REPORTS OF ORIGINAL INVESTIGATIONS

"GIOSAT": a tool to assess CanMEDS competencies during simulated crises

Le « GIOSAT »: un outil pour évaluer les compétences CanMEDS pendant des crises simulées

Victor M. Neira, MD · M. Dylan Bould, MBChB · Amy Nakajima, MD · Sylvain Boet, MD · Nicholas Barrowman, PhD · Philipp Mossdorf, MD, PhD · Devin Sydor, MD · Amy Roeske, MD · Stephen Noseworthy, MD · Viren Naik, MD · Dermot Doherty, MD · Hilary Writer, MD · Stanley J. Hamstra, PhD

Received: 28 December 2011 / Accepted: 7 December 2012 / Published online: 19 January 2013

Abstract

Purpose *Our objective was to develop and evaluate a Generic Integrated Objective Structured Assessment Tool (GIOSAT) to integrate Medical Expert and intrinsic (non-medical expert) CanMEDS competencies with non-technical skills for crisis simulation.*

Methods *An assessment tool was designed and piloted using two pediatric anesthesia scenarios (laryngospasm and hyperkalemia). Following revision of the tool, we used previously recorded videos of anesthesia residents (n = 50) who managed one of two intraoperative advanced cardiac life support (ACLS) scenarios (ventricular tachycardia or ventricular fibrillation). Four independent trained raters, blinded to the residents' level of training, analyzed the video recordings using the GIOSAT scale.*

Author contributions *Victor M. Neira* was the principal investigator and corresponding author. He was involved in data extraction and analysis, pilot study design, coordination and communication with the other investigators, and validation of the study. He contributed to the presentation, writing, editing, and correction of all drafts and the final paper. *Stanley J. Hamstra* was the senior investigator and guarantor. He was involved in the critical review and supervision in all steps of the investigation, including the development of the rating scale, pilot study, validation study, development of drafts and editions, and final paper presentation. *Victor M. Neira* and *Hilary Writer* were involved in the literature review. *Hilary Writer* was involved in the literature search and contributed to the design of the first version of the scale. *Dylan M. Bould, Sylvain Boet,* and *Viren Naik* were involved in the critical review of the pilot study results, the design of the second version of the scale, acquisition of data, and analysis and interpretation of the validation study. *Amy Nakajima, Philipp Mossdorf,* and *Dermot Doherty* were responsible for the conception and design of the first revision of the scale. *Amy Nakajima, Philipp Mossdorf, Amy Roeske, Devin Sydor, Stephen Noseworthy,* and *Dermot Doherty* were involved in the acquisition of the data (rater) and analysis and interpretation of the pilot study. *Dylan M. Bould, Sylvain Boet, Viren Naik, Amy Nakajima, Philipp Mossdorf, Amy Roeske, Devin Sydor, Stephen Noseworthy, Dermot Doherty,* and *Hilary Writer* were involved in critical revision of the paper content. *Nicholas Barrowman* was involved in the statistic analysis and the interpretation of the pilot and validation studies. He also contributed graphics for the final paper.

Electronic supplementary material The online version of this article (doi:10.1007/s12630-012-9871-9) contains supplementary material, which is available to authorized users.

V. M. Neira, MD (✉) · M. D. Bould, MBChB · P. Mossdorf, MD, PhD · A. Roeske, MD · D. Doherty, MD
Department of Anesthesia, Children's Hospital of Eastern Ontario, University of Ottawa, 401 Smyth Road, Ottawa, ON K1G 6W3, Canada
e-mail: vneira@cheo.on.ca

A. Nakajima, MD
Department of Obstetrics and Gynaecology, The Ottawa Hospital, University of Ottawa, Ottawa, ON, Canada

S. Boet, MD
Department of Anesthesia, The Ottawa Hospital, University of Ottawa, Ottawa, ON, Canada

N. Barrowman, PhD
Research Institute, Children's Hospital of Eastern Ontario, University of Ottawa, Ottawa, ON, Canada

D. Sydor, MD
Department of Anesthesiology and Perioperative Medicine, Queen's University, Kingston, ON, Canada

S. Noseworthy, MD
Department of Pediatrics, Division of Emergency Medicine, Children's Hospital of Eastern Ontario, University of Ottawa, Ottawa, ON, Canada

Inter-rater reliability was calculated using intraclass correlations (ICCs) for single raters (single measure) and the average of the four raters (average measure), and construct validity was investigated by correlating GIOSAT scores with postgraduate year of residency (PGY).

Results *Total GIOSAT scores for the ACLS scenarios had single measure ICCs of 0.62 and average measure ICCs of 0.85. Inter-rater reliability was substantial for both Medical Expert and intrinsic competencies (single measure ICCs 0.69 and 0.62, respectively; average measure ICCs 0.90 and 0.82, respectively). We found significant correlations between PGY level and total GIOSAT score (r = 0.36; P = 0.011) and between PGY level and Medical Expert competencies (r = 0.42; P = 0.003); however, correlations were not found between PGY level and intrinsic CanMEDS competencies (r = 0.24; P = 0.09).*

Conclusion *Inter-rater reliability of the total GIOSAT scores using four trained raters was substantial. Significant correlation between PGY and (i) total GIOSAT score and (ii) Medical Expert competencies supports construct validity. Evidence of validity was not obtained for intrinsic CanMEDS competencies.*

Résumé

Objectif *Notre objectif était de mettre au point et d'évaluer un Outil d'évaluation structuré objectif intégré générique (GIOSAT - Generic Integrated Objective Structured Assessment Tool) afin d'intégrer les compétences CanMEDS d'expert médical et intrinsèques (autres que expert médical) à des habiletés non techniques pour la simulation de crise.*

Méthode *Un outil d'évaluation a été conçu et mis au banc d'essai à l'aide de deux scénarios d'anesthésie pédiatrique (laryngospasme et hyperkaliémie). Après révision de l'outil, nous nous sommes servis de clips vidéo enregistrés auparavant de résidents en anesthésie (n = 50) qui avaient réussi un de deux scénarios d'ACLS (soins intensifs post-réanimation cardiaque) peropératoires (tachycardie ventriculaire ou fibrillation ventriculaire). Quatre évaluateurs formés indépendants, ne connaissant pas le niveau de formation des résidents, ont analysé les enregistrements vidéo à l'aide de l'échelle GIOSAT. La fiabilité inter-évaluateur a été calculée à l'aide de corrélations intraclasse (CIC) pour évaluateurs uniques (mesure unique) et de la moyenne des quatre évaluateurs (mesure moyenne), et la validité conceptuelle a été examinée en corrélant les scores de GIOSAT à l'année de résidence.*

Résultats *Les scores GIOSAT totaux pour les scénarios d'ACLS ont eu des CIC de 0,62 en mesure unique et des CIC de 0,85 en mesure moyenne. La fiabilité inter-évaluateur était substantielle aussi bien pour les compétences d'expert médical que pour les compétences intrinsèques (CIC en mesure unique 0,69 et 0,62, respectivement; CIC en mesure moyenne de 0,90 et 0,82, respectivement). Nous avons observé des corrélations significatives entre l'année de résidence et le score total sur l'échelle de GIOSAT (r = 0,36; P = 0,011) et entre l'année de résidence et les compétences d'expert médical (r = 0,42; P = 0,003); toutefois, aucune corrélation n'a été observée entre l'année de résidence et les compétences CanMEDS intrinsèques (r = 0,24; P = 0,09).*

Conclusion *La fiabilité inter-évaluateur des scores totaux sur l'échelle de GIOSAT en faisant appel à quatre évaluateurs formés était substantielle. Une corrélation significative entre l'année de résidence et (i) le score total GIOSAT et (ii) les compétences d'expert médical soutient la validité conceptuelle de notre outil. Aucune donnée probante de validité n'a été obtenue pour les compétences CanMEDS intrinsèques.*

Current theory in medical education emphasizes competencies that extend beyond medical knowledge and procedural skills.[1] In Canada, residency training and assessment are organized according to the CanMEDS competency framework which incorporates seven competencies: the *intrinsic* CanMEDS roles of Communicator, Collaborator, Health Advocate, Manager, Scholar, and Professional are positioned around the *Medical Expert* competency which occupies the central role.[2] Primary causes of adverse events in health care are more often related to communication and teamwork than they are to knowledge deficit or lack of technical skills (TSs).[3-6] High-reliability industries have referred to the skills required to manage critical situations as crisis resource management or non-technical skills (NTSs).[7,8] These NTSs are increasingly identified in health care and constitute a perspective on competency that parallels and overlaps the intrinsic competencies of the CanMEDS framework.

Controversy exists in the literature regarding the taxonomy, nomenclature, and definitions of the terms used for managing critical situations in health care. For high-reliability organizations, NTSs have been defined as "the cognitive, social, and personal resource skills that complement TSs and contribute to safe and efficient task

V. Naik, MD
Department of Anesthesia, The Civic Hospital, University of Ottawa, Ottawa, ON, Canada

H. Writer, MD
Department of Pediatrics, Children's Hospital of Eastern Ontario, University of Ottawa, Ottawa, ON, Canada

S. J. Hamstra, PhD
Departments of Anesthesia, Medicine and Surgery AIME, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

performance".[9] Seven skills have been well described: situation awareness, decision-making, communication, teamwork, leadership, managing stress, and coping with fatigue. The adoption of industry concepts, such as TSs and NTSs, is problematic and has recently been questioned because of inaccuracy.[10] In contrast with a dichotomous approach, the CanMEDS framework offers a more nuanced view that reflects the complex interaction and overlap of the various generic core competencies for physicians. Nevertheless, the competencies in the CanMEDS framework have not been designed specifically for crisis management.

The Medical Expert competency is traditionally assessed using written, oral, and objective structured clinical examinations (OSCEs).[11] Communication and interviewing skills are also assessed using OSCEs.[12] The intrinsic CanMEDS competencies as well as NTSs are generally assessed in the workplace training environment and are under-represented in formal assessments. Since it is challenging to assess crisis management skills using traditional assessment tools, simulation has been proposed as an alternative assessment tool to evaluate both technical and non-technical skills.[13] This proposal has stimulated the development of various assessment tools to measure the performance of NTSs during clinical crisis simulation. These scales do not explicitly mention CanMEDS, but they overlap with the Communicator, Collaborator, and Manager competencies.[8,14] The focus of other evaluation tools is almost solely on the Medical Expert role, neglecting intrinsic competencies and the NTSs discourse.[15-17] In our view, CanMEDS has the requisites to serve as the conceptual framework to facilitate translation and integration of the industry concepts of TSs and NTSs to medical and health care education.

The objective of this study was to integrate CanMEDS intrinsic and Medical Expert competencies with NTSs into a Generic Integrated Objective Structured Assessment Tool (GIOSAT) that is capable of assessing residents' performance during clinical crisis simulation and is appropriate for different specialties, scenarios, and environments. We investigated the reliability and construct validity of the GIOSAT using simulation scenarios with anesthesia residents.

Methods

Development of the tool

Using an approach similar to that of other researchers,[18] we developed the GIOSAT following the steps shown in Fig 1.[19] A group of five physicians (V.M.N., A.N., P.M., D.D., H.W.) from a variety of clinical backgrounds in anesthesia, intensive care, and obstetrics and with experience in postgraduate medical education defined the purpose of the assessment tool.

Two physicians and a senior investigator (V.M.N., H.W., and S.H.) performed a literature search in English and French using the internet search engines, PubMed, Medline, and Ovid, for the period from January 2000 to December 2010. Search terms included: assessment, assessment tools, simulation, crisis resource management, non-technical skills, anesthesia, critical care, surgery, emergency medicine, medicine, and pediatrics. We included articles that tested an assessment tool during crisis simulation and investigated construct validity as part of the study. We hand-searched reference lists from relevant papers and also included resulting key articles. The original search produced 828 publications, and the abstracts of these articles were screened, leaving 86 articles for full text review. Eighteen publications were finally selected. The retained articles were classified according to their main focus as NTSs,[8,14,20-24] Medical Expert competency,[16,17,25-30] or both.[13,15,31] The assessment tools were classified as checklists and global rating scales (GRS) (Fig. 2), and items in each assessment tool were classified according to the evidence of construct validity.[32] Assessment tool content was mapped to corresponding CanMEDS competencies and rated as 0 = not applicable or 1 = applicable (table available as Electronic Supplementary Material).

We used a modified Delphi process[33] to design an assessment tool based on the existing literature. A draft of the assessment tool was sent to the original group, and after two iterations of the first version, the GIOSAT was developed. We designed a GRS of four points (1 = poor, 2 = marginal, 3 = acceptable, 4 = good) with free text for comments divided into two sections: Medical Expert competency with 12 items and intrinsic CanMEDS competencies with eight items.

Pilot study

After obtaining Ethics Board approval at the Children's Hospital of Eastern Ontario Research Institute (Dec 3, 2008), we performed a pilot study (Dec 2008 to May 2009) with two video recorded pediatric anesthesia scenarios (laryngospasm, and hyperkalemia) based on the perioperative cardiac arrest and closed claims studies in pediatric anesthesia.[6,34] We invited anesthesia residents (postgraduate year (PGY) 3-5) to participate in the study and obtained informed consent. Ten residents participated in the laryngospasm scenario and 14 participated in the hyperkalemia scenario.

Four trained raters independently assessed the residents' video performance using the GIOSAT. Inter-rater intraclass correlation (ICC) was used to analyze reliability categorized according to Landis and Koch.[35] The

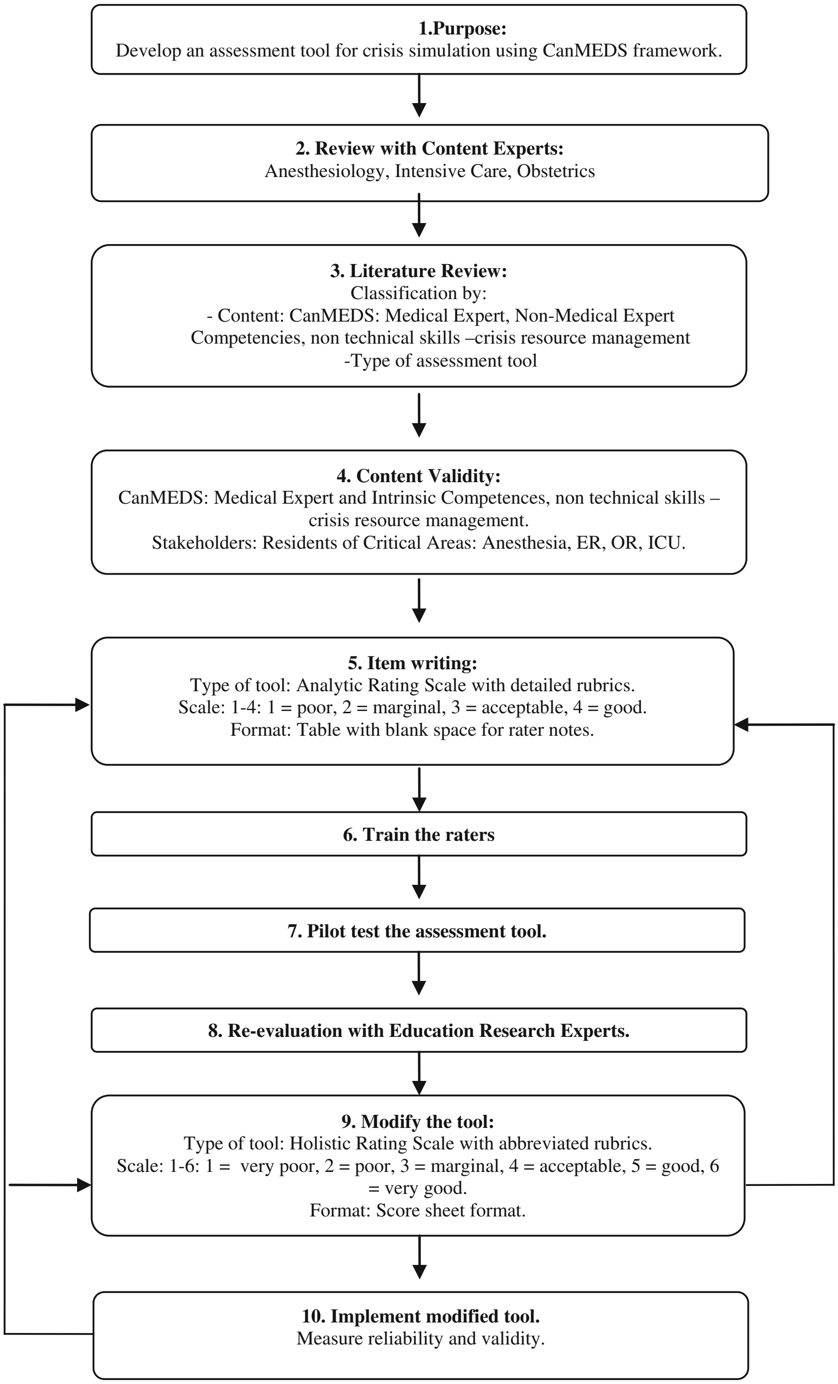
1.Purpose:
Develop an assessment tool for crisis simulation using CanMEDS framework.
2. Review with Content Experts:
Anesthesiology, Intensive Care, Obstetrics
3. Literature Review:
Classification by:
- Content: CanMEDS: Medical Expert, Non-Medical Expert Competencies, non technical skills –crisis resource management
-Type of assessment tool
4. Content Validity:
CanMEDS: Medical Expert and Intrinsic Competences, non technical skills – crisis resource management.
Stakeholders: Residents of Critical Areas: Anesthesia, ER, OR, ICU.
5. Item writing:
Type of tool: Analytic Rating Scale with detailed rubrics.
Scale: 1-4: 1 = poor, 2 = marginal, 3 = acceptable, 4 = good.
Format: Table with blank space for rater notes.
6. Train the raters
7. Pilot test the assessment tool.
8. Re-evaluation with Education Research Experts.
9. Modify the tool:
Type of tool: Holistic Rating Scale with abbreviated rubrics.
Scale: 1-6: 1 = very poor, 2 = poor, 3 = marginal, 4 = acceptable, 5 = good, 6 = very good.
Format: Score sheet format.
10. Implement modified tool.
Measure reliability and validity.

Fig. 1 Generic Integrated Objective Structured Assessment Tool (GIOSAT) development process. (Modified with permission from: *Hamstra SJ*. Keynote address: the focus on competencies and individual learner assessment as emerging themes in medical education research. Acad Emerg Med 2012; 19(12): 1336-43. Chichester, UK: Wiley)[19]

demographics of residents participating in the pilot study are shown in Table 1. The hyperkalemia scenario had moderate to substantial ICCs in all Medical Expert competencies. The Communicator, Collaborator, and Manager competencies had substantial ICCs (0.69-0.77), but there was poor reliability for the Professional competency (ICC = 0.06). Health Advocate and Scholar roles were not identified in these scenarios and were left blank by raters (Table 2). We did not undertake a formal validity analysis for this pilot study due to the small number of participants.

Validation study

The pilot study underwent critical review by an expert group of medical education researchers (D.B., S.B., V.N., and S.H.). The results of this process were used to redesign the tool, and a second study was performed to evaluate the new version. The revised version of the GIOSAT is divided into two sections: Medical Expert competencies with eight items and intrinsic competencies with six items. Each item has abbreviated descriptors and is scored with a GRS

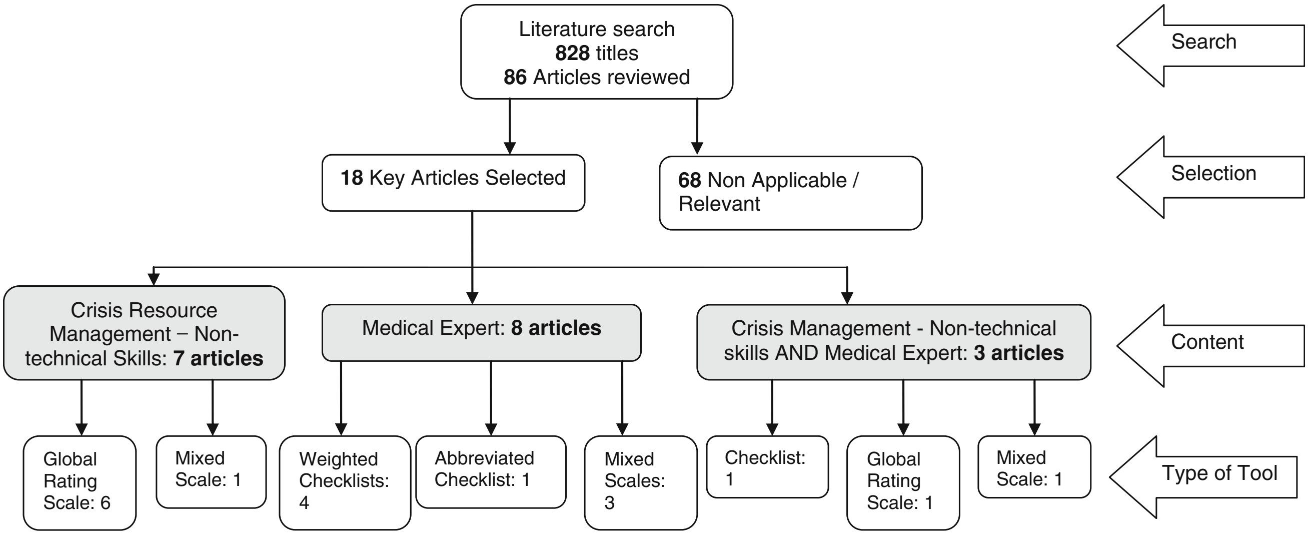

Fig. 2 Literature search and article selection process for assessment methods in clinical crisis simulation

Table 1 Pilot study

Scenario	Laryngospasm *n* (%)	Hyperkalemia *n* (%)
Total	10	14
Gender (M:F)	5 (50):5 (50)	7 (50):7 (50)
PGY 3	3 (30)	6 (43)
PGY 4	4 (40)	5 (36)
PGY 5	3 (30)	3 (21)

Demographic characteristics of residents performing two crisis pediatric anesthesia simulations: laryngospasm and hyperkalemia

n = number of subjects; PGY = postgraduate year residency; M = male; F = female. Three residents performed both scenarios

(1 = very poor to 6 = very good). Summed Medical Expert items, intrinsic items, and total GIOSAT scores are 8-48, 6-36, and 14-84, respectively (Appendix; available as Electronic Supplementary Material).

In this second study, we used video recordings from other previously reported research[36] in order to examine the reliability and validity of the newly designed GIOSAT scale. Research Ethics Board approval was obtained from St. Michael's Hospital, Toronto for the original study (October 3, 2008) comparing debriefing techniques with a pre-test/post-test design. Only pre-test video recordings were used in our analysis to avoid any bias. The sample size was calculated *a priori* for the original study. We used the method recommended by Cohen for an intraclass correlation of 0.7 using four raters for a power of 0.8 and found 35 participants.[37] We decided to include all pre-test video recordings (*n* = 50). An amendment was approved by the same Research Ethics Board to use previously collected video records for our study (February 3, 2010). Written informed consent was previously obtained from the subjects.

Fifty anesthesia residents with different levels of training (PGY 2-5) were randomized to perform in one of two intraoperative advanced cardiac life support (ACLS) scenarios lasting five minutes: ventricular fibrillation due to hyperkalemia or pulseless ventricular tachycardia secondary to myocardial infarction.[36] Four independent raters (three anesthesiologists and one emergency physician) from the University of Ottawa were trained to use the GIOSAT.[38] Raters, blinded to subjects' identity and PGY level of training, independently scored all residents' performances using GIOSAT (October- November 2010). To examine the relative difficulty of the two scenarios, a comparison of the GIOSAT scores and competency scores was analyzed with Student's *t* test.

Investigation of reliability

The ICC was determined with a consistency definition using a two-way random model for both single measures (individual rater) and average measures (the average of the four raters' scores) for total GIOSAT scores, summed Medical Expert scores, summed intrinsic scores, and individual item scores.

Investigation of construct validity

Our primary outcome was the correlation between total GIOSAT scores and PGY level, as measured using Spearman's correlation coefficient (Rho). Our secondary outcomes were correlations between PGY level and summed scores for Medical Expert items, summed scores for intrinsic items, and individual items. We hypothesized that residents' performances would improve with level of training.

Comparison of performance in the two scenarios

The potential confounding effect of gender and type of scenario on the GIOSAT score was analyzed with a two-way analysis of variance. A *Z-test* was performed to compare ICCs between scenarios corrected for multiple testing using Holm's method.[39] P values < 0.05 were considered significant. We used SPSS® version 18 (SPSS, INC. 2010, Chicago, IL, USA) for the statistical analysis.

Results

Results of the validation study

The revised GIOSAT was used in the second study. Residents' distribution according to scenario, gender, and level of training is shown in Table 3. The PGY level was similar in both scenarios, and there was an apparent imbalance in gender distribution between scenarios (Fisher's exact test $P = 0.045$). No significant differences in GIOSAT scores were found between scenarios: Medical Expert scores $P = 0.40$; intrinsic scores $P = 0.56$; and total scores $P = 0.54$ (Student's t test) (Table 4). Scholar was scored as not applicable (N/A) by raters in 69% of the ratings (139/200 ratings), and for that reason, it was excluded from the analysis.

Table 2 Inter-rater reliability of four raters using the first version of the GIOSAT

	ICC laryngospasm	ICC hyperkalemia
Medical expert competencies		
Situation awareness	0.33	0.53
Dealing with changing situations	0.55	0.85
Medical history	-	-
Examine patient/equipment	-	0.63
Diagnosis & differentials	0.5	0.62
Confirmation/investigations	-	0.85
Medical therapeutics	0.74	0.69
Procedure therapeutics	0.32	0.7
Medical overall	0.38	0.68
Intrinsic competencies		
Communicator	0.68	0.69
Collaborator	0.58	0.77
Manager	0.57	0.72
Health advocate	-	-
Scholar	-	-
Professional	0.4	0.06
Overall	0.51	0.76

Results from two pediatric anesthesia crisis scenarios, laryngospasm (10 cases) and hyperkalemia (14 cases). GIOSAT = Generic Integrated Objective Structured Assessment Tool; ICC = intraclass correlation

Table 3 Demographic characteristics of 50 anesthesia residents performing two ACLS scenarios

Scenario	Ventricular fibrillation *n* (%)	Ventricular tachycardia *n* (%)
Total	27	23
Gender (M:F)	19 (38) : 8 (16)*	9 (18) : 14 (28)*
PGY 2	7 (26)	6 (26)
PGY 3	9 (33)	5 (22)
PGY 4	8 (30)	6 (26)
PGY 5	3 (11)	6 (26)

ACLS = advanced cardiac life support; M = male; F = female. No differences in distribution between scenarios by postgraduate year of residency (PGY); n = number of subjects. *Imbalance in gender distribution (Fisher exact test $P = 0.045$)

Investigation of reliability

Inter-rater intraclass correlations with pooled results of both scenarios are shown in Fig 3. The ICCs were substantial for single measure summed scores of Medical Expert competencies (0.69), intrinsic competencies (0.62), and total GIOSAT scores (0.62). The single measure ICCs for individual Medical Expert items were moderate to substantial (0.43-0.69), except for *examine the patient and equipment* (0.29). The single measure ICCs for individual intrinsic items were fair to moderate (0.36-0.60). The average measure ICCs were substantial to almost perfect for individual items (0.61-0.90) and almost perfect for summed Medical Expert items, summed intrinsic items, and total scores (0.87-0.90).

Investigation of construct validity

For our primary outcome, we found a significant correlation between PGY and total GIOSAT scores (r = 0.36; $P < 0.011$) (Fig. 4). For our secondary outcomes, we found a significant correlation between PGY and summed Medical Expert competencies (r = 0.42; $P < 0.003$), but not for summed intrinsic competencies (r = 0.24; $P = 0.09$).

Comparison of performance in the two scenarios

The results of the analysis of variance showed that residents' GIOSAT scores were not significantly influenced by scenario type or residents' gender. We found significant differences between scenarios for single measure ICCs in two of

Table 4 GIOSAT score results in two ACLS scenarios (VF/ VT) in 50 anesthesia residents

GIOSAT / CanMEDS	VF scenario scores Mean (95% CI)	VT scenario scores Mean (95% CI)	Both scenarios scores Mean (95% CI)
Medical expert competencies			
Situation awareness	4.6 (4.3 to 4.9)	4.6 (4.2 to 5.1)	4.5 (4.2 to 5.1)
Dealing with changing. situations	4.6 (4.5 to 4.9)	4.5 (4.0 to 5.0)	4.4 (4.0 to 5.0)
Medical history	3.7 (3.3 to 4.2)	3.4 (2.8 to 3.8)	3.4 (2.8 to 4.2)
Examine patient & equipment	4.3 (4.0 to 4.6)	4.1 (3.7 to 4.3)	4.0 (3.7 to 4.3)
Diagnosis & differentials	4.1 (3.8 to 4.3)	3.8 (3.3 to 4.3)	3.8 (3.3 to 4.3)
Confirmation & investigations	3.3 (3.1 to 3.6)	3.2 (2.8 to 3.7)	3.8 (2.8 to 3.7)
Medical therapeutics	4.1 (3.8 to 4.5)	3.7 (3.2 to 4.3)	3.8 (3.2 to 4.5)
Procedure therapeutics	4.5 (4.0 to 4.8)	4.0 (3.4 to 4.5)	4.2 (3.4 to 4.8)
Intrinsic competencies			
Communicator	4.5 (4.1 to 4.7)	4.3 (3.8 to 4.7)	4.4 (3.8 to 4.7)
Collaborator	4.7 (4.4 to 4.9)	4.4 (3.9 to 4.9)	4.7 (3.9 to 4.9)
Manager	4.3 (3.9 to 4.6)	3.9 (3.3 to 4.5)	4.0 (3.3 to 4.6)
Health advocate	4.6 (4.3 to 4.8)	4.2 (3.7 to 4.7)	4.3 (3.7 to 4.8)
Professional	4.9 (4.7 to 5.0)	4.7 (4.3 to 5.0)	4.9 (4.3 to 5.0)

GIOSAT = Generic Integrated Objective Structured Assessment Tool; ACLS = advanced cardiac life support; VF = ventricular fibrillation; VT = ventricular tachycardia; CI = confidence intervals. No significant differences were found between scenarios. (Student's t test $P > 0.05$). Each item corresponds to a GIOSAT competency scored in the following scale: 1 = very poor, 2 = poor, 3 = marginal, 4 = acceptable, 5 = good, and 6 = very good

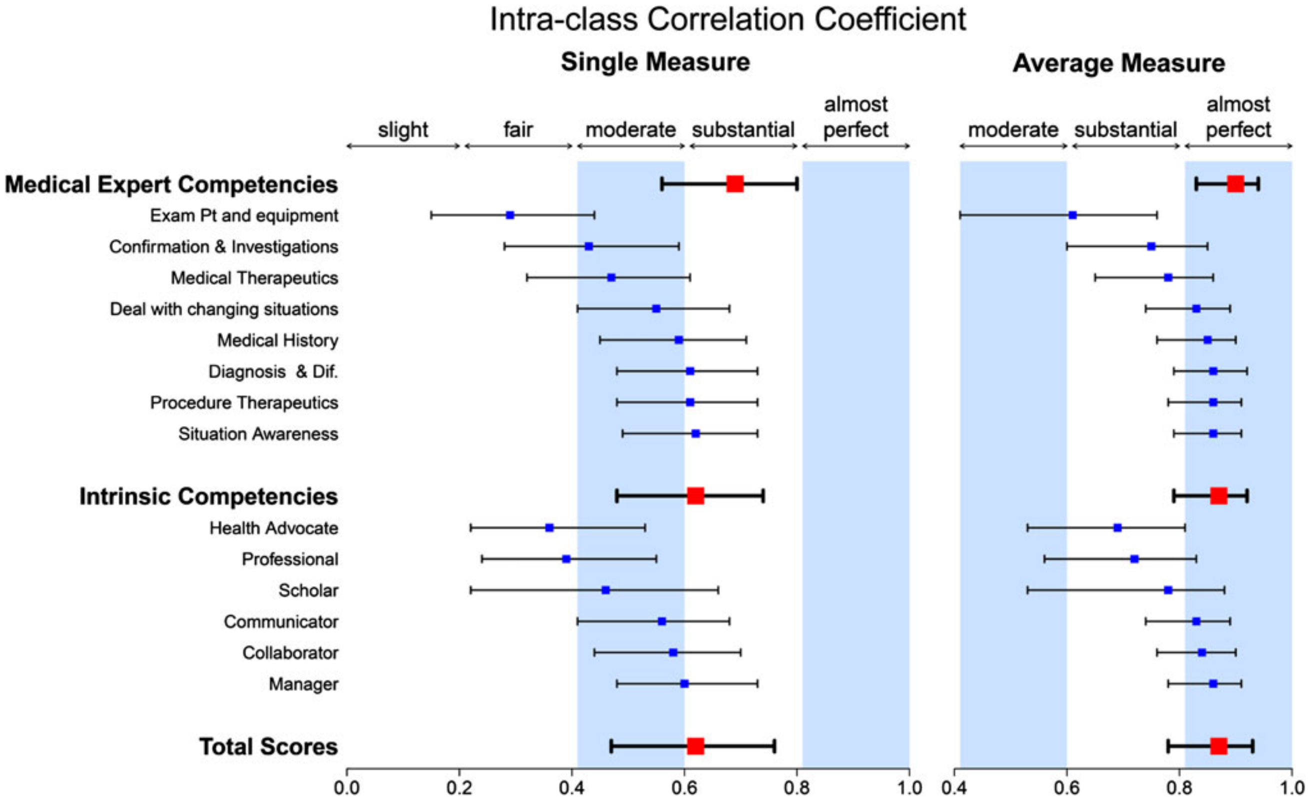

Fig. 3 Intra-class correlation coefficients (ICCs) with 95% confidence intervals (CI) for Generic Integrated Objective Structured Assessment Tool (GIOSAT) total scores and items. Dots represent ICCs and lines represent 95% CI of GIOSAT scores from four raters assessing two simulated advanced cardiac life support (ACLS) scenarios with 50 anesthesia residents. Ventricular fibrillation ($n = 27$), ventricular tachycardia ($n = 23$). Examine Pt and equipment = examine patient and equipment; Diagnosis & dif. = diagnosis and differentials

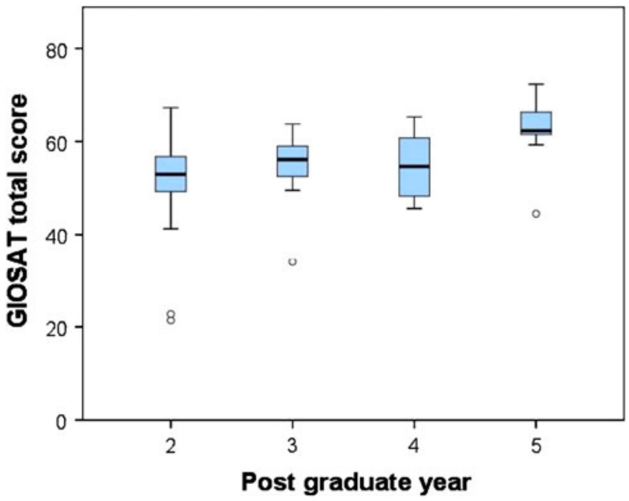

Fig. 4 Relationship between Generic Integrated Objective Structured Assessment Tool (GIOSAT) total score and postgraduate year. Spearman's correlation coefficient (Rho) = 0.36 ($P = 0.011$). GIOSAT score is the sum of all item scores with a maximum of 84. For each box, the dark horizontal line represents the median. The bottom of the box represents the 25th percentile, and the top represents the 75th percentile. A whisker extends down from the bottom of the box to the lowest data point within 1.5 times the interquartile range (IQR) of the 25th percentile. A whisker extends up from the top of the box to the highest data point within 1.5 IQR of the 75th percentile. Points beyond the whiskers are represented as open circles

the Medical Expert competencies, *diagnosis and differentials* and *dealing with changing situations* ($P = 0.03$ and 0.006, respectively), and in one intrinsic competency, *Collaborator* ($P = 0.006$). No differences between scenarios were found in average measure ICCs.

Discussion

We developed a global rating scale "GIOSAT" that integrates NTSs concepts with the CanMEDS competencies for generic assessment during crisis simulation. A pilot study was used to refine the GIOSAT through an iterative process. We found evidence of substantial reliability (single measures) for the refined tool using four raters for Medical Expert, intrinsic, and total scores. There was evidence of construct validity for the Medical Expert and total scores, but we found no evidence of construct validity for the intrinsic scores. The Medical Expert section of GIOSAT has good psychometric properties and could be used for summative assessment during simulated crises. Although the metrics for reliability and validity for the total score is acceptable, the intrinsic competencies section is problematic, which may raise concerns for total scores as well.

Non-technical skills have been described as being difficult to define[10,40] and assess.[8,14,20] The reliability of GIOSAT is comparable with previous studies of NTSs that have shown fair to substantial reliability. As would be expected, NTSs scales have greater reliability when their items are summed, as we have found with the GIOSAT. Our finding of poor reliability and construct validity of some of the intrinsic competencies not included in the NTSs discourse is in keeping with the literature on professionalism which is not often taught formally (i.e., part of the hidden curriculum) and is difficult to assess.[41,42]

Our study has a number of limitations. The data are based on a re-analysis of scenarios not designed to identify certain competencies, such as Professional, Health Advocate, and Scholar. It may also be that the ACLS scenarios were too short to identify all CanMEDS competencies fully. Increasing the number of scenarios and changing scenario design may improve construct validity of the intrinsic section. It has also been shown that reliability improves with increased testing time, for instance, it has been shown in OSCEs that several hours of testing and ten or more cases are required for high-stakes examinations.[12]

A further limitation is the possibility that intrinsic competencies are underrepresented in GIOSAT and the Medical Expert role is overrepresented. Future iterations of the GIOSAT tool may emphasize the intrinsic competencies either in terms of the number of items or the extensiveness of the descriptors. According to current concepts, reliability and validity are not properties of the instrument but properties of the instrument's scores and interpretations.[43] The same instrument used in a different setting or with different subjects may produce different results; thus, our results may not necessarily be generalizable to other populations.

The GIOSAT Medical Expert competencies section relating to psychometric properties is appropriate to be used for summative assessment. The intrinsic competencies section and, by implication, the scale as a whole are not yet appropriate for summative assessment.

The aim of future research will be to identify the number of raters, scenarios, and examinees necessary to establish reliability and generalizability. The development of scenarios specifically designed to challenge specific domains (Professional, Health Advocate and Scholar: PHAS Roles) may be required for the appropriate use of GIOSAT for summative assessment.

Acknowledgements The authors sincerely thank Kevin Thomas and Drs Michelle Chiu, Simone Crooks, George Dumitrascu, and Robert Elliott for facilitating the pilot study. We also thank the residents of the University of Toronto and the University of Ottawa for their participation in the study.

Funding This study was supported in part by an education grant from the Postgraduate Office and funds from the Department of Anesthesiology, University of Ottawa.

Competing interests None declared.

References

1. *The Royal College of Physicians and Surgeons of Canada*. CanMEDS 2005 Framework. CanMEDS: better standards, better physicians, better care. Available from URL: http://www.rcpsc.medical.org (accessed November 2012).
2. *Sherbino J*, *Frank JR*, *Flynn L*, *Snell L*. "Intrinsic roles" rather than "armour": renaming the "non-medical expert roles" of the CanMEDS framework to match their intent. Adv Health Sci Educ Theory Pract 2011; 16: 695-7.
3. *Kohn LT*, *Corrigan JM*, *Donaldson MS*. To Err is Human: Building a Safer Health Care System. Washington, DC: National Academy Press; 1999 .
4. *Canadian Institute for Health Information*. Health Care in Canada, 2004. Available from URL: www.cihi.ca (accessed November 2012).
5. *McIlvaine WB*. Human error and its impact on anesthesiology. Semin Anesth 2006; 25: 172-9.
6. *Jimenez N*, *Posner KL*, *Cheney FW*, *Caplan RA*, *Lee LA*, *Domino KB*. An update on pediatric anesthesia liability: a closed claims analysis. Anesth Analg 2007; 104: 147-53.
7. *Gaba DM*, *Howard SK*, *Fish KJ*, *Smith BE*, *Sowb YA*. Simulation-based training in anesthesia crisis resource management (ACRM): a decade of experience. Simulation & Gaming 2001; 32: 175-93.
8. *Fletcher G*, *Flin R*, *McGeorge P*, *Glavin R*, *Maran N*, *Patey R*. Anesthetists' Non-Technical Skills (ANTS): evaluation of a behavioral marker system. Br J Anaesth 2003; 90: 580-8.
9. *Flin R, O'Connor P, Crichton M*. Safety at the Sharp End - A Guide to Non-Technical Skills. Farnham: England, Ashgate Publishing Limited; 2008: 1-16.
10. *Nestel D*, *Walker K*, *Simon R*, *Aggarwal R*, *Andreatta P*. Non-technical skills: an inaccurate and unhelpful descriptor? Simul Healthc 2011; 6: 2-3.
11. *Bandiera G*, *Sherbino J*, *Frank JR*. The CanMEDS Assessment Tools Handbook. An Introductory Guide to Assessment Methods for the CanMEDS Competencies. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2006 .
12. *Hawkins RE*, *Boulet JR*. Direct observation: standardized patients. In: Holmboe ES, Hawkins RE, editors. Practical Guide to the Evaluation of Clinical Competence. Philadelphia: Mosby Elsevier; 2008. p. 102-18.
13. *Savoldelli GL*, *Nail 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.
14. *Kim J*, *Neilipovitz D*, *Cardinal P*, *Chiu M*, *Clinch J*. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High Fidelity Simulation, and Crisis Resource Management I Study. Crit Care Med 2006; 34: 2167-74.
15. *Brett-Fleegler MB*, *Vinci RJ*, *Weiner DL*, *Harris SK*, *Shih MC*, *Kleinman ME*. A simulator-based tool that assesses pediatric resident resuscitation competency. Pediatrics 2008; 121: e597-603.
16. *Morgan PJ*, *Cleave-Hogg D*, *DeSousa S*, *Tarshis J*. High-fidelity patient simulation: validation of performance checklists. Br J Anaesth 2004; 92: 388-92.
17. *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.
18. *Ringsted C*, *Hodges B*, *Scherpbier A*. 'The research compass': an introduction to research in medical education: AMEE Guide No. 56. Med Teach 2011; 33: 695-709.
19. *Hamstra SJ*. Key note address: the focus on competencies and individual learner assessment as emerging themes in medical education research. Acad Emerg Med 2012; 19: 1336-43.
20. *Graham J*, *Hocking G*, *Giles E*. Anaesthesia non-technical skills: can anaesthetists be trained to reliably use this behavioural marker system in 1 day? Br J Anaesth 2010; 104: 440-5.
21. *Yule S*, *Flin R*, *Maran N*, *Rowley D*, *Youngson G*, *Paterson-Brown S*. Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. World J Surg 2008; 32: 548-56.
22. *Jankouskas T*, *Bush MC*, *Murray B*, *et al*. Crisis resource management: evaluating outcomes of a multidisciplinary team. Simul Healthc 2007; 2: 96-101.
23. *Guise JM*, *Deering SH*, *Kanki BG*, *et al*. Validation of a tool to measure and promote clinical teamwork. Simul Healthc 2008; 3: 217-23.
24. *Malec JF*, *Torsher LC*, *Dunn WF*, *et al*. The Mayo high performance teamwork scale: reliability and validity for evaluating key crew resource management skills. Simul Healthc 2007; 2: 4-10.
25. *Boulet JR*, *Murray D*, *Kras J*, *Woodhouse J*, *McAllister J*, *Ziv A*. Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. Anesthesiology 2003; 99: 1270-80.
26. *Chopra V*, *Gesink BJ*, *de Jong J*, *Bovill JG*, *Spierdijk J*, *Brand R*. Does training on an anaesthesia simulator lead to improvement in performance? Br J Anaesth 1994; 73: 293-7.
27. *Scavone BM*, *Sproviero MT*, *McCarthy RJ*, *et al*. Development of an objective scoring system for measurement of resident performance on the human patient simulator. Anesthesiology 2006; 105: 260-6.
28. *Devitt JH*, *Kurrek MM*, *Cohen MM*, *Cleave-Hogg D*. The validity of performance assessments using simulation. Anesthesiology 2001; 95: 36-42.
29. *Murray DJ*, *Boulet JR*, *Avidan M*, *et al*. Performance of residents and anesthesiologists in a simulation-based skill assessment. Anesthesiology 2007; 107: 705-13.
30. *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.
31. *Gaba DM*, *Howard SK*, *Flanagan B*, *Smith BE*, *Fish KJ*, *Botney R*. Assessment of clinical performance during simulated crisis using both technical and behavioral ratings. Anesthesiology 1998; 89: 8-18.
32. *Kane M*, *Crooks T*, *Cohen A*. Validating measures of performance. Educational Measurement: Issues and Practice 1999; 18: 5-17.
33. *Jones J*, *Hunter D*. Consensus methods for medical and health services research. BMJ 1995; 311: 376-80.
34. *Bhananker SM*, *Ramamoorthy C*, *Geiduscheck JM*, *et al*. Anesthesia-related cardiac arrest in children: update from the Pediatric Perioperative Cardiac Arrest Registry. Anesth Analg 2007; 105: 344-50.
35. *Landis JR*, *Koch GG*. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-74.
36. *Boet S*, *Bould MD*, *Bruppacher HR*, *Desjardins F*, *Chandra DB*, *Naik VN*. Looking in the mirror: self-debriefing versus instructor debriefing for simulated crises. Crit Care Med 2011; 39: 1377-81.
37. *Cohen J*. Statistical Power Analysis for the Behavioral Sciences, Second Edition. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers; 1988.
38. *American Heart Association*. Part 4: Advanced life support. Circulation 2005; 112: III25-54.
39. *Holm S*. A simple sequentially rejective multiple test procedure. Scand J Stat 1979; 6: 65-70.
40. *Glavin RJ*. Skills, training and education. Simul Healthc 2011; 6: 4-7.

41. *Lynch DC*, *Surdyk PM*, *Eiser AR*. Assessing professionalism: a review of the literature. Med Teach 2004; 26: 366-73.
42. *Bahaziq W*, *Crosby E*. Physician professional behaviour affects outcomes: a framework for teaching professionalism during anesthesia residency. Can J Anesth 2011; 58: 1039-50.
43. *Cook DA, Beckman TJ*. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 2006; 119: 166.e7-16.