de situation : dégradation de l'état de santé du patient ?* *Changements dans l'équipe ?*						
8. L'équipe a contrôlé et réévalué la situation.						
9. L'équipe a anticipé les actions possibles. *Suggestions : préparation du défibrillateur, médicaments, équipement des voies aériennes.*						
Gestion des tâches	0	1	2	3	4	
10. L'équipe a identifié ses priorités						
11. L'équipe a suivi les standards et les directives homologués. *Suggestions : certaines dérogations peuvent être appropriées.*						
Dans l'ensemble	1	2	3	4	5	6	7	8	9	10
12. Sur une échelle de 1 à 10, donnez votre note globale sur les performances non-techniques de l'équipe.										

Commentaires : _____

_____

Annexe 3 – Notice d'information étude e-Simsit : apprenant

Paris/ Strasbourg/ Rouen, le / /

Madame, Monsieur,

Nous vous sollicitons dans le cadre du programme de recherche e-Simsit, soutenu par la Fédération Hospitalo-Universitaire IMPEC. Il s'agit d'une étude en pédagogie médicale sur l'utilisabilité et la reproductibilité de scores d'évaluation des apprenants, lors de simulation de situations d'urgences vitales, in situ, c'est-à-dire directement dans les environnements de travail des équipes qui seront formées. L'étude s'intéresse également à la faisabilité de ces entrainements in situ dans 6 services d'urgences français.

Ce programme est la continuité d'une démarche de recherche sur l'évaluation par la simulation, au cours de laquelle nous avons créé les scores grâce à la mise en œuvre de processus de consultation d'expert, puis d'analyse de contenu de chacun d'entre eux. Trois scores d'évaluation, correspondant à trois situations cliniques graves en médecine d'urgence ont pu ainsi être construit. La recherche actuelle s'inscrit dans une démarche de validation des scores dans différents contextes et notamment in situ, au sein d'un entrainement pluriprofessionnel.

Nous nous adressons à vous car votre service a accepté de participer à la recherche et vous allez pouvoir y participer au travers des sessions de formations qui vous seront proposées. Ces sessions de formations se dérouleront sur votre lieu de travail. Le déroulement de la formation ne sera pas perturbé par l'étude puisqu'il se passera selon un formation « classique » en simulation à savoir : présentation de la simulation (objectifs, matériel, environnement et conditions d'apprentissage), briefing, scénario et débriefing. Pendant le scénario, vos formateurs rempliront les scores d'évaluation, qui sont à l'étude et n'ont donc pas encore pour objectif de vous noter ou de vous évaluer. Ils sont l'objet principal de l'étude. A la fin de la session, nous recueillerons votre avis sur la session elle-même et sur votre perception d'une telle formation dans votre environnement de travail. Les sessions ne sont pas enregistrées ou filmées.

Nous vous demanderons également certaines données démographiques : votre âge, votre profession et votre expérience aux urgences. Toutes ces informations seront recueillies de manière anonyme, resteront confidentielles et seront utilisées dans le cadre strict de cette étude. C'est pourquoi, en début de session, vous recevrez un numéro d'identification qui ne nous permettra pas de relier vos réponses à votre identité. Veillez à le mémoriser au cas où vous participeriez à une deuxième session de formation. Les résultats globaux pourront vous être communiqués sur simple demande à l'adresse suivante : alphilippon@etu.unistra.fr

Si vous le souhaitez, nous pouvons aussi échanger plus directement avec vous, afin de vous expliquer l'objectif précis de cette recherche. N'hésitez pas à nous contacter en utilisant l'adresse électronique ci-dessus.

Si vous acceptez de participer à ce programme, un chercheur de notre équipe ou l'investigateur local viendra vous rencontrer, pour vous représenter l'étude et vous proposer de signer le consentement de participation. Nous vous enverrons une copie du document pour que vous puissiez vous y référer à n'importe quel moment.

Nous vous garantissons que votre participation sera totalement anonyme et que ce projet a été évalué par une commission d'éthique. Vous avez la possibilité de retirer votre consentement à tout moment sans avoir à donner de justification et vous avez également la possibilité de refuser de participer à l'étude, sans que cela ne nuise à votre formation ou à votre activité au sein du service.

En espérant une réponse positive de votre part, nous vous remercions de l'attention que vous porterez à notre demande.

Pour toutes informations relatives à ce programme, vous pouvez contacter les responsables suivants à l'adresse suivante : alphilippon@etu.unistra.fr

Anne-Laure Philippon et l'équipe e-Simsit

Annexe 4 – Consentement de participation – Etude e-Simsit

Mr /Mme (rayez la mentionne inutile)
Nom de naissance : ……
Prénom : ……
Adresse mail:……

Il m'a été proposé de participer à une étude sur l'utilisation de scores d'évaluation lors de séances de simulation in situ.
L'investigateur du site *(Prénom, Nom)* …….. *m'*a précisé que je suis libre d'accepter ou de refuser.

Afin d'éclairer ma décision, j'ai reçu et compris les informations suivantes :
1) Je pourrai à tout moment interrompre ma participation si je le désire, sans avoir à me justifier.
2) Je pourrai prendre connaissance des résultats de l'étude dans sa globalité lorsqu'elle sera achevée, à l'adresse mail suivante : alphilippon@etu.unistra.fr
3) Les données recueillies demeureront strictement confidentielles.
4) J'ai compris que si je me retire de l'étude à tout moment, cela ne nuira pas à ma formation ni à mes activités au sein du service.
Compte-tenu des informations qui m'ont été transmises :
J'accepte librement et volontairement de participer à la recherche e-simsit.

Cocher les cases appropriées en fonction de votre volonté :
OUI ☐
Non ☐

Date :
Signature du participant :

Date :
Signature de l'investigateur :

Annexe 5 – Guide entretien – Etude e-Simsit

Phrase d'introduction : « Merci beaucoup de vous joindre à ce groupe d'entretien collectif. Après vous être présentés, vous échangerez sur la simulation in situ et sur l'utilisation des scores d'évaluation à partir de vos expériences respectives. »

Première question :

Racontez-nous comment s'est passée la mise en place de la simulation in situ ainsi que l'utilisation des scores au sein de vos sessions de formation ?

Questions à poser si non abordé spontanément :

1. Comment avez-vous utilisé les scores ?
   - Qui a rempli les grilles
   - Utilisés pendant le débriefing ?
   - Les scores sont-ils facilement utilisables ?
2. Selon vous, qu'évaluent réellement les scores ?
   - Quelles performances ?
   - Le niveau de l'équipe ?
3. Comment avez-vous utilisé le score de performance globale ?
4. Avez-vous déjà utilisé la simulation pour évaluer des apprenants ? Avec quels scores ?
5. Selon votre expérience en formation et après l'utilisation des scores, quels apprenants et quelles performances pourraient être évalués par les scores ?
6. Quelle place pourraient avoir les scores dans votre pratique ?
7. Utilisez-vous la simulation pour évaluer les internes ?
8. Que pensez-vous de l'évaluation en interprofessionnel ?
9. **Après le premier entretien :**

Aborder la question de la bienveillance en simulation, paradoxale avec l'évaluation (semble être un thème qui émerge)

Annexe 6 – Questionnaire Enseignant évaluateur – Etude e-Simsit

Chers investigateurs et acteurs de l'étude e-simsit. Encore merci de votre participation à l'étude et de votre aide pour la mise en place et l'animation des sessions de formation. Voici un questionnaire qui a pour objectif de nous faire part de vos retours sur les scores d'évaluation. Pour finir, quelques questions vous concernant. Ce questionnaire est totalement anonyme et ne demande si votre nom ni votre prénom. Si toutefois vous vouliez discuter de l'étude, des scores de la simulation in situ avec nous, nous sommes à votre disposition : alphilippon@etu.unistra.fr
Merci à vous, AL Philippon

I. Pour chaque score d'évaluation ACAT :

Pour chaque question, vous pouvez répondre en choisissant un chiffre de 1 à 4 : pas du tout d'accord (1), pas d'accord (2), d'accord (3), totalement d'accord (4) ou bien si vous pensez ne pas pouvoir répondre : ne s'applique pas.

1. Le score est facile à utiliser
2. Les items sont compréhensibles
3. Les items sont pertinents et représentent bien ce qui devrait être évalué dans la prise en charge de l'ACR/Coma/DRA
4. Le score évalue la performance attendue d'un médecin urgentiste
5. Le score évalue la performance attendue d'un interne en DESMU
6. Le score évalue la performance attendue b
7. Le score évalue la performance d'équipe
8. Le score est utile pour évaluer le raisonnement clinique
9. Le score est utile pour évaluer la communication
10. Le score est un reflet des compétences requises en médecine d'urgence
11. Par rapport aux scores existants (OSCAR, ANTS….) le score a un intérêt
12. J'utiliserai ce score, en complément d'un score d'évaluation des habiletés non-techniques
13. Le score de performance globale est une aide supplémentaire pour apprécier la performance de l'apprenant
14. Le score est utile pour guider le feedback formatif délivré à l'apprenant
15. Le score aide le formateur pour le débriefing
16. J'utiliserai le score pour évaluer la performance d'un externe en MU
17. J'utiliserai le score pour évaluer la performance d'un interne DESMU
18. J'utiliserai le score pour évaluer la performance d'un médecin urgentiste

II. Quelques informations vous concernant

19. Vous êtes : aide-soignant, IDE, interne, médecin
20. Quel est votre âge ?
21. Depuis combien de temps (années) travaillez-vous aux urgences ?
22. Depuis combien de temps (années) êtes-vous formateur en simulation ?

Annexe 7 – Scores ACAT définitifs, après étude du processus de réponse

Score générique ACR

Numéro Cas :
Numéro évaluateur :
Nombre apprenants :
Rôles (nombre) : Sénior ☐ Interne ☐
Externe ☐ IDE ☐

ITEM	Complet	1	Incomplet	0,5	Non fait = 0	Total
PEC INITIALE – BLS (5 MINUTES MAX)						
1. Appeler "à l'aide" :	Immédiat		Retardé		Pas d'appel	
2. Débuter le massage cardiaque	Immédiat		30s – 2 min		> 2 min	
3. Aller chercher un DSA et mettre les patchs	Dans les 2 minutes post ACR		Dans les 2 à 3 minutes		> 3 minutes / pas de DSA	
4. Profondeur MCE	5-6cm		4-5 cm		< 4 cm	
5. Rythme	100-120/min		> 120/min		< 100/min	
6. Relaxation du thorax	Correcte		Insuffisante		Pas relâché	
7. Noter l'heure de l'arrêt	A haute voix, immédiat		Pas dite, retardée		Pas d'heure	
ALS – PEC SPECIALISEE						
8. Préparer /faire préparer les médicaments :	Demander / anticiper la préparation		confuse		Non fait	
9. Injection d'adrénaline	3 éléments : dose, délai, algorithme		1 élément incorrect		Non fait ou erreur dose	
10. Analyse du rythme +/- délivrer chocs	Toutes les 2 minutes		Mauvais intervalle		non fait	
11. Reprise MCE après choc	Immédiat		> 15 s		NON FAIT ou > 30s	
12. Alterner les compressions et la ventilation au BAVU	30/2 dès que possible. Bien communiquer		Incomplet		Pas de ventilation	
13. Intubation : savoir préparer le matériel	Complet		Incomplet		Pas de préparation	
14. Débuter / organiser le traitement étiologique	Réfléchir à l'étiologie, faire le traitement si possible		Incomplet		Non fait	
COMMUNICATION						
15. Énoncer le diagnostic au sein de l'équipe.	Complet		Incomplet		Pas d'énoncé diagnostic	
16. communication : distribuer les rôles, équipe calme	Tout fait		Manque un élément		Pas de communication	
17. Annoncer le devenir du patient, en discuter en équipe	Suite CAT, objectifs dans les 30 min		Pas d'objectif clair		Sujet non abordé	
18. communication pour les relais de massage cardiaque :	Equipe attentive au MCE : demande relais / propose		Relais confus		Pas de relais, > 2 min, pas de regard de l'équipe	
19. appeler le régulateur / le réa / coro : délai et présentation	8-10 min, présentation claire, organisée, demande d'orientation		Retard appel, présentation confuse		Pas d'appel	
20. gestion de la famille	Un membre de l'équipe se détache, parle calmement, explique ou recherche famille		Explications confuses		Pas d'attention à la famille	
TOTAL						

Aide pour la notation

1. Appeler "à l'aide" :	Dure à voix haute, « c'est un arrêt » Immédiat (dans les 30s après reconnaissance)		Appel retardé : > 30 s		Pas d'appel
2. Débuter le massage cardiaque	Immédiat		> 30 s après ACR		> 2 min
3. Aller chercher un DSA et mettre les patchs	Dans les 2 minutes post ACR, patch mis		Dans les 2 à 3 minutes		> 3 minutes / pas de DSA
4. Qualité du massage cardiaque : profondeur,	5-6cm		4-5 cm		< 4 cm
5. rythme	100-120/min		> 120/min		< 100/min
6. relaxation du thorax	Correcte		Insuffisante		Pas relaché
7. Noter l'heure de l'arrêt	Enoncée rapidement à haute voix		Pas dite ou notée, rapidement		Pas d'heure
8. Préparer /faire préparer les médicaments : **adrénaline +/- cordarone**	Demander / anticiper la préparation		Demande confuse		Pas de préparation, pas de demande
9. Injection d'adrénaline :	dose correcte, DQP si asystolie 3-5 min si FV ou 3° choc si bien faits		Délai incorrect, algorithme faux		Dose autre que 1 mg Non fait
10. Analyse du rythme +/- délivrer chocs	Toutes les 2 minutes		Mauvais intervalle : trop tot / tard, rallumer DSA		non fait
11. Reprise MCE après choc	Immédiat, dès la fin choc ++		> 15 s		NON FAIT ou > 30s
12. Alterner les compressions et la ventilation au BAVU :	Respect 30/2 dès qu'organisation possible. Bien communiquer		Pas de communication, pb algorithme		Pas de ventilation
13. Intubation : savoir préparer le matériel	Ventilateur, aspiration, sonde IOT – seringue – Eschmann sorti		Incomplet		Pas de préparation
14. Débuter / organiser le traitement étiologique	Se poser la question de l'étiologie Si identifiable et curable Ttt complet		Si cause curable : ttt incomplet		Ne se pose pas la question Pas de ttt proposé alors que cause curable identifiable
15. Enoncer le diagnostic au sein de l'équipe:	A voix haute, avec recherche étiologie en équipe				Pas d'énoncé diagnostic
16. communication : distribuer les rôles, équipe calme	Tout fait		Manque un élément, pas organisés		Pas de communication
17. Annoncer le devenir du patient, en discuter en équipe	Discussion sur suite de PEC/CAT, sur objectifs dans les 30 min		Pas d'objectif clair		Sujet non abordé
18. communication pour les relais de massage cardiaque :	Equipe attentive au MCE : demande relais / propose		Relais confus		Pas de relais, > 2 min, pas de regard de l'équipe
19. appeler le régulateur / le réa / coronarographiste : dans un délai adapté, présentation claire	Appel dans les 8-10 min, présentation claire , organisée, demande d'orientation		Retard appel, présentation confuse		Pas d'appel
20. gestion de la famille	Si présente et assez nombreux : Un membre de l'équipe se détache, parle calmement, explique Si pas de famille, se poser la question de son existence, chercher à les joindre		Explications confuses		Pas d'attention à la famille

Score générique pour évaluer une situation de Coma

Nombre apprenants :
Rôles (nombre): Sénior ☐ Interne ☐
Externe ☐ IDE ☐

	ITEMS	Complet	1	Incomplet	0.5	Non fait / erreur	Total
PEC INITIALE – 5 MINUTES	1. Mesurer la glycémie	Dans les 2 minutes		> 2 min		Pas de mesure	
	2. Calculer le score de Glasgow	Score correct, bien compris		Incomplet		Pas de mesure	
	3. Scoper le patient : mesure PA, SpO2, FR, FC	Dans les 2 minutes - 4 paramètres		incomplet		Pas de scope > 5 min	
	4. Pose d'une VVP / la demander.	Demander / poser dans les 5 min				Pas de voie > 5 min	
	5. des informations : interrogatoire : ATCD - terrain	Complet + chercher personne ressource		Recherche incomplète		Non fait	
	6. Recueillir les traitements :	Complète et adaptée		Recherche incomplète		Pas de recherche ttt	
	7. Histoire de la maladie	Mode de survenue, signes neuro, fievre....		Recherche incomplète		Pas d'HDM	
EX CLINIQUE	8. Rechercher / notifier défaillance HDQ	Prendre la PA, rechercher hypoperfusion		Manque élément		Pas de recherche	
	9. Inspection cutanée	Purpura / compression/points injection		Manque élément		Pas de recherche	
	10. Examen neuro	Pupilles / loc /RCP/ROT / nuque		Manque élément		Pas d'ex neuro	
	11. Examen le cuir chevelu, rechercher une plaie	Fait				Non fait	
	12. Résumé interrogatoire + examen : précoce	A voix haute, en équipe, résumer la situation puis émettre hypothèses diag.		Incomplet ou retardé		Ni résumé ni hypothèses	
EX COMP.	13. Bilan biologique adapté aux hypothèses diagnostiques	Complet		Incomplet		Non fait	
	14. Imagerie	Evoquer et discuter imagerie, la faire si indiquée (cf ci-joint)		Demandée, sans argumentation		Pas évoquée	
IOT	15. Protection des voies aériennes :	Evoquer le sujet, discuter, et poser l'indication d'IOT		IOT sans discussion, sans argument		Sujet pas évoqué	
	16. Pré-oxygéner le patient	Technique adaptée à la VS		Mal utilisée		Pas de pré-O2 Ventilation avec BAVU	
PEC SPE	17. traitement étiologique adapté	TTT complet (cf détails)		TTT incomplet		Pas de ttt	
COMMUNICATION	18. Compréhension	Communication adaptée, claire (cf détails)		Confus		Pas de communication	
	19. Annoncer à l'équipe le devenir du patient	Suite PEC, définir objectifs en équipe		Devenir confus		Pas abordé	
	20. appeler le régulateur / le réa / chirurgien : dès qu'une orientation est nécessaire ou si besoin d'aide	Appel pertinent, présentation et demande claires		Confus...		Pas d'appel dans les 10 minutes	
TOTAL							

Aide pour notation

ITEMS	1		0.5		0
1. Mesurer la glycémie	Dans les 2 minutes		Au-delà de 2 min		Pas de mesure
2. Calculer le score de Glasgow	Score correct, bien compris		Incorrect, manque des éléments		Pas de mesure
3. Scoper le patient : mesure PA, SpO2, FR, FC	Dans les 2 minutes - 4 paramètres		Au-delà de 2 min, manque paramètre		Pas de scope > 5 min
4. Pose d'une VVP dans les 5 minutes : la demander.	Demander / poser dans les 5 min				Pas de voie > 5 min
5. Chercher à recueillir des informations A l'interrogatoire : ATCD : analyse du terrain :	Complet (ATCD psy, neuro, toxico...) chercher personne ressource		Recherche incomplète		Pas de recherche d'info Pas d'ATCD
6. Recueillir les traitements :	Complète et adaptée (anticoag, psychotrope...)		Recherche incomplète		Pas de recherche ttt
7. Histoire de la maladie	Recherche fièvre, indices pour IMV mode de survenue et signes neuro avant coma :		Recherche incomplète		Pas d'HDM
8. Rechercher / notifier des signes de défaillance HDQ	Notifier si hypoPA, demander si marbrures/TRC abaissé		Manque élément		Pas de recherche
9. Inspection cutanée	Rechercher un purpura, des points de compression, des points d'injection		Manque élément		Pas de recherche
10. Examiner neuro	Pupilles, signes de localisation, irritation pyramidale, ROT, nuque		Manque élément		Pas d'ex neuro
11. Examen le cuir chevelu, à la recherche d'une plaie	Fait				Non fait
12. Résumé interrogatoire + examen : à faire apres anamnèse et ex clinique	A voix haute, en équipe, résumer la situation puis émettre hypothèses diag précoces.		Incomplet ou retardé		Pas de résumé, pas d'hypothèses diagnostiques
13. Bilan biologique adapaté aux hypothèses diagnostiques (sera précisé selon le cas cliniqe : Hemoc si fievre, PL (apres TDM), hémostase...	Complet		Incomplet		Non fait
14. Imagerie : demander une TDM cérébrale si coma fébrile, si signe de localisation neurologique ou si absence d'orientation diagnostique	Evoquer et discuter imagerie, puis la faire si indiquée		Demandée, sans argumentation		Pas évoquée
15. Protection des voies aériennes : poser l'indication d'intubation	Evoquer le sujet, discuter, et poser l'indication		IOT sans savoir pourquoi		Sujet pas évoqué
16. Pré-oxygéner le patient	Technique adaptée à la VS (ne pas ballonner si BAVU par exemple, MHC avec bonne utilisation)		MHC mal utilisé (mauvais débit, réserve non gonflée).		Pas de pré-O2 Ventilation avec BAVU
17. Réaction adaptée selon chaque situation clinique :	hématome sous AVK => PPSB coma fébrile => C3G en urgence coma toxique : se poser la question d'antidote et si indiqué => le débuter		TTT incomplet		Pas de ttt
Compréhension :	S'assurer de la compréhension des autres membres de l'équipe, communication calme, dirigée entre les membres		Confus		Pas de communication ou chaotique
Annoncer à l'équipe le devenir du patient	Aborder le sujet dans l'équipe, parler de l'orientation du patient		Devenir confus		Pas abordé
appeler le régulateur / le réa / chirurgien : dès qu'une orientation est nécessaire ou si besoin d'aide	Appel pertinent (avec bon interlocuteur), présentation et demande claires. Dans les 10 min		Confus, mauvais interlocuteur		Pas d'appel dans les 10 minutes

Score générique DRA

Nombre apprenants :
Rôles (nombre): Sénior ☐ Interne ☐
Externe ☐ IDE ☐

	ITEMS	COMPLET	1	INCOMPLET	0.5	NON FAIT = 0	Total
PEC initiale	1. Installer le patient semi-assis	Immédiatement		Dans les 5 min MAX		Non fait	
	2. Mesure de la SpO2 et de la FR	Immédiat et correct		> 5 min ou incomplet			
	3. Délivrer de l'oxygène dans les 5 min	Utilisation correcte et dispositif correct		Mauvais dispositif, mal utilisé, à 5 min		Non fait	
Ex clinique	4. Recherche signes de gravité respiratoire	Notifier leur identification				Non fait	
	5. Rechercher une défaillance HDQ	Prendre PA + signes I. Circulatoire		Manque élément		Pas de recherche	
	6. Examen cardio-pulmonaire	Auscultation + signes IC		Incomplet		Pas fait	
Anamnèse	7. Interroger les proches / patient	Clair, dirigé, pertinent		Confus, incohérent		Non fait	
	8. Interrogatoire : recherche d'une étiologie selon le terrain	FRCV (dont tabac), FRMTEV, ATCD cardio/pneumo/allergiques/voyage		Incomplet		Non fait. Pas de tabac cherché	
	9. Rechercher une douleur thoracique	durée, évolution, installation		Incomplet		Non fait	
	10. Rechercher une dyspnée	durée, évolution, installation		Incomplet		Non fait	
	11. Rechercher des symptômes récents	Question + toux et fièvre spécifiquement		Incomplet		Non fait	
Ex Cp	12. GDS - RP - ECG	Les demander tous dans les 5 min		Incomplet / retardés		Non fait	
Ventilation	13. Discuter la ventilation	VNI / IOT / pas de ventilation		Confus, non argumenté		Non fait	
	14. Préparer Intubation ou VNI :	IOT : matériel spécifique VNI : idem		Préparation incomplète		Non fait (cf détails)	
	15. IOT / VNI : Préparer le ventilateur	Régler les paramètres		Préparation abordée, ne connait pas paramètres		Non fait	
	16. Traitement étiologique complet à la fin du scenario.	Complet (à préciser selon étio)		Incomplet		Non fait	
Communication	17. Énoncer le diagnostic au sein de l'équipe	Complet, clair, avec orientation de PEC		Incomplet, confus		Non fait	
	18. Expliquer la mise en place du traitement au patient,	Avant de débuter la VNI ou IOT explications claires au patient, réassurance		Confus		Non fait	
	19. Distribuer les rôles, équipe calme, communication claire entre les membres de l'équipe	S'assurer de la compréhension des autres membres de l'équipe, communication calme, dirigée entre les membres		Confus		Pas de communication ou chaotique	
	20. appeler le régulateur / le réa :	dans les 10 minutes; Présenter le patient au correspondant : présentation claire et synthétique, demande claire		Présentation confuse, retardée		Non fait	
Total							

Aide Notation

ITEMS	COMPLET	1	INCOMPLET	0.5	NON FAIT = 0
1. Installer le patient semi-assis	Immédiatement		Dans les 5 min MAX		Non fait
2. Mesure de la SpO2 et de la FR	Immédiat et correct		> 5 min ou incomplet		
3. Délivrer de l'oxygène dans les 5 min	Utilisation correcte et adaptée du dispositif MHC bien gonflé, débit O2 et dispo adaptés à la SpO2		Mauvais dispositif, mal utilisé, à 5 min		Non fait
4. Recherche des signes de gravité	A voix haute, les identifier (doivent les demander à prioiri) : signes de lutte : tirage, RPA.. + cyanose		Pas assez exhaustif		Non fait
5. Rechercher une défaillance HDQ	Prendre PA, recherche signe d'insuffisance circulatoire (marbrures, TRC...)		Manque élément		Pas de recherche
6. Examen cardio-pulmonaire	Auscultation cardiaque et pulmonaire + signes IC		Incomplet		Pas fait
7. Interroger les proches / patient	Questions claires, bien dirigées, pertinentes		Confus		Non fait
8. Interrogatoire à la recherche d'une étiologie selon le terrain :.	FRCV (dont tabac), FRMTEV, ATCD cardio/pneumo/allergiques/voyage		Incomplet		Non fait
9. Rechercher une douleur thoracique :	durée, évolution, installation		Incomplet		Non fait
10. Rechercher une dyspnée :	durée, évolution, installation		Incomplet		Non fait
11. Rechercher des symptômes d'apparition récente	Poser la question et demander spécifiquement toux et fièvre		Incomplet		Non fait
12. GDS - RP - ECG	Les demander tous		Incomplet		Non fait
13. Discuter la ventilation	connaître les indications : VNI / Ventilation invasive / pas de ventilation		Confus		Non fait
14. Intubation ou VNI :	IOT : aspiration, demander les médicaments pour ISR, préparer le plateau : seringue, sonde VNI : réglage masque en expliquant au patient		Préparation incomplète		Non fait
15. IOT / VNI : Préparer le ventilateur	Régler les paramètres		Paramètres inexacts		Pas de réglage avant IOT / VNI
16. Traitement étiologique complet à la fin du scenario.	Complet (à préciser selon étio)		Incomplet		Non fait
17. Énoncer le diagnostic au sein de l'équipe	Complet, clair, avec orientation de PEC		Incomplet, confus		Non fait
18. Expliquer la mise en place du traitement au patient,	Avant de débuter la VNI ou IOT explications claires au patient, réassurance		Confus		Non fait
19. Distribuer les rôles, équipe calme, communication claire entre les membres de l'équipe	S'assurer de la compréhension des autres membres de l'équipe, communication calme, dirigée entre les membres		Confus		Pas de communication ou chaotique
20. appeler le régulateur / le réa :	dans les 10 minutes; Présenter le patient au correspondant : présentation claire et synthétique, demande claire		Présentation confuse, retardée (par ex, PEC terminée et pas d'appel)		Non fait

Evaluer les futurs médecins en situation simulée dans le contexte de l'urgence vitale : développement d'un outil et réflexions sur son apport dans le cadre d'une approche par compétence(s).

Résumé :

Introduction : il est essentiel de s'assurer que les compétences des étudiants et internes en médecine d'urgence, sont maitrisées et évaluées tout au long de leur formation. Notre objectif était de développer trois scores les évaluant par la simulation, dans trois situations cliniques d'urgence vitale : l'arrêt cardiaque (ACAT 1), le coma (ACAT 2) et la détresse respiratoire aigüe (ACAT 3).

Méthode : en suivant le processus de validation d'un test de Downing, nous avons mené une recherche en trois phases. La première s'intéressait au développent du contenu des scores, grâce à une méthode Delphi, la deuxième a consisté en une étude du processus de réponse puis en une validation de leur reproductibilité d'après des sessions de simulation filmées. La troisième phase a consisté en une évaluation de leurs propriétés psychométriques en contexte de simulation in-situ, ainsi qu'en une évaluation étude de leur utilisabilité et de leur acceptabilité auprès des enseignants.

Résultats : 3 scores ACAT (Acute Care Assessment Tool) de 20 items ont été obtenus après consultation de 51 experts. Après analyse de 43 vidéos et de 104 situations in-situ, impliquant respectivement 144 et 314 apprenants, la fiabilité des scores était bonne avec une reproductibilité excellente (coefficients de corrélation intra-classe tous > 0.88) et une bonne cohérence (coefficients alpha de Cronbach tous > 0.73). Les scores ACAT 1 et 3 avaient une bonne validité de construit. Les enseignants estimaient que les scores ACAT étaient utilisables, mais ont discuté les changements induits avec l'évaluation par la simulation, notamment la crainte de la perte d'une certaine bienveillance.

Conclusion : trois scores d'évaluation valides et fiables ont pu être développés pour mesurer la performance des étudiants et internes en médecine d'urgence, dans trois situations d'urgence vitale. L'acceptabilité de l'évaluation par la simulation est questionnée et devrait être analysée dans l'optique de son déploiement pour des évaluations à enjeu élevé.

Mots-clés : évaluation par la simulation, médecine d'urgence, étudiants en médecine, internes, approche par compétences

Abstract :

Introduction: it is critical that competency in emergency medicine is achieved and assessed during both students and residents' training. Our work aimed to develop three assessment scales thanks to a simulation-based assessment (SBA) technique, within three life-threatening situations: cardiac arrest (ACAT 1), coma (ACAT 2) and acute respiratory failure (ACAT 3).

Methods: following the Downing validation process framework, we conducted a three-phase prospective study. Phase 1 involved the development and content validation of the scores throughout a Delphi study. Phase 2 included an evaluation of their reproducibility, thanks to video-recorded simulation sessions. Phase 3 involved the psychometric testing of the tool thanks to interprofessional in situ simulations and also the usability and acceptability testing of the scores, with faculty, through a questionnaire and two focus groups.

Results: 3 20-items ACAT (Acute Care Assessment Tool) were developed after the consultation of 51 experts. After the analyze of 43 videos and 104 in situ simulation sessions, which involved respectively 144 and 314 learners, we found a good reliability with excellent reproducibility (all ICC > 0.88) and good internal consistency (all Cronbach's alpha >0.73). The usability of the ACAT was positive but the raters discussed the changes introduced by SBA, mainly a certain loss of benevolence.

Conclusion: a valid and reliable tool was developed to measure performance of both medical students and residents for three life-threatening situations. The acceptability of SBA is questioned and need to be analyzed in order to implement simulation-based high-stakes assessments.

Keywords: simulation-based assessment, emergency medicine, medical students, residents, competency-based medical education.