Testing the validity of three acute care assessment tools for assessing residents' performance during in situ simulation: the ACAT-SimSit study

Anne-Laure Philippon[a,b], Antoine Lefevre-Scelles[c,d], Xavier Eyer[e,f], Carine Zumstein[g], Aiham Ghazali[h], Simon Audibert[i], Pierrick Le Borgne[j], Emmanuel Triby[b] and Jennifer Truchot[f,k]

Background **The assessment of technical and nontechnical skills in emergency medicine requires reliable and usable tools. Three Acute Care Assessment Tools (ACATs) have been developed to assess medical learners in their management of cardiac arrest (ACAT-CA), coma (ACAT-coma) and acute respiratory failure (ACAT-ARF).**

Objective **This study aims to analyze the reliability and usability of the three ACATs when used for in situ (bedside) simulation.**

Methods **This prospective multicenter validation study tested ACATs using interprofessional in situ simulations in seven emergency departments and invited training residents to participate in them. Each session was rated by two independent raters using ACAT. Intraclass correlation coefficients (ICC) were used to assess interrater reliability, and Cronbach's alpha coefficient was used to assess internal consistency for each ACAT. The correlation between ACATs' scores and the learners' level of performance was also assessed. Finally, a questionnaire and two focus groups were used to assess the usability of the ACATs.**

Results **A total of 104 in situ simulation sessions, including 85 residents, were evaluated by 37 raters. The ICC for ACAT-CA, ACAT-coma and ACAT-ARF were 0.95 [95% confidence interval (CI), 0.93–0.98], 0.89 (95% CI, 0.77–0.95) and 0.92 (95%CI 0.83–0.96), respectively. The Cronbach's alphas were 0.79, 0.80 and 0.73, respectively. The ACAT-CA and ARF showed good construct validity, as third-year residents obtained significantly higher scores than first-year residents ($P < 0.001$; $P < 0.019$). The raters supported the usability of the tools, even though they expressed concerns regarding the use of simulations in a summative way.**

Conclusion **This study reported that the three ACATs showed good external validity and usability.** ***European Journal of Emergency Medicine*** **XXX: XXXX–XXXX**

European Journal of Emergency Medicine XXX, **XXX**:XXXX–XXXX

Keywords: competency-based medical education, emergency medicine, simulation, based assessment

[a]Emergency Department, Pitié-Salpêtrière Hospital, Sorbonne Université, GRC 14, BIOFAST, AP-HP, Paris, [b]Laboratoire Interuniversitaire des Sciences de l'Education (LISEC) – Learning Sciences Department, Strasbourg University, Strasbourg, [c]Emergency Care Training Center (CESU-76A), Department of Anaesthesiology, Intensive Care and Emergency Medical Services, Rouen University Hospital, [d]Centre d'Enseignement des Soins d'urgence, Medical Training Center (MTC), Rouen University Hospital, Rouen, [e]Emergency Department, Lariboisière Hospital, Assistance Publique – Hôpitaux de Paris (AP-HP), [f]Faculté de Médecine, Université de Paris-Cité, Paris, [g]Unité de simulation Européenne en santé (UNISIMES), Faculté de Médecine, maïeutique et science de la santé, Université de Strasbourg, Strasbourg, [h]Emergency Department, Hôpital Bichat, Assistance Publique – Hôpitaux de Paris (AP-HP), [i]Emergency Department, Georges Pompidou European Hospital, Assistance Publique – Hôpitaux de Paris (AP-HP), Paris, [j]Emergency Department, Hôpitaux Universitaires de Strasbourg, Strasbourg and [k]Emergency Department, Cochin Hospital, Assistance Publique – Hôpitaux de Paris (AP-HP), Paris, France

Correspondence to Anne-Laure Philippon, MD, PhD, Service d'accueil des Urgences, CHU Pitié-Salpêtrière, 83, bd de l'hôpital, 75013 Paris, France
Tel: +33 18 482 7651; e-mail: annelaurephi@gmail.com

Received 8 November 2023 **Accepted** 21 December 2023.

Background

The assessment of clinical competence has evolved over the last decade from knowledge-based examinations to outcome-based assessments, aiming to assess overall competencies [1]. Developing competency-based assessments is a challenge, because the need for complex assessments aligned with a realistic professional environment must be addressed [2]. Simulation-based education enables contextualized and efficient training for health care professionals regarding a full range of clinical skills, particularly for complex tasks encountered in emergency medicine [3,4]. Simulation-based assessment allows for standardized, individualized assessments in a semi-authentic context [5,6]. In acute medicine, validated assessment tools are either task-specific and nontechnical skill checklists or global rating scales [7,8].

In order to combine the assessment of both technical and nontechnical skills in emergency medicine, two previous studies developed three tools [9]. Downing [10] described it in a framework that aimed to identify the validation process' steps through the analyses of five sources

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (www.euro-emergencymed.com).

0969-9546 Copyright © 2024 Wolters Kluwer Health, Inc. All rights reserved. DOI: 10.1097/MEJ.0000000000001133

of evidence: content, response process, psychometric properties, relationships to other variables and consequences. The contents of three Acute Care Assessment Tools (ACATs) were developed following this framework, using a Delphi study [11]. Three life-threatening emergency situations were chosen: cardiac arrest (ACAT-CA), nontraumatic coma (ACAT-coma) and acute respiratory failure (ACAT-ARF). The ACATs' response process and reliability were analyzed using video-recorded simulation sessions with medical students and emergency residents. A high interrater reliability was found for each tool [12].

This study aimed to analyze the three ACATs' validity and usability, when used during in situ simulation sessions.

Methods

Setting

A multicentric prospective observational study was conducted in seven French emergency departments (EDs) from seven teaching urban hospitals. The aim was to conduct at least 10 in situ simulations in each study center and perform a minimum of 30 simulations for each ACAT, with a set of various clinical contexts, either in prehospital or in-hospital settings. The Strasbourg University Institutional Review Board (Approval Unistra/CER/2020-17) approved the study and all participants provided their written consent.

Objectives

The primary objective of this study was to analyze the three ACATs' reliability when used in an in situ simulation setting. Secondary objectives included analyses of ACATs' predictive validity and correlation with the global performance scale (GPS). They also studied the ACAT's usability. The primary endpoint was the interrater reliability. The secondary endpoints were the correlation between ACACT's grades and students' year of residency, the correlation between ACAT's grades and the GPS and a questionnaire and focus groups addressing the usability of the tools.

Participants and simulation sessions

The participants were residents during their emergency rotation (postgraduate years 1, 2, or 3), emergency nurses, caregivers and ambulance drivers. All of them belonged to the seven EDs.

In situ simulations were used to analyze the ACATs in a realistic environment and assess emergency residents within an interprofessional team context. Five standardized and reproducible scenarios were created for each ACAT (Supplementary Annex 1, Supplemental digital content 1, *http://links.lww.com/EJEM/A433*).

The go/no-go criterion was determined to ensure the safety of in situ simulation sessions. They were related to the ED environment (overcrowding, heavy clinical load and equipment needs) or the ED staff (medical or nursing understaffing, unanticipated events/threats to psychological safety) [13]. Two simulation types were used: hybrid (with simulated patients and a task trainer) and all-synthetic (with a low-fidelity mannikin). Every in situ simulation session was performed in a standardized manner, in the following order: a prebriefing describing the simulation environment, the material that the learner could use, a brief presentation of the raters, a short case-related briefing followed by the scenario and a debriefing.

Trainers and raters

Each in situ simulation required the involvement of at least two trainers to rate residents' performances. They were not blinded to participants' years of residency or specialty. They could be emergency doctors or nurses, depending on whether they were trained in simulation-based education. Prior to scoring, each rater received brief practical training regarding the ACATs. The training was brief in order to replicate normal conditions.

The three ACATs

Each ACAT comprised 20 items, behaviorally anchored as follows: 1 point = completed, 0.5 point = performed but incomplete, 0 point = not performed/performed incorrectly and N/A = not required for the scenario (Supplementary Annex 2, Supplemental digital content 2, *http://links.lww.com/EJEM/A434*) [11,12]. Each item had a detailed description of its completeness and timeliness (if it was required). To describe learners' performance, a five-level GPS was added to the ACATs.

Data analysis

Prior to the analysis, all data were anonymized and compiled in MS Excel (Windows 10, Microsoft). Statistical analysis was performed using NCSS (Kaysville, Utah, USA). Descriptive statistics were computed for demographic data and expressed as means and SDs. Reliability was assessed with interrater reliability using intraclass correlation (ICC) and internal consistency was assessed using Cronbach's alpha. ICC indicates an agreement in the scoring between the raters. The ICC is considered excellent when it is above 0.8 and good when it is between 0.6 and 0.8 [14]. Cronbach's alpha demonstrates internal consistency if it is between 0.70 and 0.90. We also analyzed the interrater reliability of the GPS using Cohen's weighted kappa.

Finally, to determine the correlation between the level of performance and total ACAT scores, the construct validity was assessed using Pearson's correlation coefficient. Residents were allocated into three categories: First- and second-year emergency residents (PGY-1,2), third-year emergency residents (PGY-3) and nonemergency residents (PGY-NE) In this construct, PGY-3 should perform as well as, or better than, PGY-1,2 if the tool accurately assesses clinicians' emergency/acute care skills. The

quantitative usability of the tools was assessed based on completeness, as illustrated by the percentage of rated items, while their relevance was assessed based on the percentage of nonrated items.

To determine whether the ACATs were usable and useful for the raters, they responded to a poststudy questionnaire that included 18 questions rated on a four-point Likert scale, ranging from strongly disagree (1) to strongly agree (4). The questionnaire was based on a previous study [15].

Thematic analysis

A qualitative analysis was conducted using the grounded theory approach [16]. The reporting and analyzing data followed the COnsolidated criteria for REporting Qualitative research guidelines [16,17]. The focus group method was chosen to allow discussions and debates between raters. Two investigators constructed a semi-directed interview guide, which was reviewed by two other researchers. After the questionnaires were completed, the main investigator moderated the focus groups' interviews, as she had experience regarding the qualitative method. The interviews were audio-recorded, transcribed verbatim the week after the interviews, translated into English and analyzed separately by two independent researchers, who followed four open-coded steps: familiarization with transcripts, identification of the main themes, exploring the transcripts using an open software and discussion between the two researchers prior to the final analysis.

Results

Seven centers, 314 learners and 37 trainers participated in 104 in situ simulation sessions, among which 13 were in a prehospital environment (Table 1). The ACAT-CA, ACAT-coma and ACAT-ARF were used 43, 30 and 31 times, respectively. Six sessions (5%) were canceled. The median duration of the scenario was 15 min, with interquartile ranges of 10–15, 13.5–15 and 12–16 for ACAT-CA, ACAT-coma, and ACAT-ARF, respectively.

Reliability

ICC between the two independent raters for each in situ simulation session showed high interrater reliability for all ACATs. Cohen's weighted kappa was found to be good and demonstrated high interrater reliability for the GPS (Table 2). The internal consistency of each ACAT, analyzed by item-total correlation, was

Table 1 Participants

Emergency department	1	2	3	4	5	6	7	Total
Participants								
Residents								
Emergency medicine								67
PGY-1,2	3	9	6	4	3	1	4	30
PGY-3	7	1	1	1		14	13	37
Other Specialty (PGY-NE)								18
First year			1	4	4		3	12
Second year	2							2
Third year	4							4
Total residents	16	10	8	9	7	15	20	85
Nurses	22	30	12	21	22	33	21	161
Caregivers	16	4	6	5	9	14	6	60
Ambulance drivers							8	8
Total	54	44	26	35	38	62	55	314
Trainers								
Doctors	1	2	1	5	5	5	7	26
Nurses	1		3		4	2		10
Caregiver					1			1
Total	2	2	4	5	10	7	7	37

NE, nonemergency residents; PGY, postgraduate year.

Table 2 Acute Care Assessment Tool and global performance score reproducibility

	ACAT-CA		ACAT-coma		ACAT-ARF	
Interrater reliability	ICC	95% CI	ICC	95% CI	ICC	95% CI
R1–R2	0.95	0.93–0.98	0.89	0.77–0.95	0.92	0.83–0.96
	GPS-CA		GPS-coma		GPS-ARF	
	Kappa[a]	95% CI	Kappa[a]	95% CI	Kappa[a]	95% CI
R1–R2	0.80	0.67–0.93	0.76	0.60–0.92	0.84	0.70–0.98

ACAT, Acute Care Assessment Tool; ARF, acute respiratory failure ; CA, cardiac arrest; CI, confidence interval; GPS, global performance score; ICC, intraclass correlation coefficient; R1, rater no. 1: R2, rater no. 2.
[a]Kappa: weighted Cohen's kappa.

acceptable, with Cronbach's alpha coefficients of 0.79, 0.80 and 0.73 for the ACAT-CA, ACAT-coma and ACAT-ARF, respectively.

Relation to other variables: predictive and construct validity

The predictive validity of the ACATs was assessed by examining their associations with the GPS. A moderately strong significant positive correlation was found between the ACATs and the GPS in the total scores, with Pearson correlation coefficients of 0.6 (95% confidence interval (CI), 0.37–0.76), 0.72 (95% CI, 0.48–0.85) and 0.83 (95% CI, 0.66–0.91) for the ACAT-CA, ACAT-coma and ACAT-ARF, respectively. The construct validity, obtained by comparing the scores between PGY-1,2, PGY-3 and PGY-NE, showed that the PGY-3 students had significantly higher mean total scores than PGY-1,2 and PGY-NE students for ACAT-CA and ACAT-ARF. Conversely, there was no significant difference between the student levels for ACAT-coma scores (Table 3).

Usability

Completeness ranged from 96.5 to 100% among rated items, except for one which was rated in 87% of the scenarios. Among the 20 incomplete items, 19 were at least 97% complete and the last was rated in 87% of the cases (Supplementary Annex 3, Supplemental digital content 3, *http://links.lww.com/EJEM/A435*).

Out of the 37 educators who participated in the study, 18 (49%) completed the questionnaire and 10 participated in the two focus groups. These lasted 70 and 83 min, respectively. Both the questionnaire respondents and focus groups included participants from each center. Most of the raters felt that the tools were easy to use (83% for ACAT-CA and ACAT-ARF, and 67% for ACAT-coma) and would add value to the preexisting tools (61–72% of the raters). Most perceived that ACATs would be useful in rating residents in emergency medicine (78%) and effective in assessing the required performance and competencies (81–94%). Finally, the majority of raters (75%) agreed that ACATs were useful in providing formative feedback and helpful for debriefings.

Five themes emerged from the qualitative data exploring the raters' experiences of using the ACATs: usability, validity, perceived effects on the learners and raters, acceptability and use within an interprofessional environment. Raters' transcripts, with their different participation numbers, were reported as follows: rater (R) + participation number (from 1 to 10).

Raters were concerned about the usability of the tools: 'there are too many items in the tools, and some of them should be removed (R9)'; 'I found the tools' items clear, concise and easy to use, without risk of interpretation errors (R4)'. All the raters considered the three ACATs to be valid for assessing the required performances for each of the three clinical situations, with valid content enabling the simultaneous assessment of technical skills and non-technical skills. They agreed that the tools were relevant to assess emergency medicine residents' performance: 'tools are focused on the team leader and his/her communication (R2)'; 'it's adequate to assess a training program for new emergency medicine residents in my ED (R1)'. Each rater used the ACAT items to guide the debriefing, as it provided a debriefing structure that improved feedback quality: 'the tool provides objective input to give areas of improvement to the learners' (R7). However, they found the grades useless, except for a normative assessment. They also found that simulation-based assessment, which is not a prevalent practice in France, could be a threat to simulation-based education. Simulation-based assessment could compromise benevolence, which is one of the pillars of simulation as a teaching tool. Hence, even though they would use the ACATs tools, the raters addressed the acceptability of the overall simulation-based assessment and pointed out a loss of benevolence: "To me, it is not within the 'simulation spirit' (R5); 'to not give grades is to stay within a benevolent approach (R9)'; 'we could use grades to certify, but not to train' (R7).

Table 3 Scores' comparison according to the learner year level

	PGT-NE		PGY-1,2		PGY-3		
Score	Mean	SD	Mean	SD	Mean	SD	*P*
ACAT-CA	9.3	3.0	11.0	2.6	13.5	2.9	<0.001
ACAT-Coma	9.6	3.1	12.2	3.1	12.3	2.6	0.114
ACAT-ARF	10.9	2.2	11.4	3.2	13.9	2.0	0.019

ACAT, Acute Care Assessment Tool; ARF, acute respiratory failure ; CA, cardiac arrest; NE, nonemergency; PGY, postgraduate year.

Discussion

The ACAT-SimSit study showed that the three ACATs were reliable and valid for measuring technical and nontechnical skills across multiple clinical settings, after a thorough validation process. Interrater reliability regarding the ACATs and GPS was high, and reproducibility appeared to be minimally influenced by a high number of raters. Good item-total correlation, with high completeness and relevance, highlights the good internal consistency and acceptability of each tool. Moreover, the interviewed raters found the ACATs usable, helpful in providing feedback to students and fit for assessment. However, they expressed concerns about simulation-based assessment acceptability, particularly the introduction of grades and evaluative judgement, within a simulation-based training activity.

The study aimed to develop a tool fit for a competency-based approach with a high degree of continuity, aided by the use of a unique assessment tool with both technical

and nontechnical skills [18]. With different identified cutoff scores, the ACATs can be used during residency, allowing trainers and raters to share improvement goals with learners and to grade within a normative assessment perspective. This would favor educational continuity through teaching, training and practicing within curriculum activities [19].

The ACAT-SimSit confirmed the results of a previous study, which showed the limit of evaluating reproducibility through only three raters and video-recorded simulation sessions [12]. The involvement of numerous raters in real-life in situ simulation-based assessments suggests that the results may be transferable to broader assessment contexts. Moreover, the use of each ACAT in at least five different contexts could be validated, which is of the utmost importance within a competency-based approach [20].

The ACAT-coma's analyses failed to demonstrate its predictive validity. As the development process was the same for the two other tools, this could be explained through the participants' characteristics. PGY-3 participants are expected to perform better than PGY-1,2. However, as they have been classified according to their year of residency, not their level of competency, this postulate could be wrong. Moreover, there may be some differences between participants' prior experiences in simulation-based training, as the curriculum is heterogeneous. As theorized and highlighted by another validation process study, that found the same result for one of its tested tools, the difference in preassessment training with a simulation setting could explain the unexpected absence of level differences between two groups with two different clinical and academic levels [21].

Competence is 'a combination of different resources, skills and attitudes, which are inherently unobservable' [22]. Clinical competencies should be determined from explicit and understandable markers, such as performance and multiple assessment tools [23]. However, even if the three ACAT validation processes follow a rigorous method, it remains a tool designed to assess learner performance and would be obtained from the correlations between various assessment tools for the same student [24].

Finally, the raters expressed concerns regarding the use of simulation-based assessment to assess student performance. They questioned the consequences of simulation-based assessment in a learning context where establishing and maintaining psychological safety is mandatory [25]. One way of improving raters' acceptance could be the use of simulation-based assessment tools through a formative assessment before using it through a normative assessment with scores, to certify the learners. It allows learners and trainers to perceive the positive effects of the assessment on learning processes and student motivation [26].

Limitations

The use of real-life in situ simulations to analyze ACATs was interesting in terms of the practical application of the tools, but it might have introduced heterogeneity. The participants were not 'standardized' and had varying levels of clinical expertise and previous exposure to simulation-based education. This may have influenced their skill levels and the absence of differentiation when using the ACAT-coma. Second, the raters were not blinded regarding participants' year of residency, which could have influenced their rating. Moreover, as only two groups of emergency residents were created without stratification, the study could have missed the differences between PGY-1 and PGY-2 levels, influencing the overall comparison between PGY-1,2 and PGY-3. Finally, further research is needed to test the consequences of the scores and ACAT thresholds in different sets of students and emergency medicine residents.

Conclusion

The three ACATs demonstrated good external validity with good interrater reliability. Moreover, the three ACATs appeared to be usable by the raters, to assess emergency medicine residents.

Acknowledgements

The authors would like to thank FHU IMPEC for the grant that allowed the outfitting of all seven EDs with the simulation devices. They also thank all the trainers and raters who participated in the design and accomplishment of the in situ simulation sessions, and the ACAT-SimSit study group: Guillaume Payan, Aurore Kolmer, Sarah Uge-Ginsberg, Clémence Bertrand, Pauline Canavaggio, Mélanie Roussel, Martin Behr, Alexandre Bitoun, Isabelle Borraccio, Walid Berrezag and Anthony Chauvin. The authors would also like to thank Pr. Yonathan Freund for providing his companionship to this study.

This work was supported by the Federation Hospitalo-Universitaire "IMProve Emergency Care" (FHU IMPEC) [grant no. 2020_01].

Conflicts of interest

There are no conflicts of interest.

References

1 Batalden P, Leach D, Swing S, Dreyfus H, Dreyfus S. General competencies and accreditation in graduate medical education. *Health Aff (Millwood)* 2002; **21**:103–111.
2 Epstein RM. Assessment in medical education. *N Engl J Med* 2007; **356**:387–396.
3 Freund Y, Philippon AL, Bokobza J, Carreira S, Riou B, Duguet A. A 1-h simulation-based course on basic life support durably enhances confidence among medical students. *Eur J Emerg Med* 2013; **20**:145–146.
4 Hazwani T, Ashraf N, Hasan Z, Antar M, Kazzaz Y, Alali H. Effect of a pediatric mock code on resuscitation skills and team performance: an in situ simulation experience over three years. *Eur J Emerg Med* 2020; **27**:e15–e16.
5 Boulet JR. Summative assessment in medicine: the promise of simulation for high-stakes evaluation. *Acad Emerg Med* 2008; **15**:1017–1024.

6 Holmboe E, Rizzolo MA, Sachdeva AK, Rosenberg M, Ziv A. Simulation-based assessment and the regulation of healthcare professionals. *Simul Healthc* 2011; **6**:S58–S62.
7 Maignan M, Koch FX, Chaix J, Phellouzat P, Binaud G, Collomb Muret R, *et al*. Team Emergency Assessment Measure (TEAM) for the assessment of non-technical skills during resuscitation: validation of the French version. *Resuscitation* 2016; **101**:115–120.
8 Oriot D, Darrieux E, Boureau-Voultoury A, Ragot S, Scépi M. Validation of a performance assessment scale for simulated intraosseous access. *Simul Healthc* 2012; **7**:171–175.
9 M`essick S. Validity. In Linn RL, editor. *Educational Measurement*. 3rd ed. American Council on Education and Macmillan; 1989: 13–104.
10 Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ* 2003; **37**:830–837.
11 Philippon AL, Hausfater P, Triby E, Freund Y. Développement d'un outil d'évaluation des compétences des étudiants en médecine d'urgence: une étude nationale par la méthode Delphi. *Ann Fr Med Urg* 2019; **9**:354–361.
12 Philippon A, Baud A, Dumont M, Remini SA, Leroy J, Truchot J, *et al*. Usability and reproducibility of three tools to assess medical students and residents in emergency medicine. *AEM Educ Train* 2021; **5**:e10704.
13 Truchot J, Boucher V, Li W, Martel G, Jouhair E, Raymond-Dufresne E, *et al*. Is in situ simulation in emergency medicine safe? A scoping review. *BMJ Open* 2022; **12**:e059442.
14 Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology: normative assessment. *Psychol Assess* 1994; **6**:284–290.
15 Liaw SY, Rashasegaran A, Wong LF, Deneen CC, Cooper S, Levett-Jones T, *et al*. Development and psychometric testing of a Clinical Reasoning Evaluation Simulation Tool (CREST) for assessing nursing students' abilities to recognize and respond to clinical deterioration. *Nurse Educ Today* 2018; **62**:74–79.
16 Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psych* 2006; **3**:77–101.
17 Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care* 2007; **19**:349–357.
18 Englander R, Carraccio C. A lack of continuity in education, training, and practice violates the 'Do No Harm' principle. *Acad Med* 2018; **93**:S12–S16.
19 Hirsh DA, Ogur B, Thibault GE, Cox M. 'Continuity' as an organizing principle for clinical education reform. *N Engl J Med* 2007; **356**:858–866.
20 Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR, *et al*. Competency-based medical education: theory to practice. *Med Teach* 2010; **32**:638–645.
21 Hall AK, Pickett W, Dagnone JD. Development and evaluation of a simulation-based resuscitation scenario assessment tool for emergency medicine residents. *CJEM* 2012; **14**:139–146.
22 Englander R, Frank JR, Carraccio C, Sherbino J, Ross S, Snell L; ICBME Collaborators. Toward a shared language for competency-based medical education. *Med Teach* 2017; **39**:582–587.
23 Farrell SE. Evaluation of student performance: clinical and professional performance. *Acad Emerg Med* 2005; **12**:302–302.
24 Schuwirth LWT, van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ* 2012; **46**:38–48.
25 Rudolph JW, Raemer DB, Simon R. Establishing a safe container for learning in simulation: the role of the presimulation briefing. *Simul Healthc* 2014; **9**:339–349.
26 Cilliers FJ, Schuwirth LW, Adendorff HJ, Herman N, van der Vleuten CP. The mechanism of impact of summative assessment on medical students' learning. *Adv Health Sci Educ Theory Pract* 2010; **15**:695–715.