

Évaluation Sommative et Simulation en Santé

TEXTE LONG

SOMMAIRE

Chapitre 1 : Résumé	p4
Chapitre 2 : Introduction	p5
Chapitre 3 : Méthodologie	p7
Chapitre 4 : La théorie de l'évaluation sommative en simulation	p7
1. Que peut-on évaluer ?	p7
a. Qu'est-ce qu'une compétence ?	p8
b. Peut-on évaluer une compétence ?	p13
c. Peut-on évaluer une compétence lors d'une séance de simulation ?	p14
d. Quelles compétences peuvent être inférées lors d'une séance de simulation ?	p15
e. Quelles compétences peuvent être certifiées ?	p16
f. Quand peut-on utiliser la simulation pour une évaluation sommative des compétences en formation initiale ? en formation continue ?	p17
g. Quelles sont les conditions nécessaires à l'évaluation d'une compétence inférée lors d'une séance de simulation ?	p18
2. Les outils d'évaluation sommative	p23
a. Quelles sont les qualités d'un outil adapté à la mesure souhaitée ?	p23
b. Comment créer un outil d'évaluation ?	p34
c. Outil d'évaluation et variation des contextes	p43
3. Conséquences de l'ES par la simulation sur les personnes formées et les formateurs	p45
a. Conséquences pédagogiques : impact sur les stratégies d'apprentissage	p45
b. Conséquences psychologiques ?	p49
Chapitre 5 : Les outils de l'évaluation sommative en simulation	p60
1. Les scénarios en ES	p60
a. Comment définir les thématiques des scénarios ?	p61
b. Quelles thématiques, objectifs, compétences ?	p62
c. Quelles sont les étapes et les éléments structurants pour la création d'un scénario d'ES ?	p62
d. Doit-on proposer des séquences courtes ou longues pour les scénarios d'ES ?	p63
e. La standardisation des scénarios est-elle évoquée dans la littérature ?	p63
f. Les outils d'évaluation doivent-ils être présents dans le scénario ?	p64
g. Peut-on utiliser les mêmes scénarios pour l'évaluation formative et l'ES ?	p65
h. Quelles sont les particularités des scénarios pour l'ES ?	p65
i. Quelle évaluation du scénario ?	p66
j. Quel scénario type pour la simulation procédurale en ES ?	p66
2. Place du Débriefing dans l'ES ?	p67

3. Place de la vidéo dans l'ES ?	p68
4. Organisation des infrastructures en simulation pour l'ES	p71
5. Place de la recherche en ES ?	p73
6. Le GBPMSS est-il utilisable en ES en simulation ?	p74
7. Formation des Formateurs pour l'évaluation sommative en simulation :	p75
a. Qui peut faire une évaluation sommative par le simulation ?	p75
b. Les formateurs doivent-ils avoir une formation spécifique pour pratiquer l'évaluation sommative ?	p76
c. Proposition de formation des formateurs en Évaluation Sommative	p79

Chapitre 6 : En pratique : mise en place de l'ES en simulation p85

1. Les grandes étapes de la mise en place	p85
a. L'équipe	p87
b. Le tableau de spécification (« blue-print ») en lien avec le référentiel de compétence	p88
c. Articulation avec le curriculum et les autres modalités d'évaluation	p89
d. Longueur / nombre de stations	p90
e. Développer une banque de scénarios	p91
f. Entraîner des évaluateurs	p94
g. Développer un pool de patients standardisés	p96
h. Déroulement de l'épreuve proprement dite	p97
i. Après l'examen	p99
j. Processus qualité	p99
k. Conclusion sur la mise en place de l'Examen	p100
2. Exemples d'Évaluation Sommative en simulation en santé	p102
a. Étudiants 1 ^{er} / 2 ^{ème} cycle des différentes filières	p102
b. Internes dans différentes disciplines médicales	p104
3. Place de la simulation Sommative dans la re-certification	p113
a. Un essor des programmes de re-certification	p113
b. Un usage de la simulation dans les processus de re-certification	p114
c. Ce qui reste identique par rapport à la démarche certificative initiale (tab. 11):	p115
d. Ce qui est particulier à la recertification (tab.11)	p115

Chapitre 7 : Lexique p120

Bibliographie p123

Groupe d'élaboration du référentiel p140

Chapitre 1 : Résumé

En santé, la simulation est, jusqu'à présent, surtout utilisée pour l'**évaluation formative** des compétences. La question se pose aujourd'hui des conditions et des modalités de son utilisation en **évaluation sommative** (ES), au terme d'un apprentissage afin d'en objectiver la maîtrise acquise.

La simulation en santé peut être utilisée pour l'ES, en complément des outils existants d'évaluation. En simulation, ce sont les compétences qui sont approchées au travers de l'observation d'une performance en situation, qui varie en fonction de facteurs humains (stress, fatigue, etc.). La disponibilité ainsi que les limites des différents outils de simulation doivent être prises en compte. Les évalués doivent avoir accès à l'outil simulation et aux outils d'évaluation pendant les phases d'apprentissage.

Ces **outils d'évaluation** doivent être valides (mesurant précisément le concept visé), reproductibles et acceptables dans leur utilisation. Cependant, leur processus de validation et/ou la formation nécessaire à l'apprentissage de ces outils sont souvent insuffisamment décrits.

Un ensemble de **scénarios** d'ES explorant différentes compétences doit être créé. Ils doivent être prévisibles, programmables, standardisables, reproductibles et inclure, là où ils sont pertinents, les outils d'ES.

Pour limiter les biais liés à tout processus d'évaluation, les formateurs en ES doivent avoir suivi eux-mêmes des formations ciblées adaptées aux 3 rôles nécessaires pour la réalisation d'ES :

- (i) des formateurs-concepteurs (conception de la situation, sélection des outils d'ES) garants de l'encadrement de la session d'ES et de la formation des formateurs-évaluateurs ;
- (ii) des formateurs-opérateurs responsables du déroulé de la situation de simulation support de l'évaluation ;
- (iii) des formateurs-évaluateurs réalisant l'ES en utilisant les outils d'ES auxquels ils auront été formés et entraînés spécifiquement.

Toute **infrastructure de simulation** peut pratiquer l'ES. Comme pour l'évaluation formative, il est probable que l'évaluation « in situ » renforce le réalisme pour l'évalué. Concernant la place du **débriefing** en ES, plusieurs approches peuvent être envisagées :

- Ne pas faire de débriefing. L'ES est factuelle, centrée sur la performance réalisée lors de la simulation, proche d'une performance réelle auprès d'un patient et moins chronophage.
- Un débriefing intégré dans l'ES : permet d'intégrer, notamment dans la notation, les intentions et les schémas cognitifs des évalués mais aussi de prendre en charge d'éventuelles conséquences psychologiques de l'ES.
- Un retour d'expérience (hors notation de l'ES) de la part du formateur-évaluateur, individuel ou en groupe, immédiat ou différé, pour spécifier a posteriori les attendus de la prise en charge.
- L'utilisation de la **vidéo** n'est pas indispensable. Cependant, dans le respect de la réglementation en vigueur, elle peut être utilisée pour réaliser une ES à distance ou servir de preuve de la performance.
- La **recherche** est nécessaire pour continuer d'affiner et d'améliorer les conditions de réalisation d'ES, d'optimiser l'alignement pédagogique entre le formatif et le sommatif, de créer et de valider les outils de simulation utilisés en ES.

Chapitre 2 : Introduction

La définition de la simulation en santé, retenue par la Société Francophone de Simulation en Santé (SoFraSimS), 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 concepts médicaux ou des prises de décision par un professionnel de santé ou une équipe de professionnels.» ¹

La simulation en santé se développe rapidement en francophonie. Jusqu'à présent, elle était surtout utilisée dans le cadre d'une évaluation formative des compétences en proposant des solutions d'amélioration en cas de déficit de performance.

La question de l'utilisation de la simulation dans le cadre d'une évaluation sommative des compétences se pose aujourd'hui. On rappellera d'emblée les distinctions qu'il convient de faire entre, d'une part l'évaluation formative (ou diagnostique, etc ...)

utilisée tout au long de la phase d'apprentissage et dont la finalité essentielle est d'analyser/dynamiser cet apprentissage et d'autre part l'évaluation sommative (ou sanctionnante ou certifiante ou finale) utilisée au terme d'un cycle/phase d'apprentissage et dont la finalité essentielle est d'objectiver/garantir l'apprentissage. L'évaluation sommative diffère donc de l'évaluation formative dans son objet qui est de « sanctionner un apprentissage » ce qui implique un surcroît d'exigence de ses qualités métrologiques pour valider, ou non, l'acquisition de la ou les compétences.

Une compétence - par définition non scorable - est soit acquise ou non acquise. Mais compte tenu de la culture de l'évaluation en France et des objectifs de sélection et de classement des étudiants, sa validation devra toutefois être associée à une note ou plus précisément un score. Le présent travail vise à guider cette double nécessité, à savoir valider une compétence et lui attribuer un score (l'équivalent d'une note pour un examen pratique) à l'aide de la simulation.

À la suite d'échanges lors du congrès de la SoFraSimS en 2019 à Strasbourg, un groupe de travail a été mis en place. Une lettre de mission adressée par le président de la SoFraSimS a permis de réunir des professionnels de santé francophones en proposant des exemples de problématiques concernant l'utilisation de la simulation en évaluation sommative :

- Les scénarios d'évaluation sont-ils les mêmes que pour les évaluations formatives ?
- Les référentiels de la SoFraSimS sont-ils utilisables ?
- Les formateurs doivent-ils avoir une formation spécifique ?
- Un équipement spécifique est-il nécessaire ?
- Une adaptation des outils de simulation est-elle nécessaire ?
- Etc...

A la demande de la SoFraSimS, les objectifs de ce groupe de travail étaient de synthétiser les problématiques et de faire une revue de la littérature permettant la mise en place de l'ES par la simulation en santé. Ce travail a été orienté principalement vers les professionnels de santé (corps médical, mais aussi pour les autres professionnels de santé pharmaciens, Infirmiers, sage femmes, dentistes, kinésithérapeutes, etc..) et ne décrit pas l'évaluation globale des équipes de professionnels de santé.

Chapitre 3 : Méthodologie

- Les items ainsi que les questions posées nécessaires à la rédaction de ce travail ont été initialement proposés puis validés par le groupe de travail ;
- La méthodologie utilisée est issue de la méthode « Recommandations par consensus formalisé » publiée par la haute autorité de santé (HAS) ² ;
- La quantification de l'accord professionnel pour la validation des items a été obtenue par la méthode du groupe nominal (proche de la méthode Delphi) ³ :
 - L'accord est obtenu si 70%, ou plus des participants sont d'accord avec la proposition, au jour de la date limite de travail ;
 - Un rejet de la proposition est obtenu si moins de 30% des participants sont d'accord avec la proposition, au jour de la date limite de travail ;
 - Un score entre 30 et 70% entraîne un débat et une reformulation de la proposition par le groupe de travail, encadré par le modérateur, puis une nouvelle soumission au vote ;
- Après cet accord, sept sous-groupes ont travaillé sur les différents items, pratiqué une analyse critique des données bibliographiques, et rédigé un texte court et un texte long des recommandations. Ces recommandations ont secondairement été soumises à un groupe expert extérieur de relecture ;
- Après relecture, le groupe a finalement proposé une version initiale des recommandations ;
- Le texte a ensuite été transmis à la SoFraSimS (Conseil d'Administration) pour en valider le contenu final.

Chapitre 4 : la Théorie de l'évaluation sommative en simulation

1. Que peut-on évaluer ?

Les enseignants et l'institution facultaire en sciences de la santé doivent s'assurer que chaque diplômé est prêt à exercer la profession. L'approche par compétences implique des définitions explicites de tous les domaines de compétences essentielles à acquérir pour être « un bon professionnel ». La compétence professionnelle pourrait se définir comme la capacité d'un professionnel à utiliser son jugement, de même que

les connaissances, les habiletés et les attitudes associées à sa profession pour résoudre des problèmes complexes.¹⁻⁴

a. Qu'est-ce qu'une compétence ?

i. Définition(s)

La définition de la compétence est polysémique et est très variable dans la littérature. Or, un manque de consensus sur les définitions et les termes est un facteur limitatif au développement d'un programme de formation et d'évaluation basé sur les compétences.

En langue anglaise, les mots COMPETENCE et COMPETENCY sont parfois utilisés indifféremment bien qu'ils n'aient pas exactement la même signification. Le mot COMPETENCE représente l'éventail des capacités dans plusieurs domaines ou aspects de la performance médicale dans un certain contexte. Les déclarations sur les COMPETENCES nécessitent des qualificatifs descriptifs pour définir les capacités pertinentes, le contexte et le stade de formation. La COMPETENCE est multidimensionnelle et dynamique. Elle change avec le temps, l'expérience et l'environnement. Le mot COMPETENCY se définit comme la capacité observable d'un professionnel de la santé, intégrant de multiples composantes telles que les connaissances, les aptitudes ou compétences (SKILLS), les valeurs et les attitudes est également utilisé. Les COMPETENCIES peuvent être rassemblées comme des blocs de construction pour faciliter un développement progressif. Elles sont observables et peuvent être mesurées et évaluées pour garantir leur acquisition.⁵

Certains auteurs définissent la compétence professionnelle comme la capacité d'un professionnel à utiliser son jugement, de même que ses connaissances, ses habiletés et les attitudes associées à sa profession pour résoudre des problèmes complexes.^{4,6}

Dans le cadre de ce travail, nous retiendrons la définition de Jacques Tardif pour qui la compétence se définit comme « 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 ». Pour Tardif, la compétence n'est pas directement observable.⁷ Les compétences s'étendent donc au-delà des connaissances médicales et des compétences procédurales.⁸ Une compétence n'est pas la simple aptitude à réaliser un geste difficile mais défini à l'avance. Il s'agit de choisir, de construire, d'adapter, voire de créer le geste approprié.

La compétence est donc souvent plus qu'un geste, mais une véritable stratégie d'action, faite d'opérations mentales et de gestes plus visibles, mais qui ne sont compréhensibles qu'en fonction du raisonnement professionnel qui les sous-tend. ⁹ La compétence est donc liée à un contexte. ⁴

Ainsi, la compétence exige deux conditions :

- des ressources cognitives acquises, notamment des savoirs et savoir-faire constitués. ⁹ Miller place les connaissances à la base de sa pyramide de l'évaluation ⁶. Il paraît évident qu'il faut posséder les connaissances pour savoir comment faire (compétence). Le « savoir » est donc la base de la compétence qui nourrit le « savoir comment »⁴. Certains auteurs proposent une modification du modèle de Miller en lui ajoutant une dimension dynamique.
- la capacité de mobiliser ces ressources, dans l'urgence et l'incertitude, en vraie grandeur, donc de les mettre en synergie, de les adapter, de les compléter, de les transposer, transférer, combiner de sorte à agir au mieux.⁹

ii. Quelles sont les compétences générales des professionnels de santé ?

Chaque profession de santé a un ou plusieurs référentiels pour définir les compétences attendues du professionnel.

A titre d'exemple, nous avons repris les référentiels des médecins, des infirmiers et des masseurs-kinésithérapeutes.

Pour les compétences professionnelles des médecins, il existe deux référentiels principaux. Le premier référentiel est celui de l'ACGME (Accreditation Council for Graduate Medical Education), publié pour la première fois en 2002 par le collège des pédagogues médicaux américains et le second est le CanMEDS, publié au Canada (Collège Royal des médecins et des chirurgiens du Canada).

L'ACGME ¹⁰ définit 6 domaines de compétences générales à atteindre par le professionnel de santé. Ces domaines de compétences sont les suivants :

1. Savoir médical
2. Prise en charge du malade
3. Apprentissage et amélioration continue
4. Relations interpersonnelles et communication
5. Professionnalisme
6. Intégration dans le système de soins

Quant au CanMEDS, il a choisi les compétences - connaissances, habiletés et attitudes observables importantes - comme concept central de la planification de l'éducation médicale. En tant qu'experts en santé, les professionnels de santé jouent tous les rôles du CanMEDS et utilisent leur savoir médical, leurs compétences spécialisées cliniques et leurs attitudes professionnelles pour dispenser des soins axés sur les patients. Le rôle d'expert en santé est le rôle pivot du professionnel de santé dans le cadre CanMEDS. ⁸

Le professionnel de santé expert est, selon le CanMEDS, : ⁵

- ▀ Un communicateur qui facilite efficacement la relation médecin-patient et les échanges dynamiques qui se produisent avant, pendant et après le contact médical.
- ▀ Un collaborateur qui travaille efficacement dans une équipe de soins de santé afin de prodiguer des soins optimaux aux patients.
- ▀ Un gestionnaire qui participe à part entière à l'organisation de soins de santé où il organise des pratiques durables, prend des décisions sur l'affectation des ressources et contribue à l'efficacité du système de soins de santé.
- ▀ Un promoteur de la santé qui utilise son expertise et son influence de façon responsable pour promouvoir la santé et le mieux-être de patients en particulier, de collectivités et de populations.
- ▀ Un érudit qui démontre un engagement de toute une vie envers l'apprentissage fondé sur la réflexion, ainsi que la création, la diffusion, l'application et l'utilisation de connaissances médicales.
- ▀ Un professionnel qui est voué à la santé et au mieux-être de la personne et de la société, à la pratique respectueuse de l'éthique, à l'autoréglementation de la profession et aux critères rigoureux de comportements personnels.

En ce qui concerne les compétences en soins infirmiers, les compétences professionnelles sont reprises dans l'annexe II de l'Arrêté du 31 juillet 2009 relatif au diplôme d'État d'infirmier.¹¹ Le référentiel reprend 10 compétences, à savoir :

1. Évaluer une situation clinique et établir un diagnostic dans le domaine infirmier,
2. Concevoir et conduire un projet de soins infirmiers,
3. Accompagner une personne dans la réalisation de ses soins quotidiens,
4. Mettre en œuvre des actions à visée diagnostique et thérapeutique,

5. Initier et mettre en œuvre des soins éducatifs et préventifs,
6. Communiquer et conduire une relation dans un contexte de soins,
7. Analyser la qualité des soins et améliorer sa pratique professionnelle,
8. Rechercher et traiter des données professionnelles et scientifiques,
9. Organiser et coordonner des interventions soignantes,
10. Informer et former des professionnels et des personnes en formation.

Concernant les masseurs-kinésithérapeutes, il existe également un référentiel de compétences (Le référentiel de la profession) qui reprend 5 compétences propres au métier de masseur-kinésithérapeute et 5 compétences communes à d'autres professions de santé.¹² Il s'agit pour les compétences propres au métier de :

1. Conduire et réaliser une évaluation clinique en kinésithérapie analytique et globale du patient,
2. Élaborer un diagnostic kinésithérapique,
3. Concevoir et conduire une stratégie d'intervention en kinésithérapie,
4. Mettre en œuvre la prise en charge kinésithérapique,
5. Concevoir et conduire une démarche de conseil, d'éducation, de prévention et de dépistage en kinésithérapie et en santé publique,

Et pour les compétences communes à d'autres professions de :

1. Évaluer, analyser et améliorer sa pratique professionnelle
2. Concevoir un projet de recherche,
3. Gérer une structure et ses ressources,
4. Participer à la continuité des soins et à des actions de santé publique,
5. Informer et former des étudiants et des professionnels.

iii. Quelles sont les compétences d'équipe dans les soins de santé ?

Les compétences d'équipe peuvent être définies comme les actions cognitives, comportementales et psychologiques dont les membres de l'équipe ont besoin pour fonctionner efficacement dans le cadre d'une équipe solidaire.

Les compétences du travail d'équipe consistent en un ensemble de cinq compétences qui sont l'adaptabilité, le leadership, le comportement de sauvegarde, la surveillance mutuelle des performances et l'orientation d'équipe. Ces

compétences sont liées par des processus de communication qui aboutissent à des modèles mentaux communs se rapportant à la tâche, à l'équipe et à la situation.¹¹ Les compétences d'équipe sont cependant difficiles à évaluer.¹³

iii. Les compétences techniques ou procédurales et les compétences non-techniques.

Il existe parfois une confusion concernant le terme skill(s) utilisé dans la littérature anglo-saxonne indifféremment pour désigner une habileté et une compétence. Il faudra donc être vigilant dans la traduction et tenir compte de la définition reprise au point 1. a. De plus, de nombreuses compétences techniques nécessitent des compétences non techniques pour être mise en œuvre.

Les compétences non techniques sont en général décrites comme difficiles à définir. Dans la littérature, sept compétences non techniques sont retrouvées ont cependant été décrites. Il s'agit de la conscience de la situation, de la prise de décision, de la communication, du travail d'équipe, du leadership, de la gestion du stress et de la gestion de la fatigue.⁸ Pour les médecins, par exemple, le rapport Medical Schools Objectives Project 14 de l'Association of American Medical Colleges énumère les procédures techniques de routine minimales qu'un étudiant en médecine devrait être capable d'effectuer avant d'obtenir son diplôme de médecine. Parmi celles-ci, nous retrouvons l'anamnèse, l'examen clinique, des gestes techniques (ponction veineuse, ponction lombaire, ...) ou encore l'interprétation d'examens comme l'ECG ou les examens radiologiques.¹⁵ Ces procédures techniques (habiletés) nécessitent des compétences non techniques comme des compétences de communication pour réaliser une anamnèse.

Un autre exemple de compétence technique spécifique aux spécialités prenant en charge des patients critiques est celui de l'intubation trachéale. Dans le cadre de l'intubation, nous pouvons distinguer le geste technique en lui-même (habileté) et la compétence technique d'intubation qui nécessitera, notamment, de pouvoir évaluer la difficulté, de choisir le matériel adapté au patient, de mettre en œuvre la sédation nécessaire à la réalisation du geste, etc.

De manière plus complexe encore, nous pouvons prendre l'exemple du choc anaphylactique proposé par Bond et al.¹⁶ pour les résidents de médecine d'urgence. Dans cet exemple, la compétence de prise en charge de cette situation critique

nécessitera d'abord que le résident ait conscience de la situation, qu'ensuite il ait des connaissances au sujet du choc et de son traitement, avant de prendre la décision d'intervenir et de finalement mettre en œuvre les habilités techniques telles que l'administration d'adrénaline ou l'intubation.

b. Peut-on évaluer une compétence ?

La compétence professionnelle est un état cognitif dynamique interne aux individus, non directement observable.¹⁷ La compétence professionnelle ne peut être décrite qu'en s'appuyant sur une construction théorique. Cette construction multivariée est composée de différents éléments qui forment un tout, ce tout représentant beaucoup plus que la somme de ses parties. Pour évaluer la compétence professionnelle, il faudra donc faire des inférences à partir d'éléments observables et mesurables,⁴ tels que l'ensemble des connaissances qui participent à cette construction, par exemple en évaluant le comportement, les connaissances, la communication, le professionnalisme, le raisonnement clinique ou encore le travail en équipe, ...¹⁸

C'est donc la performance en situation qui est observable. Il faut cependant tenir compte que la performance peut varier en fonction de facteurs humains (fatigue, stress, etc.) ou d'obstacles pouvant être rencontrés dans la situation qui empêchent que la compétence puisse pleinement être mise en œuvre.¹⁹

L'évaluation d'une compétence médicale doit donc s'attacher à apprécier le profil de fonctionnement cognitif de l'individu dans un contexte donné plus ou moins contrôlé ou aléatoire.⁴ Cependant, aucune méthode d'évaluation prise isolément ne peut donner l'ensemble des informations requises pour évaluer quelque chose d'aussi complexe que le travail d'un médecin accompli.⁶ De plus, les différentes compétences ne seraient pas évaluables de manière indépendante pour l'ACGME.¹⁰ Le niveau de formation doit également être pris en compte car la performance observée pour une même compétence sera différente s'il s'agit d'un novice ou d'un expert. Ainsi, des compétences partielles peuvent être évaluées chez les étudiants novices (en fin de cursus préclinique par exemple). L'évaluation portera sur des compétences plus achevées mais indifférenciées, en fin de cursus clinique et finalement sur des compétences abouties et différenciées chez les étudiants type « expert » en fin de 3ème cycle.¹⁷

c. Peut-on évaluer une compétence lors d'une séance de simulation ?

Il est possible d'évaluer lors de séances de simulation des compétences, telles que définies au point 1.a. via l'observation d'actions « clefs » à l'aide d'outils d'analyses spécifiques. Il faut cependant tenir compte des limites technologiques de l'environnement de simulation et des mannequins car toutes les situations ne peuvent être reproduites, ce qui peut rendre difficile l'évaluation de ces situations à l'aide de la simulation.

Ainsi, l'évaluation des compétences procédurales est variable d'une spécialité à l'autre. L'évaluation des compétences procédurales à l'aide d'un simulateur évite les préjudices potentiels pour les patients et a été décrite comme un impératif éthique bien que l'utilisation d'animaux ou de cadavres humains reste problématique.²⁰

La simulation hybride implique plus d'un type de simulation et peut être utilisée pour mettre des compétences procédurales en contexte. Par exemple, un acteur joue le rôle d'un patient pour évaluer le consentement et les compétences de communication, tandis qu'un mannequin attaché évalue le placement d'un cathéter i.v.²⁰

De plus, toutes les méthodes de simulation ne sont pas équivalentes dans l'évaluation des compétences. Par exemple, le patient simulé permet difficilement d'évaluer certaines compétences procédurales associées à des compétences managériales à la différence d'une simulation avec un mannequin de haute technicité avec un scénario.²¹

Néanmoins, aucune méthode d'évaluation à elle seule ne permet d'évaluer la compétence professionnelle.⁶ C'est pourquoi il faut combiner plusieurs méthodes d'évaluation pour apprécier au mieux, un objet aussi multiple et complexe que la compétence professionnelle. C'est la raison pour laquelle, la compétence d'expert médical du CanMEDS est traditionnellement évaluée à l'aide d'examens écrits, oraux et d'ECOS (Examen Clinique Objectifs Structurés ou OSCE en Anglais). Les compétences en communication et en entrevue médicale sont également évaluées à l'aide d'ECOS. Les autres compétences intrinsèques du CanMEDS ainsi que les compétences non techniques sont généralement évaluées dans l'environnement de formation, en milieu de travail et sont sous-représentées dans les évaluations formelles. Cela étant justifié par le fait que les compétences non techniques seraient difficiles à définir et à évaluer.⁸

d. Quelles compétences peuvent être inférées lors d'une séance de simulation ?

Les simulations médicales, en général, visent à mettre en scène des patients simulés, diverses régions de l'organisme, des mannequins ou des tâches cliniques, et / ou à refléter les circonstances réelles dans lesquelles les services médicaux sont rendus.²²

En dehors des mises en situations réelles ou de simulation, la plupart des évaluations portent principalement sur le domaine cognitif pur. La simulation (haute-technologie, patients simulés, simulateurs de tâche partielle), grâce à des mises en situation, permet d'évaluer ^{6,16,18,22-26}

- ▶ les connaissances médicales (ex. l'algorithme de prise en charge de l'activité électrique sans pouls)
- ▶ des habiletés procédurales individuelles (pose de voie veineuse mais aussi interrogatoire et examen physique de base) ainsi que les tâches sensibles (ex. examen pelvien) ou à risque de lésions pour les patients (ex. intubation)
- ▶ de manière générale, les compétences non techniques. Il existe cependant peu d'outils validés pour leur évaluation ; ²⁷
 - les soins aux patients
 - des habiletés de communication avec le patient
 - des habiletés de communication individuelle et en équipe sur des scénarios ayant recours à des mannequins « haute-fidélité ».
 - la prise de décision
 - la priorisation des soins
 - la façon d'interagir avec les patients ainsi qu'avec les autres membres de l'équipe (ex. discussion interprofessionnelle sur les cas complexes)
 - le raisonnement clinique (les situations rencontrées lors des mises en situation doivent être des symptômes et non des diagnostics)
 - le travail d'équipe et son efficacité (ex. scénario de réanimation blue code)
 - l'empathie ⁵
 - le professionnalisme
 - la construction d'une culture orientée sur la sécurité
 - l'application des sciences fondamentales
 - le bien-être au travail et le risque de burn out ? : Il n'existe pas de travail spécifique s'intéressant à cette problématique. Nous pouvons cependant

émettre l'hypothèse que cela pourrait être possible. En effet, différents éléments participant au tableau d'épuisement professionnel (ex. manque d'empathie ou de professionnalisme) peuvent être évalués à l'aide de la simulation.

- les compétences nécessaires à la prise en charge des incidents rares et/ou critiques.

Notons qu'au cours d'une évaluation, il sera possible de valider les différentes actions faites, leur minutage mais difficilement l'ordre dans lequel elles doivent être faites.²³

e. Quelles compétences peuvent être certifiées ?

Si la simulation est utilisée dans le cadre de l'évaluation formative dans de nombreux domaines des soins de santé, il existe peu de domaines qui recourent à la simulation de manière certificative.

Évaluer de manière certificative consiste, selon Stuffelbaum cité par Romainville ²⁸ à mesurer puis à apprécier, à l'aide de critères, l'atteinte des objectifs d'enseignement, en trois étapes. La première étape réside dans le recueil systématique, valide et fidèle d'informations appropriées aux objectifs d'enseignement. C'est la phase d'observation et/ou de recueil de données. La deuxième étape consiste à interpréter les informations recueillies à l'aide de critères. C'est la phase d'analyse. Celle-ci débouche sur la troisième étape, à savoir l'établissement de conclusions et/ou la prise de décisions. C'est la phase du jugement à proprement parler.

De manière théorique, toutes les compétences qui peuvent être inférées de manière formative lors de séances de simulation peuvent être évaluées de manière certificative. Néanmoins, les professionnels de santé ne devraient être évalués que sur des compétences qui sont directement pertinentes pour les soins aux patients.²⁹

De plus, la plupart des auteurs s'accordent sur la nécessité d'outils de mesure valides et fiables afin de réduire la subjectivité dans la certification des compétences. Ce sont les qualités psychométriques ou métrologiques des outils de mesure utilisés qui déterminent si la simulation peut être utilisée dans la certification d'une compétence spécifique. Notons qu'il existe, en général, une corrélation positive entre les scores obtenus lors d'une validation par simulation et les examens facultaires.²¹ Néanmoins, les examens facultaires classiques n'ont pas toujours toutes les qualités psychométriques pour l'évaluation de la compétence clinique. Des auteurs suggèrent notamment que la performance de réanimation en simulation peut être un indicateur de compétence clinique pour les résidents de médecine d'urgence ³⁰.

f. Quand peut-on utiliser la simulation pour une évaluation sommative des compétences en formation initiale ? en formation continue ?

L'objectif ultime de toute formation est d'assurer la compétence professionnelle des apprenants.³¹ L'évaluation sommative s'inscrit donc dans un continuum de la formation.³² L'évaluation ou observation formative et l'évaluation certificative seraient deux phases du même travail, fondées sur les mêmes données, mais utilisées dans des postures et à des fins différentes à savoir que lorsqu'il reste du temps, les données sont réinvesties dans des stratégies de formation ; c'est la fonction formative, qui n'exclut pas, bien au contraire, des bilans périodiques des acquis et que lorsqu'il ne reste plus de temps, que l'on s'approche de la fin d'une formation (cycle d'étude ou module), il est temps de dresser un bilan consolidé ; c'est la fonction sommative.⁹

Une évaluation sommative est donc réalisée à la fin d'une période de formation. Les évaluations sommatives attribuent généralement une note ou un échec / réussite et ont été décrites comme une « évaluation de l'apprentissage ». ²⁰ Les évaluations sommatives peuvent donc :

- ▀ Faire partie des progrès réalisés dans le cadre d'un programme de formation fondé sur les compétences,
- ▀ Être requises avant de pouvoir se voir confier des responsabilités plus importantes (ex. sélection de candidats lors de l'entrée en formation de spécialité,^{25,33} avant de débiter des gardes de première ligne, ...). Le succès de l'apprentissage par simulation axé sur les résultats suggère que l'utilisation de la simulation comme évaluation avant la pratique indépendante peut avoir un impact positif sur la sécurité des patients. À l'heure actuelle cependant, aucune recherche n'a démontré la validité prédictive d'une évaluation utilisant une simulation. Une des priorités de la recherche future sur la sécurité des patients devrait être de démontrer qu'une performance compétente sur un simulateur se traduit par une compétence réelle sur les patients.²⁰
- ▀ Faire partie de la certification d'une spécialité particulière ou de la revalidation d'une licence médicale dans le cadre du développement professionnel continu (ex. anesthésie-réanimation, chirurgie) ³⁴ ou après un incident/accident ou une période sans pratique. ³⁵
- ▀ Permettre d'évaluer la qualité de la formation dispensée.

g. Quelles sont les conditions nécessaires à l'évaluation d'une compétence inférée lors d'une séance de simulation ?

Lors de la validation d'une compétence, ce n'est pas une performance simple mais un fonctionnement cognitif complexe qui est évalué. Dans ce contexte, aucun instrument de mesure, pris isolément, n'est suffisant pour certifier la compétence professionnelle du médecin.³⁶ L'utilité de chaque méthode d'évaluation est toujours un compromis entre divers aspects de la qualité. Les cinq critères sur lesquels un tel compromis doit être fait sont : la validité ; la fiabilité ; l'impact éducatif ; la rentabilité et l'acceptabilité. De plus, il est généralement admis que la résolution d'un ou deux cas ne fournit jamais les preuves suffisantes pour étayer des conclusions générales sur l'expertise d'un candidat en matière de résolution de problèmes. De nombreux travaux ont montré que la fiabilité d'un ECOS dépend d'un échantillonnage minutieux, en particulier à travers le contenu clinique, et d'un nombre approprié de stations, ce qui signifie généralement que plusieurs heures de temps de test sont nécessaires.³⁷

L'élaboration de l'examen comprend trois étapes :

- a. la définition des conditions cliniques que les résidents sont tenus de gérer avec compétence,
- b. la définition des tâches se rapportant à chacune de ces conditions, et
- c. l'incorporation des tâches dans des postes d'examen basés sur une simulation pratique au format ECOS.³⁸

Les conditions nécessaires pour pouvoir évaluer et, le cas échéant, valider une compétence inférée lors d'une séance de simulation sont les suivantes :

- La compétence ou les compétences à évaluer seront clairement définies.³⁹ Afin d'évaluer une compétence, Scallon recommande de déterminer les savoirs, savoir-faire et savoir-être dont la mobilisation définit la compétence, puis de concevoir ou inventer des tâches concrètes qui vont fournir des indices pour inférer cette compétence.⁴⁰ Les compétences techniques devraient être évaluées si elles sont exécutées couramment ou si elles peuvent potentiellement sauver des vies. Cependant, la détermination des compétences de base pour les manœuvres permettant de sauver des vies peut être controversée, comme l'a montré un récent débat sur la question de savoir si une compétence en cricothyrotomie peut raisonnablement être attendue. Les compétences de base changent également à mesure que la

technologie et les connaissances médicales se développent.²⁰ Pour évaluer des comportements novices, les évaluations seront standardisées en utilisant des check-lists adaptées. Pour les compétences au niveau expert, il y a la nécessité de situations d'évaluations ayant une réelle complexité cognitive et professionnelle.¹⁷

- L'évaluation des compétences doit être authentique. Wiggins (1989) cité par Perrenoud définit les traits principaux d'une évaluation authentique ⁹ :
 - L'évaluation n'inclut que des tâches contextualisées. Les évaluations imiteront les tâches réelles de la profession.³⁹
 - L'évaluation porte sur des problèmes complexes.
 - L'évaluation exige l'utilisation fonctionnelle de connaissances disciplinaires
 - La tâche et ses exigences sont connues avant la situation d'évaluation.
 - L'évaluation exige une certaine forme de collaboration avec des pairs.

Les évaluations seront basées sur l'observation directe de l'évalué (et non des feedbacks des collaborateurs par exemple).³⁹

- Les évaluations seront basées sur de multiples mesures objectives et seront référencées par critères.³⁹ L'acquisition de compétences ou d'expertise dans une compétence procédurale requiert de l'expérience à travers un nombre de tentatives variable en fonction de la compétence, de la qualité de l'enseignement et des aptitudes de l'individu. Par exemple, en anesthésie, les compétences sont généralement acquises au départ dans des cas relativement simples. Lorsque la compétence est acquise dans ces situations, les apprenants sont exposés à un éventail plus large de variations normales et pathologiques nécessaires au développement d'une véritable expertise. Le maintien des habiletés motrices semble dépendre surtout du degré de perfectionnement de ces aptitudes.²⁰
- Des scénarios d'évaluation ciblée seront créés après avoir défini les compétences attendues à évaluer. Pour une évaluation à enjeux élevés, la conception de scénarios prenant en charge plusieurs évaluations plus courtes peut être plus efficace que moins de scénarios plus longs. Un scénario de situation critique ou complexe permet d'évaluer plusieurs compétences simultanément (techniques et non techniques). Il est également possible pour

évaluer plusieurs compétences de réaliser une évaluation au travers du passage en multi-stations.

- Les scénarios ayant un objectif d'évaluation sommative au cours de simulations standardisées devront notamment définir le rôle du facilitateur (briefing, voix du patient et des professionnels), la standardisation du scénario pour assurer une équité entre étudiants (ex. scénario programmé avec signes identiques), la fin de programme ou encore questions préétablies par le rédacteur que le facilitateur peut poser à l'étudiant doivent être définies dans le scénario et pourront être ajustées après test.⁴¹ En effet, le scénario utilisé doit être testé afin de s'assurer de sa validité et de sa fiabilité. (Cf chapitre scénario)
- D'après Miller, celui qui fait de l'observation formative est le mieux à même de certifier les connaissances et les compétences finales de l'apprenant. Il a eu en effet tout loisir de construire une représentation précise et analytique des ressources qu'il a accumulées aussi bien que de sa capacité de les mobiliser en situation, puisqu'il l'a vu maintes fois en train de réfléchir et d'agir, aux prises avec des tâches diversifiées, dans les conditions les plus habituelles du travail.⁶ Il paraît donc licite d'avoir dans l'équipe réalisant l'évaluation sommative, des enseignants ayant participé à l'évaluation formative des apprenants. Cependant, les étudiants peuvent également être enregistrés puis évalués à posteriori, éventuellement par des enseignants ne les connaissant pas.⁴¹ Les évaluateurs qui n'observent pas une action doivent déterminer en indépendance si c'est une erreur ou si c'est insignifiant.
- Les simulateurs doivent être suffisamment réalistes et présenter un niveau de difficulté adéquat pour être utilisés lors de l'évaluation (p.ex. simulateurs de tâches partielles adéquats validés.²⁰
- La qualité d'une méthode d'évaluation est décrite par sa fiabilité et sa validité. La faisabilité est liée à son utilité dans en pratique de soins Les autres facteurs importants dans l'évaluation des compétences procédurales sont la question de savoir s'ils testent de manière exhaustive chaque aspect de la compétence et si le retour d'informations est opportun pour promouvoir l'apprentissage ultérieur.²⁰
- Idéalement, l'évaluation des compétences (ce que l'étudiant ou le médecin est capable de faire) devrait fournir un aperçu des performances réelles (ce

qu'il fait habituellement lorsqu'il n'est pas observé), ainsi que la capacité de s'adapter au changement, de trouver et de générer de nouvelles connaissances et d'améliorer les performances globales.¹⁸

- Pour évaluer les compétences d'un étudiant, il faut le voir à l'œuvre dans des circonstances diverses et représentatives des situations professionnelles. Il faut pouvoir l'interroger pour comprendre la finalité d'une action, reconstituer ses représentations et raisonnements.
- Deux évaluateurs et huit scénarios ou plus sont nécessaires pour atteindre des niveaux élevés de fiabilité relative dans la notation des compétences clés du travail en équipe.⁸ (cf Chapitre Évaluateurs) Une augmentation de validité sera retrouvée en augmentant le nombre de scénarios avec deux juges ou en gardant le même nombre de scénarios avec un nombre de juges plus importants.
- Un consensus doit être atteint pour une notation appropriée et la définition d'une norme minimale de performance.
- Un seul candidat sera évalué par station. Cela réduit l'« effet groupe » (cf tableau 8 du présent document)

Tableau 1 : Eléments à prendre en compte pour mettre en place une évaluation sommative par simulation.

Compétences à évaluer	Définition claire	Savoir agir dans une situation professionnelle
		Identification des ressources internes à mobiliser : connaissances, habiletés mais aussi postures, attitudes, habitus, raisonnement
		Identification des ressources externes à mobiliser : équipement, publications, collègues, ...
	Nombre	Possibilité d'évaluation d'une seule compétence ou de plusieurs, simultanément
Mesure	Performance au cours de situations représentatives et diverses	
Evaluation	Authenticité du contexte	Problèmes complexes
		Complexité de la situation adaptée au niveau de formation
		Contextualisation en situation professionnelle
		Situations de collaboration (non exclusif)
	Standardisation	Tâches et exigences connues avant
		Observation directe associée à une phase d'interaction (questionnement) de l'étudiant
		Check-list
	Critères de correction	Multiples
		Objectifs
		Ajustés aux connaissances à évaluer, mises en situation
		Intégration de l'auto-évaluation
		Ne tenir compte que des erreurs importantes
		Évaluation des stratégies (cognitives et métacognitives) lors de la phase d'interaction
		Consensus préalable pour notation et définition d'un niveau attendu de développement
Scénarios	Élaboration	Après définition des compétences à évaluer
		Transposition de la réalité professionnelle
		Incorporation des compétences visées dans des tâches
	Compétences multiples	Multi stations à scénarios courts préférables à scénario long
		Situation critique
	Test préalable	Validité, fiabilité
Simulateurs (Haute-Technicité, Basse-Technicité)	Usage et niveau de difficulté validé	
Epreuve d'évaluation	Standardisation (Équité)	Rôle(s) et intervention(s) du facilitateur précisé au préalable
		Un seul candidat par station
	Conditions pratiques	Nombre minimal de scénarios (8 à 15) ¹³
		Incitation à verbaliser après l'action (raisonnement, ce qui est fait ou non fait)
	Évaluateurs	Deux évaluateurs au moins
		Idéalement, un évaluateur ayant participé à l'évaluation formative de l'apprenant

2. Les outils d'évaluation sommative

L'un des enjeux d'une évaluation de compétences réside dans la qualité des outils employés par les évaluateurs.

a. Quelles sont les qualités d'un outil adapté à la mesure souhaitée ?

i. Validité et fiabilité d'un outil : de quoi parle-t-on exactement ?

Un outil d'évaluation est considéré comme un bon outil d'évaluation si on a pu démontrer qu'il mesure réellement ce qu'il est sensé mesurer (validité) et ce d'une manière reproductible selon les variations de contexte (fiabilité).

L'objectif d'un outil d'évaluation n'est pas nécessairement d'être « parfait », il ne s'agit pas de la quête du « saint Graal de l'évaluation » qui constituerait plutôt un piège, mais quand un formateur utilise un outil d'évaluation il doit pouvoir connaître ses forces et ses faiblesses, sa valeur ajoutée au sein d'un programme de formation, toujours en le contextualisant ³⁷. En fonction du contexte, des objectifs pédagogiques attendus et des éléments qui ont permis sa validation, un outil adapté est un outil qui permet :

- de repérer les apprenants en situation d'échec,
- de déterminer un seuil de certification déterminé et donc qui a du sens par rapport aux objectifs pédagogiques
- de classer les candidats en fonction de leur performance et de leur niveau d'expertise, si l'outil a été prévu pour cela.⁴²

Ainsi un outil de qualité permet aux évaluateurs de prendre une décision adaptée à la performance observée. Le résultat du test permet donc d'identifier les candidats compétents comme l'étant réellement car le test est valide, mais il permet également de ne pas classer des candidats compétents comme étant incompetents et vice-versa. Comme pour un test diagnostique, un bon outil d'évaluation doit produire le moins possible de faux positifs et de faux négatifs.

ii. Aperçu bref de l'évolution des cadres théoriques de la validité

Plusieurs cadres théoriques se sont succédé depuis les années 60, avec initialement deux types de validité utilisés par les enseignants : la validité de contenu et la validité critériée (content and criterion validity), respectivement en rapport avec la création du contenu d'une évaluation avec la corrélation entre les scores obtenus à une nouvelle évaluation et ceux obtenus avec une évaluation « standard ». Dans

l'hypothèse d'une absence de mesure standard, les chercheurs en éducation ont ajouté une troisième notion au concept de validité : celle de validité de construit (construct validity). Cette notion ajoute la possibilité de lier entre eux des construits (notions abstraites, non palpables, non observables) avec des comportements observables, lors d'évaluations, dont l'analyse est fondée sur une relation théorique entre observation et théorie. Hormis le fait que le concept de validité était « morcelé », la principale difficulté reposait alors sur le fait que tout se rapportait à une notion abstraite (le construit) et que la reproductibilité des tests n'était pas toujours mesurée ou retrouvée.^{43,44}

Dans les années 90, ces observations ont conduit les chercheurs, avec en tête de file Messick puis Downing, à proposer un cadre alternatif, qui définit la validité comme un concept « unifié », au sein duquel, la notion de validité est celle d'une validité de construit dont la preuve s'appuie sur cinq différentes sources : le contenu, le processus de réponse, la structure interne, la relation du test avec d'autres variables et les conséquences de l'évaluation.⁴⁴⁻⁴⁶ Ce cadre est celui qui est le plus souvent utilisé dans les validations d'évaluation, même si bien souvent, il ne l'est pas sans sa globalité. Plus tard, pointant le manque de hiérarchie entre les différentes sources de preuve de validité et le manque de mise en évidence de « phases clés » pour la planification et l'évaluation des outils, Kane propose un nouveau cadre d'évaluation, la « validité argument ». Ce cadre peut s'appliquer à la fois à la création d'un outil seul, mais également à celle d'un programme d'évaluation, avec différents outils. Kane met en exergue l'importance d'identifier 4 étapes clés dans la création d'une évaluation : la création du score, sa généralisation, puis son extrapolation et son implication. Une des nouveautés de ce cadre est que son analyse se base à la fois sur des données qualitatives mais également sur des données quantitatives.⁴⁴

1. Cadre de Downing

Le cadre théorique de Downing s'appuie sur un modèle de validité dite « unifiée », dont les preuves sont collectées à partir des cinq sources énoncées plus haut. L'intérêt de cette approche consiste donc en la collecte de différentes données, pour construire une « preuve » de validité, qui permet aux concepteurs de tests ou d'évaluation de supporter ou de réfuter l'interprétation puis l'utilisation des scores créés.⁴⁶

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 le contenu de l'évaluation (différents scénarios 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(s), 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 ». ⁴⁶
- ▀ *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 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. ⁴⁶
- ▀ *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 leur capacités discriminatoires, leur difficulté, mais également sa généralisabilité. ⁴⁶
- ▀ *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). ⁴⁶
- ▀ *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. ⁴⁵

L'exposé théorique de ces concepts très facilement accessible dans un article de Downing, publié en 2003 ⁴⁶, 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 2014.⁴⁴

2. Selon Kane

La démarche de validation dépend énormément du contexte d'utilisation de l'instrument à valider.⁴⁷ Lors de la conception d'une épreuve en éducation, le concepteur ne cherche pas à valider l'épreuve en elle-même mais cherche à savoir si le score attribué représente bien le niveau d'habileté de chaque apprenant et si la décision prise sur la base de ce score est adaptée. Ce sont donc l'interprétation et l'usage des scores qui ont de l'importance.

L'intention annoncée du modèle de Kane est de rendre plus simple la démarche lorsque vient le temps de valider. Ce modèle met l'accent sur la façon de faire, plutôt que sur la définition de la « validité ».

Kane considère le processus de validation comme une accumulation de preuves structurées selon plusieurs niveaux d'inférences. Chaque inférence devant contribuer à valider les conclusions concernant un apprenant sur la base des résultats de l'évaluation.

Le modèle de validation de Kane est articulé comme une chaîne de quatre inférences qui permettent de faire le lien entre les performances observées (observations), la manière d'attribuer un score aux performances (score observé), le caractère généralisable du score (score univers), la signification du score (score cible) et la manière d'utiliser le score pour prendre une décision.⁴⁸ L'ensemble de cette chaîne peut être schématisé sur la figure ci-dessous :⁴⁹

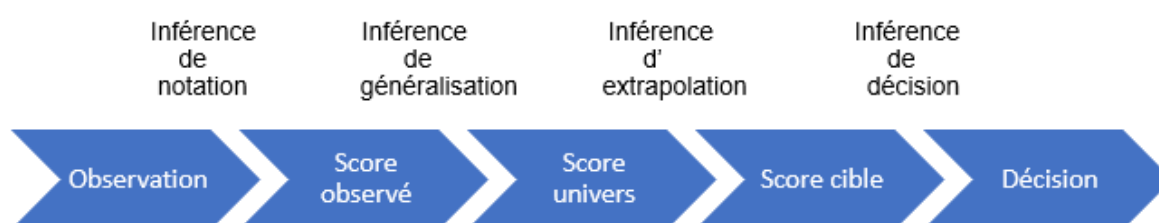


Figure 1 : Les 4 Inférences dans le modèle de Kane (adapté de Kane, Crooks et Cohen 1999).⁴⁸

Ces 4 inférences sont les suivantes :

Inférences de notation qui permettent de faire le lien entre les observations et le score observé.

L'objectif de ces inférences est de valider l'outil en s'intéressant à la méthode de construction et de notation des tests mais également de vérifier que les données issues de ces tests ont été collectées de façon appropriée.

Les méthodes les plus largement utilisées pour étayer l'inférence de l'observation à la notation sont les règles de construction des items, la structuration, les règles de notation et la "granularité", les analyses d'items, les évaluations de la pertinence, les procédures de sélection des informations, les méthodes de rapport et les questions de faisabilité.³⁷

Inférences de généralisation qui permettent de faire le lien entre le score observé et le score univers.

L'objectif de ces inférences est d'étudier et d'apporter des preuves de corrélation entre le score observé et le score univers. Elles sont basées sur l'argument selon lequel l'ensemble des scores observés est suffisamment représentatif de l'univers de tous les scores possibles. Une idée de la nature de cet «univers» est donc indispensable.

Dans un programme d'évaluation, différentes méthodes de généralisation peuvent être utilisées pour les différentes parties, parmi lesquelles la théorie des tests classiques, la théorie de la généralisabilité, les approches probabilistes, la saturation des informations, la crédibilité ou les schémas d'échantillonnage.⁵⁰

Inférences d'extrapolation qui permettent de faire le lien entre le score univers et le score cible.

L'objectif est de montrer que l'outil permet de mesurer ou d'anticiper les performances de l'étudiant dans une situation réelle. Elles doivent permettre d'étudier et de mettre en évidence des sources de variance non souhaitées. Elles sont à la fois basées sur des preuves de nature analytique en provenance de la définition du domaine à évaluer et de nature empirique. Elles doivent montrer la corrélation entre le score univers et le score cible. Les méthodes suivantes peuvent être utilisées : approches épidémiologiques ou fondées sur des critères, comparaison, corrélation et collation d'éléments, définition de paramètre standard.³⁷

L'objectif est d'évaluer les conséquences de l'évaluation tant pour l'apprenant que pour les autres parties et en particulier le patient. Elles doivent permettre d'assurer la crédibilité des interprétations des résultats et des décisions qui en découlent, mais aussi de s'interroger sur les conséquences de ces décisions.

Le travail de Tavares et al, est un exemple bien réalisé de mise en pratique du cadre de réflexion de Kane.⁵¹

Inférences de décision : leur rôle est, d'une part, de soutenir la crédibilité des interprétations des résultats et des décisions qui en découlent, mais aussi – et surtout – de s'interroger sur les conséquences de ces décisions.

iii. L'intérêt de multiplier les scores pour évaluer un même objet

« De manière similaire à l'opinion désormais généralement acceptée selon laquelle un bon test est plus qu'un ensemble aléatoire d'éléments de bonne qualité, un bon programme d'évaluation est plus qu'un ensemble aléatoire de bons instruments ».⁵⁰ Le problème de l'évaluation programmatique dépasse même cette analogie. Car, alors que des articles de bonne qualité sont réalisables, il n'existe pas d'instrument idéal.

Dès 1996, Van der Vleuten soutenait que toute évaluation unique impliquait un compromis sur les critères de qualité. Le choix du ou des critères à compromettre devrait être basé sur une décision mûrement réfléchie quant à l'élément de qualité à optimiser dans le contexte d'évaluation spécifique. Un programme d'évaluation, combinant différentes évaluations, peut atténuer les compromis sur les méthodes individuelles, rendant ainsi le total supérieur à la somme de ses parties.⁵² Selon Van der Vleuten et Schuwirth, "Dans n'importe quelle situation, un seul instrument peut ne pas être parfait". « Le processus d'évaluation programmatique résulte d'une combinaison minutieuse de divers instruments d'évaluation. En conséquence, aucun instrument unique n'est supérieur à un autre, mais chacun a ses forces, ses faiblesses et sa finalité dans un programme. Pourtant, en termes de méthodes psychométriques, une approche unique est souvent utilisée. »⁵⁰

iv. Typologie de scores : checklist, évaluation globale, métrique :

Les listes de contrôle et les échelles de notation globale sont fréquemment utilisées dans l'évaluation de la formation des professionnels de la santé, et les avantages et inconvénients relatifs de ces deux types d'outils sont débattus depuis longtemps. Les listes de contrôle incitent les évaluateurs à attester la performance ou l'omission d'actions directement observables, alors que les échelles de notation globale demandent généralement aux évaluateurs de juger la performance globale des

participants ou de fournir des impressions globales sur la performance des sous-éléments.⁵³

1. Liste de contrôle ou Check List :

Une check-list est un instrument qui comporte une liste d'éléments observables qui représentent les caractéristiques ou les comportements recherchés ou attendus. On se limite à observer la présence ou l'absence de ces caractéristiques ou de ces comportements. Ce type d'outil est fréquemment utilisé pour l'observation d'un processus ou d'un produit, notamment comme outil de vérification pour l'étudiant. Il s'agit d'une évaluation analytique qui permet de faire une analyse détaillée de tous les critères.

Les listes de contrôle sont relativement intuitives à utiliser et - en particulier pour les évaluateurs qui sont moins familiers avec la tâche clinique – et permettent de fournir les grandes lignes, étape par étape, des comportements et des conseils pour une rétroaction formative.

Chaque tâche nécessite, néanmoins, une liste de contrôle distincte et chaque liste de contrôle spécifique à une tâche nécessite une validation indépendante, en particulier dans le contexte de l'évaluation des compétences techniques. Cependant l'utilisation de listes de contrôle peut réduire les exigences de formation des évaluateurs.

La reproductibilité entre les évaluateurs est élevée et légèrement meilleure pour les listes de contrôle que pour les échelles de notation globale mais Il est également important de souligner que, malgré la perception que les listes de contrôle offrent une évaluation plus objective, la construction de ces outils nécessite souvent des jugements subjectifs et ce format ne confère pas nécessairement une plus grande validité ou fiabilité.⁵³ Les listes de contrôle seront généralement moins performantes dans la robustesse des preuves de validité que les échelles de notation globale.

2. Échelles de notation globale (GRS Global Rating Scale) :

Les échelles de notation globale permettent d'apprécier une performance d'un point de vue d'ensemble. On utilise ici aussi des critères, mais on les regroupe par niveaux auxquels on compare la production ou la performance. On structure de trois à six niveaux selon le but recherché.

- La notation globale procède différemment de la notation analytique, en ce sens que le but recherché est de porter un jugement d'ensemble plutôt que d'en faire une analyse détaillée de tous les critères.

- La notation ou l'évaluation globale est surtout utile lorsque l'on souhaite poser un jugement définitif (sommatif) auquel cas l'occasion de faire un retour sur les résultats de l'étudiant ne sera pas possible. L'évaluation analytique est quant à elle plus intéressante dans un contexte d'aide à l'apprentissage(formatif), car elle permet d'établir un diagnostic des forces et des faiblesses en vue de cibler les éléments à retravailler ou à consolider. Toutefois, l'évaluation analytique est également possible si l'on juge définitivement des apprentissages, mais il faut prévoir plus de temps.
- Les échelles de notation globale ont une fiabilité moyenne inter-articles et inter-stations plus élevée que les listes de contrôle. De plus, elles peuvent être utilisées sur plusieurs tâches, ce qui évite d'avoir à développer des instruments spécifiques à une tâche et à simplifier la validation spécifique à l'application. Elles peuvent, cependant, nécessiter une formation plus approfondie.

Les échelles de notation globale détectent les différents niveaux d'expertise de manière plus sensible que les listes de contrôle, car les réponses subjectives permettent de capturer des éléments nuancés d'expertise ou des écarts potentiellement dangereux par rapport à la pratique souhaitée, et refléter de multiples perspectives complémentaires.⁵³

Plusieurs études ont montré au cours de la dernière décennie, que les échelles de notation globale sont des instruments fiables et valides plus faciles à utiliser que les listes de contrôle.⁵⁴⁻⁵⁶ Dans certaines circonstances, les échelles de notation globale sont plus valables que les listes de contrôle, par exemple pour évaluer les performances des étudiants avancés ainsi que pour saisir les subtilités et les nuances.⁵⁷⁻⁵⁹ Pour évaluer les compétences en communication, les échelles de notation globale semblent plus appropriées.⁶⁰ Enfin, notons l'interaction inséparable entre la personne utilisant l'instrument et l'instrument lui-même : ni la liste de contrôle ni les échelles de notation globale ne supplantent le besoin d'expertise humaine et de jugement.

3. Variables et échelles d'évaluation quantitatives et qualitatives

En plus de l'évaluation subjective par l'apprenant (« j'ai bien fait, j'ai l'impression de progresser... »), qui n'a pas de valeur exploitable dans une grille d'évaluation sauf pour « tester » initialement la population d'apprenants, trois types d'évaluations sont disponibles et nous les détaillons ici :

- a) Évaluations métriques (ou mesurables par des données chiffrées) :

Ce sont des variables quantitatives pures et peuvent être, par exemple, des variables de temps, d'amplitude des mouvements, un nombre de préhension pour réaliser un geste etc.

Elles permettent de :

- Borner un exercice dans sa réalisation : limite temporelle basse et haute, notamment limite haute pour identifier des apprenants « trop » longs monopolisant le poste de travail et déstabilisant le fonctionnement de l'atelier ou ne maîtrisant pas les étapes et la gestuelle requise. (évaluation certificative) ou quand la procédure testée doit être limitée dans le temps pour des raisons de sécurité du patient (durée de la tentative d'intubation orotrachéale (IOT) lors d'une intubation avant de ventiler le patient au masque ...) (évaluation certificative)
- Identifier une absence de maîtrise des gestes étapes par étape (nombre de préhensions...)
- Evaluer la précision d'un geste (intérêt des simulateurs composites ou haute-fidélité mesurant l'amplitude des mouvements, les tractions induites...) et de permettre à l'apprenant de progresser sur la précision du geste (évaluation formative).⁶¹

Ces données métriques et leur limite haute / basse doivent être identifiées soit en s'appuyant sur les recommandations des sociétés savantes, soit sur un avis d'expert.⁶²

b) Évaluations binaires (oui/non ou fait/non fait)

Ces évaluations n'explorent pas la qualité du geste, mais uniquement les étapes de la procédure. Elles sont intéressantes dans une grille d'évaluation pour geste technique pour identifier les actions OBLIGATOIRES de SÉCURITÉ du GESTE (exemple sur une intubation orotrachéale : vérification de la lumière du laryngoscope...) qui, si non réalisées de principe font courir un risque au patient (morbidity).

Elles peuvent aussi être intéressantes dans des procédures très spécifiques qui reposent sur une succession d'étapes bien définies (exemple : réalisation d'une suture coelioscopique avec utilisation d'une check-list).

Elles permettent aussi une évaluation rapide, sous réserve que l'évaluation soit réalisée par des experts maîtrisant parfaitement étape par étape le geste (car ce type d'évaluation se limite à des critères fait/non fait).⁶³

c) Scores objectifs d'évaluation subjective (type OSATS, GOAL)

Ces scores qui incluent un certain nombre d'items évalués par l'évaluateur via une échelle de Lickert explorent la connaissance théorique de la procédure (étapes clefs), la réalisation pratique (fluidité des gestes, enchaînement des étapes), la qualité du produit final, de la procédure dans son ensemble....

Ces scores permettent aussi l'évaluation d'habiletés non techniques, telles que la communication (ANTS). Ils permettent de définir non pas la réussite ou le succès d'un geste, mais souvent le niveau de compétence d'un apprenant (de novice à expert). La possibilité qu'ils offrent de définir plusieurs niveaux de notation permet ainsi de discriminer les apprenants et de les identifier par niveau de compétence.^{64,65} L'utilisation de ces scores pour s'assurer de leur reproductibilité repose sur des évaluateurs formés à ces techniques de scores objectifs, experts dans la réalisation du geste et maîtrisant la langue utilisée dans le score. Ce type d'évaluation serait supérieure aux évaluations métriques ou binaires.⁶⁶

d) Évaluation composite

Dans la gestuelle technique, l'utilisation d'une fiche d'évaluation composite peut s'envisager, celle-ci associant :

- ▶ Une évaluation métrique qui permet :
 - Une limite dans le temps et la répétition des erreurs de l'exercice
 - l'identification rapide des apprenants trop « novices »
 - En cas de matériel disponible, des bornes hautes d'analyse du mouvement)
- ▶ Une évaluation binaire qui permet de :
 - s'assurer du déroulement de la procédure sans prise de risque pour le patient.
 - maîtriser les étapes indispensables à la procédure
 - identifier rapidement les apprenants ne maîtrisant pas ces étapes, pour les former en amont.
- ▶ Une évaluation par score objectif adapté qui permet :
 - Une évaluation globale du résultat sous réserve de la disponibilité d'un score validé et adapté au geste et à la langue de l'évaluateur.

v. Objectivité, subjectivité, jugement évaluatif

Une des difficultés rencontrées par l'évaluation en situation et donc à l'évaluation par la simulation, est l'absence d'équité apparente entre les épreuves : conditions différentes, cas cliniques de difficultés variables et enseignants qui changent. Cette

constatation semble introduire la notion de jugement, qui est, pour certains auteurs, indissociable de l'évaluation des compétences, en situation complexe.

Scallon, en soulignant que l'évaluation ne peut pas se réduire à une activité "mécanique" introduit la notion de jugement dans l'évaluation.⁶⁷ Il s'agit selon lui, d'une étape nécessaire de l'évaluation qui s'apparente à un jugement professionnel et pas à un jugement de valeurs. Il ne s'agit pas d'un jugement vague, mais d'une évaluation basée sur des faits, 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 culture.⁶⁸

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.⁶⁹ 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 jugement de la personne plutôt que de sa compétence, mais elle est impossible pour de nombreuses raisons. Les enseignants sont différents, les contextes aussi.⁶⁹ L'évaluation en situation est donc, en fait, intrinsèquement composée d'éléments subjectifs, qui ne doivent cependant pas être flous ou arbitraires.

C'est à ce moment qu'intervient la notion de 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 ». Le processus de jugement professionnel permet d'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 communiquent efficacement, accomplissent correctement d'autres gestes, dans un ensemble global de bonne prise en charge. Les évaluateurs peuvent 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 ».⁶⁹

Ainsi, il serait vain de vouloir éviter la subjectivité car elle est présente dans toutes les étapes d'une évaluation : dans le choix des objectifs, des critères d'évaluation, dans

celui des informations à recueillir ou pas, mais également dans le choix de la stratégie de recueil (nombre d'évaluateur, design des scénarios etc.) et enfin, dans la dernière étape, celle qui donne du sens et permet aux évaluateurs de confronter les informations recueillies aux critères exigés.⁶⁹ Par conséquent, plutôt que de vouloir l'éviter, au risque de créer des évaluations objectives mais non fidèles, 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.⁶⁷ Il s'agit ici d'encadrer et d'objectiver la subjectivité.²⁸

b. Comment créer un outil d'évaluation

i. Introduction : choix du cadre.

Pour illustrer le développement d'un outil d'évaluation, nous avons fait le choix de nous référer au cadre de Downing, qui identifie cinq sources de validité d'un outil d'évaluation. A ces différentes sources de validité, il faut ajouter une recherche de faisabilité afin que l'outil puisse être utilisé par tous, selon des modalités précises, qui auront elles aussi été validées.

A titre d'exemple et pour illustrer notre propos, nous avons choisi de citer le travail d'une équipe poitevine, et qui avait pour objectif de décrire et de valider le développement d'un score d'évaluation d'une pose de voie intra-osseuse.⁷⁰

ii. Contenu d'une évaluation : définition et création

La qualité du contenu d'une évaluation est un des critères de validité de cette dernière. Ainsi, dans le développement d'un outil, il faut pouvoir s'assurer 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...^{44, 71}

L'étude du contenu s'intéresse à la description rigoureuse des différentes étapes ayant permis la constitution du contenu de l'évaluation (différents scénarios 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* » (« schéma directeur », littéralement).⁴⁶

A titre d'exemple, nous détaillons ici la méthode Delphi qui est la plus répandue pour créer le contenu d'un nouvel outil est celle de la consultation d'experts, encore appelée méthode des « groupes de consensus » au sein de laquelle plusieurs outils existent. Ils ont tous pour objectif de mesurer un consensus d'experts. Les plus utilisés sont la méthode Delphi, la méthode Delphi modifiée, la Technique du Groupe Nominal et la méthode RAND.⁷²

Le point commun des groupes de consensus réside dans 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 ».⁷³

Afin d'être valide, une consultation d'experts par Delphi se doit de respecter cinq critères qui permettent d'assurer une recherche de qualité : le caractère anonyme des experts, la répétition des consultations des experts, l'utilisation d'un rétrocontrôle pour chaque réponse donnée, une analyse statistique des réponses du panel et enfin la structuration des réponses qui permet d'échanger dans un cadre d'analyse rigoureux.⁷⁴⁻⁷⁶

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 par courriel. Elle permet de regrouper un grand nombre d'experts d'horizons différents tout en restant flexible dans son utilisation, avec un coût moindre.⁷³

Enfin, il faut respecter différentes étapes pour mener une consultation d'experts dans le cadre du Delphi :

1. Établir une problématique de recherche, puis une question en rapport avec le contenu de l'évaluation
2. Effectuer une revue de la littérature : outils existants, recommandations etc...

3. Création du questionnaire
4. Consultation des experts lors de « tours » anonymes, gérés par mail
5. Feedback individuels et collectifs, entre chaque tour
6. Résultats

Le travail de Propst est un exemple de qualité d'utilisation de la méthode Delphi pour développer un outil d'évaluation de la trachéotomie en pédiatrie.⁷⁷

iii. Processus de réponse

L'évaluation du processus de réponse est la 2^e étape du processus de validation de l'outil. Cette étape va rechercher toutes les sources d'erreurs possibles dans la conception de l'outil. Le processus de réponse est défini par Downing comme une preuve de l'intégrité des données une fois que toutes les sources d'erreurs associées à l'exécution du test ont été contrôlées ou éliminées dans la mesure du possible.⁴⁶ En d'autres termes, le processus de réponse fait référence à des preuves confirmant que les sources d'erreur dans la conception ou l'utilisation de l'outil ont été minimisées.⁷⁸ Les sources d'erreurs peuvent être liées aux évaluateurs aux apprenants ou à la grille elle-même. Les évaluateurs peuvent en effet être la 1^{re} source d'erreurs potentielles. Pour Oermann, la majorité des erreurs de notation sont même dues aux évaluateurs et non à l'outil utilisé pour l'évaluation. Les formateurs ont-ils été formés à l'utilisation de l'outil et y a-t-il une cohérence dans la façon dont l'évaluation est réalisée ? Car pour minimiser le risque d'erreur, les évaluateurs doivent être entraînés à l'utilisation de l'outil dans un environnement standardisé. ⁷⁹ Pour Furman des sessions de formations régulières contribuent à une fiabilité accrue des évaluateurs. ⁸⁰ La standardisation de l'entraînement est aussi gage de validité. ⁷⁹

Pour les apprenants, il est important qu'ils soient placés dans la même situation de départ : même briefing, même informations, même environnement, même formation sur le processus d'évaluation. Par exemple : les étudiants connaissent-ils l'utilisation de la méthode d'évaluation ? Sinon, ont-ils été formés à l'utilisation de l'outil ? ⁷⁸ Selon Oermann, le fait que les évaluateurs connaissent les apprenants pourrait aussi être source d'erreur. ⁷⁹ En effet, pour Oermann, lors de l'évaluation des performances, il pourrait avoir un effet de halo dans lequel les évaluateurs laissent une impression globale des étudiants influencer leurs évaluations, d'où l'importance que les évaluateurs ne connaissent pas les étudiants évalués lorsqu'ils effectuent des évaluations sommatives surtout si les enjeux sont élevés. ⁷⁹

Enfin la grille elle-même peut être source d'erreur potentielle. En effet, la probabilité d'une réponse donnée à une question dépend de deux éléments, la compétence de l'individu et les caractéristiques de la question.⁸¹ Pour Durand, la théorie de réponse aux items peut permettre une approche de cette notion. Elle permet de mettre des mesures différentes sur une même métrique, de tenir compte que la distance entre les points d'une échelle ordinale n'est pas forcément la même pour tous les points, de tenir compte du fait que chaque item n'a pas forcément la même importance dans la mesure du construit en soi et dans la mesure des différents groupes, et d'utiliser les différences dans les diverses versions d'un questionnaire en différentes langues.

iv. Structure interne de l'outil

L'évaluation de la structure interne de l'outil est la 3^e phase du processus de validation. Elle comprend l'analyse des différents items et de leur capacités discriminatoires (cohérence), la fiabilité (analysée par la reproductibilité des différents items, des stations ou encore selon les observateurs), et la généralisabilité de l'outil.

1. L'évaluation de la cohérence de l'outil

L'évaluation de la cohérence de l'outil est le premier élément d'évaluation de la structure interne de l'outil. Elle indique à quel point les items d'un test mesurent la même dimension. Pour Downing, par exemple, les scores d'item ou groupe d'item destinés à mesurer la même variable ou le même domaine devraient être plus fortement corrélés que les scores d'items destinés à mesurer des variables ou des domaines différents.⁴⁶ La cohérence est souvent mesurée par le test Alpha de Cronbach appliqué sur les évaluations utilisant l'outil. Le résultat du test varie entre 0 (faible) et 1 (élevé). Plus le coefficient est élevé plus les items du test sont similaires dans leur contenu (c'est-à-dire homogènes). Une valeur entre 0,7 et 0,9 est considérée comme correcte. Mais un coefficient supérieur à 0,9 peut suggérer un test trop long avec une redondance dans les items suggérant de raccourcir le test.⁸²

2. L'évaluation de la fiabilité

L'évaluation de la fiabilité de l'outil est le 2^e élément d'évaluation de la structure interne de l'outil. Il correspond à la vérification de la reproductibilité de l'outil entre plusieurs mesures pour un même évaluateur (test-retest ou intra rater reliability) et entre

les mesures de plusieurs évaluateurs (inter rater reliability).

Plusieurs méthodes statistiques peuvent être utilisées comme le pourcentage de concordance, la corrélation de Pearson (paramétriques), le test de Kendal (non paramétriques), le coefficient Kappa de Cohen, ou encore le coefficient intraclass. Dans une étude destinée à mesurer la fiabilité de l'évaluation du nombre de caries et de pathologie gingivale par des dentistes, Hunt en 1986 a comparé le pourcentage de concordance, la corrélation de Pearson, et le coefficient Kappa. Pour Hunt bien que le pourcentage de concordance soit facile à utiliser, il est à décourager car il ne prend pas en compte l'accord dû uniquement au hasard.⁸³

La méthode de corrélation de Pearson est un test paramétrique (car dépend de la disposition des données) qui mesure la dépendance linéaire entre les deux variables. Mais pour Hunt, il faut être prudent dans l'interprétation de la corrélation de Pearson, car elle n'est pas affectée par la présence de biais systématiques. A noter que si les données ne proviennent pas d'une distribution normale, il convient d'utiliser le test de corrélation de Kendall et celui de Spearman basés sur un test de rang (méthodes non paramétriques). Le test de Kappa est le plus fréquemment utilisé. Il permet d'estimer l'accord entre des jugements catégoriels fournis par 2 ou plusieurs observateurs en l'absence de référence. Il permet de quantifier l'accord au-delà du hasard pour des jugements dichotomiques tels que la présence ou l'absence de maladie.⁸³

La formule du Kappa est : $\kappa = (Po - Pe) / (1 - Pe)$

Où Po est le pourcentage de concordance observé et Pe le pourcentage de concordance attendu. Ainsi le numérateur reflète le pourcentage de concordance non lié à la chance. Le dénominateur reflète l'accord maximal au-delà du hasard qui aurait été possible compte tenu des distributions marginales de la maladie. Ainsi, le kappa est interprété comme la proportion d'un éventuel accord réellement réalisé et non lié au hasard.

Il y a accord si le résultat est entre 0 et 1 mais 0 représente un accord aléatoire et 1,00 représente un accord parfait, 0,40 ou moins un accord médiocre au-delà du hasard, 0,40 à 0,75 un accord passable à bon et 0,75 et plus un accord excellent ⁸⁶. Hunt privilégie l'utilisation du coefficient Kappa mais suggère que la corrélation de Pearson et le kappa peuvent être utilisés ensemble pour découvrir une erreur d'examineur non aléatoire.

Il est également possible d'utiliser le test de student (t) pour comparer la différence entre les moyennes obtenues par chaque observateur et la variance entre chacune

des moyennes. Le test de student permet de savoir si la différence observée est statistiquement significative ou non (ce qui sera exprimé par la valeur p).

Enfin, il est possible d'utiliser un test de régression logistique linéaire qui permet d'illustrer la corrélation entre les observations de chaque évaluateur, illustré par le calcul du coefficient R^2 de la droite passant par l'origine. Cependant cette méthode est moins souvent utilisée pour évaluer la reproductibilité entre évaluateurs, son analyse s'appuyant le plus fréquemment sur des tests de concordance lorsqu'il s'agit d'évaluer un degré d'agrément entre les évaluateurs (corrélation n'étant pas toujours synonyme de concordance).

Enfin, en cas de variables qualitatives, on pourra aussi utiliser le coefficient de corrélation intra classe (ICC). Toutefois il s'utilise pour un score global (par exemple, dans le cas de Hunt sur l'évaluation du nombre de caries calculé par bouche et non par catégories de dents). Ce coefficient est calculé en utilisant l'analyse de la variance pour diviser la variance des scores (...) en plusieurs composantes : patients, examinateurs, examens répétés et interactions entre ces termes. La valeur attendue est > 0.7 et d'autant meilleure qu'elle se rapproche de 1. En général, la corrélation intra classe reflète la part de la variabilité des scores moyens qui est due aux vraies différences entre les patients et celle qui est due aux différences entre les jugements des examinateurs.

3. Possibilité de généralisabilité (generalizability)

La possibilité de généralisabilité (generalizability) est la 3^e étape de validation de la structure interne de l'outil. Elle se définit par la possibilité avec laquelle ces résultats pourraient être transférés à des situations ou des populations différentes de celles initialement étudiées. En d'autres termes, la possibilité que les résultats obtenus ne dépendent pas du sous-ensemble utilisé mais puissent être généralisables à l'ensemble de tous les éléments auxquels on aurait pu avoir recours pour élaborer l'outil.

Ces études permettent de déterminer des sources d'erreurs potentielles. En effet, pour Boulet, lors de la recherche de sources d'erreur d'une évaluation, des études de généralisabilité (G) peuvent être menées pour délimiter spécifiquement l'importance relative de diverses sources d'erreur et leurs interactions associées. Ainsi, après l'étude de généralisabilité, des études de décision peuvent être alors être entreprises pour déterminer la conception optimale de la notation (par exemple, le nombre de rencontres simulées, le nombre d'évaluateurs par rencontre donnée).²⁹

Downing estime aussi cet outil peut être utile pour estimer les diverses sources d'erreur contributive à l'examen d'un patient standardisé, telles que l'erreur due aux évaluateurs, l'erreur due aux cas et l'erreur associée au candidat.⁴⁶

La possibilité de généralisabilité a été proposée par Cronbach en 1963. Il décrit un indice (R^2) qui varie entre 0 et 1 et indique dans quelle mesure un dispositif instrumental est apte à différencier de manière fiable les individus ou les objets auxquels il est appliqué.

Une approche plus facile de la généralisabilité possible de l'outil peut être l'absence de différence significative entre les moyennes, écart type et erreur standard de mesure (écart type divisé par la racine carrée de l'effectif) lors de la comparaison des scores des différents observateurs (t test, f test et analyse de variance).

4. L'étude de la relation aux autres variables

La relation aux autres variables fait référence au concept selon lequel la note globale devrait être corrélée à d'autres méthodes validées.⁷⁹ "Les scores de votre évaluation sont-ils en corrélation avec ceux d'autres instruments qui mesurent les mêmes traits ou le même construit ?".⁷⁸

En 2016, Oermann ⁸⁴ propose aussi cette démarche de validation de l'outil utilisé pour des évaluations sommatives : « Dans le cadre de la garantie de la validité d'une simulation, des études peuvent être menées pour corréler les connaissances et les compétences évaluées dans la simulation avec d'autres évaluations, telles que les évaluations cliniques et les résultats des tests. Une autre stratégie consiste à piloter l'évaluation basée sur la simulation avec différents niveaux d'étudiants. On pourrait s'attendre à ce que les individus ayant plus d'expertise réussissent mieux dans la simulation que ceux ayant moins d'expertise ». En effet une mise en évidence de la validité peut être apportée par des scores plus élevés correspondant à des niveaux de formation avancés.⁸⁴ Enfin, pour Downing, une corrélation négative avec un outil mesurant une capacité différente peut aussi être intéressante.⁴⁶

La validité de l'outil peut donc se faire par corrélation positive avec un gold standard et/ou d'autres outils publiés mesurant tout ou une partie des compétences évaluées, ou par corrélation négative avec des outils évaluant des compétences différentes. Une autre approche est de démontrer que le résultat du score est corrélé au niveau des apprenants.

Ainsi, les différentes variables à disposition pour une comparaison avec l'outil

développé sont : une échelle d'évaluation utilisée comme le gold-standard, l'observation clinique d'échec ou de réussite d'une procédure, le statut des apprenants (novices versus expérimentés) et encore l'évolution du niveau de l'apprenant et donc de son score avant puis après une formation (dont on sait qu'elle permet la progression des apprenants). Grâce à ces comparaisons, il devient possible de déterminer la validité de l'instrument.

v. Conséquences du test

La validation de l'outil par l'évaluation de ses conséquences est la 5^e et dernière étape du processus de validation. La validité des conséquences implique l'impact des scores sur les stagiaires, les enseignants, les patients et la société.⁷⁹

Car les conséquences des évaluations sur les candidats, les évaluateurs, les patients et la société peuvent être considérables et ces conséquences peuvent être positives ou négatives, voulues ou non.⁴⁶

Trois éléments sont à considérer pour mener à bien cette évaluation : l'utilisation prévue de l'outil, l'impact de l'outil sur l'enseignement délivré et les méthodes de travail des évalués, et enfin la conséquence éventuelle de cet outil sur le raisonnement clinique :

Le premier élément à considérer est donc l'utilisation prévue de l'outil sanctionnant (sommatif) ou non (formatif) et en cas d'utilisation sommatif, les conséquences d'un échec pour l'évalué : par exemple peut-on interdire à un stagiaire d'effectuer une procédure sur un patient sans réussir le test sur un simulateur ? De plus, si l'outil est utilisé pour l'évaluation sommatif, une méthode validée doit être utilisée pour déterminer le score de réussite le plus bas.⁷⁹

Le second élément est l'impact de cet outil sur l'enseignement délivré et les méthodes de travail des évalués. En effet, les méthodes et stratégies choisies pour évaluer les étudiants peuvent avoir un impact important sur le contenu de l'enseignement, les méthodes de travail des étudiants, l'utilisation et la rétention ou non de l'apprentissage et enfin comment les étudiants voient et apprécient le processus de formation.⁴⁶ L'échelle, elle-même, peut devenir une aide cognitive.

Le troisième élément est la conséquence éventuelle de cet outil sur le raisonnement clinique. En d'autres termes, pour Downing, cela signifie que le score obtenu par l'outil puisse être en mesure de déterminer des conséquences cliniques significatives (par ex. que les scores de réussite ou d'échec de l'outil puissent différencier différents

résultats cliniques comme le succès et l'échec d'une procédure). Les preuves de cette catégorie de validité seront probablement plus développées avec l'utilisation de l'outil à la fois dans les environnements simulés et cliniques.

vi. Exemples

Nous illustrons ici le cadre théorique avec un exemple de développement de deux scores d'évaluation, l'un s'intéressant à l'évaluation d'une procédure, l'autre à l'évaluation du travail en équipe.^{70, 85}

Oriot et al se sont intéressés au développement d'un score qui a pour objectif d'évaluer la pose d'un dispositif de perfusion avec accès intra-osseux. Un des avantages de ce type de procédure est qu'on peut y définir le succès ou l'échec de l'apprenant, ce qui n'est pas le cas lorsqu'on évalue une échelle avec des items s'intéressant aux habiletés non techniques (ou CNT, voire ce qu'il a été décidé d'utiliser comme terme dans les recommandations).

Ils ont suivi le cadre de Downing pour évaluer leur outil, en cherchant à identifier des preuves de validité issues des cinq sources décrites par l'auteur.⁸⁶

Tableau 2 : Exemple de développement d'un outil d'évaluation de la réalisation d'un geste technique en situation simulée, Oriot et al, d'après le cadre de Downing.^{14, 24}

Source de Validité	Méthode	Critères de jugement	Résultats
Contenu	1. Description développement de la checklist par 2 experts 2. Révision par 2 experts extérieurs 3. Checklist définitive	Pertinence des items Illustration adaptée du geste Conditions de réalisation geste	Obtention d'une liste de 12 items (après proposition initiale de 20 items)
Processus de réponse	Etude pilote, recherche des sources d'erreur Adaptation des items Définition des unités de mesure	Reproductibilité inter-observateur Contenu des items (redondant, imprécis) Contrôle des sources d'erreurs de mesures Pondération des items	Fusion/suppression d'items redondants Minutes, degrés, centimètres Justification
Structure interne	Cohérence interne Reproductibilité Discrimination des apprenants	Coefficient alpha de Cronbach, Inter observateur : Kappa de Cohen, ICC	Résultat du Cronbach Corrélation entre 2 observateurs
Comparaison aux autres variables	Succès ou échec versus score obtenu Score obtenu versus évaluation théorique Score obtenu versus expérience antérieure/niveau d'expertise	Corrélation entre succès de la procédure ou score théorique et score obtenu avec l'échelle	Temps pour succès et score / succès et note
Conséquences	Seuil de réussite du test	Seuil d'échec réussite de procédure	14/20

Dans une étude pilote de validation, Guise et al s'intéressent au développement d'un outil d'évaluation du travail d'équipe pendant la prise en charge d'une dystocie des épaules. Ici, il n'est pas possible de dire que l'équipe a « réussi » ou « échoué » dans la communication, mais ils utilisent des situations filmées qui représentent un travail d'équipe qualifié de pauvre, intermédiaire ou excellent et ils confrontent les scores obtenus à la qualité supposée du travail d'équipe. Il s'agit d'une des façons de faire lorsque l'on s'intéresse à un construit qui n'est pas mesurable de manière binaire. Ils s'inscrivent également dans le cadre de Downing.

Tableau 3 : Exemple de développement d'un outil d'évaluation de l'observation du travail d'équipe en situation simulée ²⁵

Source de validité	Méthode/définition	Critères de jugement	Résultats
Contenu	1. Description développement de la « <i>Teamwork Scale Development</i> » (CRM)	Revue de la littérature Grille déjà utilisée dans un autre domaine (aéronautique)	15 items 5 catégories 1 score de compétence globale
Processus de réponse	1. pertinence des items 2. pondération des items 3. entraînement des évaluateurs(modéré)	1. description précise de chaque item 2. critères quantitatifs critères qualitatifs 3. principes de CRM	1. tableau d'aide à la notation 2. 0 à 10 ou 0/1 Niveaux descriptifs : non adéquat/inacceptable/ pauvre/intermédiaire/ bon/ parfait
Structure interne	1. Validité construite 2. Utilisabilité de l'échelle 3. Reproductibilité	1. Distribution des scores par rapport au niveau défini à priori 2. Nombre d'items remplis intégralement 3. inter-observateurs, concordance, corrélation entre score global et entre catégories (Kappa, Kendall, Pearsons, ICC) 4. Variance de chaque composante	1. Score adapté à chaque niveau 2. Échelle facile à utiliser (peu de perte d'information) 3. corrélation entre les observateurs 4. variation des scores entre les scénarios/sources d'erreur

c. Outil d'évaluation et variation des contextes

i. L'outil d'évaluation est-il le même entre novice et expert ?

Le principe d'un outil d'évaluation repose sur sa validité mais aussi sur sa reproductibilité. Or, le niveau des apprenants (novice / intermédiaire/ expert) étant variable, on peut se poser la question d'outils d'évaluation différents selon le niveau de l'apprenant.

Compte tenu de la lourdeur de mise en œuvre d'un outil d'évaluation qui répond à ces critères, trois approches théoriques sont possibles :

- Soit la constitution de plusieurs outils pour un même objectif d'apprentissage (procédure...) adaptés au niveau de chacun, avec des "exigences" "graduelles" ou différentes selon le niveau.

Pour : une évaluation plus précise des compétences et de la performance, en s'appuyant sur le principe que les prérequis et les étapes d'apprentissage précédentes sont acquises

Contre : la nécessité de valider chacun des outils d'une part et l'absence de dépistage via l'outil de l'origine de l'échec (notamment en l'absence de maîtrise des prérequis)

- Soit la constitution d'un seul outil dont les items peuvent être "pondérés" ou sélectionnés en fonction du niveau de l'apprenant.

Pour : un seul outil à valider plus facile à maîtriser pour les évaluateurs, une sélection d'items "pertinents" en fonction du niveau théorique de l'apprenant, une évaluation globale permettant de tester l'ensemble des connaissances.

Contre : un biais méthodologique si les items sélectionnés ne sont pas performants, une adaptation des évaluateurs dans la maîtrise de leur outil pour chaque niveau d'apprenant.

- Soit la constitution d'une grille unique sans pondération des différents items valable pour tous les apprenants.

Pour : un seul outil d'évaluation à valider, maîtrisé par tous les évaluateurs car non modifiable, simple d'utilisation car n'incluant pas de pondération ; une "corrélation" théorique entre le score obtenu et le niveau d'expertise de l'apprenant, mais à valider par des analyses psychométriques.

Contre : nécessité d'un outil d'évaluation potentiellement complexe, multimodal (métrique, check list, scores) car devant être capable de discriminer autant les apprenants novices en identifiant les lacunes que les apprenants experts, en identifiant leur niveau de performance.

ii. L'outil d'évaluation est-il le même si la profession change ?

Selon la profession, deux approches sont distinguées :

Soit on considère qu'un outil d'évaluation correspond à une procédure ou une

situation clinique, pour laquelle la profession des apprenant doit être compétente (exemple IOT pour anesthésistes, urgentistes, IADE). Dans cette situation précise, l'outil d'évaluation doit être identique car l'objectif est d'évaluer la pertinence et la réalisation de gestes, quel que soit l'opérateur.

Soit on évalue une approche pluriprofessionnelle, dans laquelle chaque "profession" doit avoir un rôle particulier et doit exécuter ce "rôle" dans la procédure évaluée.⁸⁵

Dans ce cas, il est sans doute nécessaire de s'appuyer sur un outil d'évaluation soit différent selon la profession concerné (centré sur les étapes propres à réaliser par l'apprenant), soit une grille d'évaluation unique dans laquelle certains items seront pondérés ou non relevés.^{87,88} Ces éléments n'incluent pas l'évaluation comportementale nécessaire lorsqu'il existe, en travail pluriprofessionnel, un glissement de tâche non établi entre les différentes professions présentes et qui font l'objet de briefings spécifiques et/ou d'évaluations spécifiques.⁸⁹ Dans ce contexte, de très nombreux outils ont été développés pour l'évaluation du travail d'équipe, résumés dans cette analyse récente de la littérature.⁹⁰

iii. Autres éléments à prendre en compte : la langue de validation

Chaque score, outil d'évaluation qui est validé dans une langue, devrait l'être à nouveau s'il est utilisé dans une autre langue, afin d'éviter des biais sémantiques qui pourraient nuire au sens du contenu et à sa représentation. De plus, certaines recommandations ou pratiques pourraient différer d'un pays à l'autre et être réfléchis lors de l'emploi d'un outil d'évaluation validé par des équipes étrangères.⁹¹

3) Conséquences de l'ES par la simulation sur les personnes formées et les formateurs

La littérature retrouve deux types de conséquences de la mise en place d'une évaluation sommative en simulation : des conséquences pédagogiques et des conséquences psychologiques

a. Conséquences pédagogiques : impact sur les stratégies d'apprentissage ?

i. Connaissance de l'outil et de l'environnement indispensable :

Pour que l'épreuve sommative soit fiable, il est indispensable qu'elle mesure les

compétences et non la capacité de l'étudiant à s'adapter à un environnement inconnu.

Pour De Ketele « la décision sera fiable si l'évaluation ne dépend pas de l'évaluateur ou des circonstances de l'évaluation. »⁹²

Cela induit que l'apprenant soit familiarisé avec l'environnement.

Pour Verot, « pour que la simulation clinique haute-fidélité devienne une réelle plus-value pour la formation initiale des étudiants en Sciences Maïeutique, chaque étudiant doit avoir l'opportunité de participer en tant qu'apprenant actif. Cette condition est également indispensable pour qu'elle soit envisagée comme outil d'évaluation »⁹³ Cette connaissance de l'outil et de l'environnement doit contribuer à assurer un cadre sécurisant à l'évaluation afin qu'elle soit apprenante. Par conséquent, on peut imaginer que l'apprenant ait tout intérêt à se porter volontaire lors des séances de simulation « formative » afin d'augmenter sa connaissance de l'outil lors de la séance sommative ».

La stratégie d'apprentissage serait alors impactée par l'intérêt de l'étudiant à participer activement (apprenant acteur de soin et non observateur) afin de se familiariser à l'outil et à l'environnement.

Si l'évaluation sommative vise des scénarios déjà travaillés en formatif, l'étudiant a intérêt à s'entraîner en formatif pour augmenter ses chances de réussite en normatif. Le risque est alors une dérive de l'objectif principal de l'outil simulation (l'acquisition de compétences) qui peut devenir, pour l'apprenant, la validation d'un examen normatif, ce que l'on ne retrouve pas ou moins en formatif. La conséquence serait que « l'étudiant, mis dans une situation factice, agit dans l'objectif de valider son unité d'enseignement clinique au détriment de sa compétence professionnelle [...] il agit en fonction des attentes qu'il a perçues de l'évaluateur ».⁹²

ii. Hiérarchisation des informations

Selon Anne-Laure Philippon, à propos de l'évaluation par la simulation des compétences des étudiants de 4ème année de médecine dans la prise en charge des urgences vitales « le fait d'avoir une évaluation de leur pratique leur permettait de réviser différemment et les obligeait à hiérarchiser les informations ».⁹⁴

iii. Impact motivationnel et autonomisation de l'apprenant :

L'évaluation sommative induirait un effet d'entraînement sur l'enseignement et l'apprentissage. Dans une étude menée sur des étudiants en maïeutique⁹³ les

modalités d'évaluation, notamment sommative (Examen Clinique Objectif Structuré), a induit un changement dans les méthodes de révision, les étudiants sollicitant régulièrement les formateurs pour se mettre en situation (simulation). La réussite de l'épreuve sommative entraîne une nécessité d'entraînement, avec la simulation comme outil d'évaluation sommative, on renforce l'intérêt des séances formatives préalables, avec l'idée que « l'on enseigne ce qu'on évalue et que l'on évalue ce qu'on enseigne ».⁹⁵ L'évaluation n'est pas seulement une certification, c'est aussi, en soi, un outil d'apprentissage formateur car elle augmente la motivation de l'étudiant à retenir et restituer ce qu'il a appris en lui donnant un feed-back individualisé ». ⁹⁶ Le rapport québécois de recherche expérimentation, de février 2015, abonde également en ce sens « *Quelques participants se sont rendu compte par eux-mêmes qu'ils ne disposaient pas suffisamment des connaissances pour comprendre, réagir ou intervenir adéquatement à la situation simulée. Du coup, ils ont par eux-mêmes ressenti le besoin d'aller chercher ces éléments manquants. Cela les a responsabilisés quant aux efforts intellectuels à fournir face au développement de leur pratique professionnelle.* ».⁹³

iv. Impact sur le volontariat :

En évaluation formative, les formateurs ont parfois la nécessité de faire appel à des volontaires, pour les mises en situation. En évaluation sommative, la mise en situation est obligatoire. Devant une nécessité d'augmenter les préparations à l'évaluation sommative, la mise en place des évaluations sommatives pourra donc avoir un impact positif sur le volontariat en évaluation formative.

v. Indissociabilité du formatif et du normatif

C'est la notion d'alignement pédagogique : l'articulation entre les activités d'évaluation et d'apprentissage ainsi que le choix d'outil d'évaluation sont primordiales dans toute stratégie d'enseignement.⁹⁷

Le formatif et le sommatif doivent être liés sur la forme, le fond, les intentions pédagogiques et doivent permettre l'équité entre apprenants.

Cette indissociabilité entre formatif et normatif interroge forcément sur l'utilisation que l'étudiant et le formateur vont avoir des séances de simulations formatives, ne serait-ce que du point de vue de l'équité entre étudiants. Selon cette logique si le formatif sert de base d'entraînement à la validation du normatif, alors chaque apprenant devrait pouvoir bénéficier de séances de simulation formatives

vi. L'ES assure-t-elle un bon transfert de connaissances et de compétences pour la pratique du soin ?

Cette question fait finalement référence aux conditions d'exercice de l'évaluation sommative (notion de couperet, charge émotionnelle très élevée, caractère dichotomique : bon ou mauvais, réussi ou raté) et à l'impact de la sanction potentielle sur les performances ultérieures du candidat.

Cela doit être exploré à deux niveaux :

- Si l'étudiant réussit cette évaluation sommative, a-t-il appris davantage encore pour sa pratique par ce biais-là et à cet instant là ou non ?
- Si l'étudiant échoue à cet exercice, est-ce qu'il existe un risque de dégradation de la qualité des compétences et connaissance acquises du fait de la sanction (possiblement en lien avec l'auto-dépréciation rattachée à l'échec : je suis nul, je ne vaudrai rien, qu'est-ce que ce sera avec des vrais patients, de toute façon je l'ai tué etc.) ?

S'il est admis que les patients simulés ou mannequins n'ont pas vocation à mourir dans les exercices de simulation formative ou sommative et que ce type d'évènement aurait plutôt tendance (mais cela n'est pas univoque et tient presque du principe de précaution) à diminuer la qualité des apprentissages,^{98,99} les étudiants dans leur capacité à évaluer la situation et la qualité de leurs interventions, n'ont besoin de personne pour considérer que les actions qu'ils ont menées ont conduit leur patient simulé ou leur mannequin au décès. Le danger est bien là qu'il s'agisse de l'action ou la non-action directe de l'étudiant qui provoque possiblement le décès. Par voie de conséquence, il est essentiel de définir à l'avance le cadre des scénarios de façon très bordée. Il n'apparaît pas alors impossible que le patient simulé ou le mannequin puisse mourir si l'objectif de l'évaluation comprend cet aspect de l'évolution.

Il est classique de parler du testing effect : le fait d'être testé renforce l'apprentissage. Si ce principe s'applique à l'évaluation sommative, il assurera un transfert de connaissances ? Pour cela il est important de considérer l'évaluation toute sommative soit-elle comme faisant partie intégrante du processus de formation, lui-même étant un processus progressif et évolutif. Ces processus doivent s'intégrer dans ce qui est décrit par Bédard comme étant **l'alignement pédagogique** qui notifie en substance qu'il doit exister une cohérence entre ce qui est enseigné et ce qui est évalué et que cette cohérence doit se prolonger sur les outils d'évaluation qui doivent être adaptés aux étudiants.¹⁰⁰

Il est possible d'envisager un effet contre-productif des évaluations sommatives dans le domaine de l'apprentissage. Cependant, il est difficile d'envisager l'absence totale d'évaluation sommative tout au long des études des étudiants en santé. Il s'agit donc de trouver des outils de pondération qui permettent de ne pas « enfoncer » ceux qui ne seraient pas « les étudiants qui se réalisent sur un instant précis » tout en maintenant un bon niveau de compétences et de connaissances globales. Parmi ceux-là, il est intéressant de développer des dispositifs d'évaluation avec des méthodes et des instruments dont les étudiants ont connaissance. La multiplicité et la variété des évaluations peuvent aussi être des outils pour éviter les effets négatifs de l'évaluation sommatives (ex : plusieurs évaluations sommatives avec un nombre à valider dans des domaines différents).

De la même façon, la transparence sur les critères pris en compte dans la méthode d'évaluation sommative (nature des tâches d'évaluation, modalité de correction et de détermination de la réussite ; connaissance de la grille par les étudiants etc.) est essentielle pour assurer une baisse de la charge émotionnelle propre à une meilleure réalisation de l'exercice sommatif.

b. Conséquences psychologiques :

En 2013, David Gaba publie un article sur les conséquences psychologiques de la simulation sur les personnes formées et met en avant que :

- ▶ Il est vital de considérer les conséquences psychologiques de l'enseignement par simulation sur le patient.¹⁰¹ C'est le cas de la mort ⁹⁹ ou encore de la remise en cause de la hiérarchie pendant la simulation.¹⁰²
 - La littérature sur les conséquences psychologiques de la simulation n'est pas assez fournie
 - Le formateur en simulation a un devoir sur les personnes formées mais aussi les patients et leur entourage

L'article de David Gaba est essentiellement axé sur l'évaluation formative. Qu'en est-il, des conséquences psychologiques du formé lors d'une évaluation sommative ?

- i. Conséquences psychologiques, pour les apprenants, de l'ES en simulation.

1. Connaissance de l'outil « Simulation en santé » et sécurité psychologique des participants.

La littérature est peu fournie pour répondre à la question « Faut-il avoir participé à des simulations formatives avant de participer à des simulations sommatives ? »

Gantt en 2013 ne retrouve pas, dans une étude quasi expérimentale impliquant un effectif de 39 étudiant.es infirmier.es, de bénéfice à réaliser des simulations d'entraînement avant le passage en simulation sommative.¹⁰³ Néanmoins, cette étude de faible effectif ne cible pas spécifiquement des situations critiques. A l'inverse, Wilhaus et al. dans un article de synthèse présentant les règles de bonnes pratiques pour écrire des scénarios de simulation sommative à haute-importance (*high-stakes*) insiste sur le fait que la préparation des étudiants à la connaissance des outils par lesquels ils seront évalués via une information transparente est primordiale pour assurer un climat d'évaluation « juste ».⁴¹ Rudolph et al. en 2008 exposent les situations nuisant à l'apprentissage des participants en simulation, dont le fait de ne pas « adhérer » à l'outil, trouver des failles de fidélité lors de la séance, se sentir mis en difficulté vis-à-vis de leur identité professionnelle et adopter un comportement défensif lorsqu'ils doivent discuter des performances inférieures aux standards de soin.¹⁰⁴

La réduction des menaces vis-à-vis de l'identité professionnelle et sociale des participants semble un impératif pour permettre l'apprentissage. En ce sens, l'exposition à une simulation sommative sans exposition préalable à l'outil simulation semble être porteuse de risques potentiellement dévastateurs sur les participants. Rudolph propose 4 axes pour maintenir la sécurité psychologique dans une simulation dont un est à discuter ici : « Clarifier les objectifs, l'environnement, les rôles et la confidentialité ». ¹⁰⁴ Dans le cadre de la simulation sommative, il semble essentiel d'apporter une clarté parfaite sur le quoi et comment de l'évaluation afin d'établir et de maintenir la confiance entre participants et formateurs. Une attention particulière doit être portée à la confidentialité des données recueillies, à savoir informer précisément sur qui sera informé des résultats.

Kang en 2019 dans une étude qualitative sur 15 étudiants infirmiers participant à un cursus de formation normative par simulation retrouve 4 thèmes principaux vecteurs d'une insécurité psychologique à savoir : le sentiment de manque de préparation, l'anxiété de faire des erreurs, l'inquiétude de nuire à l'action du groupe et la peur d'être évalué.¹⁰⁵

En termes épidémiologiques, Henricksen en 2017 étudie la prévalence et les déterminants des événements de détresse psychologique en simulation⁵. Dans leur cohorte de 3 900 participants, la prévalence de ces événements est <1% (27/3900)

dont un seul sévère. Néanmoins, la finalité de ces simulations n'est pas précisée et il ne peut être exclu que cette prévalence diffère en cas de simulation sommative. Cette même équipe propose un algorithme de détection et classification de sécurité psychologique du participant en simulation ainsi qu'un arsenal de techniques permettant de gérer ces états de détresse psychologique.¹⁰⁶ Ce dernier point nous semble important en ce qu'il peut renforcer l'idée que les formateurs destinés à ces évaluations en simulation doivent être formés spécifiquement, notamment afin de répondre à ce genre d'évènements. L'impact de la simulation sommative sur la sécurité psychologique des apprenants est un élément très peu exploré dans la littérature, notamment par les articles de référence et les recommandations en éducation médicale.^{95,97,107,108} Ces derniers points suggèrent que l'information transparente sur les objectifs et les méthodes d'évaluation est cruciale afin de garantir la réussite d'un tel dispositif. Il semble aussi que le parcours de certification soit important. En effet, en suivant les principes qui ont guidé l'émergence de la simulation en santé, les erreurs ne doivent pas être stigmatisées, y compris dans ce contexte. Il est ainsi pertinent de proposer un parcours de certification, à savoir un ensemble de compétences à évaluer réparties en compétences obligatoires afin de pouvoir être certifié, assorti d'un nombre de compétences optionnelles ou non-rédhibitoires. Anticiper les parcours de remédiation, en cas d'échec, centrés autour du praticien et de ses besoins est essentiel afin de ne pas rester dans une posture de jugement mais d'entrer dans une approche de compagnonnage vers l'approche d'*apprentissage tout au long de la vie* (« *lifelong learning* »), nécessaire au maintien des compétences d'un professionnel de santé au cours de sa carrière.

2. Physiologie

De multiples articles font état des conséquences d'une privation de sommeil sur la performance, en médecine comme dans d'autres domaines. En médecine, il a été montré par Howard, Neuschwander et Eastridge que la privation de sommeil impacte les compétences techniques comme non-techniques.¹⁰⁹⁻¹¹¹ même si ces différents types de compétences ne semblent pas être impactés de façon équivalente par la privation de sommeil, selon Hayter.¹¹² Pour cette raison, il est loyal de n'autoriser l'accès à ces simulations à enjeu élevé qu'à des participants dont le capital de sommeil dans les 24 à 48 heures précédentes a été le plus proche possible d'un rythme circadien normal.

En 2019, Ghazali et al. étudient les marqueurs de stress physiologique chez des participants en simulation.¹¹³ Ils choisissent d'exclure les participants ayant vécu des événements stressants dans les 36h précédentes, ainsi que ceux ayant des antécédents de PTSD ou trouble psychologique pouvant être décompensés par la simulation. Bien que ces critères tombent sous le sens d'une méthodologie visant à l'évaluation du stress généré par la simulation, il semble logique de ne pas ouvrir ces séances de simulation sommatives à des praticiens dont le stress physiologique/émotionnel est anormalement élevé et pourrait les amener à dépasser leur marge de contrôle émotionnel lors d'une simulation à enjeu élevé.

3. Respect du caractère contexte-dépendant des compétences à évaluer

La littérature nous offre quelques réflexions pertinentes sur la définition de la compétence dans le contexte professionnel et de ses implications quant à la question posée. Selon Brailovski, la compétence n'existe que dans un contexte, dans la lignée de la définition de Kane (1992) selon laquelle « La compétence professionnelle représente pour un individu le degré d'utilisation de ses connaissances, de ses habiletés et de son jugement dans des situations auxquelles il est confronté dans sa pratique professionnelle. ».⁴ Dans ce cadre, il est donc crucial de fixer les objectifs d'évaluation d'un praticien à ses compétences attendues. Ainsi, est-il pertinent d'évaluer un praticien d'anesthésie orthopédique adulte sur un cas de pédiatrie si son activité ne l'y a pas exposé depuis plusieurs années et ne semble pas pouvoir l'y exposer pendant plusieurs années supplémentaires.

Ce dernier point introduit deux éléments importants.

1/ considérer une individualisation des évaluations sommatives à des parcours professionnels plus qu'à une profession « globale ».

2/ l'évaluation et le feedback en particulier doivent être intégrés à des relations de compagnonnage ou de mentoring, comme le souligne Epstein.¹⁸

L'évaluation renvoie donc à un processus transparent, répété dans le temps, multiple en termes de compétences évaluées et de tentatives permises, visant à favoriser, encourager, promouvoir une dynamique de perfectionnement professionnel continu (*lifelong learning*) participant à la construction et au maintien d'un épanouissement des professionnels vis-à-vis de leur identité de médecin.

4. Conséquences de l'ES sur le sentiment d'efficacité personnelle ?

Le sentiment d'auto-efficacité appelé également sentiment d'efficacité personnelle (SEP), décrit initialement par Albert Bandura, est défini comme *“la croyance de l'individu en sa capacité d'organiser et d'exécuter la ligne de conduite requise pour produire des résultats souhaités”*.¹¹⁴ Dans de nombreuses situations, ce sentiment d'auto-efficacité apparaît comme un élément important dans la réalisation d'un comportement. En effet, la confiance que possède un individu en sa capacité à réaliser une tâche spécifique conditionne son engagement dans le comportement adéquat et le niveau de performance qu'il va effectivement atteindre dès lors que celui-ci dépend des actions de l'individu. Le SEP est donc un prédicteur et également un déterminant du comportement futur de l'apprenant dans une situation similaire. Le SEP a ainsi été largement étudié dans de nombreux domaines tels que l'enseignement et la formation ^{114,115} ou l'éducation thérapeutique.¹¹⁶ En situation d'urgence médicale, le sentiment d'auto-efficacité jouerait un rôle important dans la réalisation du comportement adapté ainsi que dans la mise en place des actions requises par la situation.¹¹⁷

Quatre sources principales ont été décrites par Bandura comme étant à l'origine du sentiment d'auto-efficacité : les expériences actives de maîtrise, les expériences vicariantes, la persuasion verbale et les états physiologiques et émotionnels. Compte tenu de la nature de ces sources, la simulation pleine échelle pourrait apparaître comme une situation pédagogique susceptible de renforcer le sentiment d'efficacité personnelle ^{118,119} et donc, *in fine*, d'optimiser la prise en charge des patients en situation clinique.

En recréant une situation de soins de manière extrêmement contextualisée, la simulation permet d'intervenir sur ces quatre éléments. L'expérience active de maîtrise correspond à l'expérience vécue par l'individu en situation. Une expérience positive renforce le SEP tandis qu'une expérience négative le diminue. En permettant aux individus de se confronter à des situations cliniques, la simulation permet ainsi de renforcer le SEP en lien avec la situation rencontrée dès lors que l'individu performe positivement dans la situation. A l'inverse, une situation non maîtrisée par un individu (e.g. difficultés d'intubation répétées) serait susceptible de diminuer le SEP pouvant entraîner une diminution des performances ou l'apparition de comportements inadaptés (e.g. comportement d'évitement) en situation réelle. Turner et al. ont ainsi

montré que le décès du patient simulé entraînait une baisse significative du sentiment d'auto-efficacité en lien avec la situation.⁹⁸

En situation de simulation certificative, un étudiant, et à fortiori un professionnel confirmé, ayant échoué sur une situation, par exemple d'intubation, va voir son SEP diminuer avec risque de modification comportementale s'il est confronté à la même situation en pratique clinique. Si lors d'une session de simulation à visée formative il est possible d'agir sur ce SEP, par exemple en refaisant passer l'étudiant sur une situation similaire dont le formateur contrôle le niveau de difficulté pour ré augmenter le SEP de l'étudiant. Il n'en est pas de même en situation de simulation certificative. Les formateurs encadrant les sessions de simulation certificative devront ainsi être vigilants sur ce point et il pourrait être intéressant de réfléchir à une organisation permettant de confronter de nouveau l'apprenant à cette situation afin de ne pas le laisser sur une situation d'échec (par exemple, en lui proposant une simulation formative sur le sujet).

L'expérience vicariante représente la deuxième source du SEP. Elle correspond à l'observation par le sujet des comportements et des actions réalisées par des individus similaires à soi. Dans le cas de l'évaluation certificative, les observateurs ne sont peu ou pas présents, cette source est donc faible.

Simulation certificative et résolution du conflit cognitif et socio cognitif

Selon l'approche constructiviste de l'apprentissage les connaissances se construisent progressivement à partir des structures cognitives de l'individu au travers d'une interaction permanente entre l'individu et les objets qui l'entourent.¹²⁰ Dans une perspective développementale, Piaget (1975) décrit le développement des structures cognitives de l'individu comme un processus constructif évoluant en paliers successifs qui correspondent à des niveaux d'équilibre.¹²¹ Chaque individu possède des schèmes d'action (i.e. schémas de pensée) qu'il utilise dans une situation spécifique. Ces schèmes d'action permettent alors à l'individu d'agir de façon efficace dans une situation donnée dès lors qu'elle est connue de l'individu. Face à une situation nouvelle ou inhabituelle, les schèmes d'action se relèvent inefficaces et ne permettent pas à l'individu de résoudre cette situation et l'individu se trouve alors en déséquilibre.

Dans la simulation certificative, l'apprenant a des buts de performance dans lequel

l'individu est motivé par le fait de mettre en avant ses compétences en comparaison à autrui. De façon plus précise, l'individu peut être motivé par le fait de se montrer meilleur que les autres, d'une compétence supérieure et l'on parle alors de buts de performance approche. De tels buts se traduisent par le désir de l'individu de démontrer sa compétence propre. A l'inverse, l'individu peut être motivé par le désir de ne pas se montrer moins compétent que les autres. On parle alors de buts de performance évitement. Le choix de but de performance se traduit par un traitement plus superficiel de la tâche, voire un évitement de la tâche ou le choix de tâches plus faciles garantissant l'absence de remise en cause des compétences. Dans ce cas, la recherche de l'aide d'une tierce personne peut être perçue comme un aveu de faible compétence et ainsi être évitée. Lorsque des buts de performances sont adoptés par l'individu, les autres sont alors perçus non pas comme une aide mais comme une menace.

Ainsi, dans le cadre d'une session de simulation le contexte de réalisation de cette session pourrait mettre en avant soit des buts de maîtrise correspondant au désir d'apprendre (simulation formative) soit des buts de performances centrés sur la comparaison sociale (simulation certificative) dans un but de mieux réussir que les autres (performance – approche) ou de ne pas être inférieur aux autres (performance – évitement). La survenue d'un conflit lors du débriefing serait alors susceptible d'être régulée de manière épistémique si les buts de maîtrises sont mis en avant ou de manière relationnelle si le contexte rend saillant les buts de performance. Ces éléments mettent ainsi en avant l'importance du contexte dans lequel est réalisée la session de simulation et notamment la manière dont sont formulés les objectifs de la séance.

5. Peut-on assurer la sécurité psychologique de l'apprenant sans débriefing dans une ES ?

La problématique de la place du débriefing est abordée plus loin dans ce document. Dans ce chapitre, un focus est réalisé sur la sécurité psychologique et le débriefing. Les objectifs de l'évaluation sommative sont d'attester de l'acquisition de connaissance et/ou de compétences. Il s'agit de la validation d'apprentissages qui intervient à la fin d'une séquence d'enseignement afin de juger du degré et de la valeur des apprentissages réalisés par l'étudiant. Elle est habituellement utilisée pour décider de la réussite ou de l'échec de l'étudiant.¹⁷

Habituellement, cette validation se produit à un instant donné. Ce type d'évaluation

tient peu compte de façon globale des éléments suivants : confort physique (au sens large : sommeil, besoins de base) et confort psychique du candidat (accentuation de la charge émotionnelle du fait de l'épreuve sommative). Certains évaluateurs pourraient même compter sur l'augmentation de la charge émotionnelle comme un moteur pour la réussite de l'examen. Ce type d'évaluation dans sa conception n'autorise pas ou peu à la pondération par la connaissance précise du candidat et ce d'autant plus qu'il ne s'agirait pas des mêmes évaluateurs en évaluation formative et sommative. Le débriefing est l'apanage de l'évaluation formative. Dans ces formations, tout est bon pour progresser et apprendre.

La place du débriefing dans l'évaluation sommative est rare pour deux raisons au moins :

- Manque de temps dans le cadre de l'évaluation elle-même (format, temps disponible par rapport au nombre de candidats à évaluer)
- Objectif de l'évaluation lui-même qui n'est pas d'effectuer un processus d'apprentissage mais d'évaluation sommative

Le débriefing est cependant possible même dans ce type d'évaluation sous plusieurs formes :

- Feed-back avec la possibilité de donner des résultats et un ou deux points d'amélioration (que l'évaluation soit réussie ou non)
- Vrai débriefing avec l'ensemble du groupe des candidats évalués à distance de l'évaluation avec reprise des éléments de l'évaluation et échange sur le fond, la forme et les objectifs attendus par l'équipe encadrante

Le type même de rétroaction, même sur une évaluation sommative, peut montrer un effet pauvre ou bénéfique sur les candidats : des résultats de QCM qui notifient la pondération selon les questions ou mieux encore qui restituent des commentaires sur les réponses justes ou fausses constituent des éléments de renforcement positif pour l'étudiant.¹²² C'est l'objectif attendu par exemple des commentaires des questions des dossiers progressifs inégalement effectués en pratique.

La sécurité psychologique dans le cadre d'une évaluation sommative ne tient pas évidemment au seul fait de ne pas effectuer de débriefing. Sur le fond, elle tient également à l'adéquation des objectifs de l'évaluation avec la qualité des étudiants en termes de niveau de connaissances et de compétences (Alignement pédagogique de Bédard voir plus bas). Il est donc indispensable pour assurer cette sécurité de bien connaître les capacités de performance que l'on attend des

étudiants. Sur la forme, elle nécessite que les apprenants ne découvrent pas une méthode d'évaluation qui ne leur serait pas connue. Ainsi il ne paraît pas possible de proposer une évaluation sommative sous forme d'ECOS ou de simulation haute ou basse authenticité sans qu'ils n'aient déjà effectué ce type d'exercice lors d'évaluation formative et cela plusieurs fois dans les mois ou années qui précéderont l'évaluation sommative.

Gaba, dans le cadre d'évaluation formative, propose des éléments de réponses pour éviter de majorer ou d'aggraver les problèmes psychologiques des étudiants lors de séances de simulation.¹⁰¹ Un des éléments principaux semble être de proposer des programmes de simulation adaptés à la vulnérabilité des apprenants et que, s'il existe des risques psychologiques particuliers, alors les formateurs doivent avoir une formation spécifique pour cela.

Il ne faut jamais perdre de vue que ces évaluations sommatives doivent permettre de valider des connaissances et/ou des compétences mais n'ont pas pour objectif de classer les étudiants entre eux du meilleur au moins bon. L'évaluation diagnostique ou normative pourra jouer ce rôle-là.¹⁷ Face à des étudiants dont on peut penser qu'ils ne sont pas tous les meilleurs sur un instant donné (c'est le principe de l'évaluation sommative), lors d'exercice de simulation basse ou haute authenticité, une façon d'assurer leur sécurité psychologique est aussi de permettre le recours à des facilitateurs/comparses dont le rôle est bien défini et identique à chaque intervention (équité de l'intervention).

L'évaluation sommative devrait par ailleurs, faire appel à des scénarios moins difficiles et complexes que dans les évaluations formatives ? et sur des temps suffisamment longs renvoyant là encore à la notion de validation et non de mise en compétition des étudiants. Il est indispensable de bien délimiter le périmètre de ce qui doit être évalué : la pertinence de ce qui est évalué doit être privilégiée et pas la facilité à le faire.¹⁷

Le contenu des scénarios doit permettre à l'étudiant de se raccrocher à des éléments qu'il connaît pour transformer positivement l'exercice. Des histoires cliniques « de chasse » ou des actualités très avancées dans un domaine particulier ne présentent aucun intérêt dans le cadre d'évaluations sommatives dans le parcours des étudiants en santé. Compte tenu des différents éléments à explorer lors de scénarios assez généralistes, on peut logiquement s'interroger sur le fait que les étudiants puissent même connaître à l'avance le contenu générique des exercices proposés (simulation

conversationnelle ou exercice procédural sur une ponction lombaire chez un nourrisson etc.). La connaissance du contenu ne paraît pas nuire à la qualité de l'évaluation sommative.

ii. Conséquences psychologiques, sur les formateurs, de l'ES.

La notion d'équité et de transparence sur l'évaluation sanctionnante (forme et fond) est essentielle pour éviter la contestation des étudiants. La façon dont l'évaluation sera réalisée (type d'évaluation, grille d'évaluation, débriefing ou pas, condition de validation de la simulation ou pas) doivent être très clairement précisés aux candidats. Pour autant dans le cadre sommatif, les études montrent des variabilités entre les évaluateurs partant de 18 % ou 21%¹²³ ou 21 %^{124,125} jusqu'à 40% dans une étude.¹²⁶ Malheureusement, en plus de ces facteurs cognitifs qui impliqueraient l'application stricte de la grille d'évaluation, l'évaluation implique souvent des conceptions différentes des compétences.¹²⁷ Yeales a comparé, dans une étude qualitative des évaluations faites par des médecins seniors (seniors depuis au moins 2 ans) et évaluant au moins 5 Mini CEX par an.¹²⁷ Il a montré que pour 3 vidéos différentes, les scores (sur 6) allaient de :

1,2 à 4,2 (SD 0,98) pour les vidéos « test » de niveau faible

De 2,2 à 5,7 (SD 0,96) pour la vidéo de niveau moyen

Et de 4,5 to 5,7 (SD 0,37) pour la bonne vidéo

En interrogeant les évaluateurs et en analysant leurs verbatims, l'auteur retrouve que :

- Les aspects importants pour déterminer la qualité de la compétence évaluée varient selon les évaluateurs,
- Les évaluateurs comparaient les évalués des vidéos déjà vues le même jour (jusqu'à les classer entre eux) ou encore faisaient des comparaisons avec des pratiques vues dans la vie réelle et faites par d'autres acteurs de santé.

Seulement deux évaluateurs utilisaient les termes exacts de la grille d'évaluation, les autres exprimaient plutôt des sentiments (« c'est affreux », « pas très impressionnant pour le moment », « bon début », « c'est vraiment très bon, c'est vraiment ce que je ferais si j'étais à leur place »).

Ginsburg parle d'idiosyncrasie, c'est-à-dire de différences entre évaluateurs dues à des conceptions différentes de la connaissance, du professionnalisme, du matériel, des interactions avec le patient, des interactions avec l'équipe, de la confiance et de l'impact sur le personnel. Ces impressions subjectives des évaluateurs doivent être

intégrées, au maximum, dans des méthodes d'évaluation optimisées.¹²⁸ Enfin, Gomez avance, en plus du cognitif et des conceptions différentes de l'évaluation des compétences, la présence de considérations sociales et d'une intuition.¹¹² Les émotions des évaluateurs ont trois impacts :

- Un biais dans la prise de décision
- Un bruit aléatoire lors de l'évaluation
- Une source légitime d'informations qui contribue à la décision lors de l'évaluation, notamment parfois l'absence de débriefing qui est pourtant, non seulement un élément pédagogique clef mais aussi un élément « sécurisant » de la simulation.

Néanmoins, une formation adaptée des formateurs en simulation pratiquant l'évaluation sommative rassurera les formateurs en simulation en leur donnant les moyens de placer l'évaluation en simulation dans un bon alignement pédagogique et de proposer un contexte d'évaluation sommative, notamment des scénarios et des grilles d'évaluation adaptés au niveau des évalués.

Messick dit qu'il existe plusieurs conflits éthiques pour les formateurs.¹²⁹ Ces conflits sont liés à :

- La superposition des propres items d'évaluations par le formateur lors de l'interprétation de tests.
- L'utilisation d'un test manquant de validité par l'examineur. Il est donc nécessaire de vérifier l'absence de test plus valide dans la littérature ou d'en créer un (ou de le valider), le cas échéant. Messick propose même d'augmenter le formatif plutôt que de passer en sommatif en cas de possible défaillance de l'outil d'évaluation sommative. Il insiste sur la description des outils d'évaluation (population ciblée, compétence évaluée, la façon d'utiliser le test, les conséquences sociales de l'interprétation du test).

D'autres conflits éthiques pourraient être imaginés comme :

- l'inégalité d'accès aux outils de simulation pendant la phase d'apprentissage, notamment pendant l'évaluation formative en simulation.
- une phase de formation ne mettant pas à disposition des apprenants les outils d'évaluation. Ces mêmes outils doivent être utilisés en évaluation formative et sommative.

Tableau 4 : Propositions pour diminuer l'impact lié à la mise en place de l'évaluation sommative en simulation

Participant-dépendant	Connaissance de l'outil « Simulation en santé »	Expérience en formation de type formatif ≥ 1 au cours des 2 dernières années
	Physiologie	Sommeil : pas de repos de garde en cours lors de la simulation sommative
		Sommeil : nombre limité de gardes au cours des 7 derniers jours avant l'ES
	Sécurité psychologique et émotions	Importance de débriefing l'échec avec l'étudiant (ex : rechercher le vécu d'un événement stressant en rapport avec le contexte d'évaluation)
		Si Burnout récent : feu vert reçu de la part du médecin du travail en charge
Formateurs et Participant-dépendant	Respect du caractère contexte-dépendant des compétences à évaluer	Le cas d'évaluation se situe dans le domaine d'exercice actuel ou futur du participant
	Prise en compte du Sentiment d'Efficacité Personnelle (SEP)	Information claire et transparente sur les conditions de certification
Formateurs-dépendant	Sécurité psychologique par des formateurs correctement préparés.	Les critères de sécurité psychologique peuvent-ils être assurés au mieux ? Avoir programmé des simulations formatives à une fréquence suffisamment rapprochée

Chapitre 5 : les Outils de l'Évaluation Sommative en Simulation

1. Les scénarios en ES

D'après le Guide de Bonnes Pratiques en Matière de Simulation en Santé, la pratique de la simulation en santé nécessite l'écriture préalable d'un scénario lors de l'utilisation de patients standardisés, de simulateurs de patients (mannequins grandeur nature) et de l'utilisation de la simulation hybride. La simulation hybride est l'association de plusieurs techniques de simulation. Par exemple, la combinaison d'un patient standardisé et d'une partie de mannequin (bassin d'accouchement pour l'apprentissage de gestes obstétricaux, « peau simulée » pour perfusion ou sutures, etc.) contextualisée dans un environnement adéquat permet d'apporter du réalisme

aux scénarios en associant une évaluation procédurale et une évaluation comportementale.¹

A ce sujet, la SoFraSimS a publié un référentiel pour la création de scénario et définit le scénario comme un document qui « décrit le déroulement des activités de façon chronologique, les ressources et le temps nécessaire pour chacune d'elle. Le but est d'obtenir une vue d'ensemble de l'enchaînement des séquences pédagogiques de la formation avec notamment les objectifs, la durée, les opérations à réaliser et les moyens pour leur élaboration ». Ce référentiel s'appelle « Référentiel sur l'élaboration de scénarios de simulation en immersion clinique ». Le terme immersion clinique est une modalité de simulation dans lequel l'environnement, simulé ou in situ, joue un rôle important en reproduisant l'environnement clinique.^{130,131}

Cette modalité de simulation peut faire intervenir des simulateurs patients, ou des acteurs or le précédent référentiel excluait les patients simulés en étant axé uniquement sur les scénarios avec mannequin. En partant de ces constats, la mise en place d'une séquence d'évaluation par simulation pose ainsi de nombreuses questions. Les grilles de création de la SOFRASIMS pour les scénarios en immersion clinique avec mannequin sont-elles utilisables pour l'évaluation sommative ? Faut-il créer une grille spécifique pour écrire les scénarios en immersion clinique avec patient simulé pour l'évaluation sommative ? Qu'en est-il des supports pour l'organisation de sessions évaluatives de simulations procédurales non concernées par ce référentiel ?

La grille de création de la SoFraSimS pour les scénarios en immersion clinique avec mannequin est un outil qui permet la création d'un scénario selon une thématique données des objectifs pédagogiques et une forme permettant une reproductibilité. L'évaluation sommative par la simulation nécessitant un cadre précis, les scénarios utilisés doivent également avoir une structure rigoureuse et précise. Nous confronterons ici la grille de création des scénarios de la SOFRASIMS avec les éléments structurants que l'on doit retrouver dans un scénario pour une évaluation par la simulation. Cette confrontation soulève plusieurs interrogations que nous traiterons ci-après.

a. Comment définir les thématiques des scénarios ?

La création d'un scénario d'évaluation sommative commence par le choix d'une thématique qui fait l'objet le plus souvent d'un accord entre experts au niveau local.¹³² Dans le cadre du contrat de confiance de l'évaluateur, les thématiques

doivent être envisagées le plus souvent en fonction des compétences de l'étudiant ou du professionnel à évaluer inscrites dans les référentiels-métiers aussi bien en formation initiale¹³³, qu'en formation continue.^{35,134} Une revue de littérature, avance même la nécessité d'un choix couvrant l'ensemble des compétences à évaluer.⁴⁴

b. Quelles thématiques, objectifs, compétences ?

Le choix d'une thématique et/ou la définition d'un objectif d'évaluation doit probablement dépendre des carences de connaissances des étudiants repérées en fin de processus de formation et en cohérence avec l'évaluation formative, secondairement traduites sous forme de compétences à atteindre¹³⁵ ou encore sur des séquences d'évènements rares, difficilement observables en réalité ou nécessitant (éthiquement), une prise en charge par un médecin senior.¹³⁶ Ces choix de thématiques et d'objectifs dépendent également des outils de simulation techniquement disponibles : « Les thématiques étaient choisies SI et seulement si les outils de simulation étaient capables de reproduire « une simulation réaliste » du cas ». ²⁹ Le principal critère de qualité de l'évaluation sommative en simulation est que la sélection des cas et le développement des tâches sont guidés par les objectifs d'évaluation.⁸⁰ Il est nécessaire de bien préciser les objectifs d'évaluation dans chaque scénario, afin de pouvoir choisir le bon outil d'évaluation, le concepteur du scénario d'évaluation a donc le devoir d'évaluer des objectifs de compétence clairement définis et le devoir de proposer un scénario à l'étudiant le plus proche de la réalité.

c. Quelles sont les étapes et les éléments structurant pour la création d'un scénario d'ES ?

Une fois la thématique et les objectifs d'évaluation identifiés, l'écriture du scénario peut s'envisager. Les scénarios doivent répondre à quatre grands principes en étant prévisibles, programmables, standardisables et reproductibles.³⁹ La rédaction des scénarios doit inclure un script spécifique, des indices, un timing et des évènements permettant de mettre en pratique et d'évaluer les compétences visées.¹³⁷ Ainsi, à partir d'une thématique donnée, on peut envisager des scénarios très diversifiés, mais qui permettent néanmoins de former les étudiants à plusieurs compétences transversales, (par exemple histoire de la maladie, examen clinique, communication, habiletés interpersonnelles, etc...).¹³⁸ D'autres auteurs proposent, le développement

de pools de scénarios « sommatifs » à partir de chaque thématique sélectionnée, avec des variantes sans en modifier la thématique principale et les objectifs d'évaluations, en partant du principe que les étudiants vont se communiquer les scénarios.¹³⁹ La réalisation de scénarios variables reste difficile à mettre en place. En réalité, la plupart des auteurs développent un scénario unique par thématique et objectif de compétence à évaluer.

d. Doit-on proposer des séquences courtes ou longues pour les scénarios d'ES ?

Il n'existe pas de recommandation pour fixer une durée prévisible pour un scénario. Certains auteurs insistent sur la nécessité de proposer des scénarios plutôt courts et très ciblés en soulignant que cela permet une évaluation plus simple (5 minutes).¹³² D'autres auteurs jugent qu'il est nécessaire d'adapter la durée du scénario en fonction de l'objectif d'évaluation.¹³³ On peut suggérer qu'un scénario court soit suffisant pour l'évaluation des compétences techniques pour évaluer une geste procédural simple là où l'évaluation des compétences non techniques nécessiterait un scénario plus long afin de laisser à l'étudiant ou au professionnel la possibilité de se projeter dans l'environnement simulé.¹⁴⁰ Il faut en effet parfois plus de temps pour améliorer le réalisme de la situation avec l'anticipation maximale des demandes de l'examiné.

e. La standardisation des scénarios est-elle évoquée dans la littérature ?

Plusieurs auteurs insistent sur la nécessité d'un cadre de rédaction de scénarios développé au niveau national ¹⁴¹ ou à défaut, au niveau de l'institution.¹⁴² Frey-Vogel propose même d'utiliser les mêmes guides de création de scénario en évaluation formative ou sommative.¹⁴² Furman et al. propose de faire un guide « trouble shooting » qui est en commun pour tous les sites qui utilisent les scénarios.⁸⁰ C'est-à-dire un guide recensant les principales erreurs que les formateurs pourraient commettre lors de leur évaluation sommative. Il existe également des recommandations pour la structuration de l'évaluation en soins infirmiers où les scénarios sont standardisés (International Nursing Association for Clinical Simulation and Learning (INACSL)).¹⁴³ Les Examens Cliniques Objectifs et Structurés (ECOS) consistent en une série d'ateliers dans lesquels les apprenants sont confrontés à des situations cliniques différentes, très ciblées et courtes.¹ C'est une méthode d'évaluation souvent sommative mais qui peut être

formative. Il existe également des recommandations de bonne pratique.¹⁴⁴ L'ECOS est un mode d'évaluation qui peut faire appel à la simulation. Cette approche pédagogique exige une conception robuste pour assurer la pratique et l'évaluation des soins holistiques.

Tableau 5 : Résumé des principaux éléments structurant un scénario d'ES.

Critère	Recommandation
Durée	10 – 15 min
Objectifs	Listing précis des compétences à évaluer
Items indispensables	Évaluation initiale Démarche diagnostique Prise en charge attendue Orientation attendue
Rédaction	Informatique (programmation si possible)
Échelle d'évaluation	Checklist Global Rating Scale Échelle (20 à 30 items)
Validation	Sessions pilotes (test des scénarios et entraînement des évaluateurs) 1 ou 2 cas par étudiant pendant les tests des scénarios
Évaluation	Evaluation par vidéos Test Kappa de Cohen pour différences entre évaluateurs Test t de Student pour capacité de discrimination des cas entre étudiants

f. Les outils d'évaluation doivent-ils être présents dans le scénario ?

On doit distinguer deux types d'outils, les outils adaptables à différents scénarios (type outil pour gestes ou outils pour compétences non techniques¹⁴⁵ et les outils d'évaluation d'un scénario spécifiquement créés et testés pour un scénario. Les outils d'évaluation doivent ainsi être adaptés aux objectifs et aux compétences évaluées. Ils doivent être connus des étudiants et présents dans le scénario, au moins en référence.^{146–149} On peut proposer ainsi une structuration type en fonction des objectifs et des outils d'évaluation.

Tableau 6 : Structuration type en fonction des objectifs et des outils d'ES

	Objectifs techniques	Objectifs non techniques
Scénario	Court	Long
Évaluation	Score analytique	Notation analytique et holistique (ex : ANTS)

g. Peut-on utiliser les mêmes scénarios pour l'évaluation formative et l'ES ?

Plusieurs auteurs avancent qu'il existe des différences entre des scénarios destinés à l'évaluation sommative et formative. Ainsi Rizzolo met en avant la difficulté pour connaître les schémas mentaux des étudiants lors des scénarios d'évaluation sommative et donc la nécessité de faire entrer un formateur qui va « forcer » une transmission (par exemple l'étudiant est réassigné rapidement à un autre patient et doit faire ses transmissions).¹³⁹ D'autres auteurs soulignent l'importance de réécrire (ou modifier) le scénario car les procédures médicales ou chirurgicales ne sont pas toujours standardisées au niveau local, régional, national. Par ailleurs, le matériel n'est pas toujours disponible selon le mode d'évaluation.⁴¹ On retrouve également une autre différence importante entre l'évaluation sommative et formative. En effet, les étudiants ou les professionnels sont évalués généralement seuls alors que l'évaluation formative se fait généralement en groupe avec des scénarios adaptés aux groupes. En outre, pour l'évaluation sommative, le rôle des « facilitateurs/comparses » est dans ce cas tout à fait primordial et doit nécessiter probablement une formation plus importante au préalable que pour une évaluation formative. Le scénario et les aides cognitives pour les facilitateurs doivent être très détaillés dans le document accompagnant le scénario et respecté par le formateur pour ne pas risquer d'« aider » l'examiné.

h. Quelles sont les particularités des scénarios pour L'ES ?

Dans le cadre de la recertification et l'évaluation continue, Levine décrit la nécessité de proposer des scénarios accessibles et peu difficiles pour la recertification des médecins anesthésistes : les scénarios choisis pour les simulations étaient, certes, de bon niveau (intubation difficile anticipée ou non, ischémie peropératoire, hypoxie pendant la sédation) mais fréquemment rencontrés dans la vie réelle. Les scénarios

impliquant des cas rares sont délibérément évités. (Hyperthermie maligne, etc...).¹³⁴ Cette approche peut néanmoins se discuter pour l'évaluation d'étudiants ou de professionnels sur des pathologies rares mais avec un risque de gravité très élevé. Comme cela a été précisé précédemment il est important de choisir le type de scénario en fonction des objectifs (techniques ou non). Il est particulièrement important d'être précis dans la définition des items à évaluer, il faut ainsi préciser le temps dans lequel il intervient dans le scénario et pour chaque item préciser dans le cadre d'un item de score non technique la catégorie évaluée (gestion de la tâche, travail d'équipe, conscience de la situation,...).¹⁴⁰

Il faut préciser sur les feuilles de scénario les artefacts de simulation et la « convention d'évaluation », précisant les limites du mannequin afin de ne pas le mettre en difficulté les étudiants. La validité de l'évaluation par simulation et de l'utilisation du simulateur comme outil d'évaluation dépend de la qualité de l'environnement et de la fidélité et de la reproductibilité des scénarios proposés. Il convient de s'inscrire dans une démarche qualité appuyée par une enquête qui suit la séance de simulation. Ainsi dans une étude de qualité après une évaluation des scénarios, les sujets et les évaluateurs ont noté plusieurs lacunes des simulateurs à mannequin, notamment la mauvaise qualité des sons respiratoires, les formes d'onde ambiguës de l'électrocardiogramme et l'inexactitude du capnogramme pendant l'expiration lente.¹⁵⁰ La qualité de l'évaluation passe également par la nécessité de tester et de stabiliser les scénarios.²⁹

i. Quelle évaluation du scénario ?

Il est nécessaire pour chaque scénario proposé qu'au minimum une relecture collégiale soit effectuée afin d'aboutir à un consensus. Il faut que les experts ayant rédigé les scénarios et qui sont arrivés à ce consensus puissent être identifiés dans le document source du scénario. Plusieurs méthodes peuvent être utilisées pour aider à l'évaluation d'un scénario, la méthode DELPHI peut ainsi être utilisée pour parvenir à un consensus entre plusieurs experts au sujet d'un système de notation normalisé [20].

j. Quel scénario type pour la simulation procédurale en ES ?

Dans le cadre de l'évaluation sommative d'une procédure, il faut probablement utiliser des outils d'évaluation structurés au moyen d'un formulaire d'évaluation par objectif structuré des compétences techniques. L'utilisation d'une grille paraît indispensable.¹⁵¹ Dans le cadre de l'évaluation d'un geste technique simple, l'échelle

doit être validée en amont. Par exemple, dans un modèle d'évaluation de la ponction lombaire, des experts ont évalué en aveugle et indépendamment les sujets à l'aide d'une grille validée par vidéo.¹⁵² On utilise ce mode d'évaluation dans plusieurs domaines médicaux et notamment en chirurgie pour l'évaluation de séquences de gestes plus complexes, avec la mise au point d'outils de compétence pour la chirurgie¹⁵³ ou par exemple pour le développement d'outils de mesure des compétence et d'évaluation en chirurgie laparoscopique.¹⁵⁴

En médecine d'urgence, Oriot propose également dans le cadre de l'évaluation de la pose d'un cathéter intra-osseux, un scénario pour la validation de son échelle de pose de cathéter intra-osseux. Il s'agit plus en réalité que de l'exposition d'un contexte de travail (briefing) que d'un scénario. Il propose également une grille simple d'évaluation du geste sur la base de 12 items testée au préalable.⁷⁰

2. Place du Débriefing dans l'ES ?

Dans le GBPMSS, aucune référence n'est faite sur la place du débriefing lors d'une évaluation sommative.¹

Dans la littérature, on retrouve que la plupart des centres de simulation qui assurent une évaluation de leurs apprenants effectuent un débriefing : l'usage du débriefing est indispensable en évaluation formative comme en évaluation sommative,¹⁴³ et doit faire partie des bonnes pratiques en évaluation en simulation.¹⁵⁵ Par ailleurs, pour d'autres centres, l'usage du débriefing est assuré lors de l'évaluation formative, alors que dans le cas d'une formation sommative, *a contrario*, ils n'en réalisent pas.¹⁵⁶ Quelques équipes minoritaires n'assurent pas de débriefing lors des évaluations¹³² La justification de l'usage des débriefings passe par l'amélioration des compétences,¹³⁶ même en évaluation sommative, et permet aussi de revenir sur l'analyse des compétences non techniques.^{34,134} Les débriefings sont dans la plupart des cas assurés sous forme de feedback immédiats^{80,155} soit sous forme d'entretiens individuels,⁸⁰ ou en groupes,¹³⁴ ou sous forme de compte rendu écrit.¹⁴² Ces feedbacks sont moins complets que des débriefings mais permettent aux apprenants d'avoir une comparaison par rapport aux référentiels de compétences.¹ Les centres évaluateurs proposent des débriefings à distance de la session, encadrés par les évaluateurs, soit sous forme de vidéo ou de comptes rendus écrits en tenant compte des annotations des évaluateurs.^{103,136} Au total, les débriefings sont largement utilisés en évaluation formative comme certifiante ou recertifiante. En évaluation sommative, ils sont assurés

plutôt sous forme de feedbacks courts à l'issue de la session de simulation, en groupe ou de manière individuelle. Ils peuvent être repris a posteriori en présence d'un évaluateur, avec l'aide de la vidéo ou de compte rendus écrits.

3. Place de la vidéo dans l'ES ?

Dans le GBPMSS, «l'enregistrement vidéo et/ou une indexation temps réel des événements peuvent être utilisés comme support de débriefing, mais ne doivent pas en constituer la majeure partie. En effet, il n'est pas nécessaire de revisionner entièrement le déroulement du scénario, mais plutôt, grâce au système d'indexation, d'en sélectionner rapidement les phases importantes pour appuyer au mieux les messages du formateur.» . De plus, le rôle de la vidéo pour les observateurs est abordé : « Lorsque les séances de simulation se font en groupe, la vidéo permet aux apprenants « non participants » de pouvoir observer la retransmission du scénario et ainsi de pouvoir bénéficier au mieux du débriefing ». ¹ Dans la littérature, des exemples supplémentaires d'utilisation de la vidéo sont cités :

- ▶ Boet décrit l'exemple des internes d'anesthésie pour l'apprentissage des compétences non techniques en simulation d'urgence dans une étude sur l'intérêt de l'auto-évaluation dans l'enseignement par simulation.¹⁵⁷
- ▶ Lors de la pratique de l'évaluation sommative, la vidéo est utilisée par de nombreuses équipes.
- ▶ Dans de rares cas lorsque des observateurs sont présents, elle permet, comme en évaluation formative, à ces observateurs de suivre la session.¹³⁴
- ▶ De même, certaines séquences de vidéos peuvent être utilisées lors du pré briefing ou même lors de la session de simulation (par exemple pour des enregistrements de séquences de coelioscopie per opératoire (13). Cela permet d'augmenter le réalisme d'une situation.
- ▶ Enfin, elle peut être utilisée lorsque des débriefings sont réalisés.¹⁰³
- ▶ Plus spécifiquement, la vidéo permet la création d'outils d'évaluation.¹⁵⁸
- ▶ Elle permet la formation initiale des évaluateurs¹⁴¹ mais aussi leur maintien de compétence :
 - Des vidéos de performances variables (allant du débutant à l'expert) sont à coter, permettant de se familiariser avec les outils,
 - Pour valider les formations, de nouvelles cotations sont possibles, permettant de s'assurer d'une bonne corrélation inter et intra observateurs.

- soit les évaluateurs cotent en direct, pendant le déroulement de la session de simulation^{34,153,159}
- soit ils cotent en différé, avec re-visionnage de la séquence vidéo(9).

La cotation sur vidéo peut être faite après un simple visionnage ou plusieurs, d'un ou de plusieurs évaluateurs.¹⁶⁰

La vidéo reste une preuve, auprès de l'administration et des étudiants, de la performance réalisée par les personnes évaluées.¹⁰³

La vidéo est intéressante car elle permet une anonymisation de certaines évaluations en ne montrant que les mains par exemple.¹⁶¹ La littérature avance de nombreuses limites (problèmes d'angles de vue⁴⁴ de pertes de données allant de 1 à 2,5% des vidéos.¹⁶²

Allant plus loin que le «simple» outil vidéo, certains avancent la possibilité d'un traitement des images avec une évolution vers une cotation automatique informatique, c'est notamment le cas avec la technologie du « eye tracking » qui permet, grâce au port d'une paire de lunettes, de différencier le regard d'un débutant et celui d'un expert.¹⁶³

Au total, la vidéo peut être utilisée lors du prébriefing pour l'information des participants. Elle peut être utilisée pendant la simulation pour augmenter le réalisme (boucle d'échographie, de coelioscopie, ...).

Elle permet :

- aux observateurs de visionner la session en direct.
- de revisionner les séquences parfois par l'apprenant en auto-apprentissage, ou bien sélectionnées par les formateurs lors des débriefings, pour les apprenants.
- de former les évaluateurs, (calibration et recalibration)
- de coter les apprenants en différé, le cas échéant.

Malgré des difficultés techniques à prendre en compte, la vidéo permettra probablement, dans le futur, une aide automatisée à la cotation.

Tableau 7 : Utilisation de la vidéo en simulation.

	Evaluation Formative	Evaluation Sommative
Prébriefing	Information des participants	
Simulation	Majoration du réalisme du scénario (ex : vidéo de coelioscopie)	
	Visionnage par les observateurs	
Visualisation immédiate après la simulation	Autoapprentissage	Pas d'autoévaluation (dans la littérature)
	Débriefing par formateurs (séquences sélectionnées)	
Visualisation en différé	Apprentissage de travail d'équipe ou de gestuelle à but formatif	- Formation des évaluateurs (calibration et recalibration) - Débriefing différé - Preuve administrative

Que ce soit en évaluation formative ou sommative, l'utilisation de la vidéo est soumise à des contraintes légales, variables selon les pays.

Au niveau Européen, depuis 2018, de nouvelles obligations permettent de protéger les droits des citoyens européens : de respecter les droits des personnes et de sécuriser les données.^{164,165}

En France, c'est la Commission Nationale Informatique et Liberté (CNIL) qui organise les droits, notamment en vidéo. En effet, l'article premier de la loi Informatique et Libertés précise que « "l'informatique doit être au service des citoyens. (...) Elle ne doit porter atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques".¹⁶⁶

Le fait d'enregistrer ou de conserver des données personnelles (par exemple, une photo ou une vidéo permettant d'identifier des personnes physiques qu'elles soient en groupe ou isolées), constitue, au sens de la réglementation de la protection des données, comme un traitement de données à caractère personnel.

Et de traitement de données à caractère personnel doit justifier de « finalités déterminées, explicites et légitimes"¹⁶⁷

De plus, la capture d'images et leur retransmission doivent s'effectuer dans le respect des règles relatives au droit à l'image. Toute personne pouvant s'opposer à la reproduction de son image, sur quelque support que ce soit, la prise d'une photographie et sa diffusion doivent faire l'objet d'un accord écrit de la personne concernée si elle est majeure, ou de ses deux parents s'il s'agit d'un mineur.

Lorsque des données à caractère personnel sont collectées, les personnes concernées doivent être informées de :

- l'identité et les coordonnées du responsable du traitement de données ;
- la finalité du traitement (l'objectif poursuivi nécessitant de collecter ces informations) ;
- du caractère obligatoire ou facultatif des réponses et les conséquences éventuelles d'un refus ;
- des destinataires des données ;
- la durée de conservation des données ;
- l'existence et des modalités d'exercice des droits des personnes ;
- l'existence d'un transfert des données hors de l'Union Européenne, le cas échéant ;
- les coordonnées du Délégué à la protection des données lorsqu'il en a été désigné un ;
- la base juridique du traitement de données ;
- le droit d'adresser une réclamation à la CNIL.

Il appartient donc au responsable de traitement d'obtenir toutes les autorisations utiles préalablement à l'utilisation de photographies ou de retransmission vidéo, dès lors que la personne est reconnaissable. En cas de refus d'une personne de se faire enregistrer, la loi ne stipule pas la conduite à tenir : comme il n'y a pas de vidéo possible, il n'y a pas d'enregistrement donc la loi ne s'applique pas. Il appartient donc à l'autorité organisatrice de prévoir ce cas : comme par exemple une évaluation en direct de l'étudiant.

Au total : l'utilisation de la vidéo nécessite l'accord de l'apprenant, le respect des règles locales et l'organisation de la structure mettant en place l'évaluation utilisant la vidéo pour permettre la protection des droits des personnes et la sécurisation des données.

4. Organisation des infrastructures en simulation pour ES

«Les programmes de simulation peuvent être mis en place au sein de différentes infrastructures, quelle que soit cette infrastructure, une organisation doit être identifiée et mise en place selon les éléments généraux décrits dans ce chapitre (missions, gouvernance, etc.) ».¹

Le GBPMSS revient sur la nécessité d'une bonne articulation des encadrants de la simulation (comité pédagogique, scientifique, formateurs, personnel administratif,

etc...). L'évaluation est possible, mais doit être encadrée. *“Sous réserve de prévenir les gens formés qu'ils en sont en évaluation/possibilité de leur donner les outils d'évaluation (ex.: échelle) //nécessite des formations spécifiques des formateurs à utilisation d'échelle”*

Dans le GBPMSS plusieurs infrastructures de simulation sont décrites : centres de simulation, simulation in situ, ateliers de simulation et les niveaux 1, 2 ou 3 correspondants, en fonction des équipements, du personnel, des activités de recherche, etc...

Des plateformes mutualisées de simulation sont décrites. Elles mettent à la disposition des « concepteurs » leur expérience ainsi que les moyens méthodologiques, techniques et humains à la réalisation de programmes de simulation. Des conventions, précisant les liens entre les infrastructures, la nature des prestations échangées, ainsi que le périmètre de couverture, doivent être clairement établies. Les plates-formes mutualisées de simulation peuvent être physiques ou virtuelles.»¹

Dans la littérature, l'organisation nécessaire à la mise en place de l'évaluation sommative a beaucoup de points communs avec celle nécessaire à l'évaluation formative.¹⁴³

Tout comme l'évaluation formative, la mise en place de l'évaluation sommative nécessite des évaluateurs entraînés réalisant un processus standardisé incluant des outils d'évaluation adaptés pouvant être partagés avec les personnes évaluées.¹⁶⁸ Elle nécessite, cependant, de prévenir les candidats du caractère sommatif de l'évaluation et de convenir de la présence ou non d'un feedback ou d'un débriefing à l'issue de la session d'évaluation.¹⁴³

La réalisation de sessions d'évaluations sommatives nécessite la mise en place d'un système qualité.⁸⁰ Étonnamment, une étude américaine montre que le véritable pré requis à une évaluation sommative en simulation est le niveau théorique des étudiants et que la pratique de simulation formative avec débriefing n'a pas d'impact sur les scores d'anxiété ni de performance.¹⁰³ Concernant l'infrastructure des centres, à l'instar du MOCA (Maintenance Of Certification in Anesthesiology) une édition de guidelines nationales avec cahier des charges à respecter est possible.³⁵ Comme pour l'évaluation formative, la formation in situ paraît intéressante car elle permet d'augmenter le réalisme et la projection dans la vie professionnelle.¹⁶⁹

Au total, sous réserve de respect des critères énoncés par les guidelines (formation des formateurs, matériel adéquat, présence d'un système qualité, ...), l'évaluation

sommative peut avoir lieu dans n'importe quel type de structure : centre de simulation, simulation in situ ou ateliers de simulation.

5. Place de la recherche en ES ?

Dans le chapitre 4 du guide de bonnes pratiques en matière de simulation en santé, les auteurs se sont appuyés sur la conférence de consensus internationale sur la recherche en simulation (2011) menée par la Society for Simulation in Healthcare (SSH).¹ Elle a identifié 10 axes de recherche majeurs en simulation dont un axe sur la recherche en évaluation :

Appliquées à l'évaluation sommative en simulation, des exemples de thématiques de recherche envisageables sont :

- ▶ AXE 3 : Etude sur le design des systèmes intégrés de simulation : recherche sur une automatisation des outils d'évaluation.¹⁷⁰
- ▶ AXE 5 : Recherche sur un nouvel outil d'évaluation sommative, de novo ou comparaison à un outil préexistant (cf groupe évaluation)
- ▶ AXE 5 : Recherche sur l'optimisation d'un programme de simulation.⁸⁰
 - A quel moment proposer une évaluation sommative ?
 - Comment optimiser l'articulation entre évaluation formative et sommative ?
 - Comment optimiser la formation initiale des évaluateurs, leur formation continue ?
 - Comment introduire un nouvel outil d'évaluation sommative ?
 - Comment optimiser la mise en place de programmes de formation intégrant une évaluation sommative ?
- ▶ AXE 9 : Etude des processus d'évaluation et d'accréditation basés sur la simulation en santé.

Au total, la recherche en évaluation sommative peut porter sur l'optimisation des process de simulation (programme, structure, ressources humaines...), sur la comparaison d'outils d'évaluation sommative entre eux ou encore sur l'automatisation des moyens d'évaluation

6. Le GBPMSS est-il utilisable en ES en simulation ?

Dans son chapitre 5, le GBPMSS aborde la notion d'évaluation, principalement celle des programmes de formation :

« L'évaluation est indispensable à un processus de formation, d'analyse des pratiques ou de recherche par simulation. Elle doit s'appliquer aux apprenants, aux formateurs, aux programmes de simulation proposés et à l'organisation dans son ensemble. Dans chaque infrastructure, un processus d'évaluation et d'amélioration est formalisé et mis à jour régulièrement conformément aux principes de management de la qualité définis par l'infrastructure (cf. § 2.7).

Il décrit :

- les modalités d'évaluation (indicateurs, audits et contrôles qualité) ;
- les outils d'évaluation ;
- la périodicité des évaluations ;
- les modalités d'identification des actions d'amélioration suite aux évaluations ;
- les modalités de prise en compte et d'enregistrement (traçabilité) des retours/évaluations des apprenants et des formateurs dans les futures offres de programmes de simulation ».

Le GBPMSS décrit aussi l'évaluation des formateurs (5.2) et des apprenants, réalisée lors du débriefing (5.3).

« 5.3 Évaluation des apprenants Lors du déroulement du scénario en séance de simulation, les actions réalisées par les apprenants témoignent d'un raisonnement ou de l'application de procédures connues. Lors du débriefing, les formateurs recueillent les informations utiles auprès des apprenants pour comprendre le raisonnement qui sous-tendait les actions observées. L'évaluation des apprenants concerne non seulement l'analyse de leurs comportements, mais aussi de leurs connaissances à la base des choix effectués pendant la simulation. Pour cela, il est nécessaire que le formateur fasse verbaliser à voix haute les apprenants à propos des actions qu'ils ont développées. Cela permet de favoriser le transfert des apprentissages »

« L'évaluation formative permet à l'apprenant, au cours ou au terme d'une phase d'apprentissage, de repérer ses progrès, d'identifier ses points forts, mais aussi de comprendre ses propres difficultés et d'y remédier en connaissance de cause. Le formateur doit aider à cette compréhension et, au besoin, proposer les actions d'amélioration nécessaires, par exemple, à l'aide d'un document remis en fin de session de simulation (fiche d'aide à la progression) (cf. exemple en annexe 9). Ce document présente les acquis, les axes de progrès, l'orientation des sessions de simulation à venir, l'identification des actions d'amélioration à mettre en œuvre dans la pratique professionnelle. Il peut aussi proposer le calendrier de mise en œuvre

associé et éventuellement des mesures d'impact sur la pratique à la suite de la séance de simulation. Il est conjointement validé par l'apprenant et le formateur. »

L'évaluation formative est traitée ici, avec la remise d'une fiche bilan permettant de visualiser les axes de progression

Le respect des engagements des apprenants pour suivre le programme de simulation est également évalué (assiduité, suivi des actions d'amélioration proposées, etc.). 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.*

La seule référence à l'évaluation sommative (ou sommative) est à la fin du document, avec une ouverture sur la possibilité d'utiliser des outils d'évaluation validés.

Le référentiel aborde l'évaluation par le biais de la recherche. Elle incite à l'évaluation formative par le biais des débriefings mais aussi en fin de session de simulation par le biais de fiches d'aide à la progression.

La mention de l'évaluation sommative est faite à la fin du document sous forme d'ouverture et à condition de disposer d'outils spécifiques.

Au total : même si le GBPMSS est essentiellement axé sur l'évaluation formative, l'ouverture en fin de référentiel permet l'idée que pour réaliser l'évaluation sommative, il faut

- Construire les outils spécifiques
- Assurer la validation de ceux-ci

Les chapitres concernant les programmes, l'organisation, les infrastructures et la recherche paraissent utilisables pour la mise en place de l'évaluation sommative. Concernant les débriefings et l'utilisation de la vidéo, les différences sont notables entre l'évaluation formative et l'évaluation sommative.

7. Formation des Formateurs pour l'évaluation sommative en simulation :

a. Qui peut faire une évaluation sommative par la simulation ?

Afin de répondre la question de qui peut faire une évaluation sommative en simulation, nous pouvons nous appuyer sur les trois types de feedback existant en évaluation formative en simulation lors du débriefing (p.34).¹ Au cours de ces feedbacks, la validation des bonnes pratiques et la proposition des points d'amélioration peuvent être proposées par: (i) les personnes formées elles-mêmes,

c'est-à-dire **les apprenants**, (ii) **les pairs** lorsqu'ils sont observateurs, et (iii) **les formateurs** qui animent la session. Lors d'une évaluation sommative en simulation, la littérature propose essentiellement une évaluation par des formateurs. C'est cette modalité qui est décrite dans ce référentiel.

Les expériences d'évaluation sommative par les pairs sont peu nombreuses dans la littérature et présentent des limites importantes.^{171,172} Il y est souligné l'importance de la formation préalable à l'évaluation et aux outils, échelles et scores utilisés. L'évaluation entre pairs représente un gain de temps pour les formateurs et réduit le stress des apprenants mais il y a des risques importants de biais et de faible fiabilité inter-évaluateurs.¹⁷³ Formateurs et apprenants peuvent avoir des perspectives significativement différentes sur les stratégies d'évaluations et d'apprentissages.¹⁷⁴

Les essais d'auto-évaluation « sommative » sont en nombre limité et identifient un delta entre l'évaluation faite par l'apprenant lui-même et le formateur référent. Ce delta n'étant pas toujours en faveur de l'apprenant.¹⁷⁵ L'évaluation non « sommative » montre des résultats corrélés à ceux d'évaluateurs extérieurs mais avec un delta qui souligne la nécessité d'une calibration.¹⁷⁶ Il semble actuellement compliqué de déléguer l'évaluation sommative aux apprenants ou à leurs pairs sans avoir la certitude d'un partage optimal des objectifs et du rationnel de l'évaluation. Ces limites majeures pourraient être contournées en partie par une approche globale différente de l'ensemble du cursus où évaluations formatives et sommatives se mélangent et s'étalent dans le temps.¹⁷⁷ Cette approche est loin d'être répandue actuellement, et nécessite une refonte globale de la formation pour être mise en place.

Au final, il est nécessaire que l'évaluation sommative en simulation soit valide pour la compétence évaluée, fiable, et reproductible par différents évaluateurs comme par le même évaluateur qui répèterait l'évaluation à un autre moment.⁸⁴

- b. Les formateurs doivent-ils avoir une formation spécifique pour pratiquer l'évaluation sommative ?

- i. [Une littérature très axée sur l'évaluation formative](#)

Concernant les formateurs en simulation, le guide des bonnes pratiques en matière de simulation en santé.¹ définit leur rôle et ébauche une description de leurs compétences : ceux-ci doivent avoir une expérience pédagogique, une expérience dans la thématique enseignée ainsi que des compétences spécifiques en matière

d'analyse comportementale ou de gestion des risques nécessaires en fonction des programmes de simulation proposés. *«Les formateurs, occasionnels ou réguliers, médicaux ou non médicaux, doivent détenir une formation spécifique à la simulation (diplôme universitaire, formation reconnue officiellement notamment par les conseils nationaux professionnels concernés, y compris les formations à l'étranger, stages, formations courtes) adaptée à leur implication dans la formation en simulation et au type d'infrastructure, ou à défaut une validation de l'expérience reconnue par le responsable de l'infrastructure et/ou le responsable pédagogique et/ou scientifique »* (p.21).(1) Le Guide de bonnes pratiques en matière de simulation en santé est essentiellement destiné à l'évaluation formative.¹

Dans la littérature, de nombreux auteurs avancent la nécessité d'une formation spécifique à l'évaluation sommative.¹⁷⁸ En effet, Downing avance le risque d'un jugement qui a tendance à se substituer à l'évaluation si les évaluateurs ne sont pas bien formés.¹⁷⁹ Tout comme pour l'évaluation formative, les formateurs pratiquant l'évaluation sommative en simulation doivent idéalement être des seniors avec une solide expérience clinique et une formation spécifique à l'évaluation sommative.^{38,147}

ii. Pour ou contre une formation spécifique des évaluateurs en ES ?

Pourtant plusieurs auteurs ont montré que ces mêmes seniors, experts cliniciens, avaient un risque de « résistance » à des formations à l'évaluation « sommative » et ce, d'autant plus que les schémas d'évaluations proposés différaient de leurs schémas de bonnes ou mauvaises performances.¹⁸⁰⁻¹⁸³

Pour d'autres auteurs, les évaluateurs ne doivent pas avoir obligatoirement une solide formation en évaluation formative en simulation mais surtout avoir une formation de formateurs à l'évaluation sommative.³⁸ Une formation à l'évaluation sommative permet d'obtenir jusqu'à 95% d'accord entre un jeune évaluateur et un évaluateur plus ancien.¹⁸⁴

D'autres études ont échoué à mettre en avant un quelconque bénéfice à une formation de formateurs à l'évaluation. Ainsi, dans une étude randomisée incluant des internes évalués par le score mini-CEX, Cook et al ont étudié l'impact d'un atelier d'une demi-journée de formation de formateurs sur les performances des évaluateurs en simulation quatre semaines après et n'a pas trouvé de différence significative de cet entraînement. L'auteur critique son étude en avançant un manque de puissance de l'étude en raison d'une taille d'échantillon trop petite pour montrer une différence,

des scénarios trop diversifiés pour étudier les mêmes compétences mais surtout une durée de formation trop courte avec l'idée de développer une formation de formateurs par internet.¹⁸⁰ Iramaneerat et al ont montré l'existence d'erreurs pendant les évaluations « sommative » de simulation dans des Examens cliniques Objectifs Structurés (ECOS) chez des étudiants de 4ème année de médecine et ce, malgré une formation spécifique des patients simulés à l'évaluation sommative (non décrite mais améliorable d'après l'auteur).¹³⁸

D'autres études ont aussi montré un accord initial entre évaluateurs au début des évaluations qui peuvent disparaître avec le temps. McLaughlin avance un effet fatigue mais aussi un nécessaire "échauffement" des formateurs qui deviennent plus efficaces après 15 ou 20 minutes d'évaluation.¹⁸⁵

iii. Les principales erreurs rencontrées en simulation :

La littérature décrit des erreurs qui peuvent être rencontrées lors de l'évaluation sommative :

- ▶ L'erreur de contraste : une prestation donnée se verra attribuer plus fréquemment une mauvaise note si elle fait suite à une excellente prestation. L'inverse est vrai, aussi.¹⁸⁶
- ▶ L'erreur d'homogénéisation : La tendance des évaluateurs à donner une note ni trop bonne ni trop mauvaise, rendant plus difficile la discrimination.¹⁸⁷
- ▶ L'erreur de halo : avoir tendance à tout voir bien ou mal dans une même prestation.
- ▶ L'effet temps : Les évaluations sont biaisées par des observations de performances ou de contre-performances précoces ou tardives lors la prestation.¹⁸⁷
- ▶ La clémence : éviter de donner de mauvaises notes.¹⁸⁷
- ▶ L'erreur de référentiel (« ils font comme moi ») : l'évaluateur juge en fonction de ce qu'il aurait fait et non de l'outil d'évaluation.¹⁸⁷
- ▶ L'effet groupe : les évaluations sont basées sur la prise en charge globale plutôt que sur la performance de l'évalué.¹⁸⁷

Tableau 8 : Exemples d'erreurs, d'effets et de biais d'évaluation en ES en simulation

186,187

Type d'erreur	Description de l'erreur
Erreur d'homogénéisation	Tendance des évaluateurs à donner une note ni trop bonne ni trop mauvaise, rendant plus difficile la discrimination.
Effet de halo	Tendance à tout voir bien ou mal dans une même prestation.
Effet temps	Biais lié aux observations de performances ou de contre-performances précoces ou tardives lors de la prestation.
Biais de « clémence »	Volonté de ne pas donner de mauvaises notes.
Erreur de référentiel	Jugement en fonction de ce que l'évaluateur aurait fait et non de l'outil d'évaluation.
Effet groupe	Évaluation basée sur la prise en charge globale plutôt que sur la performance de l'évalué.

C. Proposition de formation des formateurs en Évaluation Sommative

i. Les apports nécessaires lors des formations à l'évaluation sommative

Les objectifs de formation à l'évaluation « sommative » pour les évaluateurs sont les suivants :

- ▶ Apports cognitifs et sensibilisation sur le **type d'erreurs en évaluation** : sensibiliser le formateur sur les différents types d'erreurs (cf supra) en insistant sur la nécessité de réduire les erreurs autant que réussir à discriminer les prestations effectuées.¹⁸⁸
- ▶ Apport sur la dimension de la **performance évaluée** : il sera réalisé par l'étude d'une vignette ou d'une vidéo ou un jeu de rôle. A noter : plusieurs comportements peuvent être observés/attendus pour satisfaire une compétence.¹⁸⁹
- ▶ Connaissance de l'**outil d'évaluation** : les évaluateurs doivent être entraînés à évaluer de façon répétitive plusieurs vidéos avec un feedback de leur évaluation afin de diminuer la variabilité individuelle et la variabilité inter observateur.¹⁸⁹
- ▶ Développement des **comportements d'évaluateurs** : les compétences observées le sont principalement à certains moments. Les évaluateurs doivent l'avoir compris donc pouvoir anticiper ces moments-clefs pour une meilleure évaluation.¹⁸⁹
- ▶ **Calibration des évaluateurs** : les évaluateurs doivent s'entraîner afin d'avoir une évaluation reproductible entre différents évaluateurs et pour un même évaluateur à différents moments donnés. ^{104,158,178,190}

En France :

Dans le but d'homogénéiser les formateurs en simulation, la SoFraSimS a publié un référentiel de compétences transmises lors des formations courtes de formateurs en simulation en santé. (28)

Les compétences transmises lors de ces formations courtes sont essentiellement destinées à des formateurs en évaluation formative. Elles permettent de mettre en scène un scénario et de le débriefer en fonction des objectifs pédagogiques.¹⁹¹

ii. Les profils des formateurs en évaluation sommative en Simulation

Pour mettre en place une évaluation sommative en simulation, il apparaît donc que plusieurs types de formateurs sont nécessaires :

1. Des formateurs « concepteurs »

Des **formateurs « concepteurs »** capables d'intégrer la pratique de l'évaluation sommative au curriculum et d'en superviser la mise en place avec, notamment, le développement de scénarios de simulation pour l'évaluation sommative.

2. Des formateurs « opérateurs »

Des **formateurs « opérateurs »** en simulation en charge « du déroulé de scénario » capables d'encadrer la mise en scène des scénarios, nécessitant les compétences des formateurs de standard.¹⁹¹

3. Des formateurs « évaluateurs »

Des **formateurs « évaluateurs »** capables d'utiliser des outils d'évaluation sommative, entraînés et calibrés pour évaluer le niveau d'apprentissage de compétences techniques ou non techniques. (cf supra).¹⁹²

Nous proposons les formations et compétences des trois types de formateurs impliqués dans la formation « sommative » dans le tableau 9.

Tableau 9 : Formations, missions et capacités des formateurs impliqués dans l'ES.

	Formateur-concepteur	Formateur-opérateur	Formateur-évaluateur
Formation initiale du formateur	Diplôme Universitaire de formateur en simulation ou Diplôme Universitaire de pédagogie de 3 ^e cycle associé à une formation courte en simulation.	Formation courte de formateur.	Formation courte de formateur + formation spécifique à l'évaluation réalisée.
Formation continue du formateur	Réalisation de 6 sessions par an ¹	Réalisation de 6 sessions par an ¹	Réalisation de 6 sessions par an ¹ + recalibration régulière adaptée aux outils utilisés.
Missions	Intégration de l'évaluation au curriculum. Conception et validation de l'évaluation.	Encadrer la mise en scène des situations d'évaluation.	Procéder à l'évaluation.
Capacités génériques (intégrées aux formations de formateurs)	• Maîtriser les types d'erreurs en évaluation • Avoir des notions sur les performances évaluées. • Décrire les outils disponibles pour évaluer de façon sommative. • Identifier les comportements d'évaluateurs. • Expliquer l'importance de la calibration des évaluateurs. • Reconnaître l'importance du feedback/débriefing le cas échéant. • Prendre en compte les spécificités de l'utilisation de la vidéo (le cas échéant).		
Capacités spécifiques à une fonction donnée (formation spécifique)	Concevoir et valider la situation d'évaluation Comparer les outils d'évaluation. Organiser la formation du/des formateur(s)-évaluateur(s)		Identifier les performances évaluées. Maîtriser les outils d'évaluation utilisés. Identifier les comportements d'évaluateur. Utiliser fidèlement un outil d'évaluation après calibration

Selon la méthode d'évaluation sommative utilisée, un formateur pourra être amené à avoir soit une, soit plusieurs fonctions parmi celles de formateurs « formateur-concepteur » ou « formateur-opérateur » ou « formateur-évaluateur ». Les fonctions de « formateur-opérateur » et de « formateur-évaluateur » ne peuvent – a priori – pas être assumées simultanément pour une évaluation. Le « formateur-opérateur » doit être idéalement concentré uniquement sur le seul bon déroulé de la simulation. De même, le « formateur-évaluateur » doit être idéalement concentré uniquement sur la seule réalisation de l'évaluation.

Au vu des compétences spécifiques décrites pour les évaluateurs en simulation, des formations complémentaires spécifiques semblent nécessaires et seront proposées dans la version 2022 du référentiel de formation des formateurs en simulation de la SoFraSimS.

Dans ses recommandations, la SoFraSimS a précisé que des formations supplémentaires pourraient être nécessaires pour des thématiques particulières (par exemple débriefing difficile, annonce de mauvaise nouvelle, grimage...).¹⁹¹

iii. Les compétences des trois catégories de formateurs :

1. Compétences préconisées pour être formateur-opérateur :

Formation courte avec notions actualisées sur l'évaluation sommative en simulation (outils d'évaluation, scénarios cadrés avec facilitateurs stéréotypés etc...), importance de la reproductibilité. Idéalement les formateurs-opérateurs participent aux évaluations formatives et sommatives pour garantir l'alignement pédagogique.

2. Compétences préconisées pour être formateur-concepteur :

Les encadrants bénéficient préférentiellement d'une formation de type Diplôme Universitaire de formateur en simulation ou d'un Diplôme Universitaire de de pédagogie de 3e cycle associée à une formation courte en simulation :

- ▶ Notion pédagogique sur évaluation sommative :
 - Quand et comment intégrer l'évaluation sommative dans un cursus.
 - Outils : quels sont les outils validés, comment construire et valider un outil d'évaluation, etc...
 - Mise en place d'un programme intégrant l'évaluation sommative en simulation.
 - Connaître la différence entre évaluation sommative et formative; ainsi que le lien entre les deux.
 - La session de simulation « sommative : théorie et mise en place »
- ▶ Scénarios en simulation sommative :
 - Construction d'un scénario adapté à l'évaluation sommative et à ses objectifs.
 - Comprendre l'intérêt et savoir intégrer une grille d'évaluation.
 - Rôle, positionnement et encadrement du comparse/facilitateur dans le scénario.

- Spécificités du pré-briefing et du briefing en évaluation sommative.
- Connaître la place, la structure et les particularités d'un éventuel débriefing en simulation sommative.
- La conception, l'intégration et l'implémentation de l'évaluation est une compétence du formateur-concepteur :
 - Quel est l'outil de simulation le plus approprié pour la compétence à évaluer ?
 - A quel moment du parcours d'apprentissage doit-on l'utiliser ?
 - Quelle calibration du niveau de difficulté (novice/expert) ?
 - Quel choix du/des critère(s) d'évaluation ?
 - Quelles formations pour les formateurs-évaluateurs.

3. Compétences préconisées pour être formateur-évaluateur :

Formation courte avec notions actualisées sur l'évaluation sommative en simulation comprenant notamment des notions sur les outils d'évaluation, les scénarios cadrés avec facilitateurs stéréotypés etc... Pour chaque évaluation sommative en simulation, une formation spécifique à l'évaluation réalisée portant sur la performance évaluée, l'outil d'évaluation utilisé, le comportement spécifique d'évaluateur pour cette évaluation et une calibration.

- Connaissances des outils validés, de la méthodologie d'utilisation de ces outils et formation pratique à l'utilisation de ces outils incluant une calibration.
- Notion pédagogique sur évaluation sommative :
 - Quand et comment intégrer l'évaluation sommative dans un cursus.
 - Outils : quels sont les outils validés, comment construire et valider un outil d'évaluation, etc...
 - Spécificité de l'évaluation interprofessionnelle.
 - Connaître la différence entre évaluation sommative et formative ; ainsi que le lien entre les deux.
 - La session de simulation « sommative : théorie et mise en place ».
- Savoir identifier, analyser, critiquer et gérer une session de simulation difficile à évaluer et sa cause.
 - Problème vidéo.
 - Problème technique (panne de simulateur, manque de matériel...)
 - Problème lié au pré briefing ou au briefing.

- Objectif pédagogique cible de l'évaluation non atteint au cours du déroulé de scénario.
- Connaître la place, la structure et les particularités d'un éventuel débriefing en simulation sommative (entretien exploratoire).
- Maîtrise des outils d'évaluation (adaptés au scénario à évaluer) en simulation.

Pour tous :

Concernant la formation des Évaluateurs : il est nécessaire que des descriptions précises des formations suivies par les évaluateurs soient présentées lors des publications des outils validés d'évaluation. Ces descriptions incluront la durée et le contenu de la formation, les caractéristiques des évaluateurs, les objectifs et les moyens utilisés lors des formations, les moyens mis en place pour obtenir et maintenir l'homogénéisation des évaluations (calibration).¹⁸⁷

Peu d'échelles validées proposent une formation à l'utilisation de l'échelle. Voici quelques exemples :

- Échelle TAPAS (temps proposé de 2 heures d'entraînement des formateurs avant début des évaluations, le contenu est "formalisé" d'après l'auteur mais non décrit, rendant plus difficile la mise en place).¹⁹³
- Dans une traduction en Italien de l'échelle anglophone d'Ottawa, évaluant les Crisis Ressource Management, les auteurs disent qu'aucun des évaluateurs n'a eu de formation spécifique à l'utilisation de l'échelle.¹⁹⁴
- Dans la présentation d'une échelle technique d'évaluation de la mise en place d'un Cathéter Intra Osseux, l'auteur décrit tout le processus de validation de l'échelle sans pour autant décrire le temps nécessaire à la formation des futurs utilisateurs de cette échelle.⁷⁰ Enfin, il est nécessaire de développer la formation continue des évaluateurs : il faut considérer l'activité d'évaluateur comme une activité nécessitant une formation continue avec des contrôles qualité des évaluations réalisées, un suivi des formateurs et une maintenance des compétences.¹⁸⁷ Cette formation continue pourrait se faire par des évaluations de vidéos avec recalibration ou encore des retours d'expériences entre évaluateurs. Par ailleurs, il semble nécessaire que les actualisations de connaissances des formateurs intègrent les présentes recommandations.

Conclusion :

L'évaluation sommative en simulation représente des enjeux et des risques qui peuvent être optimisés par le développement de compétences spécifiques pour les formateurs en assurant l'encadrement au travers de formations initiale et continue. Nous proposons de distinguer trois types de formateurs en simulation :

- Des **formateurs-concepteurs** pour l'intégration de l'évaluation sommative dans le curriculum, pour le développement des scénarios et la mise en place dans le programme des outils d'évaluation sommative voire d'outils de recherche en évaluation sommative en simulation (titulaire Diplôme Universitaire de formateur en simulation ou d'un Diplôme Universitaire de de pédagogie de 3^e cycle associé à une formation courte en simulation).
- Des **formateurs-opérateurs** pour mettre en scène les scénarios. Les compétences nécessaires sont celles développées dans le référentiel de la SoFraSimS sur les compétences des formateurs.¹⁹¹
- Des **formateurs-évaluateurs** nécessitant l'apprentissage de compétences génériques liées à l'évaluation (principales erreurs, modalités d'utilisation des outils d'évaluation, etc...) et spécifiques à chaque formation (apprentissage et calibration de l'outil donné).
- valider le contenu final.

Chapitre 6 : En Pratique : mise en place de l'ES en Simulation

1. Les grandes étapes de la mise en place

La demande est croissante de voir la simulation prendre une place dans l'évaluation sommative, notamment dans des examens à haut enjeu (« high-stakes assessments»). Les ECOS (Examens Cliniques à Objectifs structurés) ont été décrits pour la première fois par Harden en 1975.¹⁹⁵ Ils ont été conçus pour améliorer la validité et la reproductibilité/fidélité des examens existants. Ce sont, historiquement, les premières épreuves comportant de la simulation, sous la forme de Patients Standardisés. Ils se présentent sous la forme d'une suite de stations cliniques. Initialement créés pour les étudiants en médecine, les ECOS ont été et sont utilisés actuellement dans de nombreux cursus différents, notamment infirmiers, ostéopathes, infirmiers

anesthésistes.^{31,35,196} On les a également adaptés au troisième cycle de la filière médicale, notamment en anesthésie-réanimation, en médecine d'urgence, en pédiatrie, mais également en médecine générale, en urologie et en gynécologie obstétrique.¹⁹⁷ Leur diffusion a longtemps reposé sur des pays pionniers et sur les pays Anglo-Saxons. De nombreuses études se sont attachées à débattre de leurs qualités psychométriques et de la meilleure manière d'en faire des épreuves de haute qualité. Des recommandations concernant la mise au point de ce type d'épreuve ont été émises par l'AMEE (Association for Medical Education in Europe) en 2013.¹⁹⁸ Croisées avec d'autres recommandations qui abordent l'organisation d'examen sur d'autres modalités de simulation immersive, notamment les sessions haute-fidélité, utilisant des mannequins complexes elles nous donnent un socle solide.^{80,144}

Les données scientifiques permettent aujourd'hui de donner une méthodologie pour mener un examen par simulation afin d'évaluer les compétences de professionnels de santé. Cet examen doit, notamment, satisfaire 5 axes définissant la validité de construit adoptés en 1999 par l'American Educational Research Association (AERA).¹⁹⁹⁻²⁰¹

1. Validité de contenu : le contenu recouvre ce pour quoi le curriculum est fait, les tâches sont réalistes et les bons domaines sont explorés
2. Le recueil de la performance : Le scoring est adéquat, les réponses sont bien recueillies, traitées et analysées
3. Le test a une bonne fidélité /reproductibilité
4. Relation à d'autres variables > Extrapolation : corrélation avec d'autres résultats/ avec la réalité, faible corrélation avec des variables sans rapport
5. Les conséquences de l'évaluation sont pertinentes, qu'il s'agisse de l'impact sur l'apprentissage, ou du caractère adapté de l'impact de l'échec, comme de l'acceptabilité sociale ou professionnelle de la réussite.

L'élaboration d'une épreuve valide au sens large du terme constitue donc un travail conséquent, encore décuplé si elle doit se tenir à grande échelle. Cela nécessite la collaboration de nombreux intervenants porteurs d'un panel de compétences : expertise clinique, expérience en simulation en tant que formateur, rédacteur, technicien ou acteur, rôles administratifs et logistiques. Cet effort consommateur de ressources devra s'inscrire d'emblée dans un processus continu d'évaluation et d'amélioration.

Élaborer un contenu valide nécessite de définir clairement la compétence professionnelle recherchée. L'alignement pédagogique avec le curriculum antérieur

est essentiel.

Les nécessités pédagogiques et les contraintes en ressources matérielles et humaines déterminent la technologie qui sera utilisée pour l'examen.²⁰² Sa qualité et sa faisabilité en dépendent.

La préparation, la gestion administrative et le bon déroulé d'une épreuve par simulation nécessite des connaissances, de l'expérience et un grand effort de planification. Différentes équipes vont devoir collaborer pour la bonne tenue du processus, notamment le développement d'une banque de sujets, l'entraînement des examinateurs et des patients standardisés.

Le processus qualité est continu durant tout le temps de préparation, de déroulé et d'évaluation de l'examen. Pour que les résultats soient fidèles, il faut s'assurer du bon entraînement des patients standardisés et des examinateurs. La validité sera aussi impactée si les questions ne sont pas réalistes, ou si elles s'éloignent du *blue print* ou tableau de spécification.⁴⁶

a. L'équipe

Structurer une coordination :

Il faut une équipe nombreuse pour mener à bien des examens par simulation. Le déploiement du nouveau mode d'évaluation va nécessiter un effort conséquent, un groupe doit être identifié pour le mener à bien en incluant des membres de l'équipe habituellement responsables des examens, renforcée si nécessaire. Une personne doit prendre en charge la supervision, idéalement quelqu'un ayant une expertise ou une expérience antérieure dans des examens par simulation ou par ECOS. Si ce n'est pas le cas, il ne devra pas sous-estimer l'ampleur de la tâche et le temps dont il devra disposer. Il faudra aller si possible prendre exemple dans d'autres structures, et se baser sur les références existantes dans la littérature. Le contrôle et l'évaluation continue de la qualité de cet examen doivent d'emblée être envisagés et organisés (cf infra).

Élargir l'équipe :

Dans le cadre d'un examen de grande ampleur notamment multisite, il est impératif de développer des équipes locales pour superviser les aspects pratiques comme le recrutement local d'examineurs et de patients standardisés, la préparation pratique du parcours d'examen. L'équipe devra s'assurer du bon déroulement de l'épreuve le jour J.

Si la structure est plus petite, l'équipe d'organisation de l'examen remplira ces tâches, éventuellement secondée par son personnel administratif.

Établir un calendrier pour les examens :

Idéalement uniquement guidé par le bon minutage dans le curriculum et sa place au sein des autres modalités d'évaluation pour le public donné, en pratique il sera souvent dépendant des impératifs organisationnels lourds qu'implique ce type d'épreuve, à commencer par la disponibilité des locaux, du matériel et des examinateurs.

b. Le tableau de spécification (« blue-print ») en lien avec le référentiel de compétence

Le processus de « Blueprinting » permet de s'assurer qu'un échantillon suffisamment étendu ET suffisamment pertinent des compétences pourra être évalué c'est-à-dire qu'on espère une bonne validité de contenu au regard du curriculum que l'on souhaite valider. Il faudra prendre en compte le contexte de l'examen, le stade du curriculum et la nécessité éventuelle de trianguler.²⁰³ L'évaluation pour certains domaines, c'est-à-dire d'utiliser plusieurs modes d'évaluation voire plusieurs modes de simulation (Haute-fidélité + procédural par exemple) pour la même compétence.²⁰⁴ L'examen par simulation s'inscrit comme UN outil d'évaluation certificative à combiner avec d'autres modalités de contrôle de connaissance.

Le tableau de spécification est une matrice dont le premier axe est constitué de compétences génériques à tester (recueil anamnèse, capacité de communication, examen clinique, gestion de l'organisation par exemple) et le deuxième axe comporte les contextes / terrain / situations cliniques particulières dans lesquelles ces compétences vont pouvoir s'exprimer.^{97,203,205}

Une structuration explicite et homogène du référentiel de compétences est primordiale pour pouvoir mener à bien ce travail. En France, 7 compétences génériques retenues pour le 2ème cycle des études médicales ont été fixées par l'arrêté du 8 avril 2013²⁰⁶ : Clinicien, Communicateur, Coopérateur, Acteur de santé publique, Réflexif, Scientifique et Responsable aux plans éthiques et déontologiques. Des sous compétences viennent affiner ce canevas. A charge ensuite aux différents collèges de spécialités d'exprimer les compétences spécifiques qu'ils estiment être indispensable (rang A) au 1er jour de l'interne, puis celles qui sont correspondent à un

approfondissement (rang B) et enfin celles qui ne seront pas demandées (rang C) car relevant du troisième cycle. Le référentiel pour l'examen fédéral suisse de médecine (« PROFILES ») a 3 entrées.²⁰⁷

Sept compétences génériques ou « rôles », issus des « CanMEDS » : expert médical, communicateur, collaborateur, leader, promoteur de la santé, érudit et professionnel ; 9 Activités professionnelles génériques crédibles (« Entrustable Practical Activities » ou EPAs) : anamnèse, examen clinique, démarche diagnostique, prescrire et interpréter des examens complémentaires, reconnaître et initier prise en charge des situation d'urgences, planifier une prise en charge et des prescriptions, assure la tenue du dossier, les transmissions orales et écrites, mettre en place la culture de sécurité et une démarche d'amélioration continue ; 265 situations cliniques « point de départ » (« situations as starting point » ou SSPs).

Le Blue-printing peut être fait, en local, par l'équipe responsable des examens, mais sur les examens certifiants à grand enjeu des techniques de type Delphi ou approchant doivent être utilisées pour qu'un panel d'experts se prononce sur les thèmes à inclure dans l'examen. Les questions seront ensuite développées à partir de cette trame.

c. Articulation avec le curriculum et les autres modalités d'évaluation

Le Blue-printing guide la conduite du cursus de l'évaluation formative (cf alignement pédagogique déjà défini dans ce document). La tenue d'évaluations sommatives intermédiaires a probablement un effet bénéfique sur la formation ; par un effet « catalytique » la rétro-information que fournit l'évaluation permet de mieux planifier les activités ultérieures d'apprentissage, et le fait d'être évalué renforce l'apprentissage.¹⁵⁵ Ainsi Murray montre-t-il que l'introduction d'une évaluation en 2 temps pour les résidents en anesthésie aux USA semble conduire à une amélioration de la performance finale.²⁰⁸ Isaak expose une évaluation 4 temps de 4 à 8 scénarios chacun dans une version des « milestones » d'anesthésie¹³⁶ : ce type de programme ambitieux peut se heurter à la faisabilité. En réalité selon le contexte la question est parfois modestement d'avoir pu exposer au moins une fois les étudiants au type de dispositif de simulation employé le jour de l'examen. Au-delà du respect de l'alignement pédagogique, la première simulation sur mannequin haute-fidélité nécessite l'adoption et l'intériorisation d'un certain nombre de codes pour agir d'une façon approchant la réalité.²⁹ Sur une station de 5 à 10 minutes l'étudiant peut perdre

plusieurs minutes à « rentrer » dans la situation, voire ignorer totalement la possibilité de contact verbal avec un mannequin, ce qui faussera l'évaluation de la compétence clinique réelle, notamment comportementale.

d. Longueur / nombre de stations

Les ECOS proposent des stations comprises entre 5 et 10 minutes. La fidélité / constance des résultats (« reliability ») et la validité de l'épreuve (sa capacité à être un reflet réel de l'ensemble des compétences que l'on souhaite évaluer sont influencées à la fois par la durée totale de l'épreuve et le nombre de stations.⁶⁰ La vérification d'une bonne adéquation entre la tâche et le temps imparti améliore la validité de l'épreuve.

Augmenter l'étendue du contenu testé en multipliant les stations améliore la fidélité. Une trop grande spécificité dans le contenu (trop étrié par rapport au corpus de compétences à valider) entraîne une faible reproductibilité.

Il faut donc un grand nombre de stations pour pouvoir généraliser / extrapoler de façon assez fiable le résultat obtenu à la compétence réelle des candidats.²⁰⁹ Le nombre de stations nécessaires pour une bonne fidélité peut être déterminé en utilisant des outils psychométriques : le Cronbach's alpha ou le Coefficient de Généralisabilité (G). Ceux-ci doivent être au minimum de 0,7 voire 0,8 pour un examen à haut enjeu.^{200,210,211} A l'opposé le PMETB (The Postgraduate Medical Education and Training Board) Britannique souhaite un coefficient > 0,9 pour les examens à haut enjeu. Certains outils permettent également d'éliminer dans un examen une station voire plus précisément des items inadéquats, afin d'optimiser la généralisabilité.

Il est important de comprendre qu'il n'y a pas un nombre fixe de stations valable pour tous les examens.

Il ne faut pas négliger la formation et la calibration des évaluateurs (cf infra et partie spécifique) mais c'est le nombre de stations qui aura le plus fort impact sur la généralisabilité. Il vaut mieux un évaluateur par station avec le maximum de station, la variété de cas et d'examineurs permettra d'augmenter la généralisabilité du test. Le nombre de scénarios nécessaires n'est pas fixe : il variera en fonction du nombre de compétences testées, de l'étendue du domaine d'intérêt dans lequel elles sont mises en œuvre, de la standardisation de la situation et de la qualité de la grille d'évaluation.

Certains domaines se prêtent parfois mieux à l'exploration d'un grand nombre de

compétences sur des scénarios courts (situations critiques dans les domaines de l'anesthésie-Réanimation, des urgences, de l'obstétrique ou de la pédiatrie). Il faut probablement moins de scénarios pour valider les compétences comportementales transversales du fait de leur ubiquité (elles s'expriment dans presque tous les scénarios).²⁹

Pour évaluer les compétences cliniques du 2ème cycle médical, au minimum une dizaine de stations sur plusieurs heures sont nécessaires pour espérer un Cronbach 0,85 à 0,90.^{23,195} Pour les résidents en anesthésie Boulet propose le chiffre de 8 à 12 scénarios.²⁹ Notons que le même ordre de grandeur est retrouvé lorsqu'il s'agit d'évaluer du raisonnement clinique sur cas simulé sur ordinateur. Epstein évoque 14 à 18 scénarios pour sécuriser l'objectif de fidélité..¹⁸ L'examen fédéral suisse comporte 12 stations de 13 minutes. Il faut toutefois rester humble et réaliste : Fehr décrit un examen déjà ambitieux en 10 scénarios sur de l'anesthésie pédiatrique qui ne permet pas d'obtenir une généralisabilité dépassant 0,54 à 0,57. ²¹² Un nombre plus faible encore est parfois défendu, par exemple un examen en 5 scénarios sur des résidents d'anesthésie israéliens³⁸ ou de la gynécologie-obstétrique.²¹³

L'essentiel sera d'argumenter explicitement une démarche de compromis pour maximiser la qualité et préserver la faisabilité de l'examen. Obtenir le nombre de stations idéal au détriment de leur qualité n'est probablement pas le meilleur compromis. Définir un examen trop ambitieux sans attribuer les moyens nécessaires obère la pérennité de la démarche.

e. Développer une banque de scénarios

Chaque scénario ou station proposée doit être soumise à relecture par les pairs et testé. Idéalement ce test (par exemple sur une cohorte en examen blanc, ou comme une station blanche dans un examen certifiant) devrait obtenir les données psychométriques de la station notamment sa capacité de discrimination (bonne/mauvaise performance). Une banque sécurisée de scénarios robustes ayant fait la preuve de leur qualité est un gage puissant de la reproductibilité et de la validité de l'épreuve. Voici une proposition de processus étapes par étapes pour l'élaboration de nouvelles stations ou l'adaptation / mise à jour d'anciennes.

Choisir les thèmes abordés :

Cf tableau de spécification/blue-print, en vérifiant l'alignement pédagogique et/ou réglementaire²¹⁴ la faisabilité d'une station sur ce thème en tenant compte des contraintes de temps notamment. Il est possible d'indexer les scénarios d'une banque en fonction des domaines du tableau de spécification qu'il aborde, ce qui peut faciliter la composition d'un jeu de scénarios pour une épreuve donnée. Cet assemblage de scénarios est aussi important que leur élaboration individuelle. Il pourrait être aidé par l'usage d'outil informatique dédié. Ainsi l'USMLE (United States Medical Licensing Examination) utilise un algorithme basé sur différents critères (entre autres l'ancienneté des scénarios dans la base) afin de générer automatiquement des épreuves inédites remplissant les critères de contenu.⁸⁰

Choisir les rédacteurs pour les scénarios :

Idéalement experts de la thématique, familiers de l'exercice, et comprenant les spécificités et contraintes de l'examen par simulation. Si ce n'est pas le cas, des moments doivent être dégagés pour briefer les rédacteurs. Des consignes écrites extensives doivent servir de référence pour l'ensemble des rédacteurs.

Choisir le type de station :

Le choix du type de simulation (procédurale, immersive avec PS, hybride, mannequin hifi, +/- professionnel collaborateur) dépend des besoins de la trame et des moyens disponibles. Le coordonnateur de l'examen et / ou le coordinateur de la rédaction des scénarios doivent orienter les rédacteurs en fonction.

Écrire le scénario proprement dit : en utilisant un canevas de scénario commun pour homogénéiser le type d'écriture / de questions et faciliter la relecture comparative des scénarios afin de gommer les différences de difficulté et d'aider à la construction d'une épreuve correspondant à la trame.¹³⁰

Choisir le type d'échelle de notation :

Afin d'obtenir une note qui soit, au mieux, le reflet de la performance,^{200,210,211} il est nécessaire d'utiliser une ou des échelles spécifiques, si possible validée, sinon mise au point en s'assurant qu'elle permette bien d'identifier et de discriminer la qualité des compétences attendues.

Le temps de relecture par les pairs est important. Cette relecture s'effectue entre les

différents rédacteurs, avec l'auteur présent afin de faciliter les échanges et des clarifications rapides. Des personnes expérimentées en simulation doivent être impliquées parmi ou auprès de ces rédacteurs afin de maintenir une cohérence étroite avec l'outil.²³ Au-delà de la vérification de la justesse clinique du scénario ainsi que de son caractère approprié aux contraintes d'un examen par simulation en temps limité, ce temps de révision doit permettre de soulever des problèmes éventuels de validité. Un questionnaire simple peut améliorer l'efficacité de ce temps.¹⁹⁸

Session test :

La / les session(s) test permet(tent) de débusquer des problèmes de tâches non réalisables dans la modalité de simulation choisie, ou non adaptées au temps imparti, et de corriger le scénario en fonction des éléments de la grille de notation. Des rectifications peuvent être faites pour résoudre le problème, impliquant une nouvelle session test. Pour obtenir des données psychométriques de qualité sur la nouvelle station testée, on peut utiliser des examens blancs ou des partiels à enjeu modéré. Une station peut néanmoins être testée lors d'un examen à haut enjeu, à condition que les étudiants soient informés de la présence d'une station test, et qu'elle n'aura pas d'impact sur la note. Elle ne sera cependant pas identifiée individuellement pour ne pas induire de modification de comportement nuisant à la qualité du test.

Analyse psychométrique :

Si un jeu complet de nouveaux scénarios est introduit pour un examen blanc, ces tests permettront de déterminer la reproductibilité globale de l'examen. La G theory indiquera le nombre de stations nécessaires par application d'une D ou décision study sur les données. On pourra aussi se servir de l'item « Response theory »²⁰⁰ pour identifier les sources de variabilité ou d'erreur.

Établir/choisir une grille d'évaluation

Check list

Idéalement préalablement validées, il faut préférer des checks list type Likert à 5-7 degrés attribués selon des critères clairs pour l'examineur. Les questions binaires sont associées à une mauvaise discrimination des performances. La force des checklist binaires était son attribut d'objectivité, censée sous-tendre une meilleure fiabilité inter-observateur. Mais cet argument semble petit à petit affaibli, car des publications

tendent à prouver que cette « objectivité » ne se traduit pas forcément en une bonne reproductibilité²¹⁵ en particulier si les examinateurs sont des experts du domaine.²¹⁶

Échelles globales

- Force quand la qualité de ce qui est fait est aussi importante que de savoir si cela est fait ou non.
- Utile quand le candidat ne suit pas les étapes anticipées pour parvenir au même résultat attendu.

Un exemple : les échelles portant sur l'empathie, l'organisation des connaissances, les compétences techniques.^{216,217} Les échelles globales sont en usage croissant en ECOS et en simulation avec un faisceau croissant de preuves pour une meilleure reproductibilité inter station, une meilleure validité de construit, une meilleure « concurrent validity » comparée aux checklists.^{200,210,218}

Définir le score de réussite/le seuil minimum

Dans un examen normé/un concours, le score de réussite est déterminé par les scores relatifs des candidats. En cas de groupe plus faible cela induit de valider des candidats plus faibles. C'est inacceptable pour des examens certifiants. La description précise des méthodes utilisables (Borderline Group et Contrasting groups methods...) dépasse le cadre de ce travail.¹⁷⁹ Sur un examen de cette étendue et cette lourdeur d'organisation, l'utilité pourrait être maximisée en exploitant l'aspect certifiant mais également en l'utilisant à des fins de classement et d'orientation. Plusieurs tableaux de spécifications doivent alors sous-tendre l'épreuve, regroupant ou pondérant les compétences de façon différente selon les orientations possibles. C'est une piste séduisante qui a été envisagée en France pour le futur examen de 6ème année des études médicales. L'obtention d'une validité satisfaisante pour tous ces objectifs augmente cependant encore le niveau d'exigence dans la préparation de l'épreuve.

f. Entraîner des évaluateurs

La fidélité de la notation dépend de la constance de l'examineur mais aussi de son niveau d'expertise sur la thématique du scénario.²¹⁹ Souvent l'évaluation est faite par quelqu'un de la même filière professionnelle. Mais pour les filières médicales la rareté et le coût induit a conduit à expérimenter l'évaluation par des non-médecins. Prediger

en souligne la difficulté en comparant l'évaluation d'étudiants en médecine par un médecin, une infirmière ou un autre étudiant,²²⁰ parfois par le patient standardisé ²²¹ notamment sur les aspects relationnels ; la bonne corrélation relevée ^{222,223} sur des notations en checklist semble moins forte quand il s'agit d'évaluation globale.²²⁴ Quel que soit le choix fait, l'entraînement des évaluateurs est fondamental. Il va diminuer les variations inter-observateurs^{187,225} améliorer la constance dans le comportement ce qui améliore la fidélité de l'épreuve. ²²⁶

Atelier d'entraînement des examinateurs :

Si possible en petit groupe bien en amont des épreuves en adaptant la formation au niveau et à l'expérience des futurs examinateurs.

Des objectifs explicites doivent être énoncés, l'AMEE propose, par exemple, de :¹⁹⁸

- comprendre le champ d'application et les principes de l'évaluation certificative par simulation
- maintenir un comportement professionnel durant l'examen
- Comprendre et utiliser la grille de notation pour maintenir la standardisation
- Savoir donner un Compte Rendu écrit de la performance si nécessaire sur les évaluations certificatives
- Être capable de donner un feedback oral aux étudiants sur des sessions formatives/ des examens blancs
- Assurer la confidentialité des grilles d'évaluation nominatives remplies
- Assurer la sécurité des candidats et des Patients Standardisés (comportement inapproprié ou dangereux)

En pratique l'entraînement / la calibration peut se faire en utilisant des vidéos de plusieurs performances différentes pour un même scénario, et en comparant l'évaluation à celle d'un panel de référence. Weigner et al. proposent 2 jours de formation pour les évaluateurs, et leur calibration par 6 vidéos pour chacun des 4 scénarios employés.²²⁷ Les examinateurs doivent bénéficier de mises à jour et de maintien des compétences d'examineurs par la mise en pratique d'une analyse en fin d'épreuve, et, si besoin (notamment si évolution dans les modalités ou les objectifs), des sessions de maintien de compétence en ligne ou, de nouveau, en petits groupes. Les examinateurs ne sont pas nécessairement des enseignants universitaires, il faut favoriser ceux qui ont déjà un investissement dans la pédagogie, et si possible volontaires pour améliorer la rigueur de leur prestation.²²⁸ La plupart du temps, ils ne

sont pas payés spécifiquement pour cette tâche, inclus dans leurs obligations générales, mais si ce sont de nouvelles missions cela implique d'en tenir compte dans les effectifs concernés. L'emploi d'effectifs hospitaliers de façon accrue pour cette mission universitaire peut être la clé de la faisabilité.

g. Développer un pool de patients standardisés ²²⁹

Si un vrai patient est retenu avec un réel problème de santé, il peut être recruté par son médecin s'il a une pathologie commune et stable.

D'autres recrutements de volontaires sains, d'acteurs... dont il faudra s'assurer qu'ils ne présentent pas de signes cliniques particuliers.

Il faut envisager la possibilité de rémunérations pour les patients standardisés : cela va du défraiement aux prestations à tarif standard d'acteurs pros. En France, ce type de budget est inhabituel pour un examen. Néanmoins, cela peut faire partie de la motivation de la présence de PS.

La sélection et la gestion du pool nécessite une coordination et un responsable des PS et se font plus aisément dans de grosses structures. Il faut gérer le « stock » de PS en les maintenant si possible polyvalents, afin de parer aux défis logistiques, ou aux absences de dernière minute.

Les PS doivent être entraînés mais aussi avoir intégré les impératifs de standardisation et de reproductibilité de la prestation.²²⁹ On peut parfois briefer le PS le matin même, mais la règle est une préparation en amont pour assurer la qualité et la constance du rôle.

Il est intéressant d'évaluer les PS avant de les utiliser sur des épreuves à haut enjeu, par exemple par l'évaluation de vidéos par un groupe indépendant²³⁰ ou un entraînement lors des examens blancs avec débriefing de leur prestation.²³¹

Il faudra standardiser avec la même rigueur (script, entraînement...) tous les complices, notamment le technicien responsable de la voix du mannequin, les rôles de parents en pédiatrie, ou de professionnel si travail en équipe / en interprofessionnel recherché (IDE par exemple pour les étudiants en médecine sur les ECOS Occitanie 2019 et 2020).

h. Déroulement de l'épreuve proprement dite

Tâche administrative :

Ces tâches administratives incluent toute la logistique :

- convocation, informations... commune à tout examen mais aussi le droit à l'image (en cas d'utilisation de la vidéo)
- appel des patients standardisés
- disponibilité des équipes de simulation : responsable de scénario technicien + acteur(s)
- réserve en cas d'absence de personnel formateur ou de patients simulés.
- sélection des salles, en prévoyant une « quarantaine » pour éviter que plusieurs vagues d'étudiants ne se croisent et ne risquent d'échanger à propos des sujets, autant de contraintes à anticiper pour réserver un lieu adéquat.
- sélection des examinateurs (entraînés, expert vs non expert...)

Les besoins en équipements de chaque station doivent avoir été listés bien en amont, commandés et vérifiés (par exemple le modèle et le nombre de sondes urinaires, etc...) avec des piles de rechange et du matériel de réserve si c'est un matériel susceptible de panne ou de casse.

Si les étudiants doivent se munir de leur propre matériel (stéthoscope, blouse, badge), ils doivent avoir été prévenus auparavant.

Pour le matériel de simulation haute-fidélité il peut être difficile de prévoir des dispositifs de remplacement, il faut au minimum disposer d'un technicien capable de résoudre les problèmes les plus courants sur le mannequin et sur l'environnement technique (audio-vidéo, informatique) éventuel.

Le jour J :

On peut envisager un circuit par rotation ou par flux. Prévoir dispositif de synchronisation minuté (chronomètre, applications dédiées, simple powerpoint minuté, et mode dégradé (cloche ou corne de brume audible dans l'ensemble du circuit)

Il est possible de prévoir une station « de pause » pour les étudiants au milieu du parcours, ce qui implique une salle supplémentaire suffisamment isolée.

Il est probablement nécessaire d'envisager également des pauses dans le flux de candidats afin de ménager les équipes responsables de stations.²³²

Briefings du jour J :

Importance de prévoir précisément des briefings à destination des examinateurs, des responsables et des acteurs de station et des candidats. Le briefing traditionnellement fait dans les séances formatives doit être ici standardisé et anticipé en amont : dans les conditions de stress du jour de l'examen, il faut limiter la charge cognitive liée à ces consignes à quelques points essentiels. Outre une familiarisation antérieure au dispositif par des expositions formatives, il est pertinent de créer des vidéos disponibles en amont de l'examen abordant les caractéristiques et l'usage que l'on peut faire du ou des mannequins, les instructions et limites dans le cadre des patients standardisés, la possibilité de la coopération avec un autre professionnel durant le scénario ainsi que les règles concernant l'environnement mis à disposition.

Problèmes à anticiper :

Il est nécessaire de limiter la variabilité affectant la standardisation

- ⇒ stabilité de la performance du patient standardisé, du complice, ou du déroulé d'un scénario sur mannequin HF
- ⇒ de façon préventive, la programmation des acteurs et du mannequin doit être suffisamment rigoureuse
- ⇒ l'examineur est encouragé à réagir en faisant un retour immédiat entre deux candidats à l'intervenant concerné
- ⇒ pour optimiser la standardisation et garder l'attention des examinateurs disponibles pour l'évaluation, on peut proposer sur une station de simulation haute-fidélité la distinction entre un responsable du scénario, et le/les évaluateurs.

Panne/ casse d'équipement :

- ⇒ supports papier de secours pour tout support informatique de notation.
- ⇒ équipement de remplacement prêt à l'emploi et vérifié
- ⇒ gestion concrète lors de la survenue d'un incident : il faut tenir compte du temps éventuellement perdu par l'étudiant et lui offrir la possibilité de le faire revenir sur la station en fin d'examen afin d'éviter de devoir supprimer la station pour l'ensemble des candidats. Si le problème est résolu rapidement, il peut aussi être envisageable de décaler le train d'étudiant restant en introduisant un temps d'attente de la longueur d'une station (intérêt d'avoir alors une pièce supplémentaire).

Comportements indésirables des candidats : sous l'effet du stress notamment il est possible d'observer des candidats qui se perdent dans le circuit, ou qui n'entendent pas ou ne tiennent pas compte des consignes notamment de fin de station.

Il est donc nécessaire d'informer les participants et les examinateurs et de disposer de personnels uniquement dédiés à l'orientation et à l'accompagnement éventuellement forcé des étudiants.

Disparition/dégradation en cours d'examen des consignes aux étudiants (emport accidentel)

Pour éviter ces aléas, il est intéressant de :

- ⇒ scotcher /attacher solidement les consignes qui auront été plastifiées, de prévoir des copies
- ⇒ Pour éviter la disparition ou la perte des grilles d'évaluation ou des consignes aux examinateurs / confusion dans l'identité, il est nécessaire de :
- ⇒ donner des consignes claires aux équipes, avec des consignes fixées et d'en prévoir des copies.
- ⇒ ne délivrer que les grilles correspondant au groupe effectivement en train de passer, avec pré- identification par candidat afin d'éviter des erreurs ou des oublis dans l'identité du candidat.
- ⇒ demander systématique l'identification du candidat à chaque station

i. Après l'examen

Recueil des grilles d'évaluation / vérification d'incohérences/ oublis dans la cotation
Gestion des recours

- ⇒ part du processus qualité / problème de coûts pour le candidat et-ou l'institution (ex. Suisse)

j. Processus qualité

Les processus qualité concernent de nombreuses étapes : les blueprints et scénarios revus par les pairs, l'entraînement des examinateurs, la standardisation des stations, la présence d'examineurs extérieurs observateurs, l'évaluation via les retours au sein de l'équipe d'organisation et par les candidats eux-mêmes et l'usage des outils psychométriques.

Un focus sur les outils psychométriques (Cronbach alpha) permet de déterminer la

fidélité globale de l'examen et éventuellement d'identifier une station particulièrement source d'erreur à éliminer tout en garantissant la pertinence des stations restantes. La G theory permet des manipulations plus poussées des données permettant de s'intéresser à une station, à des items, ou à des interactions selon le candidat / l'examineur. L'item « response theory » permet un certain nombre d'analyses similaires mais ne permet pas de conclure si le nombre de stations a été altéré.

La sécurité des données de l'examen doit faire l'objet d'une attention particulière, qu'il s'agisse de la sécurisation de la banque de sujet ou de la sécurisation et de l'anonymisation des données individuelles des candidats, y compris documents vidéo, résultats, et éléments du curriculum. Les engagements à cet égard doivent être explicites et réels.^{80,233,234}

k. Conclusion sur la mise en place de l'Examen

La définition du processus « qualité » de l'examen l'est également, l'analyse psychométrique doit s'intégrer dans la démarche.

L'ACGME (Accreditation Council for Graduate Medical Education), responsable de nombreux programmes d'évaluation sommative reprend dans son guide sur la mise en place des « Milestones » la formule de Norcini et Van der Vleuten.¹⁵⁵

Utilité évaluation sommative = Validité X Fiabilité X Impact Formatif X Acceptabilité X Faisabilité.

Cette formule doit permettre de situer et de questionner régulièrement la place « utile » de chaque modalité de simulation dans l'évaluation : malgré une pertinence indiscutable, la faisabilité de l'examen idéal par simulation reste suspendue à l'obtention de moyens conséquents.

Il faudra donc argumenter explicitement les compromis afin de fixer aux équipes des objectifs réalistes et soutenables dans la durée.

Tableau 10 : La mise en place de l'ES en simulation.

Item	But	Modalités
Equipe	Définir le personnel formateur	Structurer une coordination Dimensionner l'équipe : compétences, temps disponible, stabilité (projet sur plusieurs mois / années)
Tableau de spécification	Créer la matrice des compétences à évaluer	Panels d'experts. Dicte le nombre et les types d'épreuves nécessaires. Doit être connu des étudiants
Curriculum	Articuler l'ES avec le curriculum	Alignement pédagogique : le sommatif guide la conduite du cursus formatif. Pas d'ES sans exposition préalable à la simulation Des ES intermédiaires pourraient être utiles ⁸⁵
Examen	Définir les modalités d'ES par la simulation	Longueur / nombre de stations-scénarios ^{86 87} Plus un examen est complet (fidélité) et moins il est aisé à mettre en place (faisabilité)
Scénarios	Développer une banque de scénarios et de grilles d'évaluation ⁸⁸	Choisir les rédacteurs pour les scénarios. Écrire les scénarios Relecture par les pairs et test. Établir/choisir une grille d'évaluation (Checklist ou échelle globale) Définir le score de réussite/la barre Les scénarios de la banque doivent permettre de générer des épreuves couvrant les compétences du tableau de spécification.
Entraîner des évaluateurs	Limitier les variations pour une performance donnée	Choix des évaluateurs Atelier d'entraînement des examinateurs
Patients Standardisés	Développer un pool de patients standardisés	Sélection/recrutement/entraînement/standardisation ⁸⁹
Jour J	Déroulement de l'épreuve proprement dite	Logistique (dates, salles, patients simulés...), droit à l'image/RGPD... Circuit des évalués/pauses Matériels nécessaires sur place/à amener par les étudiants (ex. stéthoscope) Briefings du jour J Problèmes à anticiper (stabilité de la standardisation, panne ou de casse de matériel, supports papier de secours, personnel dédié pour les « stressés », etc..)
Après l'examen immédiat	Finaliser l'épreuve	Recueil des grilles d'évaluation ⇔ incohérences/ oublis de cotation ? Gestion des recours / plaintes des évalués
Processus qualité	Préparer les épreuves à venir	Modifications du contenu de certains scénarios ? Elimination d'un scénario (temps trop long, scénario source d'erreur avec manque de cohérence des évaluations, etc.) Modifications de formations de PS Modifications de formations des évaluateurs

2. Exemples d'Évaluation Sommative en simulation en santé

a. Etudiants 1^{er} / 2^{ème} cycle des différentes filières

Etudiants en Médecine

a) Pour la sélection à l'entrée des études médicales :

En Israël, l'intégration de la simulation dans les examens d'entrée modifie environ 20% des admissions par rapport à des tests cognitifs seuls.^{38,235}

L'examen est basé sur des critères de communication interpersonnelle, de gestion du stress, du sens de l'initiative et des responsabilités, de la connaissance de soi au moyen d'une évaluation multimodale :

- Pré-test identifiant des compétences / traits de personnalités préexistant à une orientation
- Simulations comportementales (3 situations de 6 à 9 minutes), des tâches de groupe,
- Entretien individuel.

b) A l'issue du cursus médical initial :

Les 5 facultés suisses ont mis au point un examen fédéral comprenant 12 stations de 15 minutes de simulation/ECOS et un volet d'évaluation classique par 300 QCM. Depuis l'élaboration du référentiel PROFILES (cf supra) jusqu'au processus d'évaluation de l'examen, les étapes visant à une bonne validité ont toutes été respectées. Le jour J, 1000 étudiants sont répartis sur 3 plateaux simultanés, animés par 70 Patients Simulés, et 60 examinateurs formés. La notation de chaque station est assurée par 3 généralistes issus des 5 facultés Suisses. Les étudiants bénéficient d'un defusing émotionnel en sortie d'examen. Le coût était toutefois estimé en 2018 à 4500 euros par candidat. (nb: aucun recours n'avait été enregistré).^{207,236}

Etudiants en Soins infirmiers

b) Rarell et al. (Barcelone) ont réalisé une évaluation sommative de 120 étudiants(es) sur 6 stations d'ECOS en (répartis en 12 sessions avec légère modification des scénarios pour éviter les fuites). Les 2 observateurs se calibraient sur les deux premiers candidats puis cotaient de façon séparée pour les autres candidats. Les ECOS sont vus comme un diagnostic individuel par identification des gaps de formation, grâce à l'évaluation sommative elle-même mais aussi grâce à l'espace créé pour un échange et des prises

de décision en équipe pédagogique complète (Tuteurs de stage et Equipe académique).²³⁷

c) Oermann et Jeffries pour la National League for Nursing Project ont publié en 2015 le détail d'une démarche de mise en œuvre d'une évaluation sommative de fin d'étude dans un réseau de 10 établissements. Un processus rigoureux sur plusieurs années avec l'élaboration de scénarios, l'usage d'une grille d'évaluation éprouvée (The Creighton Competency Evaluation (CCEI) tool, 23 items, 4 domaines), et le test des stations dans les futurs centres examinateurs donnant lieu à de nombreuses corrections et amélioration (scénarios, prebriefing des étudiants), la formation d'évaluateurs recrutés parmi des experts en simulation et formés. La corrélation intra classe inter évaluateurs (évaluations sur vidéos) malgré plusieurs temps de concertation n'a pas dépassé les 70% par scénario, même régulièrement inférieure à 50%. La principale explication proposée par les auteurs est un problème de blueprint: il n'y avait pas un consensus complet sur les attendus, problème que l'équipe allait donc s'attacher à résoudre.¹³⁹

Etudiants Infirmiers Anesthésistes

a) Wunder et al. (Miami, Floride) ont décrit la mise en place d'une évaluation sommative semestrielle. 35 étudiants passent sur 5 stations en 2 jours avec des évaluations sur le contrôle des équipements d'anesthésie, l'évaluation préanesthésique, le transfert et positionnement des patients, l'induction générale en anesthésie, y compris l'induction à séquence rapide. S'ils concluent à la faisabilité, ils mettent en évidence des difficultés liées à la haute technicité. Nécessité de prévoir une équipe de support technique très réactive, des solutions et modes dégradés : guidage du candidat à la seule voix, interruption, patient standardisé, mannequin de remplacement, temps de la pause dans le flux.¹⁵⁶

b) L'examen final du curriculum Suisse comprend en routine une partie pratique avec une situation réelle et une simulation pour reconnaître et traiter les situations d'urgence. La validation se fait par compétence. Chaque compétence est divisée en 4 sous compétences, elles-mêmes décomposées en 3 actions à effectuer. 2 actions effectuées sur trois valident une sous compétence et 3 sous compétences validées sur 4 valident une compétence.

Exemple : intubation sélective lors d'une laparoscopie, la sous compétence « évalue l'efficacité des mesures destinées à sauver la vie et les complète au besoin » sera

divisée en :

- Ausculte à nouveau le patient après la manœuvre
- Observe la normalisation des paramètres respiratoires (pressions intrathoraciques et saturation)
- Réalise un contrôle systématique ABCDE

Étudiants en Maïeutique

La Haute École de Santé de Genève (HEdS) propose, en 1ère et 2è année, de la simulation qui compte pour 50% de la note annuelle. Cette même Haute École de Santé de Genève (HEdS) propose des ECOS en fin de dernière année. Les étudiant(e)s ont 4 situations différentes et sont évalué(e)s par un barème fédéral.

Sur impulsion nationale, un certain nombre de départements de maïeutique français comme celui de Nîmes ont également intégré évaluations formatives et sommatives par simulation à toutes les étapes de la formation.

b. Internes dans différentes disciplines médicales

Résidents de Médecine générale

4 Universités Belges ont mis en place des ECOS les résidents de Médecine générale Année après année ils constatent des scores de plus en plus hauts (>75/100). Malgré le maintien d'une bonne validité d'après les tests psychométriques, il a été suspecté un défaut d'adaptation au niveau des étudiants (en hausse) donc une baisse de la validité de contenu. Pour augmenter la complexité au moins 2 demandes par patient dans chaque scénario. 32 scénarios ont été élaborés, 8 ont été éliminés après révision. La promotion de 340 étudiants a joué 8 stations de 12 minutes. Le score moyen est revenu à 70/100, avec 5% d'échec (Seuil fixé à 55/100) tout en maintenant de bonnes données psychométriques : l'ajout de complexité semble avoir produit l'effet recherché.²³⁸

Résidents Médecine d'Urgence

Au Queens, Canada, Dagnone l'évaluation des compétences de résidents de médecine d'Urgence par une grille non spécifique de scénario, le « Queen's Simulation Assessment Tool" (QSAT): Items sur l'évaluation initiale, les actions diagnostiques, les actions thérapeutiques, la communication, la performance globale. 103 étudiants inclus ont été évalués dans 5 centres, sur des scénarios d'arrêt

cardiorespiratoire, avec 3 stations (par limite de moyens). L'évaluation est faite par 3 évaluateurs par centre (de plus de 5 ans d'expérience) sur vidéo. Ces formateurs ont reçu 3 h d'entraînement sur des vidéos type du scénario. Les étudiants bénéficient d'un débriefing formatif. Bonne qualité psychométrique sur la variabilité inter observateur et la capacité de discrimination, mais généralisabilité moins favorable (seulement 3 scénarios) : la piste d'amélioration retenue est de réorganiser les moyens pour réaliser un examen monocentrique (en gardant l'équipe pédagogique multicentrique) de 6 stations avec 1 seul examinateur par station.²³⁹

Tableau 11 : Exemple de grille d'évaluation sommative utilisée en simulation (QSAT)
(cas d'une station d'hémorragie méningée): (d'après l'article de Dagnone239, Traduction française non validée)

Nom et prénom de la personne évaluée :				
Date de l'évaluation :				
Nom de l'examineur :				
Evaluation initiale				
Constantes (dont dextro)/monitorage/voie d'abord veineuse/Evaluation neurologique/ Evaluation des voies aériennes supérieures/vérification du rythme				
1 <u>Inférieur</u> Performance différée ou incomplète sur tous les items	2 <u>Novice</u> Performance différée ou incomplète sur de nombreux items	3 <u>Compétent</u> Performance différée ou incomplète sur de quelques items	4 <u>Avancé</u> Performance adaptée sur la plupart des items	5 <u>Supérieur</u> Efficient et rapide sur tous les items
Actions Diagnostiques				
Recueil des antécédents/Examen clinique/ECG/demande d'examens complémentaires/Demande de scanner cérébral.				
1 <u>Inférieur</u> Performance différée ou incomplète sur tous les items	2 <u>Novice</u> Performance différée ou incomplète sur de nombreux items	3 <u>Compétent</u> Performance différée ou incomplète sur de quelques items	4 <u>Avancé</u> Performance adaptée sur la plupart des items	5 <u>Supérieur</u> Efficient et rapide sur tous les items
Actions thérapeutiques				
Intubation en séquence rapide (ISR) avec neuroprotection/proclive 30°/monitorage et contrôle de la tension/Drogues utilisées (drogues de reversion des anticoagulants/ Drogues pour baisser l'Hypertension Intra cérébrale, drogues pour l'isr...)/ paramètres du ventilateur avec hyperventilation				
1 <u>Inférieur</u> Performance différée ou incomplète sur tous les items	2 <u>Novice</u> Performance différée ou incomplète sur de nombreux items	3 <u>Compétent</u> Performance différée ou incomplète sur de quelques items	4 <u>Avancé</u> Performance adaptée sur la plupart des items	5 <u>Supérieur</u> Efficient et rapide sur tous les items
Communication				
Se présente et explique la situation clinique/donne des ordres clairs et concis/leadership efficace en situation de crise/Prioriser les tâches et anticiper/Appel approprié au neurochirurgien/Demande la présence de la famille				
1 <u>Inférieur</u> Performance différée ou incomplète sur tous les items	2 <u>Novice</u> Performance différée ou incomplète sur de nombreux items	3 <u>Compétent</u> Performance différée ou incomplète sur de quelques items	4 <u>Avancé</u> Performance adaptée sur la plupart des items	5 <u>Supérieur</u> Efficient et rapide sur tous les items
Compétence globale				
1 <u>Inférieur</u> Toutes les items nécessitent des améliorations significatives	2 <u>Novice</u> La plupart des items nécessitent des améliorations modérées ou significatives	3 <u>Compétent</u> Quelques items nécessitent des améliorations modérées.	4 <u>Avancé</u> Compétent Quelques items nécessitent des améliorations mineure	5 <u>Supérieur</u> Peu si ce n'est aucun item ne nécessite d'amélioration mineure.

Résidents d'Anesthésie

a) Blum et al. (USA) ont décrit en 2018 le test de la validité d'un dispositif d'évaluation par simulation de résidents en anesthésie. Il visait notamment à confirmer cette validité en multicentrique (3 centres accrédités par l'ACGME à la Harvard medical School), et à renforcer les preuves de son caractère discriminant en incluant 2 populations (1ère et 3ème années). 7 scénarios de 15 minutes étaient joués par chaque participant on enregistrait également après chaque scénario la réponse de l'apprenant à 3 questions stéréotypées permettant de mieux comprendre la performance. L'évaluation se faisait sur vidéo par 20 évaluateurs différents. 10% des vidéos étaient évaluées 2 fois pour calculer la fiabilité inter observateurs. L'outil "Harvard Assessment of Anesthesia Resident Performance" (HARP) préalablement validé était utilisé. Il évalue 5 compétences clé pour les résidents d'anesthésie : (1) Synthétise l'information pour formuler un plan anesthésique clair ; (2) Met en œuvre un plan adapté aux modifications du patient; (3) Démonstre des compétences interpersonnelles et de communication efficaces avec les patients et le personnel; (4) Identifie les moyens d'améliorer les performances; et (5) reconnaît ses propres limites. Les 3ème années avaient une performance globale meilleure que les 1ère années. Les tests psychométriques montrent un ICC de 0,76 inter-observateurs et un score de généralisabilité à 0,70 : malgré la robustesse de la démarche, les auteurs considèrent cela comme encore insuffisant pour des examens à haut enjeu.¹³⁷

b) Le Groupe de travail sur le Programme national de simulation en anesthésiologie au Canada (CanNASc) compte 368 experts participants (sur 958 sollicités) Il a permis d'élaborer de façon nationale le dispositif d'évaluation : le blue print, les scénarios, leurs grilles d'évaluation spécifique, le choix d'outils d'évaluation des compétences génériques notamment non techniques (Managing Emergencies in Pediatric Anesthesia Global Rating Scale (MEPA GRS)). Les formateurs disposent d'un guide pour apprendre à évaluer et de deux vidéos d'entraînement pour se calibrer. Les auteurs concluent à la faisabilité de cette démarche d'homogénéisation nationale, et sur l'intérêt suscité par le processus pour la formation continue et la maintenance de certification.¹⁴¹

c) En Caroline du Nord, Isaak a comparé l'évaluation en direct ou par vidéo à distance de 41 résidents d'années différentes. Le dispositif comprenait l'élaboration

de checklists d'évaluation critériées en 5 niveaux, dérivées des compétences Milestones. L'épreuve comprenait 3 scénarios dont un basé sur la compétence "professionnalisme" dans une situation d'entretien avec un patient standardisé. C'est sur ce dernier que la corrélation intra classe entre évaluateurs (1 en direct et 2 en vidéo) est la moins bonne (0,5 vs 0,65 et 0,85) avec la part de variance attribuée au résident la plus faible. La réponse à la question vidéo VS direct semble donc dépendre du type de scénario.²⁴⁰

d) Everett et al. ont mené une étude sur la validité d'une épreuve d'anesthésie pédiatrique. Cette étude multicentrique a inclus 391 sessions évaluant des apprenants de différents niveaux y compris des seniors. Les évaluateurs étaient formés en 3 heures de séances de cotation suivies de discussions collégiales sur les vidéos sur les critères utilisés (checklist ou GRS). 9 centres de simulation ont été inclus dans 2 pays, 6 scénarii, jusqu'à 5 observateurs par vidéo. Après formation, les résultats ont montré que les outils étaient valides, reproductibles (ICC inter évaluateurs 0,85-0,94) et discriminants, qu'il y avait une bonne corrélation (0,74) entre l'évaluation par l'échelle globale (Managing Emergencies in Pediatric Anesthesia (MEPA)) et par la checklist de scénario. L'étude identifie certains écarts de performance récurrents y compris parmi les seniors, identifiant de nouveaux besoins de formation.¹⁶²

Résidents en pédiatrie

a) Frey-Vogel a intégré 6 scénarios formatifs appariés à 6 scénarios sommatifs comparés à l'évaluation de la compétence clinique au lit du malade. 3 fois par an, les internes passaient sur 2 scénarios formatifs de 10 minutes suivis d'un débriefing. Évaluation par une échelle des 6 compétences génériques des Milestones (Fig ci-dessous) couplée à des Liste de Contrôle spécifiques des cas. En fin d'année, les étudiants passaient sur 6 scénarios sommatifs de 10 min suivis d'un feedback bref sur 90 minutes. Une association significative a été retrouvée entre l'évaluation au lit du patient et l'évaluation sommative en simulation. 24 internes sont passés et les 3 observateurs des vidéos avaient un agrément modéré (Kappa 0,54). Hypothèses : absence de validation préalable de l'outil d'évaluation, problème de calibration. (NB: perte de données sur des vidéos inexploitable).¹⁴²

Tableau 12 : Exemple d'une grille d'évaluation sommative utilisée en simulation
(d'après l'article de Frey-Vogel¹⁴² (Traduction française non validée)

Prise en charge du patient				
1	2	3	4	5
- incapable d'obtenir l'historique du patient - incapable de prioriser les hypothèses diagnostiques - issue de session non adaptée		- Raisonne par analogie pour poser les bonnes questions, proposer des hypothèses diagnostiques puis une prise en charge adaptée. - capable d'échafauder un plan de prise en charge du patient - issue de session organisée et adaptée		- Raisonnement intuitif avec rassemblement des informations essentielles qui permet d'échafauder des hypothèses diagnostiques puis une analyse fine permet une prise en charge adaptée et efficiente. - issue de session fluide et adaptée.
Connaissances médicales				
1	2	3	4	5
- comprend les concepts de base mais ne peut les appliquer lors d'une mise en situation médicale		- comprend les concepts de base et peut les appliquer dans des mises en situation médicale		- Analyse la situation et la compare aux expériences et connaissances passées prendre en charge de façon adaptée cette situation clinique
Professionnalisme				
1	2	3	4	5
- ignore les proches - ne reconnaît pas/ne répond pas aux besoins du patient ou des proches		- Fait preuve d'empathie, compassion et respect et s'efforce de répondre aux besoins du patient /famille		- Défense proactive des patients, de leurs proches et prise en compte des besoins sociétaux.
Habiletés interprofessionnelles et communication				
1	2	3	4	5
- Utilise un jargon médical avec les proches - ne sait pas adapter son type de communication aux patients et aux proches		- Etablit des rapports avec toutes les personnes présentes lors de la prise en charge - instaure un climat de confiance		- connexions authentiques aux patients et leurs familles - Eduque les patients, les familles et le public - prend en charge de façon intuitive les scénarios difficiles avec facilité.
Apprentissage et amélioration fondés sur la pratique				
1	2	3	4	5
- Ne peut élaborer aucune force ou faiblesse		- élabore au moins une force, une faiblesse et une action qu'il ferait différemment		- prévoit des scénarios hypothétiques qui combleraient les lacunes cernées et des questions pour explorer les lacunes et les forces
Utilisation des ressources disponibles				
1	2	3	4	5
- Pas d'appel à l'aide ou appel trop précoce - agit individuellement et ne prend pas en compte l'importance de tous les membres de l'équipe		- appel correctement à l'aide au bon moment - reconnaît les rôles de tous les membres de l'équipe		- reconnaît que les soins de qualité aux patients ne se produisent que dans le contexte d'une équipe interprofessionnelle et servent de modèle à cet égard.

b) Brett en 2008 a utilisé 2 observateurs calibrés par discussion point par point autour d'une vidéo pour tester un outil d'évaluation standardisé, le TRACS (Tool for Resuscitation Assessment Using Computerized Simulation) . Ces observateurs ont ensuite évalué des groupes de niveaux différents sur 3 scénarios. 7 des 72 items de la grille se sont avérés inopérants : Il était notamment difficile d'évaluer des aspects cognitifs ou des prises d'informations (qui sont non verbalisés pendant la simulation). Le coefficient de corrélation intra classe était acceptable pour les 4 domaines de compétence évalués (0.70 à 0.76) et meilleur pour le score global (0.80). L'agrément moyen item par item était de 82 %. En revanche une simple tendance est retrouvée quant à la corrélation positive au niveau d'ancienneté.²⁴¹

Résidents en chirurgie

L'évaluation sommative fait partie intégrante de la formation des internes de chirurgie. L' OSATS (Objective Structured Assessment of Technical Skills ;^{242,243} est un outil conçu pour une évaluation objective de la compétence technique. Il comprend une échelle globale et une checklist procédurale spécifique. C'est un outil pour donner un feedback et mesurer la progression dans les cursus chirurgicaux.

a) En chirurgie générale : Le programme FLS (Fundamentals of Laparoscopic Surgery) comprend 5 tâches dont la validation est obligatoire pour la certification par l' American Board of Surgery. Les critères de performance et les critères de réussite ont été bien établis. La détermination du seuil de réussite s'est faite selon la méthode des "contrasting groups" suivie d'une optimisation au moyen d'une courbe ROC.²⁴⁴ L' "Operating Room black-box Program" est en cours d'évaluation et comprend des critères techniques et non techniques.²⁰²

b) Les résidents de Chirurgie Colorectale des USA ont été la première spécialité à développer une certification de compétences techniques spécifiques : COSATS. Cet examen comporte les critères de validité classique attendus avec un seuil de réussite déterminé conformément aux standards. Il met régulièrement en évidence des insuffisances dans les compétences techniques qui seraient, sinon, passées inaperçues.²⁴⁵

c) Les Résidents de chirurgie orthopédique.

Pour l'orthopédie, voici quatre exemples d'évaluations sommatives :

1. Arthroscopic Surgical Skill Evaluation Tool (ASSET) à Rochester sur cadavre avec des psychométriques évalués et acceptables.²⁴⁶
2. Un programme supporté par l'American Academy of Orthopaedic Surgeons (AAOS) évalue les étudiants sur l'arthroscopie diagnostique en situation réelle mais a montré qu'un entraînement en réalité virtuelle augmente les performances. ²⁴⁷
3. FAST The Fundamentals of Arthroscopic Surgery Training (6 modules)²⁴⁸
4. Copernicus (en lien avec Arthroscopy Association of North America (AANA)) propose un entraînement puis une évaluation finale de la procédure complète sur le cadavre (épaule).²⁴⁹

d) Les résidents de chirurgie urologique

Quelques programmes évoluent vers une évaluation sommative (Basic Laparoscopic Urologic Skills (BLUS) et le Fundamental of Robotics Surgery program (FRS). Par ailleurs, certains développent des stations ECOS pour évaluer les compétences CanMeds (Canadian Medical Education Directives for Specialists), communicateur, promoteur de la santé, leader et expert médical. Ils prennent en charge un PS avec une infirmière.²⁵⁰

e) Dans le Diplôme inter Universitaire de Chirurgie infantile de Nice (DIU Coelioscopie pédiatrique, Bréaud, Nice, France), 40% de la note est obtenue sur simulateur de tâche par la Grille F-OSATS (OSATS adapté à la suture endoscopique) incluant 7 items évalués par une échelle de Likert de 1 à 5.

f) Résidents de Gynécologie Obstétrique

- Winkel et al (New York) , rapportent un examen hybridant simulation synthétique et patient standardisé. Les 24 résidents de 4 années différentes sont évalués de façon bidirectionnelle : évaluation par le PS lui-même (pour les aspects relationnels et cognitifs) et par un observateur (pour les capacités techniques) à l'aide de Checklists critériées, le "Test of Integrated Professional Skills" (TIPS), une évaluation de Compétences professionnelles intégrées (compétences Accreditation Council for Graduate Medical Education - ACGME). L'examen comprend 5 stations de 16 minutes et 5 minutes de feedback/debriefing. La

validité interne est défendue par un Cronbach estimé entre 0.69 et 0.82 selon le scénario. Des écarts de performances ont été mis en évidence sur 12% des étudiants, qui n'avaient pas été détectés par les examens classiques. Un temps de débriefing a permis d'envisager les remédiations.²⁵¹

Utilisation de serious games :

Der Klerk et Kato ont exposé le futur de l'évaluation sommative par les serious games.²⁵² Le comportement des évalués produit de très nombreuses données (clic de souris, pages consultées, temps, eyetracking, etc...) qui permettent de surveiller, évaluer et éventuellement valider l'étudiant.²⁵³ Mais malgré les aspects prometteurs, peu de publications ont à ce jour montré de façon rigoureuse la validité d'un serious game spécifique en évaluation sommative.

a) Gerard et al. (Saint Louis, USA) ont exploré la validité d'une évaluation par un Serious Game de situations critiques pédiatriques, selon les 4 premiers critères de validité de Messick (Contenu, processus de réponse, Validité interne, Corrélation à d'autres variables) sur 3 groupes de 20 apprenants d'expertise différentes (Étudiant en médecine, Résidents, Senior).²⁵⁴ Le résultat est favorable :

- ▶ Validité du contenu : Serious Game 3D à la première personne scénarisant des situations d'urgence sur des patients de 3 mois à 10 ans. Conçu initialement pour des internes et des étudiants en médecine. Blueprint à partir des recommandations et sur collaboration d'un panel d'experts multicentrique. Sept scénarios : anaphylaxie, bronchiolite, acido-cétose, détresse respiratoire, convulsions, choc septique, tachycardie supraventriculaire.
- ▶ Processus de réponse : Grilles d'évaluations des scénarios pondérées, 55% pour les actions critiques, 28% pour l'évaluation et les réévaluations du patient, 17 % sur les actions moins critiques, avec double évaluation sur un échantillon aléatoire de 20% des scénarios joués (n=84). Coefficient de Corrélation Intra classe entre évaluateur 0.91 (IC 0.89–0.92)
- ▶ Validité interne : bonne avec un Cronbach α à 0,82.
- ▶ Corrélation à d'autres variables : forte corrélation avec l'évaluation écrite par QCM sur les mêmes thèmes, réalisée avant le jeu et discrimination nette des groupes dans l'ordre attendu d'expertise.

b) Sur un design similaire Blanie et al. ont évalué la validité selon Messick d'un Serious Game testant la capacité d'étudiants infirmiers à détecter et prendre en charge une détérioration clinique.

71 sujets répartis en 3 groupes d'expertise différente (étudiant, récemment diplômé et expérimenté) ont été inclus. 3 scénarios étaient proposés : Traumatisme crânien, Hémorragie, Occlusion. Le score évaluait des éléments du raisonnement clinique, de conscience de la situation et de communication (modèle SBAR).

Il est intéressant de noter que les critères de validité se sont révélés décevants, avec par exemple une absence de corrélation avec le niveau d'expertise des groupes. Les auteurs envisagent des biais dans le cours du jeu, notamment une proportion de bonnes réponses trop élevées parmi les propositions. Cette étude rigoureuse les contraint donc à déconseiller l'usage sommatif de ce jeu dans la version actuelle.²⁵⁵

c) A Bogota, en Colombie, Botezatu et al. ont décrit l'utilisation de Patient Virtuels ("web-Sp", développé par les suédois du Department of Learning, Informatics, Management and Ethics (LIME) du Karolinska Institutet) pour de l'évaluation sommative.²⁵⁶ Les cas cliniques durent de 45 minutes à 1h explorant le raisonnement clinique. 216 étudiants en médecine de 4ème année ont été inclus dont certains ont eu une phase formative sur le jeu avec travail en autonomie et accès à un module de feed-back en fin de cas. L'examen comprenait un cas sur Web-Sp et un cas papier classique. Le groupe ayant eu accès en formatif à Web-SP » a présenté de meilleurs résultats aux 2 composantes. Les résultats étaient dans les deux groupes meilleurs pour le cas web-sp, ce qui fait là aussi suspecter aux auteurs un biais facilitation par design du jeu (déroulé trop linéaire et accès à trop d'information).

3. Place de la simulation sommative dans la recertification

a. Un essor des programmes de re certification

Des programmes de re-certification de différentes filières de santé sont aujourd'hui en cours de déploiement ou en projet dans de nombreux pays. On peut citer le processus de revalidation en Grande-Bretagne, le suivi des compétences médicales au Canada, ou encore le processus de maintien des certifications : MOC.²⁵⁷ C'est la suite du mouvement cherchant à promouvoir le maintien de compétence avec par exemple, en France, la création d'une agence du Développement Professionnel

Continu (DPC).²⁵⁸ La certification des établissements de santé voire des équipes est aussi étudiée.¹⁹ Certaines structures australiennes ou anglo-saxonnes proposent des certifications et recertifications intra hospitalières essentiellement pour des professions indépendantes souvent médicales ou des dentistes. Le **credentialing** est une évaluation régulière des certificats, de l'expérience et des compétences des cliniciens. C'est un processus interne à l'établissement réalisé par des pairs. On peut citer l'*American Nurses Credentialing Center*²⁵⁹ ou le dispositif australien ²⁶⁰ , le **privileging** est une autorisation accordée par l'autorité compétente (par exemple la gouvernance) à un praticien pour délivrer des soins spécifiques, des traitements ou des services bien définis dans l'organisation sur des critères de formation, de qualification, d'expérience, de compétence, et d'état de santé. Ces processus réalisés lors du recrutement sont ensuite renouvelés à périodicité définie (*recredentialing* et *reprivileging*), par exemple tous les 3 ans aux États-Unis, et si une situation nouvelle apparaît, comme la survenue d'un événement indésirable (processus continu).

b. Un usage de la simulation dans les processus de recertification :

La simulation est présente ou encouragée dans ces processus. C'est ainsi que le rapport Uzan de 2018 pour le développement de la recertification des médecins en France exprime l'idée qu'on y fera « largement appel ».²³³

Actuellement même si les formations par simulations peuvent être obligatoires (processus de Maintien des Compétences en Anesthésie aux USA par exemple « MOCA ») il s'agit le plus souvent d'une démarche formative.^{34,134}

L'usage « sommatif » ou « sanctionnant » au sens d'évaluation de la performance dont dépend directement une décision contraignante pour le professionnel (qu'il s'agisse d'une obligation de formation ou d'une modulation de son habilitation à exercer) reste peu décrit.

De Maria et al. ont décrit leur expérience d'un programme évaluation + /- recyclage et réévaluation conditionnant le retour après interruption d'activité de médecins anesthésistes aux USA. La composante évaluation par simulation durait deux jours et comprenait un briefing d'une demi-journée (familiarisation avec l'environnement de simulation) puis 3 ou 4 scénarios/jour avec 2 évaluateurs. Les scénarios du premier jour étaient identiques pour tous les participants et visaient à établir un niveau de

référence de compétence globale. Les scénarios présentés le jour 2 étaient choisis et modifiés en fonction des performances du participant le jour 1. Le système de notation est basé sur l'ANTS. Cette démarche hautement personnalisée (20 professionnels en 10 ans) faisait l'objet d'une contractualisation entre le professionnel et l'établissement de formation.³⁵

Weinger et al. se sont servis du processus MOCA pour une étude impliquant 263 anesthésistes diplômés. Les scénarios ciblaient des écarts de performances critiques et/ou récurrents décrits dans la littérature. L'évaluation comprenait un check liste d'éléments clés, des échelles holistiques techniques et comportementales assorties d'une décision binaire (le professionnel a-t-il le niveau attendu pour un anesthésiste certifié ?). Environ 25% des anesthésistes produisaient une performance insuffisante, avec des écarts de performance recoupant ceux identifiés dans des études cliniques, soulignant la pertinence des thématiques choisies sur ces bases. En pratique outre la recertification, la simulation a, là encore, encore un objectif finalement formatif, contribuant à l'objectif de remédiation de lacunes collectives.²²⁷

c) Ce qui reste identique par rapport à la démarche certificative initiale (tab. 11):

- L'objectif est de garantir que les professionnels entretiennent une compétence clinique assurant la sécurité des patients et leur prise en charge conforme aux données actualisée de la science.
- La simulation a sa place dans des évaluations multimodales (dans un seul des 4 axes du MOCA par exemple)
- La rigueur de la mise en place des processus est la même, dans une démarche qualité.

d) Ce qui est particulier à la recertification (tab. 11) :

- Quel degré de contrainte :
90% des médecins USA certifiés mais seuls 30% à jour de leur recertification
L'acceptabilité d'une modalité de recertification est cruciale pour qu'un corps professionnel s'y soumette. Le rapport Uzan précise ainsi de nombreuses conditions pour permettre une approche en confiance.
- Objectifs annexes :
Pour le rapport Uzan, cette démarche continue constitue un mode

d'accompagnement qui permet aussi de dépister et d'aider des professionnels en difficulté (burn out...)

► Quel organisme responsable de la démarche :

Si pour les formations initiales les rôles universitaires ont assez clairs, la démarche de recertification peut faire intervenir des autorités de régulation ou de gouvernance de la santé (DGOS, HAS), des organes représentatifs comme les conseils de l'ordre de différentes professions, Sociétés savantes, Collège d'Enseignants, « Board » de spécialité, structures de coordination de la formation continue (ANDPC)...

► Quel « Blue print », quelle est l'étendue du domaine évalué ?

- Les compétences sont-elles évaluables de la même façon chez les experts ? Il a par exemple été montré que des grilles d'évaluation détaillant la démarche diagnostique pas à pas peuvent défavoriser une démarche d'expert (raisonnement rapide par « pattern »).²⁶¹
- L'échantillon de compétences évaluables lors d'une recertification est nécessairement plus étroit que lors de la certification initiale car contraintes de temps pour les structures d'évaluation (dont les évaluateurs) comme pour les professionnels.
- Privilégier un panel de compétences correspondant à des écarts de pratiques ou des événements indésirables souvent rapportés ? Cela permet en théorie de contribuer à une démarche territoriale de qualité des soins efficiente car centrée sur des besoins identifiés.²²⁷
- Compétences liées à la communication et au travail en équipe particulièrement importantes au vu des données connues sur les causes d'événements indésirables.
- Le mécanisme du choix des compétences sur lesquelles centrer ces évaluations est donc à définir (quels principes, quelle(s) autorité(s) pour les arbitrages et les évolutions).

► Quel « mix évaluatif »

- Pré-requis / renvoi à des sources en lignes pour préparer
- Outre des épreuves cognitives et de la simulation technique (MOC chirurgie) ou immersive, une évaluation de l'activité clinique réelle est proposée avec diverses modalités (revue de dossiers, accompagnement au lit du malade/au bloc opératoire). Le programme de recyclage

exposé par De Maria et al. comprend ainsi plusieurs semaines mélangeant simulation formative et suivi sur le terrain.

► Quelle équipe pour évaluer ?

- Les auteurs insistent sur la nécessité de formateurs expérimentés en simulation (plus de 5 d'expérience pour le programme de Mount Sinai NY).
- Équipe d'évaluation restreinte (confidentialité particulièrement importante en termes de sécurité émotionnelle et d'implication légale). Ratio formateur/apprenant minimisé (1 : 5 évoqué par exemple en anesthésie). Évaluateurs formés (Weinger proposait 2 jours de formation et calibration des évaluateurs, à partir de 24 vidéos montrant des performances variables pour chacun des 4 scénarios choisis).
- Faut-il une accréditation d'évaluateurs (par le collège de spécialité par exemple, mais possible rôle de la société savante de simulation ?)

► Quel centre pour évaluer ?

- Un réseau de centres affilié ou accrédité d'une façon ou d'une autre (par ex. Simulation Education Network (SEN) en lien avec l'American Board of Anesthesia aux USA)
- Pour le MOCA par exemple, un système de « cahier des charges » pour homologuer le programme.
- Place d'une société savante en simulation dans une évaluation/une accréditation de centre pour cette mission particulière ?

► Quelle organisation :

- Nombre de candidats : pas d'épreuve « de masse » quel que soit le programme. Privilégier l'individualisation de l'évaluation et de l'accompagnement.
- Nœud du problème : disponibilité des professionnels et des structures de formation. Par exemple, pour un MOCA d'une périodicité de 10 ans le réseau de 27 centres du Simulation Education Network (SEN) de l'American Society of Anesthesiology (ASA) projette 100 à 160 personnes à évaluer par centre et par an. Une charge de travail importante, qui nécessite un temps de travail dédié non négligeable dans des professions déjà parfois en tension. Et certains programmes évoquent une périodicité de 3 ans ou de 6 ans.

- ▶ Quel financement :
 - La question du financement du dispositif est incontournable, intégrant éventuellement, le manque à gagner pour chaque praticien :
 - Faut-il faire payer le praticien, comme dans le cas du MOCA, évalué à 1500 \$ la journée aux états Unis. ¹³⁴
 - Faut-il un financement par l'état, par les assurances professionnelles ou par un organisme centralisé type ANDPC ?
- ▶ Quelles conséquences de l'évaluation re certificative
 - Faut-il une accréditation individuelle ? une accréditation d'équipe ? une autorisation d'exercer ?
 - Quel est l'impact financier et psychologique de toute décision modifiant l'exercice professionnel
 - Comment assurer le maintien de la confidentialité sur les contenu des évaluations (sécurisation des données - « coffre-fort numérique», absence de production de données permettant « classer » les professionnels).²³³
 - Le programme décrit par De Maria, bien que conduisant à l'absence de retour au travail de 2 professionnels ne donne pas de conclusion quant à une aptitude clinique. Un rapport exprime le détail du déroulé de chaque session de simulation, une liste des lacunes identifiées et des mesures correctrices envisagées.

Tableau 13 : Points communs et divergences entre certification et recertification

Points communs	Divergences
Process multimodal (cours, simulation, etc..) ^{1 42 93}	Faible pourcentage de recertification existant ^{42 93} Quelle acceptabilité et faisabilité ?
Centres accrédités (Cahier des charges) ^{42 93} Même rigueur dans la mise en place	Les universités, en charge de la certification, peuvent-elles assurer la recertification ?
Niveau de compétence visé	Difficultés de sélection efficiente des compétences à évaluer pour la recertification : • Contraintes multiples (temps/moyens) • Communication/travail d'équipe, écarts de pratiques, événements indésirables fréquents ? Scénarios et grilles d'évaluation adaptés pour attendus d'apprentissage ⁹⁴
Possibilités de suivi sur le terrain ¹²	La certification individuelle doit-elle continuer ou être remplacée par une recertification d'équipe ?
Échec possible de la certification ou de la recertification	L'impact d'un refus de recertification est majeur pour un professionnel Discrétion du processus nécessaire Dépistage possible des professionnels en difficulté (burn out...) ^{1 81}
Difficultés du financement	Nombreuses pistes de financements possibles en recertification (Etat, assurances professionnelles, etc.)

Au total

La simulation sommative a un intérêt dans les processus de recertification, mais de nombreux écueils sont à éviter et il est à court terme plus aisé d'intégrer des sessions formatives. Le consensus actuel semble être qu'il ne doit pas y avoir simulation sanctionnante re-certificative sans un accompagnement global personnalisé professionnel, dans un processus qualité plus que dans une logique binaire aptitude/inaptitude.

Chapitre 7 : Lexique

Construits : concepts et principes abstraits déduits du comportement observés et expliqués par une théorie. Par exemple, la réussite universitaire est un construit déduit des résultats obtenus lors des évaluations telles que des tests écrits, des examens oraux sur un domaine de connaissances bien défini ou encore des résultats lors des OSCE, pour un domaine de compétence défini.¹⁴

Cohérence interne : L'évaluation de la consistance interne de l'outil est le premier élément d'évaluation de la cohérence de l'outil. Elle indique à quel point les items d'un test mesurent la même dimension. Pour Downing, par exemple, les scores d'item ou groupe d'item destinés à mesurer la même variable ou le même domaine devraient être plus fortement corrélés que les scores d'items destinés à mesurer des variables ou des domaines différents.¹⁴

Cohérence externe : La cohérence externe permet de s'assurer que le résultat n'est pas unique mais qu'il s'intègre dans un cadre logique : confirmation par d'autres essais, d'autres mesures, cohérence avec les connaissances fondamentales, épidémiologiques, etc.^{20-22 95}

ECOS - Examen Clinique Objectif Structuré (*Objective Structured Clinical Examination-OSCE*) : Outil visant une évaluation standardisée et objective du comportement et de la performance de futurs professionnels de santé. Il a été décrit pour la première fois en 1975.⁸⁰ Les candidats parcourent une série de "stations" reproduisant des situations cliniques. Dans chacune d'elle l'étudiant est évalué et noté selon une grille de notation standardisée par des évaluateurs entraînés.

Évaluation formative, normative et certificative : Selon Perrenoud, l'évaluation formative « aide l'étudiant à apprendre et à se développer ». Un de ses intérêts pour l'étudiant est de s'adapter au feedback reçu lors de l'évaluation, pour améliorer ses connaissances mais également de savoir s'autoévaluer et dans le cas de la constatation d'un manque de connaissance, savoir y remédier.⁹⁶ L'étudiant « apprend à apprendre », il pratique une autorégulation de ses apprentissages.⁹⁷

L'évaluation normative (ou sommative) permet à l'apprenant de recevoir une note

qui le situe par rapport à d'autres apprenants alors que l'évaluation formative lui permet de se situer par rapport à sa progression propre et de se fixer des objectifs pour la suite du programme. L'évaluation certificative est une forme d'évaluation sommative et elle permet de délivrer un diplôme, une certification, en décidant de l'échec ou de la réussite d'un apprenant au sein d'un programme de formation.⁹⁸ Les deux modèles d'évaluation ne s'excluent pas, l'évaluation formative aidant l'étudiant à progresser par étapes pour valider une évaluation sommative.⁹⁹

Fiabilité : est le reflet de l'objectivité de la mesure et donc de sa reproductibilité. Il s'agit du degré de reproductibilité des notes des apprenants, sur des échantillons différents d'évaluateurs, de lieux d'examens, de temps etc.¹⁰⁰

Fidélité : représente le degré avec lequel l'évaluation par la simulation mesure réellement ce qu'elle prétend mesurer.^{101 102}

Inférence : action de tirer une conclusion d'un fait ou d'un événement donné. Dans le cadre de l'évaluation des compétences, il s'agit de valider l'acquisition d'une compétence définie sur base des performances observées en situation.¹⁰³

Outil / instrument d'évaluation : dispositif d'une évaluation, qui permet de « collecter des données et, dans une certaine mesure, de guider l'interprétation et le jugement [...] Leur nature et leur forme peuvent varier selon la nature de ce qui est à évaluer, la personne à qui incombe la responsabilité de l'évaluation, la forme des résultats à obtenir et même selon l'expertise en évaluation de ceux qui les conçoivent ou les utilisent. [...] L'instrumentation est au cœur de la démarche évaluative.¹⁰⁴

Programme d'évaluation : programme qui regroupe différentes méthodes d'évaluation, afin d'adapter au mieux son contenu aux objectifs d'évaluation. Il répond à trois objectifs majeurs : optimiser l'apprentissage (concept d'évaluation pour l'apprentissage), permettre aux évaluateurs de prendre des décisions avec un enjeu majeur (certification/promotion des apprenants), enfin fournir des informations pour améliorer l'enseignement et le programme de formation.¹⁰⁵

Reproductibilité : Un outil d'évaluation est considéré comme doté de reproductibilité s'il permet d'obtenir des résultats comparables (ou avec des variations minimales) lors

de son utilisation, qu'il soit utilisé par le concepteur de l'outil ou un évaluateur formé à l'utilisation de celui-ci, et quel que soit le niveau de l'apprenant dans des conditions d'apprentissages identiques. Plus précisément, on peut distinguer la reproductibilité intra-individuelle (résultats comparables lors d'évaluation récurrentes par un même évaluateur) de la reproductibilité inter-individuelle (résultats comparables pour plusieurs évaluateurs).¹⁰⁶

Schémes d'action : ce qui, dans une action, est ainsi transposable, généralisable ou différenciable d'une situation à la suivante, autrement dit ce qu'il y a de commun aux diverses répétitions ou applications de la même action.¹⁰⁷

Sources de validité : preuves de validité d'un score, utilisées pour valider chaque étape d'un processus de validation, afin de rendre un test fidèle à ce qu'il veut mesurer.

Validité : la validité d'un test fait référence au « degré avec lequel les preuves de validité et la théorie utilisées étayent les interprétations des résultats d'un test » *The Standards*. Elle renvoie toujours à la pertinence de l'interprétation des scores et au sens qui leur est donné plutôt qu'aux scores eux-mêmes.^{14 108} Elle est le « sine qua non » de l'évaluation car sans preuve de validité, l'évaluation médicale n'a que peu ou pas de signification intrinsèque.¹⁴ *Ainsi un outil n'est pas valide ou invalide, mais il a fait preuve ou pas de validité, pour un contexte donné, une interprétation spécifique des résultats.* En bref, il s'agit du degré de fidélité avec lequel un test ou un outil mesure précisément le concept visé.

Validation: processus de recueil des preuves de validité, afin d'argumenter l'interprétation et l'utilisation des scores.^{14 18}

Bibliographie

1. HAS. Guide de bonnes pratiques en simulation en santé. 2012. Available at: https://www.has-sante.fr/upload/docs/application/pdf/2013-01/guide_bonnes_pratiques_simulation_sante_guide.pdf. Accessed February 2, 2020.
2. Haute Autorité de Santé. Recommandations par consensus formalisé (RCF). Haute Aut Santé 2011. Available at: https://www.has-sante.fr/jcms/c_272505/fr/recommandations-par-consensus-formalise-rcf. Accessed October 29, 2020.
3. McMillan SS, King M, Tully MP. How to use the nominal group and Delphi techniques. *Int J Clin Pharm* 2016;38:655–62.
4. Brailovsky CA, Miller F, Grand'Maison P. L'évaluation de la compétence dans le contexte professionnel. *Serv Soc* 2005;47:171–89.
5. Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR, Harris P, Glasgow NJ, Campbell C, Dath D, Harden RM, Iobst W, Long DM, Mungroo R, Richardson DL, Sherbino J, Silver I, Taber S, Talbot M, Harris KA. Competency-based medical education: theory to practice. *Med Teach* 2010;32:638–45.
6. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med J Assoc Am Med Coll* 1990;65:S63-67.
7. Tardif J, Fortier G, Préfontaine C. L'évaluation des compétences: documenter le parcours de développement. Montréal: Chenelière Éducation, 2006.
8. Neira VM, Bould MD, Nakajima A, Boet S, Barrowman N, Mossdorf P, Sydor D, Roeske A, Noseworthy S, Naik V, Doherty D, Writer H, Hamstra SJ. "GLOSAT": a tool to assess CanMEDS competencies during simulated crises. *Can J Anesth Can Anesth* 2013;60:280–9.
9. Perrenoud P. Évaluation formative et évaluation certificative : postures contradictoires ou complémentaires ? *Form Prof Suisse* 2001;4:25–8.
10. Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of the accreditation council for graduate medical education: a systematic review. *Acad Med J Assoc Am Med Coll* 2009;84:301–9.
11. Burke CS, Salas E, Wilson-Donnelly K, Priest H. How to turn a team of experts into an expert medical team: guidance from the aviation and military communities. *Qual Saf Health Care* 2004;13 Suppl 1:i96-104.
12. Ministère des affaires sociales, de la santé et des droits des femmes. Arrêté du 2 septembre 2015 relatif au diplôme d'État de masseur-kinésithérapeute., 2015:192–207. Available at: https://solidarites-sante.gouv.fr/fichiers/bo/2015/15-08/ste_20150008_0000_p000.pdf. Accessed July 6, 2021.
13. Stone J. Moving interprofessional learning forward through formal assessment. *Med Educ* 2010;44:396–403.
14. Learning objectives for medical student education--guidelines for medical schools: report I of the Medical School Objectives Project. *Acad Med J Assoc Am Med Coll* 1999;74:13–8.
15. Sanders CW, Edwards JC, Burdenski TK. A Survey of Basic Technical Skills of Medical Students. *Acad Med* 2004;79:873–5.

16. Bond WF, Spillane L, The CORD Core Competencies Simulation Group. The Use of Simulation for Emergency Medicine Resident Assessment. *Acad Emerg Med* 2002;9:1295–9.
17. Jouquan J. L'évaluation des apprentissages des étudiants en formation médicale initiale. *Pédagogie Médicale* 2002;3:38–52.
18. Epstein RM. Assessment in medical education. *N Engl J Med* 2007;356:387–96.
19. Haute Autorité de Santé. Évaluation des compétences des professionnels de santé et certification des établissements de santé., 2015;45. Available at: https://www.has-sante.fr/upload/docs/application/pdf/2015-12/rapport_l_evaluation_des_compences_des_professionnels.pdf. Accessed November 2, 2020.
20. Bould MD, Crabtree NA, Naik VN. Assessment of procedural skills in anaesthesia. *Br J Anaesth* 2009;103:472–83.
21. Boulet JR, Murray DJ. Simulation-based Assessment in Anesthesiology: Requirements for Practical Implementation. *Anesthesiology* 2010;112:1041–52.
22. Scalese RJ, Obeso VT, Issenberg SB. Simulation technology for skills training and competency assessment in medical education. *J Gen Intern Med* 2008;23 Suppl 1:46–9.
23. Boulet JR, Swanson DB. Psychometric Challenges of Using Simulations for High-Stakes Assessment. :12.
24. Scoles PV, Hawkins RE, LaDuca A. Assessment of clinical skills in medical practice. *J Contin Educ Health Prof* 2003;23:182–90.
25. Gale TCE, Roberts MJ, Sice PJ, Langton JA, Patterson FC, Carr AS, Anderson IR, Lam WH, Davies PRF. Predictive validity of a selection centre testing non-technical skills for recruitment to training in anaesthesia. *Br J Anaesth* 2010;105:603–9.
26. Waisel DB, Ruben MA, Blanch-Hartigan D, Hall JA, Meyer EC, Blum RH. Compassionate and Clinical Behavior of Residents in a Simulated Informed Consent Encounter. *Anesthesiology* 2020;132:159–69.
27. Gordon M, Farnan J, Grafton-Clarke C, Ahmed R, Gurbutt D, McLachlan J, Daniel M. Non-technical skills assessments in undergraduate medical education: A focused BEME systematic review: BEME Guide No. 54. *Med Teach* 2019;41:732–45.
28. Romainville M. Objectivité versus subjectivité dans l'évaluation des acquis des étudiants. *Rev Int Pédagogie L'enseignement Supér* 2011;27. Available at: <https://ripes.revues.org/499>. Accessed June 28, 2017.
29. Boulet JR, Murray D, Kras J, Woodhouse J. Setting performance standards for mannequin-based acute-care scenarios: an examinee-centered approach. *Simul Healthc J Soc Simul Healthc* 2008;3:72–81.
30. Weersink K, Hall AK, Rich J, Szulewski A, Dagnone JD. Simulation versus real-world performance: a direct comparison of emergency medicine resident resuscitation entrustment scoring. *Adv Simul* 2019;4:9.
31. Ziv A, Rubin O, Sidi A, Berkenstadt H. Credentialing and certifying with simulation. *Anesthesiol Clin* 2007;25:261–9.
32. Ellaway RH, Kneebone R, Lachapelle K, Topps D. Practica continua: Connecting and combining simulation modalities for integrated teaching, learning and assessment. *Med Teach* 2009;31:725–31.

33. Lois F, De Kock M. L'introduction d'une séance de simulation in situ standardisée dans un dispositif de sélection en anesthésie-réanimation peut-elle aider à mieux sélectionner les futurs candidats spécialistes ? *Pédagogie Médicale* 2017;18:7–15.
34. Gallagher CJ, Tan JM. The current status of simulation in the maintenance of certification in anesthesia. *Int Anesthesiol Clin* 2010;48:83–99.
35. DeMaria S, Samuelson ST, Schwartz AD, Sim AJ, Levine AI. Simulation-based assessment and retraining for the anesthesiologist seeking reentry to clinical practice: a case series. *Anesthesiology* 2013;119:206–17.
36. Vleuten CPM van der, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ* 2005;39:309–17.
37. Schuwirth LWT, Vleuten CPM van der. Programmatic assessment and Kane's validity perspective. *Med Educ* 2012;46:38–48.
38. Berkenstadt H, Ziv A, Gafni N, Sidi A. Incorporating simulation-based objective structured clinical examination into the Israeli National Board Examination in Anesthesiology. *Anesth Analg* 2006;102:853–8.
39. Wong AK. Full scale computer simulators in anesthesia training and evaluation. *Can J Anaesth J Can Anesth* 2004;51:455–64.
40. Scallon G. L'évaluation des apprentissages dans une approche par compétences. Bruxelles: De Boeck Université-Bruxelles, 2007.
41. Willhaus J, Burleson G, Palaganas J, Jeffries P. Authoring Simulations for High-Stakes Student Evaluation. *Clin Simul Nurs* 2014;10:e177–82.
42. Brailovsky C, Charlin B, Beausoleil S, Côté S, Van der Vleuten C. Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test. *Med Educ* 2001;35:430–6.
43. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ* 2015;49:560–75.
44. 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* 2014;19:233–50.
45. Messick S. The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educ Res* 1994;23:13–23.
46. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ* 2003;37:830–7.
47. Kane MT. Validating the Interpretations and Uses of Test Scores. *J Educ Meas* 2013;50:1–73.
48. Kane M, Crooks T, Cohen A. Validating Measures of Performance. *Educ Meas Issues Pract* 2005;18:5–17.
49. Loye N. Et si la validation était plus qu'une suite de procédures techniques ? *Mes Éval En Éducation* 2018;41:97–123.
50. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. *Med Teach* 2011;33:478–85.

51. Tavares W, Brydges R, Myre P, Prpic J, Turner L, Yelle R, Huiskamp M. Applying Kane's validity framework to a simulation based assessment of clinical competence. *Adv Health Sci Educ* 2018;23:323–38.
52. Vleuten CPM van der, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, Tartwijk J van. A model for programmatic assessment fit for purpose. *Med Teach* 2012;34:205–14.
53. Ilgen JS, Ma IWY, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ* 2015;49:161–73.
54. Cohen DS, Colliver JA, Marcy MS, Fried ED, Swartz MH. Psychometric properties of a standardized-patient checklist and rating-scale form used to assess interpersonal and communication skills. *Acad Med J Assoc Am Med Coll* 1996;71:S87-89.
55. Van der Vleuten CP, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ* 1991;25:110–8.
56. Cushing A. Assessment of Non-Cognitive Factors. In: Norman GR, Vleuten CPM van der, Newble DI, Dolmans DHJM, Mann KV, Rothman A, Curry L, eds. *International Handbook of Research in Medical Education. Vol 7. Springer International Handbooks of Education*. Dordrecht: Springer Netherlands, 2002:711–55. Available at: http://link.springer.com/10.1007/978-94-010-0462-6_27. Accessed July 6, 2021.
57. Regehr G, Freeman R, Robb A, Missiha N, Heisey R. OSCE performance evaluations made by standardized patients: comparing checklist and global rating scores. *Acad Med J Assoc Am Med Coll* 1999;74:S135-137.
58. Schwartz CE, Powell VE, Rapkin BD. When global rating of change contradicts observed change: Examining appraisal processes underlying paradoxical responses over time. *Qual Life Res Int J Qual Life Asp Treat Care Rehabil* 2017;26:847–57.
59. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Acad Med J Assoc Am Med Coll* 1999;74:1129–34.
60. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. *Med Educ* 2004;38:199–203.
61. Fahy AS, Fok K-H, Gavrilovic B, Farcas M, Carrillo B, Gerstle JT, Azzie G. Refinement in the analysis of motion within low-cost laparoscopic simulators of differing size: Implications on assessing technical skills. *J Pediatr Surg* 2018;53:2480–7.
62. Bréaud J, Chevallier D, Benizri E, Fournier J-P, Carles M, Delotte J, Venissac N, Myx A, Ianelli A, Levraut J, Jones D, Benchimol D. The place of simulation in the surgical resident curriculum. The pedagogic program of the Nice Medical School Simulation Center. *J Visc Surg* 2012;149:e52-60.
63. Faudeux C, Tran A, Dupont A, Desmontils J, Montaudie I, Bréaud J, Braun M, Fournier J-P, Bérard E, Berlengi N, Schweitzer C, Haas H, Caci H, Gatin A, Giovannini-Chami L. Development of Reliable and Validated Tools to Evaluate Technical Resuscitation Skills in a Pediatric Simulation Setting: Resuscitation and Emergency Simulation Checklist for Assessment in Pediatrics. *J Pediatr* 2017;188:252-257.e6.
64. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system†. *BJA Br J Anaesth* 2003;90:580–8.

65. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997;84:273–8.
66. Robertson RL, Park J, Gillman L, Vergis A. The impact of rater training on the psychometric properties of standardized surgical skill assessment tools. *Am J Surg* 2020;220:610–5.
67. Scallon G. Des savoirs aux compétences: explorations en évaluation des apprentissages. Bruxelles: De Boeck Supérieur, 2015.
68. Frenay M, Romainville M. L'accompagnement des mémoires et des thèses. Louvain-la-Neuve: UCL Presses universitaires de Louvain, 2013.
69. Gerard F-M. Évaluer des compétences: guide pratique. 2. éd., 3. tir. Bruxelles: De Boeck, 2009.
70. Oriot D, Darrieux E, Boureau-Voultoury A, Ragot S, Scépi M. Validation of a performance assessment scale for simulated intraosseous access. *Simul Healthc J Soc Simul Healthc* 2012;7:171–5.
71. André N, Loyer N, Laurencelle L. La validité psychométrique : un regard global sur le concept centenaire, sa genèse, ses avatars. *Mes Éval En Éducation* 2015;37:125–48.
72. Humphrey-Murto S, Varpio L, Gonsalves C, Wood TJ. Using consensus group methods such as Delphi and Nominal Group in medical education research. *Med Teach* 2017;39:14–9.
73. Humphrey-Murto S, Varpio L, Wood TJ, Gonsalves C, Uffholz L-A, Mascioli K, Wang C, Foth T. The Use of the Delphi and Other Consensus Group Methods in Medical Education Research: A Review. *Acad Med J Assoc Am Med Coll* 2017;92:1491–8.
74. Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CF, Askham J, Marteau T. Consensus development methods, and their use in clinical guideline development. *Health Technol Assess Winch Engl* 1998;2:i–iv, 1–88.
75. Linstone HA, Turoff M, eds. The Delphi method: techniques and applications. Reading, Mass: Addison-Wesley Pub. Co., Advanced Book Program, 1975.
76. Jones J, Hunter D. Consensus methods for medical and health services research. *BMJ* 1995;311:376–80.
77. Propst EJ, Wolter NE, Ishman SL, Balakrishnan K, Deonarain AR, Mehta D, Zalzal G, Pransky SM, Roy S, Myer CM, Torre M, Johnson RF, Ludemann JP, Derkay CS, Chun RH, Hong P, Molter DW, Prager JD, Nguyen LHP, Rutter MJ, Myer CM, Zur KB, Sidell DR, Johnson LB, Cotton RT, Hart CK, Willging JP, Zdanski CJ, Manoukian JJ, Lam DJ, et al. Competency-Based Assessment Tool for Pediatric Tracheotomy: International Modified Delphi Consensus. *The Laryngoscope* 2020;130:2700–7.
78. Association of American Medical Colleges. MedEdPORTAL Assessment Instrument Guidelines. In: MedEdPORTAL Author Handbook. Washington (DC): AAMC, 2011:36. Available at: https://www.adea.org/uploadedFiles/ADEA/Content_Conversion_Final/mededportal/Documents/authorhandbook2012.pdf.
79. Oermann MH, Saewert KJ, Charasika M, Yarbrough SS. Assessment and grading practices in schools of nursing: national survey findings part I. *Nurs Educ Perspect* 2009;30:274–8.

80. Furman GE, Smee S, Wilson C. Quality Assurance Best Practices for Simulation-Based Examinations: Simul Healthc J Soc Simul Healthc 2010;5:226–31.
81. Durand C. Application de la théorie de réponse aux items à l'analyse d'échelles d'attitude. Bull Sociol Methodol Méthodologie Sociol 1998;59:27–48.
82. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ 2011;2:53–5.
83. Hunt RJ. Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability. J Dent Res 1986;65:128–30.
84. Oermann MH, Kardong-Edgren S, Rizzolo MA. Summative Simulated-Based Assessment in Nursing Programs. J Nurs Educ 2016;55:323–8.
85. Undre S, Sevdalis N, Healey AN, Darzi A, Vincent CA. Observational teamwork assessment for surgery (OTAS): refinement and application in urological surgery. World J Surg 2007;31:1373–81.
86. Guise J-M, Deering SH, Kanki BG, Osterweil P, Li H, Mori M, Lowe NK. Validation of a tool to measure and promote clinical teamwork. Simul Healthc J Soc Simul Healthc 2008;3:217–23.
87. Sevdalis N, Davis R, Koutantji M, Undre S, Darzi A, Vincent CA. Reliability of a revised NOTECHS scale for use in surgical teams. Am J Surg 2008;196:184–90.
88. Baker VO, Cuzzola R, Knox C, Liotta C, Cornfield CS, Tarkowski RD, Masters C, McCarthy M, Sturdivant S, Carlson JN. Teamwork education improves trauma team performance in undergraduate health professional students. J Educ Eval Health Prof 2015;12:36.
89. Mishra A, Catchpole K, McCulloch P. The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. Qual Saf Health Care 2009;18:104–8.
90. Pires S, Monteiro S, Pereira A, Chaló D, Melo E, Rodrigues A. Non-technical skills assessment for prelicensure nursing students: An integrative review. Nurse Educ Today 2017;58:19–24.
91. Breaud J, Talon I, Fourcade L, Podevin G, Rod J, Audry G, Dohin B, Lecompte J-F, Bensaid R, Rampal V, Azzie G. The National Pediatric Surgery Simulation Program in France: A tool to develop resident training in pediatric surgery. J Pediatr Surg 2019;54:582–6.
92. De Ketele J-M, Gérard F-M. La validation des épreuves d'évaluation selon l'approche par les compétences. Mes Éducation En Éval 2005;28:1–26.
93. Verot C. Place de la simulation clinique dans la validation des compétences des étudiants sages-femmes pour l'obtention du Certificat de Synthèse Clinique et Thérapeutique. 2018. Available at: <http://www.medesim.fr/wp-content/doc/memoire/diu2018cverotmemoire.pdf>. Accessed July 8, 2021.
94. Philippon A-L. Evaluation des compétences des étudiants de 4ème année de médecine dans la prise en charge des urgences vitales : quelle place pour la simulation ? <http://www.adjectif.net/spip> 2017. Available at: <http://www.adjectif.net/spip/spip.php?article451&lang=fr>. Accessed October 24, 2020.
95. Howe R. La note de l'évaluation finale d'un cours dans l'approche par compétences : quelques enjeux pédagogiques /. Pédagogie Collégiale 2006;20:10–5.
96. Roulin V, Allin-Pfister A-C, Berthiaume D. Comment évaluer les apprentissages dans l'enseignement supérieur professionnalisant?, 2020.

97. Pelaccia thierry. Comment mieux former et évaluer les étudiants en médecine et en sciences de la santé. DeBoeck., 2016.
98. Turner NM, Lukkassen I, Bakker N, Draaisma J, Cate OThJ ten. The effect of the APLS-course on self-efficacy and its relationship to behavioural decisions in paediatric resuscitation. *Resuscitation* 2009;80:913–8.
99. Corvetto MA, Taekman JM. To Die or Not To Die? A Review of Simulated Death: *Simul Healthc J Soc Simul Healthc* 2013;8:8–12.
100. Bédard D, Bécharard JP. L'innovation pédagogique dans le supérieur : un vaste chantier. In: *Innover dans l'enseignement supérieur*. Paris: Presses Universitaires de France, 2009:29–43.
101. Gaba DM. Simulations That Are Challenging to the Psyche of Participants: How Much Should We Worry and About What? *Simul Healthc J Soc Simul Healthc* 2013;8:4–7.
102. Calhoun AW, Boone MC, Miller KH, Pian-Smith MCM. Case and commentary: using simulation to address hierarchy issues during medical crises. *Simul Healthc J Soc Simul Healthc* 2013;8:13–9.
103. Gantt LT. The Effect of Preparation on Anxiety and Performance in Summative Simulations. *Clin Simul Nurs* 2013;9:e25–33.
104. Rudolph JW, Simon R, Raemer DB, Eppich WJ. Debriefing as Formative Assessment: Closing Performance Gaps in Medical Education. *Acad Emerg Med* 2008;15:1010–6.
105. Kang SJ, Min HY. Psychological Safety in Nursing Simulation. *Nurse Educ* 2019;44:E6–9.
106. Henricksen JW, Altenburg C, Reeder RW. Operationalizing Healthcare Simulation Psychological Safety: A Descriptive Analysis of an Intervention. *Simul Healthc J Soc Simul Healthc* 2017;12:289–97.
107. Froger O. Comment évaluer les étudiants avec des patients standardisés en situation simulée? In: *Comment évaluer les apprentissages dans l'enseignement supérieur professionnalisant? 1ère*. Guides pratiques., 2017:115–32.
108. Deschênes M-F. Le développement du jugement clinique infirmier lors d'activités en simulation clinique haute fidélité (SCHF). Montréal: Pôle de spécialisation et d'innovation en santé, 2015:66. Available at: <https://cdc.qc.ca/pdf/deschenes-jugement-clinique-infirmier-simulation-haute-fidelite-rapport-EDU-MTL-2015.pdf>.
109. Howard SK, Gaba DM, Smith BE, Weinger MB, Herndon C, Keshavacharya S, Rosekind MR. Simulation study of rested versus sleep-deprived anesthesiologists. *Anesthesiology* 2003;98:1345–55; discussion 5A.
110. Neuschwander A, Job A, Younes A, Mignon A, Delgoulet C, Cabon P, Mantz J, Tesniere A. Impact of sleep deprivation on anaesthesia residents' non-technical skills: a pilot simulation-based prospective randomized trial. *Br J Anaesth* 2017;119:125–31.
111. Eastridge BJ, Hamilton EC, O'Keefe GE, Rege RV, Valentine RJ, Jones DJ, Tesfay S, Thal ER. Effect of sleep deprivation on the performance of simulated laparoscopic surgical skill. *Am J Surg* 2003;186:169–74.
112. Hayter MA, Friedman Z, Katznelson R, Hanlon JG, Borges B, Naik VN. Effect of sleep deprivation on labour epidural catheter placement. *Br J Anaesth* 2010;104:619–27.
113. Ghazali DA, Breque C, Sosner P, Lesbordes M, Chavagnat J-J, Ragot S, Oriot D. Stress response in the daily lives of simulation repeaters. A randomized controlled trial

assessing stress evolution over one year of repetitive immersive simulations. *PloS One* 2019;14:e0220111.

114. Bandura A. Much Ado Over a Faulty Conception of Perceived Self-Efficacy Grounded in Faulty Experimentation. *J Soc Clin Psychol* 2007;26:641–58.
115. Galand B, Vanlede M. Le sentiment d'efficacité personnelle dans l'apprentissage et la formation : quel rôle joue-t-il ? D'où vient-il ? Comment intervenir ? *Savoirs* 2004;Hors série:91–116.
116. Marks R. Self Efficacy Attributes and Arthritis Disability: A Call for More Standardized and Thoughtfully Applied Measurement Approaches. *EC Orthop* 2017;7:215–9.
117. Maibach E, Murphy DA. Self-efficacy in health promotion research and practice: conceptualization and measurement. *Health Educ Res* 1995;10:37–50.
118. Durand C, Abraham L, Secheresse T, Dumas J, Leconte M, Dibisceglie A, Dinatale P, Théry G, Deiber M, Bouchon N, Dessieux E, Ughetto A, Berger C, Vandenberg A. Simulation haute-fidélité et sentiment d'efficacité personnelle. Application à l'évaluation d'un programme de formation à la réanimation du nouveau-né en salle de naissance. *Rev Médecine Périnatale* 2014;6:134–43.
119. Secheresse T, Usseglio P, Joriot C, Habold D. Simulation haute-fidélité et sentiment d'efficacité personnelle. Une approche pour appréhender l'intérêt de la simulation en santé. *Anesth Réanimation* 2016;2:88–95.
120. Raynal F, Rieunier A. *Pédagogie, dictionnaire des concepts clés: apprentissage, formation, psychologie cognitive*. 10e ed. ESF Editeur, 2014.
121. Piaget J. *L'équilibration des structures cognitives: problème central du développement*. Paris: Presses universitaires de France, 1975.
122. Sargeant J, Eva KW, Armson H, Chesluk B, Dornan T, Holmboe E, Lockyer JM, Loney E, Mann KV, Vleuten CPM van der. Features of assessment learners use to make informed self-assessments of clinical performance: Informed self-assessment in learners. *Med Educ* 2011;45:636–47.
123. Alves de Lima A, Conde D, Costabel J, Corso J, Van der Vleuten C. A laboratory study on the reliability estimations of the mini-CEX. *Adv Health Sci Educ* 2013;18:5–13.
124. Wilkinson JR, Crossley JGM, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Med Educ* 2008;42:364–73.
125. Margolis MJ, Clauser BE, Cuddy MM, Ciccone A, Mee J, Harik P, Hawkins RE. Use of the Mini-Clinical Evaluation Exercise to Rate Examinee Performance on a Multiple-Station Clinical Skills Examination: A Validity Study. *Acad Med* 2006;81:S56–60.
126. Weller JM, Jolly B, Misur MP, Merry AF, Jones A, Crossley JGM, Pedersen K, Smith K. Mini-clinical evaluation exercise in anaesthesia training. *Br J Anaesth* 2009;102:633–41.
127. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. *Adv Health Sci Educ Theory Pract* 2013;18:325–41.
128. Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G. Toward Authentic Clinical Evaluation: Pitfalls in the Pursuit of Competency. *Acad Med* 2010;85:780–6.
129. Messick S. Evidence and Ethics in the Evaluation of Tests. *Educ Res* 1981;10:9–20.

130. Der Sahakian G, Lecomte F, Buléon C, Guevara F, Jaffrelot M, Alinier G. Référentiel sur l'élaboration de scénarios de simulation en immersion clinique. Paris: Société Francophone de Simulation en Santé, 2017:22. Available at: <https://sofrasims.org/wp-content/uploads/2019/10/R%C3%A9f%C3%A9rentiel-Scenario-Simulation-Sofrasims.pdf>. Accessed February 15, 2021.
131. Gasia V. Aide cognitive pour scénariser une session de simulation en santé : proposition du Groupe de Travail. :3.
132. Adler MD, Trainor JL, Siddall VJ, McGaghie WC. Development and evaluation of high-fidelity simulation case scenarios for pediatric resident education. *Ambul Pediatr Off J Ambul Pediatr Assoc* 2007;7:182–6.
133. Boulet JR, Murray D, Kras J, Woodhouse J, McAllister J, Ziv A. Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. *Anesthesiology* 2003;99:1270–80.
134. Levine AI, Flynn BC, Bryson EO, Demaria S. Simulation-based Maintenance of Certification in Anesthesiology (MOCA) course optimization: use of multi-modality educational activities. *J Clin Anesth* 2012;24:68–74.
135. Blum RH, Boulet JR, Cooper JB, Muret-Wagstaff SL, Harvard Assessment of Anesthesia Resident Performance Research Group. Simulation-based assessment to identify critical gaps in safe anesthesia resident performance. *Anesthesiology* 2014;120:129–41.
136. Isaak RS, Chen F, Martinelli SM, Arora H, Zvara DA, Hobbs G, Stiegler MP. Validity of Simulation-Based Assessment for Accreditation Council for Graduate Medical Education Milestone Achievement: *Simul Healthc J Soc Simul Healthc* 2018;13:201–10.
137. Blum RH, Muret-Wagstaff SL, Boulet JR, Cooper JB, Petrusa ER. Simulation-based Assessment to Reliably Identify Key Resident Performance Attributes. *Anesthesiology* 2018;128:821–31.
138. Iramaneerat C, Yudkowsky R. Rater errors in a clinical skills assessment of medical students. *Eval Health Prof* 2007;30:266–83.
139. Rizzolo MA, Kardong-Edgren S, Oermann MH, Jeffries PR. The National League for Nursing Project to Explore the Use of Simulation for High-Stakes Assessment: Process, Outcomes, and Recommendations: *Nurs Educ Perspect* 2015;36:299–303.
140. Mudumbai SC, Gaba DM, Boulet JR, Howard SK, Davies MF. External validation of simulation-based assessments with other performance measures of third-year anesthesiology residents. *Simul Healthc J Soc Simul Healthc* 2012;7:73–80.
141. Chiu M, Tarshis J, Antoniou A, Bosma TL, Burjorjee JE, Cowie N, Crooks S, Doyle K, Dubois D, Everett T, Fisher R, Hayter M, McKinnon G, Noseworthy D, O'Regan N, Peachey G, Robitaille A, Sullivan M, Tenenbein M, Tremblay M-H. Simulation-based assessment of anesthesiology residents' competence: development and implementation of the Canadian National Anesthesiology Simulation Curriculum (CanNASc). *Can J Anesth Can Anesth* 2016;63:1357–63.
142. Frey-Vogel AS, Scott-Vernaglia SE, Carter LP, Huang GC. Simulation for Milestone Assessment: Use of a Longitudinal Curriculum for Pediatric Residents. *Simul Healthc J Soc Simul Healthc* 2016;11:286–92.
143. INACSL Standards Committee. INACSL Standards of Best Practice: Simulation. *Simulation Design. Clin Simul Nurs* 2016;12:S5–12.

144. Kelly MA, Mitchell ML, Henderson A, Jeffrey CA, Groves M, Nulty DD, Glover P, Knight S. OSCE best practice guidelines—applicability for nursing simulations. *Adv Simul* 2016;1:10.
145. Scavone BM, Sproviero MT, McCarthy RJ, Wong CA, Sullivan JT, Siddall VJ, Wade LD. Development of an objective scoring system for measurement of resident performance on the human patient simulator. *Anesthesiology* 2006;105:260–6.
146. Brydges R, Hatala R, Zendejas B, Erwin PJ, Cook DA. Linking Simulation-Based Educational Assessments and Patient-Related Outcomes: A Systematic Review and Meta-Analysis. *Acad Med* 2015;90:246–56.
147. Cazzell M, Howe C. Using Objective Structured Clinical Evaluation for Simulation Evaluation: Checklist Considerations for Interrater Reliability. *Clin Simul Nurs* 2012;8:e219–25.
148. Maignan M, Viglino D, Collomb Muret R, Vejux N, Wiel E, Jacquin L, Laribi S, N-Gueye P, Joly L-M, Dumas F, Beaune S, IRU-SFMU Group. Intensity of care delivered by prehospital emergency medical service physicians to patients with deliberate self-poisoning: results from a 2-day cross-sectional study in France. *Intern Emerg Med* 2019;14:981–8.
149. Adler MD, Vozenilek JA, Trainor JL, Eppich WJ, Wang EE, Beaumont JL, Aitchison PR, Pribaz PJ, Erickson T, Edison M, McGaghie WC. Comparison of Checklist and Anchored Global Rating Instruments for Performance Rating of Simulated Pediatric Emergencies. *Simul Healthc J Soc Simul Healthc* 2011;6:18–24.
150. Schwid HA, Rooke GA, Carline J, Steadman RH, Murray WB, Olympio M, Tarver S, Steckner K, Wetstone S, Anesthesia Simulator Research Consortium. Evaluation of anesthesia residents using mannequin-based simulation: a multiinstitutional study. *Anesthesiology* 2002;97:1434–44.
151. Alcaraz-Mateos E, Jiang X “Sara,” Mohammed AAR, Turic I, Hernández-Sabater L, Caballero-Alemán F, Párraga-Ramírez MJ, Poblet E. A novel simulator model and standardized assessment tools for fine needle aspiration cytology training. *Diagn Cytopathol* 2019;47:297–301.
152. Gerard JM, Kessler DO, Braun C, Mehta R, Scalzo AJ, Auerbach M. Validation of Global Rating Scale and Checklist Instruments for the Infant Lumbar Puncture Procedure: *Simul Healthc J Soc Simul Healthc* 2013;8:148–54.
153. Ghaderi I, Vaillancourt M, Sroka G, Kaneva PA, Vassiliou MC, Choy I, Okrainec A, Jacob Seagull F, Sutton E, George I, Park A, Brintzenhoff R, Stefanidis D, Fried GM, Feldman LS. Evaluation of surgical performance during laparoscopic incisional hernia repair: a multicenter study. *Surg Endosc* 2011;25:2555–63.
154. IJgosse WM, Leijte E, Ganni S, Luursema J-M, Francis NK, Jakimowicz JJ, Botden SMBI. Competency assessment tool for laparoscopic suturing: development and reliability evaluation. *Surg Endosc* 2020;34:2947–53.
155. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach* 2011;33:206–14.
156. Wunder LL, Glymph DC, Newman J, Gonzalez V, Gonzalez JE, Groom JA. Objective Structured Clinical Examination as an educational initiative for summative simulation competency evaluation of first-year student registered nurse anesthetists' clinical skills. *AANA J* 2014;82:419–25.

157. Boet S. Etude de l'Intérêt de l'Auto-Evaluation Dans l'Enseignement par Simulation. 2010. Available at: http://shs-app.univ-rouen.fr/civiic/memoires_DEA/textes/T_BOET.pdf. Accessed July 8, 2021.
158. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA* 2009;302:1316–26.
159. Atesok K, Hurwitz S, Anderson DD, Satava R, Thomas GW, Tufescu T, Heffernan MJ, Papavassiliou E, Theiss S, Marsh JL. Advancing Simulation-Based Orthopaedic Surgical Skills Training: An Analysis of the Challenges to Implementation. *Adv Orthop* 2019;2019:1–7.
160. Hultin M, Jonsson K, Härgestam M, Lindkvist M, Brulin C. Reliability of instruments that measure situation awareness, team performance and task performance in a simulation setting with medical students. *BMJ Open* 2019;9. Available at: <https://bmjopen.bmj.com/content/9/9/e029412>. Accessed February 2, 2020.
161. Evaluation and management of cauda equina syndrome in the emergency department | Elsevier Enhanced Reader. Available at: <https://reader.elsevier.com/reader/sd/pii/S0735675719305352?token=A58E58F938DB02EA7D3A195688597096EB4BE1D7BB8600EB4AD4E14B29D811F3FD5B670AF7575571F7E4B70C85F3A133>. Accessed November 12, 2020.
162. Everett TC, McKinnon RJ, Ng E, Kulkarni P, Borges BCR, Letal M, Fleming M, Bould MD, Marsh C, Heather D, Colovic V, Kulcsar Z, Boyle R, for the MEPA Collaborators. Simulation-based assessment in anesthesia: an international multicentre validation study. *Can J Anesth Can Anesth* 2019;66:1440–9.
163. Matsuda T, McDougall EM, Ono Y, Hattori R, Baba S, Iwamura M, Terachi T, Naito S, Clayman RV. Positive correlation between motion analysis data on the LapMentor virtual reality laparoscopic surgical simulator and the results from videotape assessment of real laparoscopic surgeries. *J Endourol* 2012;26:1506–11.
164. Commission Nationale de l'Informatique et des Libertés. Règlement européen sur la protection des données : ce qui change pour les professionnels., 2018. Available at: <https://www.cnil.fr/fr/reglement-europeen-sur-la-protection-des-donnees-ce-qui-change-pour-les-professionnels>. Accessed July 8, 2021.
165. Commission Nationale de l'Informatique et des Libertés. RGPD : passer à l'action. 2021. Available at: <https://www.cnil.fr/fr/rgpd-passer-a-laction>. Accessed July 8, 2021.
166. Commission Nationale de l'Informatique et des Libertés. Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés., 2019. Available at: <https://www.cnil.fr/fr/la-loi-informatique-et-libertes>.
167. Commission Nationale de l'Informatique et des Libertés. Article 6 - Licéité du traitement., 2018. Available at: <https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre2#Article6>.
168. Sando CR, Coggins RM, Meakim C, Franklin AE, Gloe D, Boese T, Decker S, Lioce L, Borum JC. Standards of Best Practice: Simulation Standard VII: Participant Assessment and Evaluation. *Clin Simul Nurs* 2013;9:S30–2.
169. Sørensen JL, Vleuten C van der, Rosthøj S, Østergaard D, LeBlanc V, Johansen M, Ekelund K, Starkopf L, Lindschou J, Gluud C, Weikop P, Ottesen B. Simulation-based multiprofessional obstetric anaesthesia training conducted in situ versus off-site leads to similar individual and team outcomes: a randomised educational trial. *BMJ Open* 2015;5:e008344.

170. Henneman EA, Cunningham H, Fisher DL, Plotkin K, Nathanson BH, Roche JP, Marquard JL, Reilly CA, Henneman PL. Eye Tracking as a Debriefing Mechanism in the Simulated Setting Improves Patient Safety Practices. *Dimens Crit Care Nurs* 2014;33:129–35.
171. Khan R, Payne MWC, Chahine S. Peer assessment in the objective structured clinical examination: A scoping review. *Med Teach* 2017;39:745–56.
172. Hegg RM, Ivan KF, Tone J, Morten A. Comparison of peer assessment and faculty assessment in an interprofessional simulation-based team training program. *Nurse Educ Pract* 2019;42:102666.
173. Wikander L, Bouchoucha SL. Facilitating peer based learning through summative assessment - An adaptation of the Objective Structured Clinical Assessment tool for the blended learning environment. *Nurse Educ Pract* 2018;28:40–5.
174. Li S-TT, Favreau MA, West DC. Pediatric resident and faculty attitudes toward self-assessment and self-directed learning: a cross-sectional study. *BMC Med Educ* 2009;9:16.
175. Plant JL, Corden M, Mourad M, O'Brien BC, Schaik SM van. Understanding self-assessment as an informed process: residents' use of external information for self-assessment of performance in simulated resuscitations. *Adv Health Sci Educ Theory Pract* 2013;18:181–92.
176. Weller J, Henderson R, Webster CS, Shulruf B, Torrie J, Davies E, Henderson K, Frampton C, Merry AF. Building the Evidence on Simulation Validity: Comparison of Anesthesiologists' Communication Patterns in Real and Simulated Cases. *Anesthesiology* 2014;120:142–8.
177. Vleuten CPM van der, Schuwirth LWT. Assessment in the context of problem-based learning. *Adv Health Sci Educ Theory Pract* 2019;24:903–14.
178. Pelgrim E a. M, Kramer AWM, Mokkink HGA, Elsen L van den, Grol RPTM, Vleuten CPM van der. In-training assessment using direct observation of single-patient encounters: a literature review. *Adv Health Sci Educ Theory Pract* 2011;16:131–42.
179. Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teach Learn Med* 2006;18:50–7.
180. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *J Gen Intern Med* 2009;24:74–9.
181. Fowlkes J, Dwyer DJ, Oser RL, Salas E. Event-Based Approach to Training (EBAT). *Int J Aviat Psychol* 1998;8:209–21.
182. Ludbrook J, Marshall VR. Examiner training for clinical examinations. *Br J Med Educ* 1971;5:152–5.
183. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;15:270–92.
184. Evans LV, Morse JL, Hamann CJ, Osborne M, Lin Z, D'Onofrio G. The development of an independent rater system to assess residents' competence in invasive procedures. *Acad Med J Assoc Am Med Coll* 2009;84:1135–43.
185. McLaughlin K, Ainslie M, Coderre S, Wright B, Violato C. The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. *Med Educ* 2009;43:989–92.

186. Gaugler BB, Rudolph AS. The influence of assessee performance variation on assessors' judgments. *Pers Psychol* 1992;45:77–98.
187. Feldman M, Lazzara EH, Vanderbilt AA, DiazGranados D. Rater Training to Support High-Stakes Simulation-Based Assessments. *J Contin Educ Health Prof* 2012;32:279–86.
188. Hedge JW, Kavanagh MJ. Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *J Appl Psychol* 1988;73:68–73.
189. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: A quantitative review. *J Occup Organ Psychol* 1994;67:189–205.
190. Carlson J, Tomkowiak J, Knott P. Simulation-based examinations in physician assistant education: A comparison of two standard-setting methods. *J Physician Assist Educ Off J Physician Assist Educ Assoc* 2010;21:7–14.
191. Lecomte F, Berton J, Blanié A, Herbreteau F, Perrot C, Sécheresse T, Der Sahakian G. Compétences transmises lors des Formations Courtes: Formateurs en Simulation en Santé. Paris: Société Francophone de Simulation en Santé, 2018. Available at: <http://www.sofrasims.fr/medias/files/recos-formation-sofrasims-c-nov-2018-post-db-post-experts.pdf>. Accessed November 3, 2019.
192. Eppich W, Nannicelli AP, Seivert NP, Sohn M-W, Rozenfeld R, Woods DM, Holl JL. A rater training protocol to assess team performance. *J Contin Educ Health Prof* 2015;35:83–90.
193. Oriot D, Bridier A. Development and Assessment of an Evaluation Tool for Team Clinical Performance: The Team Average Performance Assessment Scale (TAPAS). *Health Care Curr Rev* 2016;4. Available at: <http://www.esciencecentral.org/journals/development-and-assessment-of-an-evaluation-tool-for-team-clinicalperformance-the-team-average-performance-assessment-scale-tapas-2375-4273-1000164.php?aid=72394>. Accessed February 2, 2020.
194. Franc JM, Verde M, Gallardo AR, Carenzo L, Ingrassia PL. An Italian version of the Ottawa Crisis Resource Management Global Rating Scale: a reliable and valid tool for assessment of simulation performance. *Intern Emerg Med* 2017;12:651–6.
195. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *Br Med J* 1975;1:447–51.
196. Navas-Ferrer C, Urcola-Pardo F, Subirón-Valera AB, Germán-Bes C. Validity and Reliability of Objective Structured Clinical Evaluation in Nursing. *Clin Simul Nurs* 2017;13:531–43.
197. Nguyen LN, Tardioli K, Roberts M, Watterson J. Development and incorporation of hybrid simulation OSCE into in-training examinations to assess multiple CanMEDS competencies in urologic trainees. *Can Urol Assoc J J Assoc Urol Can* 2015;9:32–6.
198. 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:e1447–63.
199. Kane MT. Current Concerns in Validity Theory. *J Educ Meas* 2001;38:319–42.
200. Tavakol M, Dennick R. Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations – a commentary on two AMEE Guides. *Med Teach* 2012;34:245–8.

201. Messick S. Validity. In: Education measurement. 3rd ed. American Council on Education and Macmillan American Council on Education and Macmillan. New York: R. L. Linn, 1989:13–103.
202. Amin Z, Boulet JR, Cook DA, Ellaway R, Fahal A, Kneebone R, Maley M, Ostergaard D, Ponnampuruma G, Wearn A, Ziv A. Technology-enabled assessment of health professions education: Consensus statement and recommendations from the Ottawa 2010 conference. *Med Teach* 2011;33:364–9.
203. Coderre S, Woloschuk W, McLaughlin K. Twelve tips for blueprinting. *Med Teach* 2009;31:322–4.
204. Boulet JR, Murray D. Review article: assessment in anesthesiology education. *Can J Anaesth J Can Anesth* 2012;59:182–92.
205. Vargas AL, Boulet JR, Errichetti A, Zanten M van, López MJ, Reta AM. Developing performance-based medical school assessment programs in resource-limited environments. *Med Teach* 2007;29:192–8.
206. Ministère des affaires sociales et de la santé, Ministère de la défense, Ministère de l'enseignement supérieur et de la recherche. Arrêté du 8 avril 2013 relatif au régime des études en vue du premier et du deuxième cycle des études médicales., 2013. Available at: <https://www.legifrance.gouv.fr/loda/id/JORFTEXT000027343762/>.
207. Michaud P-A, Jucker-Kupper P. Principal Relevant Objectives and Framework for Integrated Learning and Education in Switzerland. Bern: Joint Commission of the Swiss Medical Schools, 2017:44. Available at: http://www.profilesmed.ch/doc/Profiles_2017.pdf.
208. Murray DJ, Boulet JR, Avidan M, Kras JF, Henrichs B, Woodhouse J, Evers AS. Performance of residents and anesthesiologists in a simulation-based skill assessment. *Anesthesiology* 2007;107:705–13.
209. Roberts C, Newble D, Jolly B, Reed M, Hampton K. Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Med Teach* 2006;28:535–43.
210. Pell G, Fuller R, Homer M, Roberts T, International Association for Medical Education. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. *Med Teach* 2010;32:802–11.
211. Tavakol M, Dennick R. Post-examination analysis of objective tests. *Med Teach* 2011;33:447–58.
212. Fehr JJ, Boulet JR, Waldrop WB, Snider R, Brockel M, Murray DJ. Simulation-based assessment of pediatric anesthesia skills. *Anesthesiology* 2011;115:1308–15.
213. Winkel AF, Gillespie C, Hiruma MT, Goepfert AR, Zabar S, Szyld D. Test of Integrated Professional Skills: Objective Structured Clinical Examination/Simulation Hybrid Assessment of Obstetrics-Gynecology Residents' Skill Integration. *J Grad Med Educ* 2014;6:117–22.
214. General Medical Council (Great Britain). Tomorrow's doctors. London: GMC, 2009.
215. Wilkinson TJ, Frampton CM, Thompson-Fawcett M, Egan T. Objectivity in Objective Structured Clinical Examinations: Checklists Are No Substitute for Examiner Commitment. *Acad Med* 2003;78:219–23.
216. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. *Med Educ* 2003;37:1012–6.

217. Morgan PJ, Cleave-Hogg D, Guest CB. A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. *Acad Med J Assoc Am Med Coll* 2001;76:1053–5.
218. Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. *Fam Med* 2008;40:574–8.
219. Morgan P, Cheung R, LeBlanc V, Haley M. Expert versus non-expert raters in simulation performance evaluation. *Can J Anaesth* 2008;55:4718651–2.
220. Prediger S, Fürstenberg S, Berberat PO, Kadmon M, Harendza S. Interprofessional assessment of medical students' competences with an instrument suitable for physicians and nurses. *BMC Med Educ* 2019;19:46.
221. Zanten M van, Boulet JR, Norcini JJ, McKinley D. Using a standardised patient assessment to measure professional attributes. *Med Educ* 2005;39:20–9.
222. Cohen R, Reznick RK, Taylor BR, Provan J, Rothma A. Reliability and validity of the objective structured clinical examination in assessing surgical residents. *Am J Surg* 1990;160:302–5.
223. Mann KV, MacDonald AC, Norcini JJ. Reliability of objective structured clinical examinations: Four years of experience in a surgical clerkship. *Teach Learn Med* 1990;2:219–24.
224. Humphrey-Murto S, Smee S, Touchie C, Wood TJ, Blackmore DE. A Comparison of Physician Examiners and Trained Assessors in a High-Stakes OSCE Setting: *Acad Med* 2005;80:S59–62.
225. Vleuten CP van der, Luyk SJ van, Ballegooijen AM van, Swanson DB. Training and experience of examiners. *Med Educ* 1989;23:290–6.
226. Tan CPL, Azila NMA. Improving OSCE examiner skills in a Malaysian setting. *Med Educ* 2007;41:517.
227. Weinger MB, Banerjee A, Burden AR, McIvor WR, Boulet J, Cooper JB, Steadman R, Shotwell MS, Slagle JM, DeMaria S, Torsher L, Sinz E, Levine AI, Rask J, Davis F, Park C, Gaba DM. Simulation-based Assessment of the Management of Critical Events by Board-certified Anesthesiologists. *Anesthesiology* 2017;127:475–89.
228. Humphrey-Murto S, Wood TJ, Touchie C. Why do physicians volunteer to be OSCE examiners? *Med Teach* 2005;27:172–4.
229. Lewis KL, Bohnert CA, Gammon WL, Hölzer H, Lyman L, Smith C, Thompson TM, Wallace A, Gliva-McConvey G. The Association of Standardized Patient Educators (ASPE) Standards of Best Practice (SOBP). *Adv Simul* 2017;2:10.
230. Williams RG. Have Standardized Patient Examinations Stood the Test of Time and Experience? *Teach Learn Med* 2004;16:215–22.
231. Stillman PL. Technical issues: logistics. *AAMC. Acad Med* 1993;68:464–8.
232. Humphris GM, Kaney S. Examiner fatigue in communication skills objective structured clinical examinations. *Med Educ* 2001;35:444–9.
233. Uzan S. Mission de recertification des médecins - Exercer une médecine de qualité | Vie publique.fr. Ministère des Solidarités et de la Santé - Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, 2018. Available at: <https://www.vie-publique.fr/rapport/37741-mission-de-recertification-des-medecins-exercer-une-medecine-de-qualit>. Accessed October 29, 2020.

234. Edgar L, McLean S, Hogan SO, Hamstra S, Holmboe ES. The Milestones Guidebook. Accreditation Council for Graduate Medical Education (ACGME), 2020;41. Available at: <https://www.acgme.org/Portals/0/MilestonesGuidebook.pdf?ver=2016-05-31-113245-103>.
235. Ziv A, Rubin O, Moshinsky A, Gafni N, Kotler M, Dagan Y, Lichtenberg D, Mekori YA, Mittelman M. MOR: a simulation-based assessment centre for evaluating the personal and interpersonal qualities of medical school candidates. *Med Educ* 2008;42:991–8.
236. Gessel E van, Picchiottino P, Doureradjam R, Nendaz M, Mèche P. Interprofessional training: Start with the youngest! A program for undergraduate healthcare students in Geneva, Switzerland. *Med Teach* 2018;40:595–9.
237. Raurell-Torredà M, Romero-Collado À, Bonmatí-Tomàs A, Olivet-Pujol J, Baltasar-Bagué A, Solà-Pola M, Mateu-Figueras G. Objective Structured Clinical Examination: An Assessment Method for Academic-Practice Partnerships. *Clin Simul Nurs* 2018;19:8–16.
238. Schoenmakers B, Wens J. The objective structured clinical examination revisited for postgraduate trainees in general practice. *Int J Med Educ* 2014;5:45–50.
239. Dagnone JD, Hall AK, Sebok-Syer S, Klinger D, Woolfrey K, Davison C, Ross J, McNeil G, Moore S. Competency-based simulation assessment of resuscitation skills in emergency medicine postgraduate trainees - a Canadian multi-centred study. *Can Med Educ J* 2016;7:e57-67.
240. Isaak R, Stiegler M, Hobbs G, Martinelli SM, Zvara D, Arora H, Chen F. Comparing Real-time Versus Delayed Video Assessments for Evaluating ACGME Sub-competency Milestones in Simulated Patient Care Environments. *Cureus* 2018;10:e2267.
241. Brett-Fleegler MB, Vinci RJ, Weiner DL, Harris SK, Shih M-C, Kleinman ME. A simulator-based tool that assesses pediatric resident resuscitation competency. *Pediatrics* 2008;121:e597-603.
242. Bodle JF, Kaufmann SJ, Bisson D, Nathanson B, Binney DM. Value and face validity of objective structured assessment of technical skills (OSATS) for work based assessment of surgical skills in obstetrics and gynaecology. *Med Teach* 2008;30:212–6.
243. Hove PD van, Tuijthof GJM, Verdaasdonk EGG, Stassen LPS, Dankelman J. Objective assessment of technical surgical skills. *Br J Surg* 2010;97:972–87.
244. Gardner AK, Scott DJ, Choti MA, Mansour JC. Developing a comprehensive resident education evaluation system in the era of milestone assessment. *J Surg Educ* 2015;72:618–24.
245. Montbrun SL de, Roberts PL, Lowry AC, Ault GT, Burnstein MJ, Cataldo PA, Dozois EJ, Dunn GD, Fleshman J, Isenberg GA, Mahmoud NN, Reznick RK, Satterthwaite L, Schoetz D, Trudel JL, Weiss EG, Wexner SD, MacRae H. A novel approach to assessing technical competence of colorectal surgery residents: the development and evaluation of the Colorectal Objective Structured Assessment of Technical Skill (COSATS). *Ann Surg* 2013;258:1001–6.
246. Szasz P, Grantcharov TP, Sweet RM, Korndorffer JR, Pedowitz RA, Roberts PL, Sachdeva AK. Simulation-based summative assessments in surgery. *Surgery* 2016;160:528–35.
247. Weber A, Domes C, Christian M, Coale M, Griffith C, O'Hara NN, Henn RF, O'Toole RV, Sciadini MF. Effect of Training Modules on Hip Fracture Surgical Skills Simulation Performance: Early Validation of the AAOS/OTA Simulator. *J Bone Joint Surg Am* 2019;101:2051–60.

248. Cychosz CC, Tofte JN, Johnson A, Gao Y, Phisitkul P. Fundamentals of Arthroscopic Surgery Training Program Improves Knee Arthroscopy Simulator Performance in Arthroscopic Trainees. *Arthrosc J Arthrosc Relat Surg Off Publ Arthrosc Assoc N Am Int Arthrosc Assoc* 2018;34:1543–9.
249. Lubowitz JH, Provencher MT, Brand JC, Rossi MJ. Learning the Language of Copernicus. *Arthrosc J Arthrosc Relat Surg Off Publ Arthrosc Assoc N Am Int Arthrosc Assoc* 2015;31:1423–5.
250. Aydin A, Raison N, Khan MS, Dasgupta P, Ahmed K. Simulation-based training and assessment in urological surgery. *Nat Rev Urol* 2016;13:503–19.
251. Winkel AF, Gillespie C, Uquillas K, Zabar S, Szyld D. Assessment of Developmental Progress Using an Objective Structured Clinical Examination-Simulation Hybrid Examination for Obstetrics and Gynecology Residents. *J Surg Educ* 2016;73:230–7.
252. Klerk S de, Kato PM. The Future Value of Serious Games for Assessment: Where Do We Go Now? *J Appl Test Technol* 2017;18:32–7.
253. Rupp M, McConnell DS, Smither JA. Determining skill transferability of action games as a method to reduce in-vehicle phone distractions. *Work Read Mass* 2012;41 Suppl 1:5880–1.
254. Gerard JM, Scalzo AJ, Borgman MA, Watson CM, Byrnes CE, Chang TP, Auerbach M, Kessler DO, Feldman BL, Payne BS, Nibras S, Chokshi RK, Lopreiato JO. Validity Evidence for a Serious Game to Assess Performance on Critical Pediatric Emergency Medicine Scenarios. *Simul Healthc J Soc Simul Healthc* 2018;13:168–80.
255. Banié A, Amorim M-A, Meffert A, Perrot C, Dondelli L, Benhamou D. Assessing validity evidence for a serious game dedicated to patient clinical deterioration and communication. *Adv Simul Lond Engl* 2020;5:4.
256. Botezatu M, Hult H, Tessma MK, Fors UGH. Virtual patient simulation for learning and assessment: Superior results in comparison with regular course exams. *Med Teach* 2010;32:845–50.
257. Board of Directors of the American Board of Medical Specialties (ABMS). Standards for the ABMS Program for Maintenance of Certification (MOC). American Board of Medical Specialties, 2014:13. Available at: <https://www.abms.org/media/1109/standards-for-the-abms-program-for-moc-final.pdf>. Accessed November 2, 2020.
258. ANDPC. Agence nationale du Développement Professionnel Continu: Engagée pour un DPC de qualité. Available at: <https://www.agencedpc.fr/>. Accessed November 2, 2020.
259. Weeks S, Robinson D, Tilley DS. The ABCs of organizational credentialing. *Nurs Manag (Harrow)* 2007;38:28–32, 43–4.
260. Bedford G, Allen S, Slawomirski L. From the Australian Commission on Safety and Quality in Health Care. *Med J Aust* 2012;196:156.
261. Hodges B, McNaughton N, Regehr G, Tiberius R, Hanson M. The challenge of creating new OSCE measures to capture the characteristics of expertise. *Med Educ* 2002;36:742–8.

Groupe d'élaboration du référentiel

Conduite des travaux

François LECOMTE, coordinateur du projet.

Médecin Urgentiste, APHP, CHU COCHIN-HOTEL DIEU,

Responsable du Diplôme Universitaire de Formateurs à l'Enseignement de la Médecine par la Simulation, Université Paris Descartes

A réalisé de nombreuses formations courtes de formateurs en simulation en Santé dans le monde et participé à la mise en place de centres de simulation, membre de la SOFRASIMS; francois.lecomte@cch.aphp.fr

Groupes de travail

Groupe 1 : **« Que peut-on évaluer ? »** : Fernande LOIS, Anne BELLOT, Isabelle CRUBLÉ, Guillaume PHILIPPOT, Thierry VANDERLINDEN.

Groupe 2 : **« Outils d'évaluation pour l'évaluation sommative »** : Anne-Laure PHILIPPON, Sébastien BATRANCOURT, Claire BOITHIAS-GUEROT, Jean BRÉAUD, Philine de VRIES, Louis SIBERT.

Groupe 3 : **« Conséquences de l'évaluation sommative »** : Olivier BRISSAUD, Thierry SÉCHERESSE, Virginie BOULANT, Louis DELAMARRE, Laurent GRILLET.

Groupe 4 : **« Scénarios pour l'évaluation sommative »** : Antoine LEFEVRE-SCELLES, François LECOMTE, Marianne JUND.

Groupe 5 : **« Débriefing, vidéo et recherche et évaluation sommative »** : Erwan GUILLOUËT, François LECOMTE.

Groupe 6 : **« Formateurs pour l'évaluation sanctionnante »** : Clément BULÉON, François LECOMTE, Christophe MATHURIN.

Groupe 7 : **« Mise en place de l'évaluation sommative en simulation en santé »** : Laurent MATTATIA, Jacques BERTHOD, Blaise DEBIEN, Olivier GACIA.



CHU du Kremlin-Bicêtre
78, rue du Général Leclerc
94270 LE KREMLIN-BICÊTRE
<https://sofrasims.org/>