Can J Anesth/J Can Anesth (2019) 66:1431–1436
https://doi.org/10.1007/s12630-019-01489-3

EDITORIALS

High-stakes assessment in anesthesia via simulation: Are we there yet?

Melanie C. Wright, PhD

Received: 5 September 2019 / Revised: 5 September 2019 / Accepted: 5 September 2019 / Published online: 27 September 2019

For assessing competence in managing intraoperative anesthetic emergencies, it is difficult to imagine a better alternative than simulation. Knowledge-based tests or evaluation of routine practice are likely to be inadequate.[1] While simulation has been widely embraced for formative assessment in anesthesiology, there remains hesitation to incorporate simulation into summative and, in particular, high-stakes summative assessment.[2-5] There is good reason for this—high-stakes assessment involves significant professional and financial consequences for examinees.[6] It is critical that the high-stakes assessment methodology attains a level of validity and fairness above and beyond that which may be appropriate for formative assessment.

In this issue of the *Journal*, Everett *et al*.[7] report on a multi-site validation of simulation-based assessment using checklist scoring of key actions in managing pediatric anesthesia emergencies. The Managing Emergencies in Pediatric Anesthesia (MEPA) scenarios (anaphylaxis, equipment failure, hypovolemia, local anesthesia toxicity, laryngospasm, retained throat pack, malignant hyperthermia) and scoring checklists were developed from literature review and iterative refinement with over 30 pediatric anesthesiologists participating in the MEPA national committee.[8] This expert-based development process, with iterative content refinement after pilot testing, was identified by the authors as the primary source of evidence for content validity of the assessment exercise.

M. C. Wright, PhD (✉)
Saint Alphonsus Regional Medical Center Research Institute, Trinity Health, 1055 N. Curtis Rd, Boise, ID 83706, USA
e-mail: Melanie.wright@trinity-health.org

The authors conducted a prospective, observational, research trial drawing on ongoing MEPA course training of junior residents, senior residents, and staff at nine centres in two countries to validate the MEPA simulation scenario checklists. They sought to: 1) establish construct validity of the MEPA scenarios and checklists with respect to differentiating the performance of junior trainees, senior trainees, and staff and 2) evaluate inter-rater reliability in scoring. They set a high bar, hypothesizing that the MEPA assessment would differentiate competence and be used to signify preparation for independent practice (high-stakes assessment).

Guidance on the use of simulation for assessment of managing critical incidents in anesthesiology promotes: 1) attention to rater training and inter-rater reliability, 2) scenarios and scoring designed to differentiate strong *vs* weak performers, 3) scenarios that progress in a consistent fashion independent of learner actions to ensure the same opportunity to gain marks, and 4) attention to inter-scenario reliability.[2,9] With respect to differentiating expertise, Boulet and Murray argue that checklist-based scoring typically fails to consider timing and sequence of actions.[9] A key difference between experts and novices may not be whether a specific action is undertaken, but instead how long it takes to identify a problem and take the appropriate action, particularly in emergencies.

With respect to evaluating inter-scenario reliability, McIndoe describes the need to evaluate Objective Structured Clinical Examination (OSCE) simulation stations for association with scores on other OSCE stations, overall scores, and subjective assessment. They promote a process of iterative improvement of stations to achieve high reliability.[2] Boulet and Murray assert that multiple performance samples are needed to ensure reliability of simulation scenarios and scoring and that

there is a need to evaluate consistency of scoring over multiple encounters.[9] As an example, Weller *et al.*[10] reported that 12-15 scenarios were needed to reliably rank trainees in the management of anesthesia incidents. In my own research with colleagues, we found that two raters and eight or more scenarios were needed to attain high levels of relative reliability in scoring key teamwork skills.[11] For high-stakes assessment, scenario design supporting multiple, shorter evaluations may be more effective than fewer, longer scenarios which may be more typical of formative assessment simulation scenarios.[11]

The scenarios evaluated by Everett *et al.* in this issue of the *Journal* were adapted from scenarios used in the context of formative assessment for MEPA training.[7] They were scripted, pre-programmed, and actors gave timed prompts. Timed arrival of help was defined. This ensured consistency between centres and a standardized scoring opportunity for all participants. Although the rationale for timed arrival of help was to avoid psychological distress in the event of a "patient death", this may also have served as a sort of time limit (presuming checklist scoring ended when help arrived). If well-timed, this could have resulted in faster performers (presumed to be more expert in managing emergencies) achieving higher scores.

The assessment checklists included a list of desirable actions for each scenario. The authors did not describe whether the action list was intended to be comprehensive or whether there was a focus on selecting items expected to differentiate strong from weak performers. For each action, participants were given a score of 0 (not done), 1 (done poorly or late), or 2 (done well). The ability to select 1 *vs* 2 based on timing of response provides an opportunity for raters to account for timeliness in scoring. There was a different number of actions for each scenario and scores were interpreted as the percentage of maximum possible points that could be obtained for each scenario. Raters also scored participants on a global rating scale informed by the checklist score, but the rater could adjust the score at their discretion, based on observations not captured by the checklist.

Although a power analysis projected that 40 staff anesthesiologists (the smallest participant group) would be necessary to detect a medium effect size difference between trainees and staff, Everett *et al*. ultimately enrolled 154 (89 junior, 65 senior) trainees and 21 staff. While individuals participated in all seven scenarios, they acted as the primary anesthesiologist (scored performer) in half. About 16% of simulation videos were not analyzed for technical reasons. Thus, each participant was scored on two, three, or four scenarios. The smallest sample size by scenario was nine (nine staff in three scenarios and nine senior trainees in one scenario).

The authors evaluated inter-rater reliability from five raters on a sample of 20 of each of the seven scenarios (140 total). They concluded that reliable rating of checklist and global rating scores can be achieved with two raters. The remaining 251 videos were scored by a single rater. Thus the expected reliability for the scores that were subsequently analyzed with respect to differentiating levels of expertise is expected to be 0.5 or higher, not the expected 0.8 that would be achieved had they used two raters.

Analysis of these scores revealed that junior trainees exhibited lower scores than both senior trainees and staff. There was no difference between scores of senior trainees and those of staff. While the authors were surprised that staff did not achieve higher levels of performance than senior trainees, this finding, and the authors' assertion of a ceiling effect, fits with my personal experience (unpublished data) of comparing resident and staff anesthesiologist performance in simulated adult anesthesia emergencies (checklists did not differentiate resident and staff performance, but time to a critical response in one of two scenarios did). Nevertheless, inadequate sample size for staff in this present study also raises the question of whether there was sufficient power to detect performance differences between senior trainees and staff.

The analysis also identified scenario-related differences in performance, which is expected given that scenarios were not designed to be similar in difficulty. The authors assert that the lack of a training level by scenario interaction suggests that there was a consistent training level effect across all scenarios. On close inspection of Figs 3 and 4,[7] and considering the small sample size by training level within each scenario, I question this assertion. I would argue that the only clearly supported finding of the analysis of variance evaluating performance by training level is that, for the full set of seven scenarios, junior trainees exhibited lower scores than both senior trainees and staff. The study is likely underpowered to support assertions of a null effect between senior trainees and staff or to assert that any one scenario would effectively differentiate performance between junior trainees and higher-level clinicians.

The study was not able to show inter-scenario reliability. In test design, it is expected that some questions (or scenarios) will be more difficult than others, supporting the ability to differentiate different levels of expertise. Nevertheless, a comprehensive analysis of the internal structure of an assessment should include an analysis of whether a person who performs well on a given scenario also performs well on other scenarios. Given the limitations that each participant performed only two to four scenarios each in the current study, the authors are not able to

establish internal structure validity of either the seven scenarios as a set or any individual scenario.

The authors report low performance scores across all training levels for a subset of scenarios. For example, a low-pass global rating scale score was not achieved by a median of staff anesthesiologists for managing hypovolemia, anaphylaxis, or equipment failure (see Fig. 4).[7] The authors argue that this suggests a practice performance deficit. Given their failure to establish internal structure validity of the scenarios, this conclusion may be premature. The finding does, however, raise important questions worthy of further study. For example, are these scenarios simply harder? Was the timing of help arrival too early in the scenario progression or was the deterioration of the patient too rapid for competent anesthesiologists to attain satisfactory performance? The authors suggest that for certain scenarios, skills may deteriorate over time. This raises other important questions. For example, is current training inadequate for these emergencies? Does the rarity of these emergencies lead to performance decrements due to lack of exposure, suggesting a need for more frequent re-training?

This study adds to the existing literature: 1) with good rater training, structured rating tools, and scenarios designed for consistency of scoring, it is feasible to establish high levels of inter-rater agreement with as few as two raters, and 2) a multi-scenario training set can differentiate performance of junior trainees from senior trainees and staff (and, by corollary, is believed to differentiate strong from weak candidates).[7,10,11] Nevertheless, the presentation of only one scenario for each emergency type, lack of evaluation of inter-scenario reliability, and low staff performance scores on a subset of scenarios raise questions about the validity of the MEPA scenarios and scenario scoring.

Is anesthesia ready for the use of simulation in high-stakes assessment of readiness for independent practice? I believe the answer is a qualified "Yes", as part of a more comprehensive assessment of breadth of knowledge and structured observation of supervised real-world clinical practice.[1,5] Moreover, I would argue that simulation-based assessment of critical incident management should be a necessary component of assessing readiness for independent practice.

Do we have the tools needed for high-stakes, simulation-based assessment in anesthesia? We know how to develop and validate scenarios and scoring tools for high-stakes assessment,[2,9] but evidence is lacking regarding a defined set of scenarios with structured scoring tools that meet the level of validity evidence needed for applying high-stakes pass-fail cutoffs. The American Board of Anesthesiology, in their ten-year Maintenance of Certification Requirements, addresses this limitation by requiring evidence of two simulation-based formative performance improvement activities, but not relying on any pass-fail scoring.[7] The Fellowship of the Royal College of Anesthesia addresses this limitation in their use of simulation-based OSCE by: 1) eliminating examiners' global ratings of performance from final scoring, 2) ensuring that an examinee cannot fail based on a single simulation assessment, and 3) scoring for ranking as opposed to pass-fail determination.[2]

For formative assessment, the MEPA scenarios and scoring evaluated in the *Journal* by Everett *et al*.,[7] are likely to be useful. Nevertheless, questions about the inter-scenario reliability and internal structure validity of MEPA scenarios and checklists suggest that more work is needed to refine and establish the validity of this scenario set before they are used for higher-stakes assessment, including trainee ranking.

Les évaluations à enjeux majeurs en anesthésie par la simulation : Sommes-nous prêts?

Pour évaluer la capacité d'une personne à prendre en charge des urgences anesthésiques peropératoires, il est difficile d'imaginer une meilleure option que la simulation. Les tests fondés sur les connaissances ou l'évaluation de la pratique de tous les jours ne suffiront probablement pas.[1] Alors que la simulation est aujourd'hui un outil répandu pour l'évaluation formative en anesthésiologie, nous hésitons encore à intégrer la simulation dans l'évaluation sommative et, en particulier, dans l'évaluation sommative à enjeux majeurs.[2-5] Il y a de bonnes raisons à cela – l'évaluation à enjeux majeurs a d'importantes conséquences professionnelles et financières pour les individus examinés.[6] Il est impératif que, lorsqu'il est question d'enjeux majeurs, la méthodologie d'évaluation atteigne un niveau de validité et d'impartialité bien au-delà de ce qui est considéré suffisant pour une évaluation formative.

Dans ce numéro du *Journal*, Everett *et coll*.[7] rapportent la validation multicentrique d'un outil d'évaluation fondé sur la simulation, outil qui s'appuie sur l'évaluation d'une liste de contrôle comportant les gestes clés pour la prise en charge d'urgences anesthésiques pédiatriques. Les auteurs ont conçu des mises en situation de Prise en charge des urgences en anesthésie pédiatrique (ou MEPA, pour *Managing Emergencies in Pediatric Anesthesia*) (anaphylaxie, bris d'équipement, hypovolémie, toxicité par anesthésique local, laryngospasme, rétention d'une

compresse pharyngée, hyperthermie maligne) et des listes de contrôles à noter à partir de revues de littérature et d'un processus itératif avec la participation de plus de 30 anesthésiologistes pédiatriques au niveau national de MEPA.[8] Ce processus de mise au point par les experts, suivi du perfectionnement itératif du contenu après avoir réalisé des tests pilote, a été décrit par les auteurs comme étant la source principale de données probantes appuyant la validité du contenu de l'exercice d'évaluation.

Les auteurs ont réalisé une étude de recherche prospective et observationnelle s'appuyant sur la formation en MEPA pour les résidents en début et en fin de résidence ainsi que pour les patrons de neuf centres dans deux pays afin de valider les listes de contrôle des mises en situation de simulation de MEPA. Leurs objectifs étaient : 1) d'établir la validité des mises en situation de MEPA et des listes de contrôle en matière de différenciation de la performance des résidents junior, senior et des patrons; et 2) d'évaluer la fiabilité inter-évaluateur en matière d'attribution de notes. La barre qu'ils se sont fixée était haute, en émettant l'hypothèse que l'évaluation de la MEPA permettrait de distinguer la compétence et pourrait ainsi être utilisée pour indiquer le degré de préparation d'un individu pour la pratique indépendante (une évaluation à enjeux majeurs).

Les suggestions sur l'utilisation de la simulation pour l'évaluation de la prise en charge des incidents critiques en anesthésiologie recommandent : 1) une attention à la formation de l'évaluateur et à la fiabilité inter-évaluateur, 2) des mises en situation et un système de notation conçus de façon à différencier les exécutants forts des faibles, 3) des mises en situation qui progressent de manière constante et indépendamment des gestes des exécutants afin de garantir à chacun les mêmes occasions de marquer des points, et 4) une attention à la fiabilité entre les mises en situation.[2,9] En ce qui touche à l'expertise en matière de différentiation, selon Boulet et Murray, les systèmes de notation fondés sur des listes de contrôle ne tiennent en général pas compte du moment où un geste est posé ni de la séquence de gestes.[9] En d'autres termes, la différence clé entre un expert et un novice pourrait résider non pas dans la pose d'un geste spécifique, mais plutôt dans le délai avant l'identification du problème et la prise de la mesure adaptée, particulièrement lorsqu'il s'agit d'une urgence.

En ce qui touche à l'évaluation de la fiabilité d'un mise en situation par rapport à une autre, McIndoe décrit la nécessité d'évaluer les stations de simulation d'Examen clinique objectif structuré (ECOS) pour déterminer leur association avec les notes obtenues à d'autres stations d'ECOS, avec la note globale et avec l'évaluation subjective. Ils promeuvent un processus d'amélioration itérative des stations afin d'atteindre une fiabilité élevée.[2] Selon Boulet et Murray, de nombreux échantillons de performance sont nécessaires afin de garantir une fiabilité des mises en situation de simulation et du système de notation; en outre, il est nécessaire d'évaluer la cohérence du système de notation sur plusieurs rencontres.[9] Par exemple, Weller *et coll.*[10] ont rapporté que 12-15 mises en situation étaient nécessaires afin de classer de façon fiable les résidents selon leur prise en charge des incidents anesthésiques. Dans le cadre de mes recherches avec mes collègues, nous avons observé que deux évaluateurs et huit mises en situation ou plus constituaient le minimum nécessaire pour obtenir des niveaux élevés de fiabilité relative lors de l'évaluation de compétences clés de travail en équipe.[11] En matière d'évaluation à enjeux majeurs, la conception de mises en situation compatibles avec des évaluations nombreuses et plus courtes pourrait être plus efficace que des mises en situation plus longues mais moins nombreuses, lesquelles seraient plus typiques des mises en situation de simulation destinées aux évaluations formatives.[11]

Les mises en situation évaluées par Everett *et coll.* dans ce numéro du *Journal* sont l'adaptation de mises en situation utilisées dans le contexte d'évaluations formatives pour la formation en MEPA.[7] Elles ont été orchestrées, préprogrammées, et des acteurs donnaient des répliques à des moments prédéterminés. L'arrivée programmée d'aide était prédéfinie. Ainsi, les auteurs ont pu garantir la cohérence entre les centres et ont donné une occasion de notation standardisée à tous les participants. Alors que la justification d'une arrivée programmée de l'aide était d'éviter toute détresse psychologique en cas de « décès du patient », cela a également pu servir de limite temporelle (en présumant que la notation sur la liste de contrôle prenait fin lors de l'arrivée de l'aide). Si l'aide arrivait au bon moment, cela pourrait avoir eu comme résultat des notes plus élevés pour les exécutants plus rapides (présumés comme étant plus experts dans la prise en charge des urgences).

Les listes de contrôle utilisées pour l'évaluation comportaient une liste de gestes désirables pour chaque mise en situation. Les auteurs n'ont pas mentionné si la liste de gestes avait pour objectif d'être exhaustive ou si l'emphase était plutôt mise sur la sélection des items qui pourraient différencier les intervenants forts des plus faibles. Pour chaque geste, les participants recevaient une note de 0 (pas fait), 1 (mal fait ou trop tard), ou 2 (bien fait). La capacité à choisir 1 *vs* 2 selon le délai de réponse de l'exécutant était une occasion pour les évaluateurs de tenir compte du délai d'action dans leur notation. Chaque mise en situation comportait un nombre différent de gestes et les notes étaient interprétées en tant que pourcentage des points possibles maximaux pouvant être obtenus pour chaque mise en situation. Les évaluateurs ont également noté les participants sur une échelle d'évaluation globale selon la note de la liste de contrôle, mais ils pouvaient également ajuster cette note à leur discrétion, selon des observations que la liste de contrôle ne pouvait pas saisir.

Bien qu'une analyse de puissance ait projeté que 40 anesthésiologistes en pratique (le plus petit groupe de participants) seraient nécessaires pour détecter une différence moyenne de taille d'effet entre les résidents et les patrons, Everett *et coll*. ont finalement enrôlé 154 résidents (89 junior, 65 senior) et 21 patrons. Bien que les intervenants aient participé aux sept mises en situation, ils n'ont joué le rôle d'anesthésiologiste principal (c'est-à-dire d'intervenant noté) que dans la moitié des mises en situation. Environ 16 % des vidéos de simulation n'ont pas été analysées pour des raisons techniques. Ainsi, chaque participant était noté sur deux, trois ou quatre mises en situation. La plus petite taille d'échantillon par mise en situation était de neuf personnes (neuf patrons dans trois mises en situation et neuf résidents senior dans une mise en situation).

Les auteurs ont évalué la fiabilité inter-évaluateur de cinq évaluateurs sur un échantillon de 20 simulations de chacune des sept mises en situation (soit 140 simulations au total). Ils ont conclu qu'une évaluation fiable des listes de contrôle et des notes d'évaluation globales pouvait être obtenue avec deux évaluateurs. Les 251 vidéos restantes ont été notées par un seul évaluateur. Ainsi, la fiabilité anticipée pour les notes analysées par la suite en matière de différenciation des niveaux d'expertise devrait être de 0,5 ou plus, et non de 0,8, comme cela aurait été le cas s'il y avait eu deux évaluateurs.

L'analyse de ces notes a révélé que les résidents junior ont obtenu des notes plus basses que les résidents plus avancés et les patrons. Aucune différence n'a été observée entre les notes obtenues par les résidents senior et les patrons. Alors que les auteurs étaient surpris que les patrons n'obtiennent pas des niveaux plus élevés de performance que les résidents senior, cette constatation, ainsi que la déclaration des auteurs d'un effet plafond, reflète mon expérience personnelle (données non publiées) lorsque j'ai comparé la performance des résidents et des patrons en anesthésiologie dans des urgences simulées en anesthésie adulte (les listes de contrôle n'ont pas différencié la performance des résidents et des patrons, mais le délai avant une réponse critique dans l'une des deux mises en situation a permis de les différencier). Toutefois, la trop petite taille d'échantillon pour les patrons dans l'étude présentée ici soulève également la question de savoir si l'étude disposait de suffisamment de puissance pour discerner des différences de performance entre les résidents senior et les patrons.

L'analyse a également identifié des différences de performance en fonction de la mise en situation, ce qui était à prévoir étant donné que les mises en situation n'étaient pas conçues pour être de difficulté semblable. Selon les auteurs, l'absence du niveau de formation pour l'interaction entre les mises en situation indiquerait qu'il y avait un effet du niveau de formation constant dans toutes les mises en situation. En examinant de plus près les Figures 3 et 4,[7] et étant donné la petite taille d'échantillon par niveau de formation au sein de chaque mise en situation, je me questionne sur cette affirmation. Selon moi, le seul résultat clairement étayé de l'analyse de la variance évaluant la performance par niveau de formation est que, dans l'ensemble des sept mises en situation, les résidents junior ont affiché des scores plus bas que les résidents senior et les patrons. L'étude manque probablement de puissance pour soutenir l'affirmation d'un effet nul entre les résidents senior et les patrons ou pour affirmer qu'une quelconque mise en situation puisse effectivement différencier la performance des résidents junior de celle de cliniciens de plus haut niveau.

L'étude n'a pu démontrer de fiabilité d'une mise en situation par rapport à une autre. Dans la conception d'un test, on s'attend à ce que certaines questions (ou mises en situation) soient plus difficiles que d'autres, ce qui permet de différencier les différents niveaux d'expertise. Toutefois, une analyse exhaustive de la structure interne d'une évaluation devrait inclure une analyse examinant si une personne qui réussit bien dans une mise en situation réussit également bien dans d'autres mises en situation. Étant donné les limites inhérentes à l'étude présentée ici, imposées par le fait que chaque exécutant n'a participé qu'à deux à quatre mises en situation, les auteurs n'ont pas été en mesure de déterminer la validité de la structure interne des sept mises en situation en tant qu'ensemble, ni individuellement.

Les auteurs rapportent des notes de performance basses, tous niveaux de formation confondus, pour un sous-ensemble de mises en situation. Par exemple, une note de passage basse sur l'échelle d'évaluation globale n'a pas été obtenue par une médiane de patrons anesthésiologistes pour la prise en charge de l'hypovolémie, de l'anaphylaxie ou de bris d'équipement (voir Figure 4).[7] Selon les auteurs, cela suggèrerait un déficit de performance de la pratique clinique. Cependant, si l'on considère qu'ils n'ont pas réussi à établir la validité de la structure interne des mises en situation, cette conclusion semble prématurée. Par contre, cette observation soulève des questions importantes dignes d'être approfondies. Par exemple, ces mises en situation sont-elles tout simplement plus difficiles? Est-ce que le moment programmé pour l'arrivée d'aide était trop précoce dans la progression de la mise en situation, ou la détérioration du patient était-elle trop rapide pour qu'un anesthésiologiste compétent obtienne une performance satisfaisante? Les auteurs suggèrent que, dans certaines des mises en situation, les compétences pourraient se détériorer au fil du temps. Cela soulève d'autres questions importantes. Par exemple, la formation actuelle est-elle mal adaptée à ces urgences? Est-

ce que la rareté de telles urgences entraîne une diminution de la performance en raison d'un manque d'exposition, ce qui suggèrerait le besoin de recyclage plus fréquent?

Voici quelques éléments que cette étude ajoute à la littérature existante : 1) si l'on dispose d'évaluateurs bien formés, d'outils d'évaluation structurés et de mises en situation conçues de manière à favoriser une cohérence dans la notation, il est possible d'établir des niveaux élevés de concordance inter-évaluateur avec seulement deux évaluateurs; et 2) un ensemble de formation disposant de multiples mises en situation peut distinguer la performance des résidents junior de celle des résidents senior et des patrons (et, par extension, pourrait probablement différencier les candidats forts des plus faibles).[7,10,11] Toutefois, la présentation d'une seule mise en situation pour chaque type d'urgence, le manque d'évaluation de la fiabilité d'une mise en situation à l'autre, et les faibles notes de performance des patrons dans un sous-ensemble de mises en situation soulèvent plusieurs questions quant à la validité des mises en situation en MEPA et du système de notation des mises en situation.

L'anesthésie est-elle prête à utiliser la simulation pour les évaluations à enjeux majeurs, examinant le niveau de préparation à la pratique indépendante? Selon moi, la réponse est un « oui » nuancé, c'est-à-dire dans le cadre d'une évaluation plus exhaustive de l'ampleur des connaissances et d'une observation structurée d'une pratique clinique réelle et supervisée.[1,5] En outre, selon moi, l'évaluation fondée sur la simulation de la prise en charge des incidents critiques devrait constituer une composante obligatoire de l'évaluation du niveau de préparation à la pratique indépendante.

Disposons-nous des outils nécessaires pour réaliser une évaluation fondée sur la simulation à enjeux majeurs en anesthésie? Nous savons comment mettre au point et valider des mises en situation et des outils de notation destinés à l'évaluation à enjeux majeurs,[2,9] mais nous manquons de données probantes concernant un ensemble défini de mises en situation possédant des outils de notation structurés et atteignant le niveau de validité nécessaire pour l'application des limites de passage ou d'échec lors d'enjeux majeurs. *L'American Board of Anesthesiology*, dans ses Exigences pour le maintien de la certification aux dix ans, résout ce problème en exigeant la preuve de deux activités formatives d'amélioration de la performance en simulation, mais ne s'appuie pas sur une note de passage ou d'échec.[7] Le Collège royal en anesthésie contourne cet écueil en utilisant des ECOS fondés sur la simulation en : 1) éliminant les notes globales de performance des évaluateurs de la note finale, 2) garantissant qu'un candidat ne peut échouer sur la base d'une seule évaluation en simulation, et 3) attribuant des notes visant à déterminer le rang plutôt qu'un passage ou un échec.[2]

En ce qui a trait à l'évaluation formative, les mises en situation de MEPA et les systèmes de notation évalués par Everett *et coll.*[7] dans le *Journal* seront sans doute utiles. Ceci étant, les questions concernant la fiabilité d'une mise en situation par rapport à une autre et la validité de structure interne des mises en situation de MEPA et des listes de contrôle suggèrent que des travaux supplémentaires sont nécessaires afin de perfectionner et d'établir la validité de cet ensemble de mises en situation avant de les utiliser pour réaliser des évaluations à enjeux majeurs, incluant le classement des résidents.

Conflicts of interest None declared.

Editorial responsibility This submission was handled by Dr. Steven Backman, Associate Editor, *Canadian Journal of Anesthesia.*

Conflit d'intérêt Aucun.

Responsabilité éditoriale Cet article a été traité par Dr Steven Backman, rédacteur adjoint, *Journal canadien d'anesthésie*.

References

1. *Tetzlaff JE*. Assessment of competency in anesthesiology. Anesthesiology 2007; 106: 812-25.
2. *McIndoe A*. High stakes simulation in anesthesia. Continuing Education in Anaesthesia Critical Care & Pain 2012; 12: 268-73.
3. *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.
4. *Cook DA*, *Brydges R*, *Zendejas B*, *Hamstra SJ*, *Hatala R*. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. Acad Med 2013; 88: 872-83.
5. *Ryall T*, *Judd BK*, *Gordon CJ*. Simulation-based assessments in health professional education: a systematic review. J Multidiscip Healthc 2016; 9: 69-82.
6. *Downing SM*. Validity: on meaningful interpretation of assessment data. Med Educ 2003; 37: 830-7.
7. *Everett TC, McKinnon RJ, Ng E, et al.* Simulation-based assessment in anesthesia: an international multicentre validation study. Can J Anesth 2019; 66. DOI: https://doi.org/10.1007/s12630-019-01488-4.
8. *Everett TC*, *Ng E*, *Power D*, *et al*. The Managing Emergencies in Paediatric Anaesthesia global rating scale is a reliable tool for simulation-based assessment in pediatric anesthesia crisis management. Paediatr Anaesth 2013; 23: 1117-23.
9. *Boulet JR*, *Murray DJ*. Simulation-based assessment in anesthesiology: requirements for practical implementation. Anesthesiology 2010; 112: 1041-52.
10. *Weller JM*, *Robinson BJ*, *Jolly B*, *et al*. Psychometric characteristics of simulation-based assessment in anaesthesia and accuracy of self-assessed scores. Anaesthesia 2005; 60: 245-50.
11. *Wright MC*, *Segall N*, *Hobbs G*, *Phillips-Bute B*, *Maynard L*, *Taekman JM*. Standardized assessment for evaluation of team skills: validity and feasibility. Simul Healthc 2013; 8: 292-303.