Can J Anesth/J Can Anesth (2019) 66:1440–1449
https://doi.org/10.1007/s12630-019-01488-4

REPORTS OF ORIGINAL INVESTIGATIONS

Simulation-based assessment in anesthesia: an international multicentre validation study

L'évaluation basée sur la simulation en anesthésie: une étude multicentrique internationale de validation

Tobias C. Everett, MBChB · Ralph J. McKinnon, MBChB · Elaine Ng, MD · Pradeep Kulkarni, MD · Bruno C. R. Borges, MD · Michael Letal, MD · Melinda Fleming, MD · M. Dylan Bould, MBChB · for the MEPA Collaborators

Received: 12 September 2018 / Revised: 14 June 2019 / Accepted: 14 June 2019 / Published online: 26 September 2019

Abstract

Purpose *Simulated clinical events provide a means to evaluate a practitioner's performance in a standardized manner for all candidates that are tested. We sought to provide evidence for the validity of simulation-based assessment tools in simulated pediatric anesthesia emergencies.*

Methods *Nine centres in two countries recruited subjects to participate in simulated operating room events. Participants ranged in anesthesia experience from junior residents to staff anesthesiologists. Performances were video recorded for review and scored by specially trained, blinded, expert raters. The rating tools consisted of scenario-specific checklists and a global rating scale that allowed the rater to make a judgement about the subject's performance, and by extension, preparedness for independent practice. The reliability of the tools was classified as "substantial" (intraclass correlation coefficients ranged from 0.84 to 0.96 for the checklists and from 0.85 to 0.94 for the global rating scale).*

Results *Three-hundred and ninety-one simulation encounters were analysed. Senior trainees and staff significantly out-performed junior trainees ($P = 0.04$ and $P < 0.001$ respectively). The effect size of grade (junior vs senior trainee vs staff) on performance was classified as "medium" (partial $\eta^2 = 0.06$). Performance deficits were observed across all grades of anesthesiologist, particularly in two of the scenarios.*

Conclusions *This study supports the validity of our simulation-based anesthesiologist assessment tools in several domains of validity. We also describe some residual challenges regarding the validity of our tools,*

The MEPA Collaborators are listed in "Acknowledgements".

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12630-019-01488-4) contains supplementary material, which is available to authorized users.

T. C. Everett, MBChB (✉) · E. Ng, MD
Department of Anesthesia and Pain Medicine, The Hospital for Sick Children, University of Toronto, 555 University Avenue, Toronto, ON M5G 1X8, Canada
e-mail: tobias.everett@sickkids.ca

R. J. McKinnon, MBChB
Department of Anesthesia, Royal Manchester Children's Hospital, Manchester, United Kingdom

P. Kulkarni, MD
Department of Anesthesia, Stollery Children's Hospital, University of Alberta, Edmonton, AB, Canada

B. C. R. Borges, MD
Department of Anesthesia, McMaster Children's Hospital, McMaster University, Hamilton, ON, Canada

M. Letal, MD
Department of Anesthesia, Alberta Children's Hospital, University of Calgary, Calgary, AB, Canada

M. Fleming, MD
Department of Anesthesia, Queens University, Kingston, ON, Canada

M. D. Bould, MBChB
Department of Anesthesia, Children's Hospital of Eastern Ontario, University of Ottawa, Ottawa, ON, Canada

Table 1 Messick's domains of assessment validity evidence[15]

Content	The extent to which all components of the assessment (stimuli, challenges, instruments, etc.) are representative of the intended construct
Response process	The extent to which the combination of participant behaviours and associated rater scores integrate to quantify the intended construct
Internal structure	The robustness of the performance of the rating instruments, when used by multiple raters or across multiple stimuli
Relation to other variables	The extent to which the instruments under investigation align with other tools thought to represent the same construct
Consequences	Arguments to justify a given "pass mark" and proposed inferences/future opportunities resulting from the candidate's performance in the assessment

some notes of caution in terms of the intended consequences of their use, and identify opportunities for further research.

Résumé

Objectif *Les événements cliniques simulés offrent la possibilité d'évaluer de façon standardisée la performance de tous les praticiens mis à l'épreuve. Notre objectif était de fournir des données probantes concernant la validité d'outils d'évaluation basés sur la simulation dans le contexte d'urgences anesthésiques pédiatriques simulées.*

Méthode *Neuf centres situés dans deux pays ont recruté des praticiens pour prendre part à des événements simulés en salle d'opération. L'expérience anesthésique des participants à la simulation allait de résidents juniors à patrons. Les simulations étaient enregistrées en format vidéo et ont été passées en revue et notées en aveugle par des évaluateurs experts spécialement formés. Les outils d'évaluation comprenaient des listes de contrôle spécifiques à chaque cas et une échelle d'évaluation globale, qui permettait à l'évaluateur de donner son avis sur la performance d'un sujet et, par extension, sur son état de préparation pour une pratique indépendante. La fiabilité des outils a été classée comme étant « substantielle » (les coefficients de corrélation intraclasse allaient de 0,84 à 0,96 pour les listes de contrôle et de 0,85 à 0,94 pour l'échelle d'évaluation globale).*

Résultats *Trois cent quatre-vingt-onze séances de simulation ont été analysées. Les résidents plus avancés et les patrons étaient clairement meilleurs que les résidents moins avancés ($P = 0,04$ et $P < 0,001$, respectivement). La taille d'effet de l'expérience (résident junior vs senior vs patron) sur la performance a été classée comme « moyenne » ($\eta 2$ partielle = 0,06). Des déficits de performance ont été observés dans tous les sous-groupes d'anesthésiologistes, particulièrement dans deux scénarios.*

Conclusion *Cette étude confirme la validité de nos outils d'évaluation de l'anesthésiologiste fondés sur la simulation dans plusieurs domaines de validité. Nous décrivons également quelques défis résiduels concernant la validité de nos outils, certaines mises en garde en termes des conséquences voulues de leur utilisation, et identifions diverses pistes de recherches futures.*

Simulation-based summative assessments of physician competence are currently a high research priority. In anesthesia, there have been multiple investigations of simulation-based assessments[1–11] and in some jurisdictions simulation is already a component of certification examinations[12–14] and maintenance of certification. If the consequences of a given assessment are important (e.g., licensing examinations), we must be confident that the inferences derived from the assessment tools are valid and that the scores reflect the candidate's level of competence.

Making arguments for the validity of an assessment is complex and multi-faceted. In his unified theory of validity, Messick proposes that all arguments attempt to establish the construct validity of a tool (i.e., whether it actually shows, quantifies, or delineates the defined construct) and that there are five domains in which evidence can support the construct validity of a given instrument (Table 1).[15] Although there are other validity frameworks, Messick's has been adopted by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education.

The Managing Emergencies in Pediatric Anesthesia (MEPA) collaborative[16] is a global community of pediatric anesthesia educators. Originally, MEPA was a one-day simulation course aimed at teaching senior anesthesia trainees the medical management of pediatric operating room crises. The process by which the course content was constructed is described elsewhere.[17] Designed as an educational intervention, we have been studying the utilization of the standardized MEPA content for the purposes of simulation-based assessment. We previously

considered the reliability of assessment tools associated with the MEPA simulation scenarios,[17] but that study was limited in scope to assessing trainee anesthesiologists at a single institution. In the current study, we extended recruitment to include anesthesiologists with a wide range of experience, from nine MEPA institutions. Our objective was to provide evidence that would support or refute the construct validity (considered under Messick's domains) of our simulation-based assessment tools. Our principle construct was defined as "competence in the management of clinical emergencies in pediatric anesthesia". The purpose of the instruments was to provide evidence of readiness for independent practice in each of the topics covered. Our primary hypothesis was that our combination of simulations and rating tools would be able to identify an anesthesiologist who is competent in the management of these operating room emergencies and, by extension, is adequately prepared for independent practice in that regard.

Methods

Nine tertiary-level university-affiliated pediatric hospitals at which MEPA is available (seven in Canada and two in the United Kingdom) participated in this prospective observational trial. Research Ethics Board approval was obtained at each of the participating institutions (available as Electronic Supplementary Material, eTable).

Investigators at each centre recruited anesthesia providers of all grades to participate in the study, from junior trainees to long-qualified staff anesthesiologists. Trainees and staff participated in separate sessions. At each of the centres, the MEPA course is provided regularly to trainees as a component of their educational curriculum. During the two-year study period (July 2014 to June 2016), trainees were approached by an investigator (with no influence on their program evaluation or career progression) to seek their consent for their simulation to be video recorded and included in the study. Staff participants were all pediatric anesthesiologists in university-affiliated children's hospitals and were invited by standard email from a study investigator unknown to them. The scenarios provided to the staff were identical in every way to the trainees' scenarios. No participant had previous experience of the MEPA course (as instructor or participant). Trainees were classified by their years of postgraduate training rather than by a country-specific grade/rank. Participants signed an informed consent form and were briefed on the objectives of the education session and the study. The participants also signed a confidentiality agreement to not discuss the scenarios outside of the session, to help ensure naiveté of the participant pool. They were then orientated to the simulation environment, equipment, and mannequin. The simulation environment was standardized to include an operating room table, an anesthesia workstation and supply cart (consistent with the participants' home institution) and all other relevant operating room supplies and equipment. The audiovisual capture configuration was standardized. The computerized mannequin used at each site was a SimBaby (Laerdal Medical, Stavanger, Norway). During the study period we wanted to maintain equitable access to the educational benefit of the MEPA course, within the confines of available simulation laboratory time. To allow all trainees to pass through a MEPA course (and thus all derive an educational benefit), participants would attend in pairs and be the primary anesthesiologist in half the scenarios and a passive (non-contributing) observer in the other half. As we were studying seven scenarios, each participant was the primary anesthesiologist in three or four scenarios (anaphylaxis, equipment failure, hypovolemia, local anesthetic toxicity, laryngospasm, retained throat pack, malignant hyperthermia). For consistency, staff were similarly scheduled. The order of scenarios assigned to each participant as the primary anesthesiologist was randomized using an online random number generator stored in a password-protected spreadsheet and only accessed by simulation laboratory instructors on the day of the activity. The scenarios were scripted and pre-programmed for consistency between centres. Actors in the scenario gave timed prompts as dictated in the scenario timeline. Any request for help was acknowledged but aid was withheld until a pre-determined point in the scenario, so that all participants had an equal opportunity to complete key actions and accumulate points. The arrival

Table 2 Managing Emergencies in Pediatric Anesthesia (MEPA) global rating scale[17]

1	2	3	4	5	6
Very poor (Appears to be a novice)	Poor	Borderline and unsatisfactory	Borderline but satisfactory	Good	Excellent (appears to be highly expert)

Please rate the overall performance in this simulation scenario as 1–6.

Scores of 1–3 are unsatisfactory for an anesthesiologist in independent practice and would constitute a failing performance in a high-stakes examination such as a Royal College or board examination. Scores of 4–6 are satisfactory for an anesthesiologist in independent practice.

Table 3 Interrater reliability for each scenario for checklists (CL) and global rating scale (GRS). Statistics derived from the 140 videos that were rated by all fiver raters. 95% confidence intervals shown in brackets. $P < 0.001$ for every coefficient. (ICC = intraclass correlation coefficient)

Scenario	Scenario-specific CL ICC		GRS ICC	
	Individual measures	Average measures	Individual measures	Average measures
Anaphylaxis	0.51 (0.24 to 0.81)	0.84 (0.61 to 0.95)	0.53 (0.26 to 0.81)	0.85 (0.63 to 0.96)
Equipment failure	0.84 (0.68 to 0.94)	0.96 (0.91 to 0.99)	0.75 (0.54 to 0.91)	0.94 (0.85 to 0.98)
Hypovolemia	0.54 (0.25 to 0.83)	0.85 (0.63 to 0.96)	0.55 (0.25 to 0.84)	0.86 (0.62 to 0.96)
Local anesthesia toxicity	0.66 (0.39 to 0.88)	0.91 (0.76 to 0.97)	0.65 (0.39 to 0.87)	0.90 (0.76 to 0.97)
Laryngospasm	0.69 (0.39 to 0.90)	0.92 (0.77 to 0.98)	0.70 (0.45 to 0.90)	0.92 (0.80 to 0.98)
Retained throat pack	0.78 (0.56 to 0.93)	0.95 (0.87 to 0.99)	0.61 (0.34 to 0.85)	0.89 (0.72 to 0.97)
Malignant hyperthermia	0.55 (0.28 to 0.83)	0.86 (0.66 to 0.96)	0.61 (0.35 to 0.86)	0.89 (0.73 to 0.97)

Table 4 Predicted reliability. Average measures and individual measures values were generated using all the raters' data in the mixed effect model. The values for 2, 3, and 4 raters were generated using the Spearman–Brown prophecy formula

	Global rating scale reliability				
	Individual measures	Predicted 2-rater	Predicted 3-rater	Predicted 4-rater	Average measures (5 rater)
Intraclass correlation coefficient	0.66	0.80	0.86	0.89	0.91

of help was timed to avoid the psychologically distressing event of the patient "dying" and to allow the patient to be "saved". All participants received a 30-min debrief following each scenario during which trained instructors facilitated a reflective process to maximize their learning experience.

Sample size

Our previous work formed the foundation of our sample size calculation[17]. We expected our tools to show a "medium" effect size in distinguishing performances between grades of practitioner. We also anticipated that we would be able to recruit more trainees than staff. Accordingly, we made the sample size calculation based on a trainee:staff ratio of 4:1. With a two-tailed α error of 0.05, 160 trainee subjects/scenarios and 40 staff subjects/scenarios yielded an 80% power of detecting a partial η^2 effect size of 0.06.

Outcome measures

We used the checklists (CL) and global rating scale (GRS) published previously in our pilot study.[17] The CL comprise a list of desirable actions, each of which could be scored by the raters as 0 = not done, 1 = done poorly or late, or 2 = done well. As there are a different number of action points on each scenario (ranging from 10 to 18), the CL score was converted to a percentage of maximum score available for that scenario. Raters were instructed to use the GRS to give their overall impression of the participant's performance. This gestalt judgement scored the quality of the participant's performance from novice to expert (Table 2). The GRS score was to be informed by, but not necessarily proportional to, the participant's CL score (i.e., raters were free to assign a higher GRS to a lower CL scoring participant or assign a lower GRS score to a participant who actually accumulated a reasonably high CL score).

Raters

It was important to offer participants anonymity with respect to the raters. We used five expert raters (three in UK, one in Canada, one in New Zealand) unknown to participants with no influence on their professional standing or career progression. In this context, we defined *expert* as a practicing pediatric anesthesiologist with experience in both simulation-based medical education and clinical performance rating. We conducted a total of four three-hour rater-training sessions using voice over internet protocol videoconferencing (Skype™, Luxemburg City, Luxemburg). Raters simultaneously viewed multiple sample videos of MEPA scenarios, scored them with the CL and GRS then discussed their rating rationale and factors which influenced their scores.

The raters shared ideas regarding the tools' fitness-for-purpose and agreed to rating rules for specific circumstances. Raters did not have access to the contents of the post-simulation debrief because we wanted scores to be based only on observable behaviours, as is the case with other assessment tools.

Following training, all five raters rated a subset of 140 videos (20 of each of the seven scenarios). Interrater reliability for the CL and the GRS for each scenario was described with intraclass correlation coefficients (ICC). Average and individual measures ICCs were calculated (Table 3). The average measures ICC is the observed reliability averaged across the five raters. The individual measures ICC is a calculated index of reliability for a single "typical" rater and represents the expected reliability if one rater was to solo-rate. We used a two-way, random effects model (and absolute agreement) to accommodate the repeated measures of conducting more than one simulation for each participant and considering our raters to be five of a large pool of potential raters with the same characteristics. We included the items used for each simulated scenario in the mixed effect model when this calculation was performed. Our selection in this regard was informed by the seminal paper on the topic by Shrout and Fleiss.[18] To predict the reliability of different numbers of raters scoring a given scenario, we used the Spearman–Brown prophecy formula. For ICC description of agreement, we used terms suggested by Landis and Koch[19,20] where an ICC > 0.80 suggests "near-perfect" agreement, 0.61–0.80 = "substantial" agreement, 0.41–0.60 = "moderate" agreement, 0.21–0.40 = "fair" agreement, 0.00–0.20 = "slight" agreement, and less than 0.00 = "poor" agreement.

Table 3 shows the interrater reliability by scenario, derived from the 140 videos assessed by all five raters. The individual measures ICC for the GRS was "substantial" in all but two scenarios. The correlation between scores on the GRS and the scenario-specific CL was very good ($r^2 = 0.74$). Based on these analyses, we progressed to solo-rating of the remaining videos, mindful of the possibility of rater attrition. Table 4 shows the projected reliability of the GRS for different numbers of raters using the Spearman–Brown prophecy formula. To achieve reliability on the GRS with ICC > 0.8, at least two raters are required.

Statistical analysis

Considering experience in anesthesia to be a categorical variable of three grades (junior trainee, senior trainee and staff), we looked at the mean and standard deviation of performance by grade and sought between-group differences using analysis of variance (ANOVA) (overall and by scenario). We defined junior trainee as up to three completed years in anesthesia training, senior trainee as more than three completed years (i.e., postgraduate year 4 and higher), and staff as holding an independent license to practice anesthesia. We used Bonferroni pairwise comparisons between grades to describe between-pair differences. We used a two-way mixed-model ANOVA to describe variation in GRS by grade and scenario, with a defined interaction term of scenario by grade. The mixed-model allows for the lack of independence between data points, given that there would be within-subject correlation when a subject participated in more than one scenario. We used partial η^2 to describe the effect size of grade on performance, because this accommodates the possibility that the nature of the scenario could also influence the performance and so first excludes the variance due to other variables (e.g., scenario). The magnitude of effect size (as suggested by Cohen) was determined by η^2: 0.01 = "small" effect size, 0.06 = "medium" effect size, and > 0.14 = "large" effect size.[21] We also used linear regression to describe the correlation between duration of anesthesia experience (expressed as a continuous variable: months) and score on the CLs and GRS. On visual inspection of these data, we found an inflection point in performance plotted against months' experience (i.e., a non-linear relationship, or plateau) so we also analyzed these data in two split-range categories: early career (< 100 months total experience in anesthesia, including as a trainee) *vs* established career (> 100 months total experience in anesthesia, including as a trainee). All statistical analyses were performed using Stata software (version 12.1, College Station, TX, USA).

Results

Across nine centres we collected data from 154 trainees and 21 staff who participated in 469 simulations. The shortfall from the target sample of 160 trainees and 40 staff was due to exhaustion of available resources at some of the study centres (simulation laboratory time, faculty time, saturation of participants). Of the 469 simulations, we included 391 videos in the final analysis (Table 5). Reasons for this decrement were multi-factorial. In some circumstances the participant form-based data were incomplete and could not be retrieved retrospectively. The commonest reason for excluding encounters was unsatisfactory audiovisual capture: video not captured or only partially captured; camera angles not conforming to standard; audio asynchrony; and poor audio quality. These losses were shared in similar proportions across the nine centres, so the suspicion of information bias was low. The losses meant that each participant contributed two, three, or

Table 5 Number of scenarios per grade included in the final analysis. Participants contributed 1–4 scenarios to the final data set

Scenario	Junior trainees (n = 89)	Senior trainees (n = 65)	Staff (n = 21)	Total (n = 175)
Anaphylaxis	31	18	10	59
Equipment failure	30	17	9	56
Hypovolemia	33	17	10	60
Local anesthetic toxicity	33	15	11	59
Laryngospasm	28	9	9	46
Retained throat pack	32	19	10	61
Malignant hyperthermia	27	14	9	50
Total	214	109	68	391

four simulations to the final analysis. Participants ranged in clinical experience from nine months in anesthesia training to over 30 years in anesthesia practice. Expressed as median [interquartile range (IQR)], junior trainees had 2 [1–3] years of experience in anesthesia, senior trainees had 4 [4–4.5] years of experience in anesthesia and staff had 15 [9.5–24] years of experience in anesthesia.

Bonferroni pairwise comparison showed that there was a difference between the performance (as rated by the GRS) of junior and senior trainees ($P = 0.04$) and between junior trainees and staff ($P < 0.001$). No difference in performance was observed between senior trainees and staff ($P = 0.33$). Regression analysis of performance (as rated by the GRS) by grade of practitioner revealed a "moderate" correlation ($r^2 = 0.21$). The effect size of grade on performance as rated by the GRS was "medium" (partial $\eta^2 = 0.06$). There was a degree of variation in performance by scenario, but, when we repeated the two-way ANOVA by grade and scenario with an interaction term for grade + scenario, there was no interaction ($P = 0.51$), suggesting that the association between grade and performance was similar across scenarios. Figure 1 shows scenario-specific performance by grade, as scored with the scenario-specific CL. Figure 2 shows performance

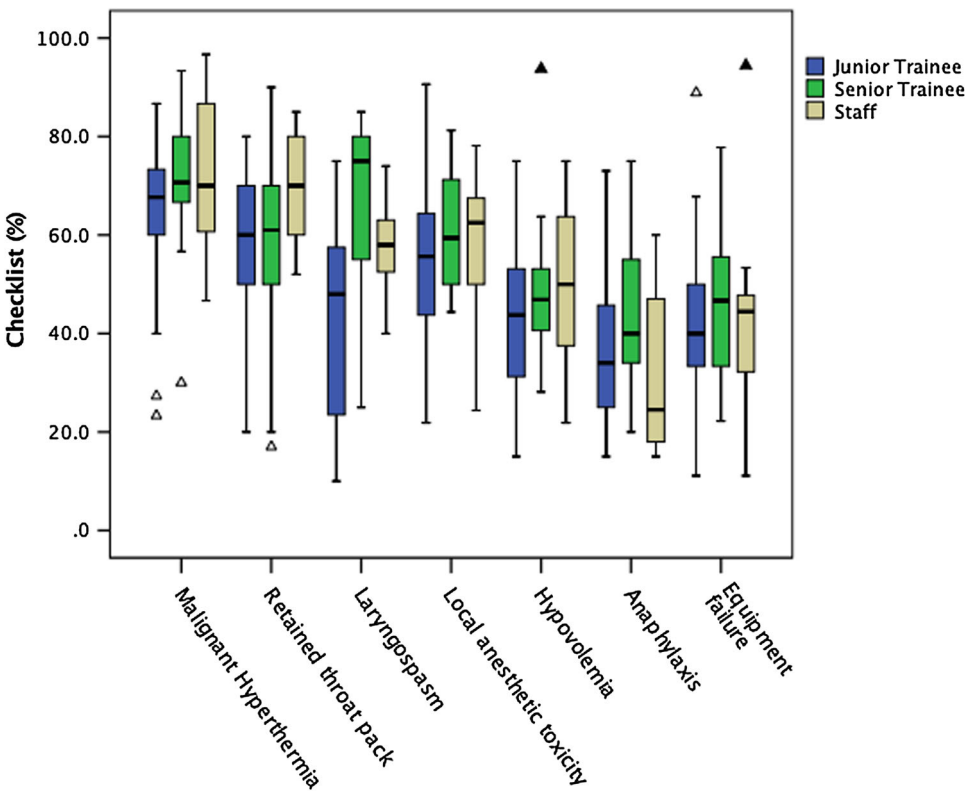

Fig. 1 Performance as rated by scenario-specific checklist (CL), displayed by grade of practitioner and scenario. Junior trainees in blue, senior trainees in green, staff in beige. As each scenario has a different number of action points, the CL score is expressed as a percentage of maximum score available for that scenario. Boxes delimited by interquartile range [IQR] with median marked as line within box. Whiskers show 1.5 × IQR, with triangles showing outliers

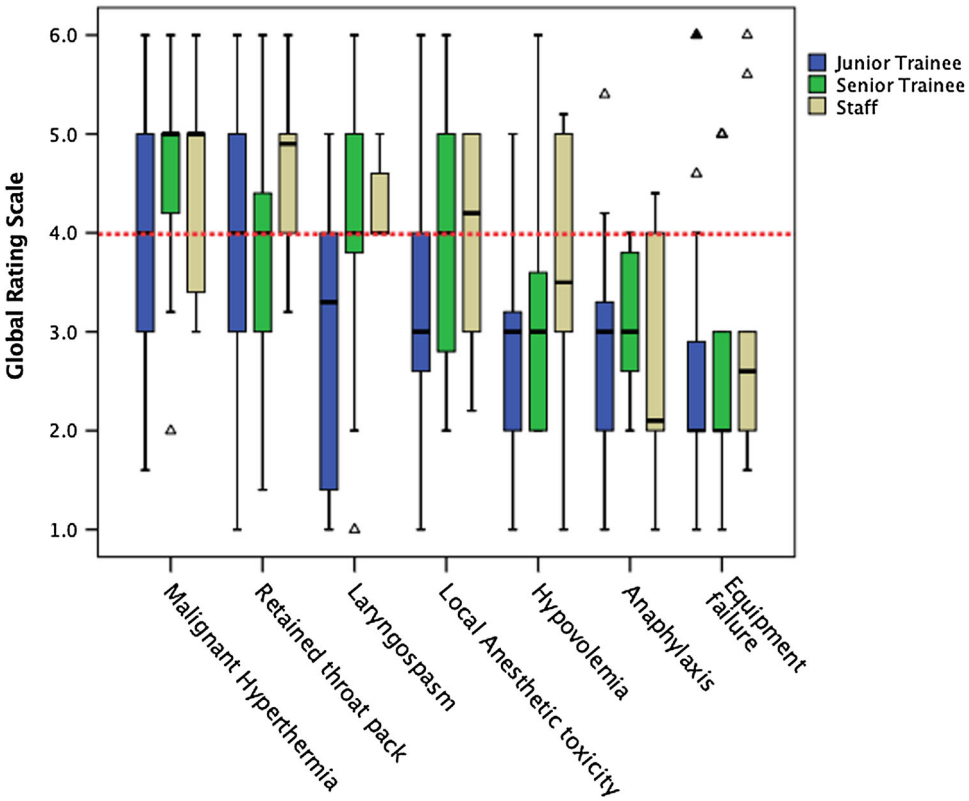

Fig. 2 Performance as rated by the global rating scale (GRS), displayed by grade of practitioner and scenario. Junior trainees in blue, senior trainees in green, staff in beige. A GRS of 4 (dotted line) signifies the standard expected of an independent anesthesiologist. Boxes delimited by interquartile range [IQR] with median marked as line within box. Whiskers show 1.5 × IQR, with triangles showing outliers

by scenario and by grade, as scored with the GRS. In the split-range analysis accounting for duration of anesthesia experience, we found that in the early-career stage (< 100 months' total experience in anesthesia), there is a weak correlation between months' experience in anesthesia and performance as graded by CL or GRS ($r^2 = 0.079$ and 0.081 respectively, $P < 0.001$). In the established-career phase (> 100 months in anesthesia), the regression analysis on the relationship between experience and performance as rated by the CLs and GRS showed no significant correlation ($r^2 = 0.012$, $P = 0.21$ and $r^2 = -0.002$, $P = 0.35$, respectively).

Discussion

In this study, evidence from nearly 400 simulation encounters in nine centres supports the reliability and validity of our simulation-based performance assessment tools. Using the framework proposed by Messick[15] and endorsed by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, we will now consider evidence of validity in five domains: content, response process, internal structure, relation with other variables, and consequences.

Content

The content of the MEPA course, the list of desired management points, and the rating tools themselves are based on assessment of needs, literature review, published guidelines, peer-review by a national committee of anesthesia educators, and piloting.[17] Based on what we learned in a pilot study, we refined the content of the simulation scenarios and the rating tools. As such, we consider the construct to be well-represented by the study assessments.

Response process

Both the examinees and the examiners had ample opportunity to contribute their opinions of the assessment process. The raters were experts in both clinical anesthesia and trainee education and assessment. They also contributed to the refinement of the tools. We undertook two analyses of interrater reliability during training: near the start and in the middle of the process. The sequential improvement of the ICC we observed confirmed the impact of the rater training on reliability. By these measures, we were satisfied with the integrity and efficacy of the rater-trainer process, an important component of response process validity. Nevertheless, the loss of data due to

problems with audiovisual capture represents a threat to validity in this domain. Despite familiarity with all components of the video capture process, and the efforts made to ensure technical success, we were unable to avoid these losses.

Internal structure

The combination of our raters, rating instruments, and rater-training processes yielded substantial interrater reliability. This is the second published study showing the reliability of these instruments. The raters were instructed to consider the CL and GRS scores independently. It is encouraging that there was a close correlation between CL and GRS scores ($r^2 = 0.74$) indicating that strong performers scored well using different methods of assessment. In 140 scenarios, all five raters scored the performances. These data, along with the Spearman–Brown formulaic prophecies indicate that an optimal combination of feasibility and reliability may be achieved with two raters per performance, as we have suggested previously.[17]

Relation with other variables

It is not surprising that staff and senior trainees outperformed the junior trainees. More puzzling is that performance of senior trainees and staff was similar. This may be explained by the learning trajectory of senior trainees, who, towards the end of postgraduate training, focus on competence in high-stakes, low-frequency events. Once in practice, the rarity of these emergencies is unlikely to confer additional experience and therefore staff might not outperform senior trainees in this dimension. We do not regard the inability to distinguish staff from senior trainees as a weakness (or failure) of the tool; rather, that alternative scenarios should be developed that can measure experience accrued with years in practice. It should be borne in mind that the tools are designed to allow the participant to demonstrate competence in these domains, not mastery of the specialty. The correlation between the CL scores and the GRS (two assessments designed to show the same construct) adds further validity in this regard.

Consequences

There were no consequences for the participants of our study, but based on our findings, these tools could potentially be used as part of the examination process at the end of residency training such that competence in these domains would be required to transition into independent practice. As there were no consequences for the participants in our study, we can make no further comment on our tools' validity in this domain.

The validity of simulation-based assessment has been subject to extensive investigation,[22,23] but the quality of published evidence has been variable.[23,24] Simulation-based assessments have been shown to correlate with real-life clinician performance and, in a few studies, with tangible patient outcomes.[25] In the discipline of anesthesia, there have been over twenty published investigations of the validity of simulation-based assessment tools. These studies have approached validity arguments in different ways, some using frameworks, some not. For example, a study by Blum and the Harvard Assessment of Anesthesia Resident Performance Research Group[9] used the Kane framework[26] as the basis for their validity arguments. They compared junior and senior scores on simulation scenarios to make inferences about construct representation, and generalizability theory to comment on reliability and applicability. They found an interaction between practitioner grade and scenario indicating that in their trial, performance not only depended on grade but also on the scenario encountered (as they varied in level of difficulty).[9] In our study, practitioner grade predicted performance; there was no significant interaction between grade and scenario ($P = 0.51$). This suggests that although there is a performance variation (by scenario) within in each grade, the effect is distributed uniformly across all levels of practitioner. In our study, the scenarios and tools were designed to be at a uniform level of difficulty (i.e., passable by end-of-training anesthesia residents). The fact that participants across all grades failed so frequently on the anaphylaxis and equipment failure indicates either a deficit in the rating tool, the rater, or the performance of the participant. We believe the scenario and rating tools to be robust and have therefore exposed a genuine deficit in practitioner performance in these emergencies, at least within the limitations of the simulated environment. Our original statistical analysis plan did not include provision for a generalizability analysis and indeed our final data set would not have satisfied the assumptions necessary to conduct one. Blum *et al.* presented some convincing arguments for the validity of their tools, but their participant group and intended construct was different than ours in that their objective was to identify unsafe gaps in resident performance rather than to identify staff-level performance.

Authorities responsible for assessing and certifying physician competence require a range of evidence to support or withhold the granting of a license to practice.

This can be accumulated from a variety of sources, each with associated merits and pitfalls. Certainly, with simulation-based assessment, there are feasibility and practicality considerations, but this should not prohibit its implementation. A common limitation among validation studies of simulation-based assessment is that they involve only trainees. To evaluate if a given tool can distinguish between a broader range of practitioners, validation studies must include licensed practitioners as a benchmark. We showed a ceiling effect of our scenarios and rating instruments in that established-career staff did not outperform early-career staff. We propose that this plateau in performance is unimportant insofar as licensing bodies are not looking to establish that practitioners have achieved mastery, simply that they have maintained the passing standard of competence of an independent anesthesiologist. We acknowledge that established-career staff may be less accustomed to, and more uncomfortable with simulation as a modality, and this may influence test scores unpredictably. Moreover, community or office-based anesthesiologists may have even less access to simulated or real crises, and so limit the applicability of our work in those contexts.

In 2015, The Royal College of Physicians and Surgeons of Canada moved to "Competence by Design", which involves the evolution of assessment tools that include simulation-based milestones. Our GRS has been adopted as the principal outcome measure for assessing residents in Canada-wide simulation-based milestones.[14] In the UK, simulation forms a component of the primary credentialing examinations[13] and in the US, simulation features in the Accreditation Council for Graduate Medical Education milestones for anesthesia;[27] although simulation is not yet being employed for scenario-based assessment purposes. Evaluating trainee competence is only one potential application of simulation-based assessment. There is precedent in several jurisdictions for using simulation as a component of maintenance of certification processes.

With a large sample size, a statistically significant result can be shown without much generalizable relevance (analogous to statistical *vs* clinical significance in clinical trials). For this reason (among others), it is important that effect sizes are considered alongside *P* values. In the current study, we showed a "medium" effect size of practitioners' grade on performance as rated by the GRS, which provides some reassurance in this regard. Although the simulated patients in our scenarios were pediatric, the scenarios are also plausible in adult patients. A universal concern with investigations about simulation is the extent to which assessment reflects real-life clinical performance and the impact on patient outcome. Similar criticisms may be made of other modes of practitioner assessment.

Conclusions

This study provides further evidence that the thoughtful combination of simulations, rating tools, and trained raters can be a useful instrument in the complex challenge of defining a practitioner's competence. We propose that simulation-based assessment can comprise a useful, informative component of multi-modal physician evaluation. Further research is required to reveal whether performance on multi-modal evaluations predicts future performance in the clinical realm and whether this affects patient care.

Acknowledgements Neil Cowie MD, Site lead investigator, Department of Anesthesia, University of Saskatchewan, Saskatoon, Canada. The MEPA Collaborators: Christopher Marsh, Department of Anesthesia, Royal United Hospital, Bath, Somerset, UK. David Heather, Department of Anesthesia, Middlemore Hospital, Auckland, New Zealand. Vesna Colovic, Department of Anesthesia, Royal Manchester Children's Hospital, Manchester, UK. Zsuzsanna Kulcsar, Department of Anesthesia, Royal Manchester Children's Hospital, Manchester, UK. Riley Boyle, Department of Anesthesia, Stollery Children's Hospital, Edmonton, Canada.

Conflict of interest The authors have no conflict of interest to declare. TE sits on a committee responsible for implementing the Canadian National Anesthesia Simulation Curriculum (CanNASC). Portions of this study have been presented at the International Meeting for Simulation in Healthcare (2016), Orlando, Florida, USA and the Canadian Anesthesiologists Society annual meeting (2016), Vancouver, British Columbia, Canada.

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

Author contributions *Tobias C. Everett* and *M. Dylan Bould* were involved in study conception and design, data cleaning, statistical analysis, and manuscript drafting. *Tobias C. Everett*, *Ralph J. McKinnon*, *Elaine Ng*, *Pradeep Kulkarni*, *Bruno C. R. Borges*, *Michael Letal*, *Melinda Fleming*, and M. *Dylan Bould* were involved in data acquisition, and in critical review and editing of the manuscript.

Funding This work was supported by: Canadian Anesthesiologists Society (Toronto, Canada) (CAS-2011-060); Royal College of Physicians and Surgeons of Canada Medical Education Research Grant (Ottawa, Canada) (2011MERG); Academy for Innovation in Medical Education (Ottawa, Canada); and Department of Anesthesia of the University of Ottawa (Canada) (#9680).

References

1. *Morgan PJ*, *Cleave-Hogg D*, *Guest CB*. A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. Academic Medicine 2001; 76: 1053-5.
2. *Morgan P*, *Cleave-Hogg D*, *DeSousa S*, *Tarshis J*. High-fidelity patient simulation: validation of performance checklists. Br J Anaesth 2004; 92: 388-92.

3. *Morgan PJ*, *Cleave-Hogg DM*, *Guest CB*, *Herold J*. Validity and reliability of undergraduate performance assessments in an anesthesia simulator. Can J Anesth 2001; 48: 225-33.
4. *Murray DJ*, *Boulet JR*, *Avidan M*, *et al*. Performance of residents and anesthesiologists in a simulation-based skill assessment. Anesthesiology 2007; 107: 705-13.
5. *Murray DJ*, *Boulet JR*, *Kras JF*, *Woodhouse JA*, *Cox T*, *McAllister JD*. Acute care skills in anesthesia practice: A simulation-based resident performance assessment. Anesthesiology 2004; 101: 1084-95.
6. *Schwid HA*, *Rooke GA*, *Carline J*, *et al*. Evaluation of anesthesia residents using mannequin-based simulation: a multiinstitutional study. Anesthesiology 2002; 97: 1434-44.
7. *Fehr JJ*, *Boulet JR*, *Waldrop WB*, *Snider R*, *Brockel M*, *Murray DJ*. Simulation-based assessment of pediatric anesthesia skills. Anesthesiology 2011; 115: 1308-15.
8. *Savoldelli GL*, *Naik VN*, *Joo HS*, *et al*. Evaluation of patient simulator performance as an adjunct to the oral examination for senior anesthesia residents. Anesthesiology 2006; 104: 475-81.
9. *Blum RH*, *Boulet JR*, *Cooper JB*, *Muret-Wagstaff SL*. Simulation-based assessment to identify critical gaps in safe anesthesia resident performance. J Am Soc Anesthesiol 2014; 120: 129-41.
10. *Murray DJ*, *Boulet JR*, *Kras JF*, *McAllister JD*, *Cox TE*. A simulation-based acute skills performance assessment for anesthesia training. Anesth Analg 2005; 101: 1127-34.
11. *Waldrop WB*, *Murray DJ*, *Boulet JR*, *Kras JF*. Management of anesthesia equipment failure: a simulation-based resident skill assessment. Anesth Analg 2009; 109: 426-33.
12. *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.
13. *McIndoe A*. High stakes simulation in anesthesia. Continuing Education in Anesthesia, Critical Care & Pain 2012; 12: 268-73.
14. *Chiu M*, *Tarshis J*, *Antoniou A*, *et al*. Simulation-based assessment of anesthesiology residents' competence: development and implementation of the Canadian National Anesthesiology Simulation Curriculum (CanNASC). Can J Anesth 2016; 63: 1357-63.
15. *Validity Messick S*. In: Linn R, editor. Educational Measurement. NY: Macmillan; 1989. p. 13-103.
16. *Everett TC*, *MacKinnon R*, *de Beer D*, *Taylor M*, *Bould MD*. Ten years of simulation-based training in pediatric anesthesia: The inception, evolution, and dissemination of the Managing Emergencies in Pediatric Anesthesia (MEPA) course. Pediatr Anesth 2017; 27: 984-90.
17. *Everett TC*, *Ng E*, *Power D*, *et al*. The Managing Emergencies in Pediatric Anesthesia global rating scale is a reliable tool for simulation-based assessment in pediatric anesthesia crisis management. Pediatr Anesth 2013; 23: 1117-23.
18. *Shrout PE*, *Fleiss JL*. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 1979; 86: 420-8.
19. *Landis JR*, *Koch GG*. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977; 33: 363-74.
20. *Landis JR*, *Koch GG*. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-74.
21. *Cohen J*. Statistical Power Analysis for the Behavioural Sciences. Hillsdale, NJ: Lawrence Erlbaum; 1988 .
22. *Ryall T*, *Judd BK*, *Gordon CJ*. Simulation-based assessments in health professional education: a systematic review. Journal of multidisciplinary healthcare 2016; 9: 69-82.
23. *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.
24. *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.
25. *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.
26. *Kane MT*. Current concerns in validity theory. J Educ Measurement 2006; 38: 319-42.
27. *Culley D*, *Cohen N*, *Hall S*, *et al*. The Anesthesiology Milestone Project. Chicago, US: Accreditation Council for Graduate Medical Education; 2015 .