

A Comparison of Global Rating Scale and Checklist Scores in the Validation of an Evaluation Tool to Assess Performance in the Resuscitation of Critically Ill Patients During Simulated Emergencies (Abbreviated as "CRM Simulator Study IB")

John Kim, MD, MEd, FRCPC;
David Neilipovitz, MD, FRCPC;
Pierre Cardinal, MD, FRCPC;
Michelle Chiu, MD, FRCPC

Background: Crisis resource management (CRM) skills are a set of nonmedical skills required to manage medical emergencies. There is currently no gold standard for evaluation of CRM performance. A prior study examined the use of a global rating scale (GRS) to evaluate CRM performance. This current study compared the use of a GRS and a checklist as formal rating instruments to evaluate CRM performance during simulated emergencies.

Methods: First-year and third-year residents participated in two simulator scenarios each. Three raters then evaluated resident performance in CRM using edited video recordings using both a GRS and a checklist. The Ottawa GRS provides a seven-point anchored ordinal scale for performance in five categories of CRM, and an overall performance score. The Ottawa CRM checklist provides 12 items in the five categories of CRM, with a maximum cumulative score of 30 points. Construct validity was measured on the basis of content validity, response process, internal structure, and response to other variables. T-test analysis of Ottawa GRS scores was conducted to examine response to the variable of level of training. Intraclass correlation coefficient (ICC) scores were used to measure inter-rater reliability for both scenarios.

Results: Thirty-two first-year and 28 third-year residents participated in the study. Third-year residents produced higher mean scores for overall CRM performance than first-year residents ($P < 0.05$), and in all individual categories within the Ottawa GRS ($P < 0.05$) and the Ottawa CRM checklist ($P < 0.05$). This difference was noted for both scenarios and for each individual rater ($P < 0.05$). No statistically significant difference in resident scores was observed between scenarios for both instruments. ICC scores of 0.59 and 0.61 were obtained for Scenarios 1 and 2 with the Ottawa GRS, whereas ICC scores of 0.63 and 0.55 were obtained with the Ottawa CRM checklist. Users indicated a strong preference for the Ottawa GRS given ease of scoring, presence of an overall score, and the potential for formative evaluation.

Conclusion: Construct validity seems to be present when using both the Ottawa GRS and CRM checklist to evaluate CRM performance during simulated emergencies. Data also indicate the presence of moderate inter-rater reliability when using both the Ottawa GRS and CRM checklist.

(*Sim Healthcare* 4:6–16, 2009)

Key Words: Crisis resource management, Mannequin-based simulation, Evaluation of performance, Medical education, Medical evaluation, Computer simulation, Validation, Evaluation of resuscitation skills

The acute management of critically ill patients is not limited to Critical Care Medicine physicians, but is pertinent to all

acute care specialties.¹ Despite this fact, instruction of skills required to successfully manage such crises is rarely provided in undergraduate or postgraduate education.^{2,3} Although medical knowledge forms the foundation of resuscitation skills, it alone is insufficient. Studies reviewing Operating Room critical events indicate that human error is responsible for half of these complications.^{4,5} A more important finding is that most of these events are not due to a lack of medical knowledge, but are attributable to errors in crisis resource management (CRM). CRM refers to a set of skills outside of medical knowledge that are required to effectively manage the actual crisis itself. Errors that occurred were found to be similar to those discovered during reviews of emergencies in other nonmedical professions such as the aviation, aerospace, and nuclear industries.^{6–11} The original crew resource man-

From the Division of Critical Care Medicine and Department of Medicine (J.M.), University of Ottawa/The Ottawa Hospital, Gloucester, ON, Canada; Division of Critical Care Medicine and Departments of Anesthesiology (D.N., M.C.) and Medicine (P.C.), University of Ottawa/The Ottawa Hospital, Ottawa, ON, Canada.

Supported by 2001 Royal College of Physicians and Surgeons of Canada Research in Medical Education Award.

Reprints: John Kim, MD, MEd, FRCPC, Division of Critical Care Medicine and Department of Medicine, University of Ottawa/The Ottawa Hospital, 3017 Quail Run Avenue, Gloucester, ON, Canada K1T 3S3 (e-mail: jkim@aserty.com or jkim@ottawahospital.on.ca).

The authors have indicated that they have no conflicts of interest to disclose.

Copyright © 2009 Society for Simulation in Healthcare
DOI: 10.1097/SIH.0b013e3181880472

agement skills identified in the nonmedical professions were then adapted for use in the medical setting as CRM skills.¹

Given the universal nature of critically ill patients, the instruction of CRM skills is essential for any acute care specialty. Similarly, the ability to formally evaluate performance in CRM seems equally important.

Recent advances in computer technology have enabled innovative medical education programs to incorporate computer-based patient simulation as an instruction tool for CRM skills. Although simulations provide a safe environment in which evaluation of resuscitation skills can be performed, it cannot be equated with evaluation of real-life performance. However, real-life resuscitation often takes place in an uncontrolled environment, with many factors influencing both the process and result of the attempted resuscitation. Because patient welfare takes precedence over the requirements to provide a controlled evaluation setting, many confounding factors are introduced in the evaluation of performance during a real-life emergency. These include supervisor intervention, variability in support staff assistance, and differences in the presentation and/or severity of each case. For all of the above reasons, while simulation can only provide an approximation of real-life resuscitation, it remains the only viable means to formally evaluate performance of resuscitation skills in a real-time and dynamic environment.

Few studies exist that formally validate performance evaluation during simulation.^{12–24} The lack of a gold standard for CRM performance represents the greatest obstacle in formally validating CRM evaluation and evaluation of performance during simulated emergencies. One rating instrument used in the Operating Room setting, the Anesthetists' Non-Technical Skills tool, has undergone formal evaluation and shows considerable promise.²⁵ A recent study at the University of Ottawa examined the role of a global rating scale (GRS) for the formal evaluation of CRM performance during common simulated emergencies.²⁶ The rating instrument used in the study was the University of Ottawa CRM GRS (hereafter referred as the "Ottawa GRS").

It is unclear what type of instrument should be used to measure CRM performance. While most of the studies in high-fidelity simulation have used checklists, these have been for evaluation of crises where specific solutions or "best actions" are recognized. In contrast, emergencies that commonly occur in the intensive care unit and Emergency Room (ER) (respiratory failure, shock, etc.) do not have a single "best action" or specific remedy. Some studies in medical education in other settings suggest checklists are unable to detect increases in levels of expertise, while GRS seem to be superior in detecting such differences.^{27–29} Other studies have demonstrated that checklists can discriminate between differing levels of performance, especially in settings where clearly accepted "best actions" are present.^{30,31} Given the conflicting data, a comparison of a CRM checklist (hereinafter referred to as the "Ottawa CRM checklist") and GRS (Ottawa GRS) in the evaluation of CRM performance was conducted.

The American Educational Research Association and American Psychologic Association Standards for Educational and Psychologic Testing have outlined five components of

evidence to support the presence of construct validity.^{32,33} The following four areas were examined for this study.

Content Validity

Content validity refers to whether or not the rating instrument covers all the relevant domains of CRM.

Response Process

Response process refers to the integrity of data and the maximum control and/or elimination of error associated with test administration.^{32,33} Therefore, both case delivery and scoring of performance were examined.

Relationship to Other Variables

Relationship to other variables was examined using the variable of residency experience and/or training through the hypothesis that postgraduate third-year (PGY-3) residents will have higher CRM scores than postgraduate first-year (PGY-1) residents. A corollary hypothesis that exists is that as residents gain more experience in CRM, they will have higher CRM scores.

Internal Structure

Internal structure refers to the statistical or psychometric properties of the instrument itself.^{32,33} Internal structure was therefore examined by the evaluation of both reliability and discriminatory ability measures for each simulator scenario and for each category within the Ottawa GRS and each category and/or item within the Ottawa CRM checklist.

METHODS

Target Population

The study was conducted with consenting PGY-1 and PGY-3 residents at the University of Ottawa, after receiving ethics approval from The Ottawa Hospital Research Ethics Board. The University of Ottawa holds accredited residency programs in all major medical, surgical, and anesthesiology specialties. Participants with prior simulator experience in residency were excluded. This excluded all PGY-3 residents from the Department of Anesthesiology. Recruitment took place via an anonymous e-mail advertisement.

Study Design

Participating residents were allocated into PGY-1 or PGY-3 groups. Each group of residents then participated in three separate half-day sessions—the simulator tutorial session, and two individual simulator scenario sessions.

Simulator Tutorial Session

The simulator tutorial session took place 2 days before the first simulator session and covered topics in acute resuscitation of critically ill patients. This was done to ensure that a minimal foundation of medical knowledge required for the acute management of critically ill patients was provided to all residents. The tutorial session also provided a brief orientation to the simulator patient and simulator room environment.

Simulator Environment

Each case took place, using the MedSim simulation mannequin, in a dedicated simulation room that recreated the intensive care unit physical environment.

Simulator Sessions

The two simulator sessions took place within a 2–3 week period. Each resident performed as the lead physician for each simulator scenario. The order of scenarios was identical for each resident. The first scenario involved cardiac ischemia and arrhythmias in a postoperative patient, whereas the second scenario involved a patient presenting in acute shock and hypoxemic respiratory failure after severe trauma from a fall. Both cases were developed from real-life cases and were reviewed by simulator instructors and staff intensivists from across Canada for realism of case content and timing. Both cases were reviewed and/or used by instructors during the Canadian Local and National Acute Critical Events Simulation (ACES) courses—a skills-based Continuing Medical Education (CME) course using high-fidelity patient simulation and task trainers to train critical care residents in resuscitation skills. These cases were reviewed and/or used during the 2001 Local and 2002 National ACES courses. The programmed sequence of events for both scenarios required that residents re-evaluate and react to new problems in each scenario. Before each scenario, residents received a brief synopsis of the case, similar to a phone consultation for assessment.

The initial setting and the clinical events in each case were identical for all residents. For each case, support staff was present to assist the resident. Trained actors assumed the roles of registered nurse and respiratory therapist in each case. In the event that residents failed to intervene unaided to specific events during the scenario, support staff would give preset cues to assist residents in recognizing these events. These cues included repeated observation of abnormal vital signs or aberrant clinical signs. These cues were determined during the scenario development and were peer-reviewed by simulation instructors during the Canadian National ACES course and local ACES courses for realism and timing. The cues thus enabled the scenario to progress in a realistic fashion, as support staff interaction could assist residents of different skill levels.

Resident performance during the simulator scenario sessions was videotaped. The videotapes were digitally edited to superimpose a black dot over the image of the participant's face to completely hide their identity. Residents were also instructed to avoid using their own names. Three attending physicians, with experience in Critical Care Medicine and/or CRM skills instruction, evaluated each resident simulator scenario session using the Ottawa GRS and the Ottawa CRM checklist.

Development of Evaluation Instrument—Ottawa GRS

The Ottawa GRS is divided into five categories of CRM skills based on recognized CRM literature—problem solving, situational awareness, leadership, resource utilization, and communication (Appendix 1).^{2,3} An overall rating category for CRM performance was also provided. Each category was measured on a seven-point anchored ordinal scale with descriptive anchors to provide guidelines on alternating points along the scale. These descriptors were added to reduce personal bias in interpreting performance. The scoring system was designed so that a score of one corresponded to the performance judged to be that of a complete novice, and a score

of three corresponded to the performance of a novice with some CRM and resuscitation experience. A score of five corresponded to the performance of a physician with sufficient CRM and resuscitation experience to manage critical events competently, whereas a score of seven corresponded to the performance of a physician with expertise in the area of resuscitation and CRM. The amount of cueing necessary for residents to act was taken into account in the Ottawa GRS descriptive anchors.

The individual Ottawa GRS categories, scoring system, and descriptive anchors for each category were developed using a Delphi process³⁴ to collect expert data and opinion about CRM evaluation. Input was received from Critical Care Medicine physicians, acute care specialty staff physicians (Anesthesiology, Emergency Medicine) and CRM experts from the University of Ottawa. Input was also received from other high-fidelity simulation instructors across Canada. An initial GRS was circulated to each expert in Ottawa and feedback received. After input was received and initial revisions made, the GRS was circulated to national simulation and/or CRM instructors during the 2001–2003 National ACES courses for feedback. Further modifications were made after receiving the feedback from national instructors. Before administration of the GRS, the local group of instructors met and made final revisions after a test run of GRS use on a pilot group of residents performing in simulated emergencies.

Development of Evaluation Instrument—Ottawa CRM Checklist

The Ottawa CRM checklist incorporates the CRM categories used with the Ottawa GRS (Appendix 2). However, the categories are further subdivided into individual items, each representing important actions or behaviors within that category. Each item was scored on a two-point scale, with two points for a successfully completed action/behavior, one point for a partially completed behavior (or if it required cueing), and no points for an omitted or inadequately completed behavior. A total of 12 items were identified; three items were given double weight (four, two, and zero) based on expert opinion. These were given extra weighting after unanimous input from both CRM and acute care specialty physicians indicating that these elements formed the foundation of successful CRM. A Delphi process³⁴ identical to the Ottawa GRS development (as described earlier) was used in the development of the Ottawa CRM checklist categories, scoring system, and descriptors.

Rater Training

Three raters (A, B, and C) were chosen to evaluate each resident case scenario session. Each rater was chosen for their expertise as an acute care physician and CRM instructor. The simulator instructor present for all simulator sessions was excluded from being a rater to preserve the integrity of the blinding process.

The simulator instructor and two of the three raters participated in the original Delphi process used in the development of the Ottawa GRS and Ottawa CRM checklist. Each simulator instructor reviewed video files of a substandard, standard, and near-expert level of performance for each scenario. Each rater individually rated the video files using the Ottawa GRS and Ottawa CRM checklist. All raters and the

simulator instructor then met again to review the scoring by each rater for each session. The Ottawa GRS and Ottawa CRM checklist scoring system and use of descriptive anchors were revised, and statistical consensus was obtained from all three raters for each scoring category. Raters then individually rated all sessions. Raters were not instructed on order of instrument used and were allowed to review the videotape if necessary to complete scoring with either instrument.

Measurement of Construct Validity

Construct validity was measured with the four American Educational Research Association and American Psychologic Association Standards for Educational and Psychologic Testing variables of content validity, response process, internal validity, and the relationship to the variable of training.^{32,33} Content validity and response process were not measured statistically, but through the analysis as described in the Discussion section. Response to the variable of training was thus measured by comparing PGY-1 and PGY-3 Ottawa GRS and Ottawa CRM checklist scores. Analysis of performance between groups was conducted by *t* test analysis of a mean of the raters' scores. A comparison of PGY-1 and PGY-3 scores between scenarios was also performed for both the Ottawa GRS overall scores and Ottawa CRM checklist total scores using an analysis of variance for each rater to examine if any differences in scores were present in both cases, and with each rater. Internal consistency was measured by assessed by measures of interobserver reliability using type III intraclass correlation coefficient (ICC) for the overall CRM performance score on the Ottawa GRS and individual GRS category scores. A type III ICC was also calculated for the summation score on the Ottawa CRM checklist and for the individual CRM checklist category scores.

Sample Size and Timeline

Sample size was based on the primary hypothesis that PGY-3 residents will outperform PGY-1 residents. The sample size was based on Cohen's³⁵ definition of moderate effect size of 0.5 population standard deviation units. Since the hypothesis is directional, a one-tailed significance test was used with a power of 0.8. Based on these parameters, the required sample size was 50 for each group of residents.

In May 2002, when the study was presented at the 2002 Ontario Network of Medical Education, expert opinion indicated that an effect size calculation of 0.8 units would be more appropriate. This translated to a sample size target of 20 PGY-1 and 20 PGY-3 residents per group. By that time, 32 PGY-1 residents and 19 PGY-3 residents had been successfully completed participation in the study, with an additional nine PGY-3 residents already recruited for sessions in July and August 2002. Based on expert opinion, enrollment was terminated after the August 2002 resident group completed participation in the study.

RESULTS

Thirty-two PGY-1 and 28 PGY-3 residents were recruited (Table 1). One PGY-1 and one PGY-3 resident withdrew from the second simulator session after participating in one simulator session due to an inability to attend the second

Table 1. Resident Demographics (PGY-1 and PGY-3)

Demographic Information	PGY-1 Group	PGY-3 Group
No. participants	32	28
Age (mean + standard deviation)	27.75 ± 3.73	30.39 ± 3.58
Gender (male/female)	13/19	18/10
Specialty		
Internal medicine	8	13
General surgery	1	5
Other surgery*	9	0
Anesthesia	3	0
Emergency medicine	1	2
Family medicine	8	8†
Other‡	2	0
Medical school (Canadian/Other)	28/4	26/2
Prior ICU weeks (mean + SD)	1.22 ± 1.68	8.61 ± 5.23
Prior ICU on-call nights	1.19 ± 2.34	12.96 ± 10.12

*Other surgery—ear/nose/throat (ENT) (2), obstetrics (3), ophthalmology (2), and orthopedics (2).

†PGY-3 in family medicine in combined anesthesia (2)/emergency medicine (6).

‡Other—neurology (1) and psychiatry (1).

PGY-1, Post-Graduate Year-1; PGY-3, Post-Graduate Year-3.

simulator session. During the video recording process, three videos of PGY-1 simulator sessions and three videos from PGY-3 simulator sessions encountered technical problems and were thus excluded from the analysis. The remaining 112 sessions were analyzed, with 28 PGY-1 and 27 PGY-3 residents having recorded videos of both scenarios. All videos were included for analysis of construct validity by comparison of PGY-1 and PGY-3 scores. All videos were also used for analysis of inter-rater reliability. However, videos from residents with only one session for scoring were excluded from analysis of intercase reliability.

Construct validity through relationship to the variable of training for both the Ottawa GRS and Ottawa CRM checklist was examined by comparison of PGY-1 and PGY-3 scores (Tables 2 and 3). A significant difference between PGY-1 and PGY-3 residents was noted in both overall CRM performance scores on the Ottawa GRS ($P < 0.05$) and cumulative CRM checklist scores ($P < 0.05$).

Individual categories within the Ottawa GRS and Ottawa CRM checklist underwent analysis for construct validity with respect to relationship to the variable of training (Tables 4 and 5). Significant differences were present in all Ottawa GRS categories and all Ottawa CRM checklist categories ($P < 0.05$). An analysis of variance for overall CRM scores in the Ottawa GRS and cumulative CRM scores for the Ottawa CRM checklist was conducted for each case and with each rater (Tables 6 and 7). Significant differences were present in both cases with all raters ($P < 0.05$).

Table 2. Ottawa GRS Scores PGY-1 vs. PGY-3: Overall CRM Performance

Session No.	PGY-1 Overall CRM Score	PGY-3 Overall CRM Score	Mean Difference* (95% CI)	<i>P</i>
Overall	4.13 ± 0.87	5.54 ± 0.85	1.41 (0.92–1.90)	<0.05
1	3.84 ± 1.67	5.42 ± 1.28	1.57 (0.97–2.18)	<0.05
2	4.33 ± 1.14	5.79 ± 1.00	1.45 (0.87–2.04)	<0.05

*Mean difference = PGY-3 – PGY-1 score.

GRS, Global Rating Scale; CRM, Crisis Resource Management.

Table 3. Ottawa CRM Checklist Scores PGY-1 vs. PGY-3 Cumulative Checklist Scores

Session No.	PGY-1 CRM Cumulative Score	PGY-3 CRM Cumulative Score	Mean Difference (95% CI)	P
Overall	20.72 ± 3.61	25.51 ± 2.55	4.84 (3.03–6.64)	<0.05
1	19.34 ± 4.80	25.15 ± 4.08	5.81 (3.58–8.04)	<0.05
2	21.77 ± 4.17	26.19 ± 2.94	4.42 (2.45–6.39)	<0.05

A comparison of resident scores for overall GRS category scores and summation scores on the Ottawa CRM checklist from Scenario 1 to 2 was conducted (Table 8). No significant difference in CRM performance was noted for either the Ottawa GRS or the Ottawa CRM checklist for PGY-1 and PGY-3 residents.

Inter-rater reliability for both the Ottawa GRS and Ottawa CRM checklist was also examined by a calculation of ICC scores. The overall CRM performance score on the Ottawa GRS demonstrated an ICC score of 0.59 and 0.61 for Scenarios 1 and 2, respectively (Table 9). The cumulative checklist score on the Ottawa CRM checklist demonstrated an ICC score of 0.63 and 0.55 for Scenarios 1 and 2, respectively. Further analysis of individual categories within the Ottawa GRS revealed similar reliability scores for problem solving, leadership, and situational awareness, with ICC scores ranging from 0.48 to 0.63. The Ottawa GRS categories of Resource Utilization and Communication demonstrated lower reliability scores, with ICC scores ranging from 0.24 to 0.38. The pattern of reliability was again noted for individual items within the Ottawa CRM checklist, with ICC scores ranging from 0.46 to 0.60 for the categories of problem solving, leadership, and situational awareness. Lower inter-rater reliability measurements were again observed for Resource Utilization and Communication, with ICC scores ranging from 0.24 to 0.38.

DISCUSSION

The scientific validation of an evaluation device represents a daunting task. For high-fidelity simulation, the challenge is even greater, especially given the lack of an accepted gold standard for comparison.³⁶ In the case of CRM skills, it was also necessary to create and validate a formal instrument to

measure CRM skill performance. Given the lack of gold standard, the role of high-fidelity simulation as a potential evaluation device was first examined by measures of construct validity and reliability for both instruments.

The Ottawa GRS and Ottawa CRM checklist demonstrated evidence for the presence of construct validity in multiple domains.^{32,33} Both the Ottawa GRS and Ottawa CRM checklist categories are based on the work by Gaba et al.² Both rating instruments and simulator cases were reviewed and modified through a Delphi process by both simulation and CRM instructors from across Canada, therefore content validity seems to be present.

Residents received an orientation session to familiarize themselves with the simulator environment, and each resident participated in identical scenarios. The use of scripted cues from support staff for each case also ensured uniformity of case delivery. An extensive Delphi rater training process was undertaken to complete the Ottawa GRS and the Ottawa CRM checklist, and that all videotapes were digitally converted to a uniform viewing format for rater evaluation. Therefore, both the case delivery and the rating process itself also seem to meet optimal criteria to minimize error in the scoring process itself, and therefore meet the criteria of response process.

Relationship to the variable of training was examined by comparing PGY-1 and PGY-3 scores. The study results indicate that both the Ottawa GRS and Ottawa CRM checklist could differentiate between PGY-1 and PGY-3 performance during simulator scenarios. Some readers may question whether the difference in performance was simply due to a difference in medical knowledge between the PGY-1 and PGY-3 groups. However, as the second case involved a trauma scenario, a difference in case-specific knowledge alone would seem insufficient to explain this difference, given that over half the PGY-3 residents did not train in surgical or trauma-related specialties. More importantly, regardless of the reason for the difference in observed performance between groups, all three raters consistently and reliably noted a significant difference in overall CRM scores between the PGY-1 and PGY-3 resident groups, as well as cumulative scores for the Ottawa CRM checklist (Tables 2 and 3). These

Table 4. Ottawa GRS Scores PGY-1 vs. PGY-3 Individual CRM Categories

Session No./GRS Category	PGY-1 Score	PGY-3 Score	Mean Difference (95% CI)	P
Session 1				
Overall	3.84 ± 1.67	5.42 ± 1.28	1.57 (0.97–2.18)	<0.05
Leadership	4.13 ± 1.03	5.39 ± 1.02	1.26 (0.71–1.82)	<0.05
Problem solving	3.84 ± 1.26	5.40 ± 0.91	1.56 (0.95–2.17)	<0.05
Situational awareness	3.85 ± 1.19	5.23 ± 0.96	1.38 (0.79–1.97)	<0.05
Resource utilization	4.46 ± 0.90	5.43 ± 0.76	0.97 (0.51–1.43)	<0.05
Communication	4.91 ± 0.83	5.57 ± 0.63	0.66 (0.26–1.07)	<0.05
Session 2				
Overall	4.33 ± 1.14	5.79 ± 1.00	1.45 (0.87–2.04)	<0.05
Leadership	4.56 ± 1.09	5.86 ± 0.81	1.31 (0.73–1.89)	<0.05
Problem solving	4.35 ± 1.11	5.79 ± 1.12	1.44 (0.81–2.08)	<0.05
Situational awareness	4.30 ± 1.12	5.74 ± 0.99	1.44 (0.87–2.02)	<0.05
Resource utilization	4.73 ± 0.88	5.90 ± 0.71	1.17 (0.74–1.61)	<0.05
Communication	5.33 ± 0.87	5.99 ± 0.57	0.65 (0.25–1.06)	<0.05

Table 5. Ottawa CRM Checklist Scores PGY-1 vs. PGY-3 Individual CRM Categories

Session No./Checklist Category (Max. score)	PGY-1 Score	PGY-3 Score	Mean Difference (95% CI)	P
Session 1				
Summation (30)	19.34 ± 4.80	25.15 ± 4.08	5.81 (3.58–8.04)	<0.05
Leadership (8)	5.03 ± 1.46	6.90 ± 0.98	1.87 (1.18–2.56)	<0.05
Problem solving (4)	2.19 ± 1.03	3.38 ± 0.58	0.67 (0.32–1.04)	<0.05
Situational awareness (6)	3.34 ± 1.24	4.75 ± 0.92	1.41 (0.80–2.01)	<0.05
Resource utilization (4)	2.45 ± 0.78	3.13 ± 0.46	0.67 (0.32–1.04)	<0.05
Communication (8)	6.33 ± 0.94	7.00 ± 0.61	0.67 (0.23–1.10)	<0.05
Session 2				
Summation (30)	21.77 ± 4.17	26.19 ± 2.94	4.42 (2.45–6.39)	<0.05
Leadership (8)	5.57 ± 1.33	6.89 ± 1.12	1.32 (0.65–1.99)	<0.05
Problem solving (4)	2.80 ± 0.86	3.58 ± 0.55	0.78 (0.38–1.17)	<0.05
Situational awareness (6)	3.72 ± 1.20	5.00 ± 0.78	1.28 (0.73–1.84)	<0.05
Resource utilization (4)	3.12 ± 0.48	3.55 ± 0.32	0.43 (0.21–0.66)	<0.05
Communication (8)	6.56 ± 1.00	7.16 ± 0.79	0.61 (0.17–1.04)	<0.05

results would suggest that both the Ottawa GRS and Ottawa CRM checklist differentiated groups not based on differences in medical knowledge alone, but on some other skill set, such as CRM performance. The fact that all CRM categories in the Ottawa GRS and Ottawa CRM checklist demonstrated this effect would suggest that from a relationship-to-other-variables perspective, construct validity is present (Tables 4 and 5). The fact that these differences were also observed with each case and with each rater (Tables 6 and 7) further strengthens the notion that construct validity is present.

Construct validity was also measured by the effect of simulator session participation on CRM scores from the second simulator session (Table 8). The fact that resident overall CRM scores on the Ottawa GRS and cumulative scores on the Ottawa CRM checklist did not improve is not surprising, given the short-time interval between the first and second sessions. A more robust analysis of the effect of experience on CRM would involve PGY-1 residents again performing in the simulator later in training, when they would have presumably acquired a significant amount of CRM experience. Future studies will incorporate this design change.

Both the Ottawa GRS and the Ottawa CRM checklist seemed to demonstrate discriminatory ability for each case, as statistically significant differences between PGY-1 and PGY-3 residents were observed both in overall and individual categories. From a reliability perspective, the Ottawa CRM overall performance, problem solving, situational awareness, and leadership style categories demonstrated moderate ICC

scores (Table 9). Similar ICC scores were seen in the same categories for the Ottawa CRM checklist. Therefore, construct validity seems to be present from the perspective of internal structure.

Although prior studies in high-fidelity simulation mostly used checklists of recognized “best” actions and behaviors for problems with recognized solutions, data from other settings in medical education suggest that the GRS is superior in detecting differences from the novice and expert performer.^{27,28,29} However, in this case, both the Ottawa GRS and Ottawa CRM checklist demonstrated good measures of construct validity. The fact that all Ottawa GRS and Ottawa CRM checklist categories demonstrated this effect is also important. These findings would support the premise that differences in CRM performance during simulated emergencies can be detected using both the Ottawa GRS and Ottawa CRM checklist.

The Ottawa CRM checklist differs from many of the checklists used in prior studies is that it was designed for generic use and not as a checklist of “correct” actions for specific emergencies where a “best solution” is present. Given that most emergencies in medicine do not have such “best solutions,” this difference may account for the similar properties observed for both the Ottawa GRS and Ottawa CRM checklist.

From the perspective of reliability, while an ICC score of 0.60 represents a moderate level of inter-rater reliability, it is not considered to be ideal for high stakes (or “summative”)

Table 6. Ottawa GRS scores PGY-1 vs. PGY-3 by Individual Rater ANOVA

Session No./Rater	PGY-1 Score	PGY-3 Score	Mean Difference (95% CI)	P
Rater A				
Session 1	3.97 ± 1.51	5.42 ± 1.14	1.45 (0.71–2.19)	<0.05
Session 2	4.00 ± 1.33	5.63 ± 0.97	1.63 (1.00–2.27)	<0.05
Rater B				
Session 1	4.21 ± 1.86	5.58 ± 1.28	1.37 (0.48–2.25)	<0.05
Session 2	4.93 ± 1.66	6.11 ± 1.37	1.18 (0.35–2.02)	<0.05
Rater C				
Session 1	3.34 ± 1.23	5.25 ± 0.97	1.91 (1.29–2.52)	<0.05
Session 2	4.07 ± 1.30	5.63 ± 1.18	1.56 (0.88–2.23)	<0.05

Table 7. Ottawa CRM Checklist Scores PGY-1 vs. PGY-3 by Individual Rater ANOVA

Session No./Rater	PGY-1 Score	PGY-3 Score	Mean Difference (95% CI)	P
Rater A				
Session 1	17.97 ± 5.60	23.17 ± 4.09	5.20 (2.48–7.92)	<0.05
Session 2	19.56 ± 4.01	24.00 ± 3.50	4.44 (2.06–6.83)	<0.05
Rater B				
Session 1	21.06 ± 6.64	26.67 ± 3.52	5.60 (2.61–8.60)	<0.05
Session 2	23.48 ± 6.64	27.85 ± 3.53	4.37 (1.46–7.27)	<0.05
Rater C				
Session 1	19.00 ± 4.82	25.63 ± 3.93	6.63 (4.21–9.04)	<0.05
Session 2	22.26 ± 4.24	26.70 ± 3.57	4.44 (2.30–6.59)	<0.05

Table 8. Overall CRM Scores: Case 1 vs. Case 2

Level of Training	CRM Scores		Mean Difference (95% CI)	P
	Session 1	Session 2		
Ottawa GRS*				
PGY-1	3.93 ± 1.27	4.33 ± 1.14	0.40 (−0.27–1.07)	0.232
PGY-3	5.43 ± 0.93	5.65 ± 1.01	0.22 (−0.20–0.63)	0.288
Ottawa CRM checklist†				
PGY	19.67 ± 5.03	21.77 ± 4.17	2.10 (−0.19–4.39)	0.07
PGY-3	25.29 ± 2.98	25.81 ± 3.00	0.52 (−0.83–2.42)	0.43

*Fifty residents: 27 PGY-1 and 23 PGY-3 residents in overall CRM category scores used for analysis.

†Fifty residents: 27 PGY-1 and 23 PGY-3 residents in summation CRM checklist scores used for analysis.

evaluation. Given the fact that both rating instruments are unproven as evaluation tools, it is difficult to determine if reliability was decreased due to errors in the rating instrument design or due to errors in the rater training process, or a combination of the two. In reviewing the individual rater scores, the so-called “dove/hawk” effect was observed—some raters may consistently score performances lower than other raters (hawks), whereas other raters may consistently score performances higher than other raters (doves).³⁷ Rater B clearly demonstrated mean overall CRM performance scores and cumulative checklist scores higher than Raters A and C in both simulator scenarios. This finding suggests that the rater training process may be at least partially responsible for some of the variability between raters. Given the above results, it is likely that revisions both to the original instrument design and rater training process will be required before greater inter-rater reliability will be observed.

It is important to recognize that the results of this study indicate both the Ottawa GRS and Ottawa CRM checklist seem equivalent at this time in terms of construct validity and inter-rater reliability. This was based on a comparison of scores from two different sets of residents at the PGY-1 and PGY-3 level of training. However, the study does not specif-

Table 9. Ottawa GRS and Ottawa CRM Checklist Intraclass Correlation Coefficient (ICC) Scores

GRS Category	Ottawa GRS Type-III ICC (ICC31)	Ottawa CRM Checklist Type-III ICC (ICC31)
Session 1		
Overall*	0.590	0.633
Leadership	0.491	0.603
Problem solving	0.551	0.546
Situational awareness	0.475	0.502
Resource utilization	0.346	0.439
Communication	0.236	0.272
Session 2		
Overall*	0.613	0.545
Leadership	0.626	0.548
Problem solving	0.567	0.456
Situational awareness	0.544	0.506
Resource utilization	0.355	0.156
Communication	0.384	0.236

*For the Ottawa CRM checklist, the ICC31 score refers to that of the cumulative score.

ICC, intraclass correlation coefficient.

ically determine which instrument may be able to detect changes in CRM skills expertise as residents’ progress through training; such comparisons are premature given the scope of this study.

While the rating instruments are being compared in their potential role as a summative evaluation tool, they may also serve a formative role in providing valuable feedback and direction for self-improvement to residents as they progress in training. This study did not specifically examine this issue. The Ottawa CRM checklist provides more items for each CRM category, which may be helpful in identifying specific strengths and weaknesses within a resident’s CRM skill set. However, the GRS provides a greater range of scores within each category, which may allow a resident to track their progress as they become more proficient in both acute resuscitation and CRM skills. The Ottawa GRS also provides descriptive anchors that incorporate many of the ideal behaviors and actions for each CRM category. Ultimately, if revisions to instrument design and/or rater training fail to separate the Ottawa GRS and Ottawa CRM checklist in terms of reliability and validity, issues of feasibility and formative feedback will likely play a pivotal role in determining which rating scale is preferable for CRM evaluation. Furthermore, while construct validity and reliability are key characteristics of any potential evaluation device, feasibility is a key educational consideration. Given that both instruments demonstrate excellent measures of construct validity and similar measures of inter-rater reliability, feasibility may ultimately determine which device is preferable. Each rater indicated a strong preference for the Ottawa GRS, given its simplicity, overall CRM performance score, ease of use, and greater flexibility in allocating scores for CRM performance. Each instructor noted that the Ottawa GRS was easier to score, and allowed for greater flexibility in judgment. They also commented that the Ottawa GRS took less time to administer while scoring. While the Ottawa CRM checklist provides more categories for scoring, the narrow range for scoring and extra time required to complete the checklist were both noted by each rater as significant drawbacks compared with the Ottawa GRS. The narrow range of scoring was also identified by each rater as an obstacle for use as a formative evaluation tool. However, the checklist’s comprehensive list of actions may also provide more information to a resident in training. As revisions to the Ottawa GRS and Ottawa CRM checklist are underway, issues of reliability and validity will need to be re-examined to ensure equivalence between both instruments still exists.

High-fidelity simulation is fast becoming an integral component of training in academic centers.³⁸ This is likely in recognition of the fact that high-fidelity simulation represents the only safe alternative to real-life practice for the acquisition of experience in acute resuscitation and CRM. For this reason alone, further examination of its potential as a formal evaluation device for CRM and other skill sets in acute resuscitation is warranted. The results of this publication suggest that high-fidelity simulation may indeed play a key role in the evaluation of CRM. A second follow-up study with resident participation at the University of Ottawa is already underway to assess revisions to both instrument design and

rater training for both the Ottawa GRS and Ottawa CRM checklist. Resident recruitment has also been redesigned to allow PGY-1 residents to participate in their PGY-2 and PGY-3 year of training. The results from this study will provide a more robust comparison on the validity, reliability, and feasibility of both instruments. Until then, given the similarity in measures of reliability and validity, the choice of rating instrument for the assessment of CRM will likely reflect differences in feasibility and personal preference for each instrument.

CONCLUSION

High-fidelity simulation offers the opportunity to evaluate a skill set previously no gold standard for evaluation of CRM current exists. This study examined the use of a CRM skills GRS and checklist in formally evaluating CRM performance. Construct validity seemed to be present for both instruments from the perspective of content validity, response process, internal structure, and relationship to the variable of training. The Ottawa GRS and Ottawa CRM checklist demonstrated statistically significant differences in PGY-1 and PGY-3 scores, indicating that both instruments demonstrate the ability to discriminate between participants of differing levels of ability. Both the Ottawa GRS and Ottawa CRM checklist demonstrated moderate measures of inter-rater reliability, with poor reliability demonstrated in both instruments in several CRM categories, indicating that future revisions in design for both instruments are necessary. From a feasibility perspective, users indicated a strong preference for the Ottawa GRS, given the ease of scoring, presence of overall score, and potential use as a tool for formative evaluation. A second study is underway to examine further issues of validity and reliability once instrument design and rater training revisions are completed.

ACKNOWLEDGMENTS

The authors acknowledge the Royal College of Physicians and Surgeons of Canada (RCPSC) for their financial support for this study through the 2001 RCPSC Research in Medical Education Award. The RCPSC reviewed the original study design before funding CRM Simulator Study IA. The RCPSC did not participate in collection, management, analysis, and interpretation of the data. The RCPSC has received a final report on this study, but did not participate in preparation or approval of the final manuscript.

The authors also acknowledge Ms. Jennifer Clinch's (Biostatistician, Ottawa Health Research Institute) contributions in providing statistical support in study design and data analysis of the results for this study.

REFERENCES

- Issenberg SB, McGaghie WC, Hart IR, et al. Simulation technology for health care professional skills training and assessment. *JAMA*. 1999; 282:861–866.
- Gaba DM, Fish KJ, Howard SK. *Crisis Management in Anesthesia*. London: Churchill Livingstone; 1994.
- Schwid HA, Rooke GA, Michalowski P, Ross BK. Screen-based anesthesia simulation with debriefing improves performance in a mannequin-based anesthesia simulator. *Teach Learn Med*. 1996;13: 92–96.
- Cooper JB, Newbrow RS, Kitz RJ. An analysis of major errors and equipment failures in anesthesia management: considerations for prevention and detection. *Anesthesiology*. 1984;60:34–42.
- Gaba DM. Dynamic decision-making in anesthesiology: cognitive models and training approaches. In Evans DA, Patel VI, eds. *Advanced Models of Cognition for Medical Training and Practice*. Berlin: Springer-Verlag; 1992:123–147.
- Hays RT, Jacobs JW, Prince C, Salas E. A meta-analysis of the flight simulation training literature. *Mil Psychol*. 1992;4:63–74.
- Bell HH, Waag WL. Evaluating the effectiveness of flight simulators for training combat skills: a review. *Int J Aviat Psychol*. 1998;3:223–242.
- Koonce J, Bramble WJ. Personal computer-based flight training devices. *Int J Aviat Psychol*. 1998;8:277–292.
- Rolfe JM, Staplese KJ. *Flight Simulation*. Cambridge, England: Cambridge University Press; 1986:232–249.
- Wachtel J. The future of power plant simulation in the United States. In: Walton DG, ed. *Simulation for Nuclear Reactor Technology*. Cambridge, England: Cambridge University Press; 1985:339–349.
- Saliterman SS. A computerized simulation for critical care training: new technology for medical education. *Mayo Clin Proc*. 1990;65: 968–978.
- Byrne AJ, Greaves JD. Assessment instruments used during anaesthetic simulation: review of published studies. *Br J Anaesth*. 2001;86:445–450.
- Gaba DM, DeAnda A. A response of anesthesia trainees to simulated critical incidents. *Anesth Analg*. 1989;68:444–451.
- DeAnda A, Gaba DM. Role of experience in the role of simulated critical incidents. *Anesth Analg*. 1991;72:308–315.
- Schwid HA, O'Donnell D. Anesthesiologists' management of simulated critical incidents. *Anesthesiology*. 1992;76:495–501.
- 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;73:293–297.
- Byrne AJ, Jones JG. Responses to simulated anesthetic emergencies by anesthetists with different duration of clinical experience. *Br J Anaesthesia*. 1997;78:553–556.
- Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R. Assessment of clinical performance during simulated crises using both technical and behavioural ratings. *Anesthesiology*. 1998;89:8–18.
- Devitt JH, Kurrek MM, Cohen MM, Fish P, Murphy PM, Szalai JP. Testing the raters: inter-rater reliability of standardized anesthesia simulator performance. *Can J Anaesth*. 1997;44:924–928.
- Devitt JH, Kurrek MM, Cohen MM, et al. Testing internal consistency and construct validity during evaluation of performance in a patient simulator. *Anesth Analg*. 1998;86:1160–1164.
- Kurrek MM, Devitt JH, Cohen M. Cardiac arrest in the OR: how are our ACLS skills? *Can J Anaesth*. 1998;45:130–132.
- Morgan PJ, Cleave-Hogg D. Evaluation of medical students' performances using the anesthesia simulator. *Acad Med*. 1999;74:202.
- Morgan PJ, Cleave-Hogg D. Performance evaluation using the anesthesia simulator. *Med Educ*. 2000;34:42–45.
- Morgan PJ, Cleave-Hogg D, DeSousa S, Tarshis J. High-fidelity patient simulation: validation of performance checklists. *Br J Anaesth*. 2004; 92:388–392.
- Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Anaesthetists' non-technical skills (ANTS): development and evaluation of a behavioural marker system. *Br J Anaesthesia*. 2003;90: 580–588.
- 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;34:2167–2174.
- Morgan PJ, Cleave-Hogg D, Guest C. A comparison of global ratings

- and checklist scores from an undergraduate assessment using an anesthesia simulator. *Acad Med.* 2001;76:1053–1055.
28. Hodges B, Regehr G, McHaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Acad Med.* 1999;74:1129–1134.
 29. Dreyfus HL, Dreyfus SE. *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer.* New York: The Free Press; 1986.
 30. Issenberg SB, Gordon MS, Greber AA. Bedside cardiology skills training for the osteopathic internist using simulation technology. *J Am Osteopath Assoc.* 2003;103:603–607.
 31. Wayne DB, Fudala MJ, Butter J, et al. Comparison of two standard-setting methods for advanced cardiac life support training. *Acad Med.* 2005;80:S63–S66.
 32. Adler M, Ziglio E. *Gazing into the Oracle: The Delphi Method and its Application to Social Policy and Public Health.* London: Jessica Kingsley Publishers; 1996.
 33. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ.* 2003;37:830–837.
 34. American Educational Research Association, American Psychological Association, National Council on Measurements in Education. *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association; 1999.
 35. Cohen J. *Statistical Power Analysis for the Behavioural Sciences.* Hillsdale: Lawrence Erlbaum Associates, Publishers; 1988.
 36. Stanislaw H. Tests of computer simulation validity: what do they measure? *Simul Games.* 1986;17:173–191.
 37. Muzzin LJ, Hart L. Oral examinations. In: Neufeld VR, Norman GR, eds. *Assessing Clinical Competence.* New York: Springer; 1985: 71–93.
 38. Morgan PJ, Cleave-Hogg D. A worldwide survey of the use of simulation in anesthesia. *Can J Anesth.* 2002;49:659–662.

APPENDIX 1 – OTTAWA CRISIS RESOURCE MANAGEMENT (CRM) GLOBAL RATING SCALE

EVALUATION CRITERIA:

This evaluation scale is directed towards assessing competence in crisis management (CM) skills and care of critically ill patients. The standard of competence has been set at the senior resident level, i.e. the third-year resident who has had prior ICU experience, and through experience as a senior housestaff physician, has previous experience in managing crises. As there exists a requisite base of medical knowledge required to effectively manage crises, this will also be evaluated. However, the focus of evaluation will be on crisis management skills. The skills listed below comprise essential aspects of crisis management. In the simulator case scenario sessions, performance in each of these areas will be assessed, in addition to the amount of prompting or guidance required during the case scenario sessions.

The following criteria will be evaluated:

LEADERSHIP SKILLS

Stays calm and in control during crisis
 Prompt and firm decision-making
 Maintains global perspective (“Big picture”)

SITUATIONAL AWARENESS

Avoids fixation error
 Reassesses and re-evaluates situation constantly
 Anticipates likely events

COMMUNICATION SKILLS

Communicates clearly and concisely
 Uses directed verbal/non-verbal communication
 Listens to team input

PROBLEM SOLVING

Organized and efficient problem solving approach (ABC’s)
 Quick in implementation (Concurrent management)
 Considers alternatives during crisis

RESOURCE UTILIZATION

Calls for help appropriately
 Utilizes resources at hand appropriately
 Prioritizes tasks appropriately

OVERALL

Resident #: _____

Date: _____

Staff: _____

Time: _____

OVERALL PERFORMANCE

1	2	3	4	5	6	7
Novice; all CM skills require significant improvement		Advanced novice; many CM skills require moderate improvement		Competent; most CM skills require minor improvement		Clearly superior; few, if any CM skills that only require minor improvement

I. LEADERSHIP SKILLS

1	2	3	4	5	6	7
Loses calm and control for most of crisis; unable to make firm decisions; cannot maintain global perspective		Loses calm/control frequently during crisis; delays in making firm decisions (or with cueing); rarely maintains global perspective		Stays calm and in control for most of crisis; makes firm decisions with little delay; usually maintains global perspective		Remains calm and in control for entire crisis; makes prompt and firm decisions without delay; always maintains global perspective

II. PROBLEM SOLVING SKILLS

1	2	3	4	5	6	7
Cannot implement ABC's assessment without direct cues; uses sequential management despite cues; fails to consider any alternative in crisis		Incomplete or slow ABC assessment; mostly uses sequential management approach unless cued; gives little consideration to alternatives		Satisfactory ABC assessment; without cues; mostly uses concurrent management approach with only minimal cueing; considers some alternatives in crisis		Thorough yet quick ABC without cues; always uses concurrent management approach; considers most likely alternatives in crisis

III. SITUATIONAL AWARENESS SKILLS

1	2	3	4	5	6	7
Becomes fixated easily despite repeated cues; fails to reassess and re-evaluate situation despite repeated cues; fails to anticipate likely events		Avoids fixation error only with cueing; rarely reassesses and re-evaluates situation without cues; rarely anticipates likely events		Usually avoids fixation error with minimal cueing; reassesses re-evaluates situation frequently with minimal cues; usually anticipates likely events		Avoids any fixation error without cues; constantly reassesses and re-evaluates situation without cues; constantly anticipates likely events

IV. RESOURCE UTILIZATION SKILLS

1	2	3	4	5	6	7
Unable to use resources and staff effectively; does not prioritize tasks or ask for help when required despite cues		Able to use resources with minimal effectiveness; only prioritizes tasks or asks for help when required with cues		Able to use resources with moderate effectiveness; able to prioritize tasks and/or ask for help with minimal cues		Clearly able to use resources to maximal effectiveness; sets clear task priority and asks for help early with no cues

V. COMMUNICATION SKILLS

1	2	3	4	5	6	7
Does not communicate with staff; does not acknowledge staff communication, never uses directed verbal/non-verbal communication		Communicates occasionally with staff, but unclear and vague; occasionally listens to but rarely interacts with staff; rarely uses directed verbal/non-verbal communication		Communicates with staff clearly and concisely most of time; listens to staff feedback; usually uses directed verbal/non-verbal communication		Communicates clearly and concisely at all times, encourages input and listens to staff feedback; consistently uses directed verbal/non-verbal communication

APPENDIX 2 – SIMULATOR SESSION CRISIS MANAGEMENT SKILLS CHECKLIST

ACTION	YES (2 points)	With Prompting (1 point)	NO (0 points)
PROBLEM SOLVING			
Prompt ABC assessment			
Implements concurrent management approach (4 points)			
SITUATIONAL AWARENESS			
Avoids fixation error (4 points)			
Re-assesses and re-evaluates situation (4 points)			
RESOURCE UTILIZATION			
Calls for help when indicated			
Delegates and directs appropriately			
LEADERSHIP			
Maintains calm demeanor			
Acts decisively and maintains control of crisis			
Maintains global perspective			
COMMUNICATION			
Communicates clearly and concisely			
Closes the loop and uses names			
Listens to team input			
TOTAL SCORE (30 points)			

Resident #:

Scenario #:

Staff #:

Date: