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

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

Objective: Resuscitation of critically ill patients requires medical knowledge, clinical skills, and nonmedical skills, or crisis resource management (CRM) skills. There is currently no gold standard for evaluation of CRM performance. The primary objective was to examine the use of high-fidelity simulation as a medium to evaluate CRM performance. Since no gold standard for measuring performance exists, the secondary objective was the validation of a measuring instrument for CRM performance—the Ottawa Crisis Resource Management Global Rating Scale (or Ottawa GRS).

Design: First- and third-year residents participated in two simulator scenarios, recreating emergencies seen in acute care settings. Three raters then evaluated resident performance using edited video recordings of simulator performance.

Setting: A Canadian university tertiary hospital.

Interventions: The Ottawa GRS was used, which provides a 7-point Likert scale for performance in five categories of CRM and an overall performance score.

Measurements and Main Results: Construct validity was measured on the basis of content validity, response process, internal structure, and response to other variables. One variable measured in this study was the level of training. A t-test analysis of Ottawa GRS scores was conducted to examine response to the variable of level of training. Intraclass correlation coefficient scores were used to measure interrater reliability for both scenarios. 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 < .0001) and in all individual categories within the Ottawa GRS (p = .0019 to p < .0001). This difference was noted for both scenarios and for each individual rater (p = .0061 to p < .0001). No statistically significant difference in resident scores was observed between scenarios. Intraclass correlation coefficient scores of .59 and .61 were obtained for scenarios 1 and 2, respectively.

Conclusions: Data obtained using the Ottawa GRS in measuring CRM performance during high-fidelity simulation scenarios support evidence of construct validity. Data also indicate the presence of acceptable interrater reliability when using the Ottawa GRS. (Crit Care Med 2006; 34:2167–2174)
ing actual real-life emergencies, high-fidelity patient simulation may provide a unique medium to evaluate CRM performance in a medical crisis.

To improve CRM evaluation, a pilot study was conducted at the University of Ottawa. The High Fidelity Simulation, Critical Care Medicine and Crisis Management I (abbreviated as the Ottawa GRS). Downing (5) described an approach in seeking evidence of construct validity in medical education using five distinct sources: content, response process, internal structure, relationship to other variables, and consequences. This study seeks to demonstrate evidence of construct validity for the Ottawa Crisis Resource Management Global Rating Scale (abbreviated as the Ottawa GRS).

Materials and Methods

Target Population

After receiving ethics approval, first-year and third-year postgraduate (PGY-1 and PGY-3) residents at the University of Ottawa were recruited by e-mail advertisement into the study. The university holds accredited residency programs in medical, surgical, and anesthesiology specialties, with >500 residents enrolled annually. Residents with prior simulator experience in residency were excluded. All study residents indicated they had received no previous formal CRM training before study participation.

Fifty-nine residents from several different training programs were recruited: 21 residents were from internal medicine, 15 residents from general surgery or other surgical specialties, 16 from family medicine, and three from anesthesiology and emergency medicine. Of the 16 family medicine residents, eight PGY-3 residents were in an overlap training year, with six residents in emergency medicine and two in anesthesiology. Further information is provided in Table 1.

Study Design

Participating residents were allocated into PGY-1 or PGY-3 groups. Each PGY-1 and PGY-3 group of residents participated in three separate half-day sessions—the simulator tutorial session and two simulator sessions. Each participant performed as the lead physician for each simulator scenario. The same two scenarios were given to all participants in the same sequence—cardiac events in a post-operative patient, and a patient presenting in acute shock and respiratory failure after severe trauma from a fall. The first simulator session took place 2 days after a tutorial session, whereas the second simulator session took place 2–3 weeks after the first simulator session.

Resident performance during the simulator scenario sessions was videotaped. To maintain anonymity, the videotapes were digitally altered to conceal the identities of the participating residents. The digitally altered videotapes of each simulator session were then converted to Video CD (VCD) format and stored on recordable DVD discs (DVD-r) in a randomized sequence for rater evaluation. Three raters then scored each session using the Ottawa GRS.

Simulator Environment

The University of Ottawa Patient Care Simulation Center provided a physical recreation of both an operating room and ICU environment. The simulator mannequin, Eagle Medsim, provided a simulation of a critically ill patient (6, 7). The simulator instructor performed both the operation of the mannequin software and the coordination of actor responses from an adjacent room with a two-way mirror.

In addition to the physical environment and patient, an ICU registered nurse (RN) and respiratory therapist (RT) were present during the simulator sessions to provide the assistance normally given during a medical emergency. One RN participated in all scenarios, whereas four RTs were used. The staff consisted of experienced professionals from the Ottawa Hospital ICU and received prior training to standardize their performance for each case. In both scenarios, both the RN and RT gave pre-scripted responses to resident action and/or inaction. Both the RN and RT actors received extensive pre-case training and also received instructions from the simulator instructor via headset communication to ensure standardization of responses.

Simulator Tutorial Session

Residents participated in a tutorial session 2 days before the first simulator session. Residents received a brief orientation to the simulator room and mannequin. Tutorial topics covered included the management of acute respiratory failure, basic airway management, and the management of shock. CRM instruction was not provided. The tutorial session was implemented to ensure that lack of knowledge was not responsible for differences in performance and to minimize the lack of familiarity with the simulator environment as a possible influence on simulator performance.

Simulator Case Scenario Sessions

The simulator scenarios represented recreations of common emergencies seen in the ICU, postanesthesia care unit, or emergency room environment. 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
Evaluation of Global Rating Scale—Construct Validity

The American Educational Research Association and American Psychological Association Standards for Educational and Psychological Testing have identified several components of evidence to support the presence of construct validity (5, 8). They are as follows:

Content Validity. Content validity refers to whether the rating instrument covers all the relevant domains of CRM. The Ottawa GRS was divided into five subsets of CRM skills, based on the original work by Gaba (1), and an overall CRM performance category (Appendix). Each category was measured on a 7-point Likert scale with descriptive anchors to provide guidelines on alternating points along the scale. The scoring system was designed so that a score of 1 corresponded to the performance of a physician with sufficient CRM performance categories, whereas a score of 7 corresponded to the performance of a novice with some CRM and resuscitation experience. A score of 5 corresponded to the performance of a physician with sufficient CRM and resuscitation experience to manage critical events competently, whereas a score of 7 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. These descriptors were added to reduce personal bias in interpreting performance.

The initial setting and the clinical events in each case were identical for all residents. 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 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 to resuscitate critically ill patients.

RESULTS

A total of 32 PGY-1 and 28 PGY-3 residents were recruited into the study from August 2001 to August 2002. One PGY-1 and one PGY-3 resident withdrew from the study after participating in one simulator session. During the video recording process, three PGY-1 sessions and three PGY-3 sessions were damaged, leaving 112 resident sessions were available for analysis. Additional videos where only one simulator session was available for scoring were included for analysis of construct validity and interrater reliability. Videos from residents with only one session for scoring were excluded from analysis of intercase reliability.

Table 1 outlines the demographic makeup of both groups. As expected, the PGY-3 group had a higher amount of previous ICU experience and were older.

Evaluation of Global Rating Scale—Reliability

Reliability of the Ottawa GRS was assessed by measures of internal consistency and stability. Internal consistency was assessed by calculating Cronbach’s alpha (9) for the entire scale (five areas of assessment plus the overall rating). Intra-observer reliability was also measured, comparing resident performance from the first to the second scenario. Evaluation of interobserver reliability was also performed, as all three raters rated each individual case scenario session. A type III intraclass correlation coefficient (ICC) to determine interrater reliability for the overall CRM performance score and individual CRM category scores for the Ottawa GRS was used.
Table 2. Ottawa Crisis Resource Management Global Rating Scale scores postgraduate year (PGY)-1 vs. PGY-3 overall crisis resource management (CRM) performance

<table>
<thead>
<tr>
<th>Session No.</th>
<th>PGY-1 Overall CRM Score</th>
<th>PGY-3 Overall CRM Score</th>
<th>Mean Difference* (95% Confidence Interval)</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>4.13 ± 0.87</td>
<td>5.54 ± 0.85</td>
<td>1.41 (0.92–1.90)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>1</td>
<td>3.84 ± 1.67</td>
<td>5.42 ± 1.28</td>
<td>1.57 (0.97–2.18)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>2</td>
<td>4.33 ± 1.14</td>
<td>5.79 ± 1.00</td>
<td>1.45 (0.87–2.04)</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

*Mean difference: PGY-3 minus PGY-1 scores.

Table 3. Ottawa Crisis Resource Management Global Rating Scale (Ottawa GRS) scores postgraduate (PGY)-1 vs. PGY-3 individual crisis resource management categories

<table>
<thead>
<tr>
<th>Session No/GRS Category</th>
<th>PGY-1 Score</th>
<th>PGY-3 Score</th>
<th>Mean Difference (95% Confidence Interval)</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1—overall</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leadership</td>
<td>3.84 ± 1.67</td>
<td>5.42 ± 1.28</td>
<td>1.57 (0.97–2.18)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Problem solving</td>
<td>3.84 ± 1.26</td>
<td>5.40 ± 0.91</td>
<td>1.56 (0.95–2.17)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Situational awareness</td>
<td>3.85 ± 1.19</td>
<td>5.23 ± 0.96</td>
<td>1.38 (0.79–1.97)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Resource utilization</td>
<td>4.46 ± 0.90</td>
<td>5.43 ± 0.76</td>
<td>0.97 (0.51–1.43)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Communication</td>
<td>4.91 ± 0.83</td>
<td>5.57 ± 0.63</td>
<td>0.66 (0.26–1.07)</td>
<td>.0019</td>
</tr>
<tr>
<td>Session 2—overall</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leadership</td>
<td>4.33 ± 1.14</td>
<td>5.79 ± 1.00</td>
<td>1.45 (0.87–2.04)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Problem solving</td>
<td>4.56 ± 1.09</td>
<td>5.86 ± 0.81</td>
<td>1.31 (0.73–1.89)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Situational awareness</td>
<td>4.35 ± 1.11</td>
<td>5.79 ± 1.12</td>
<td>1.44 (0.81–2.08)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Resource utilization</td>
<td>4.30 ± 1.12</td>
<td>5.74 ± 0.99</td>
<td>1.44 (0.87–2.02)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Communication</td>
<td>4.73 ± 0.88</td>
<td>5.90 ± 0.71</td>
<td>1.17 (0.74–1.61)</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Table 4. Ottawa Crisis Resource Management Global Rating Scale scores session 1 vs. session 2 overall crisis resource management (CRM) performance

<table>
<thead>
<tr>
<th>Level of Training</th>
<th>Session 1 Overall CRM Score</th>
<th>Session 2 Overall CRM Score</th>
<th>Mean Difference (95% Confidence Interval)</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall*</td>
<td>4.52 ± 1.36</td>
<td>5.06 ± 1.29</td>
<td>0.51 (−0.08 to 0.71)</td>
<td>.115</td>
</tr>
<tr>
<td>PGY-1*</td>
<td>3.93 ± 1.27</td>
<td>4.33 ± 1.14</td>
<td>0.40 (−0.27 to 1.07)</td>
<td>.233</td>
</tr>
<tr>
<td>PGY-3*</td>
<td>5.43 ± 0.93</td>
<td>5.65 ± 1.01</td>
<td>0.22 (−0.20 to 0.63)</td>
<td>.288</td>
</tr>
</tbody>
</table>

*PGY, postgraduate year.

Table 5. Ottawa Crisis Resource Management Global Rating Scale (GRS) interrater reliability intraclass correlation coefficient (ICC)

<table>
<thead>
<tr>
<th>GRS Score</th>
<th>Type III ICC (ICC31)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1—Overall</td>
<td>.590</td>
</tr>
<tr>
<td>Leadership</td>
<td>.491</td>
</tr>
<tr>
<td>Problem solving</td>
<td>.551</td>
</tr>
<tr>
<td>Situational awareness</td>
<td>.475</td>
</tr>
<tr>
<td>Resource utilization</td>
<td>.346</td>
</tr>
<tr>
<td>Communication</td>
<td>.436</td>
</tr>
<tr>
<td>Session 2—Overall</td>
<td>.613</td>
</tr>
<tr>
<td>Leadership</td>
<td>.626</td>
</tr>
<tr>
<td>Problem solving</td>
<td>.567</td>
</tr>
<tr>
<td>Situational awareness</td>
<td>.544</td>
</tr>
<tr>
<td>Resource utilization</td>
<td>.355</td>
</tr>
<tr>
<td>Communication</td>
<td>.384</td>
</tr>
</tbody>
</table>

A comparison of PGY-1 and PGY-3 Ottawa GRS scores served as the first measure of construct validity. A statistically significant difference between PGY-1 and PGY-3 residents was noted, with PGY-3 residents achieving higher scores than PGY-1 residents (p < .0001). This difference was also noted in both scenario 1 and scenario 2 (Table 2).

Individual CRM categories within the Ottawa GRS underwent analysis for construct validity (Table 3). The scores between PGY-1 and PGY-3 residents for all items were significantly different, with PGY-3 residents achieving higher scores in all categories.

Intercase reliability and construct validity were examined by a comparison of resident scores from scenario 1 to scenario 2 (Table 4). No statistically significant differences in overall CRM performance for both the PGY-1 and PGY-3 groups were found.

The Ottawa GRS overall CRM score demonstrated an ICC score of .590 and .613 for scenarios 1 and 2, respectively (Table 5). Further analysis of Ottawa GRS categories revealed similar reliability scores for problem-solving, leadership, and situational awareness, with ICC scores ranging from .475 to .626. The categories of resource utilization and communication demonstrated much lower reliability scores, with ICC scores ranging from .236 to .384.

Individual rater scores for overall CRM performance were also examined. The analysis also demonstrated a statistically significant difference between PGY-1 and PGY-3 residents with each rater, with PGY-3 residents scoring higher in each case (Table 6).

**DISCUSSION**

Validation of any potential evaluation device in medical education is a daunting task. With a few notable exceptions (such as the objective-structured case examinations, multiple choice questions, etc) (11–16), many instruments and/or mediums used in medical evaluation have not undergone formal validation. Such is the case with high-fidelity human patient simulation (17, 18). Most of the prior studies with high-fidelity simulation were focused on OR cases with specific, recognizable solutions; none directly measured CRM performance (19–35). One other measuring instrument, called the ANTS system (36), is based on a taxonomy of anesthesiologists’ nontechnical skills and also provides categories to measure nontechnical skills. It has demonstrated considerable promise when used in the OR setting (36, 37). This study represents the initial step in the validation process in using human patient simulation as a medium for specific assessment of CRM performance and the Ottawa GRS as the tool for CRM assessment during “generic” acute care emergencies.

Reliability (are the measurements reproducible?) and validity (are the measurements actually measuring what they are supposed to?) are two essential characteristics of any evaluation device. For this reason, both characteristics were examined in this study.

There are many aspects to validity. Streiner and Norman (38) originally identified several key characteristics in the validation of evaluation scales, including content validity, concurrent validity, and construct validity. The American Educational Research Association and American Psychological Association...
Standards for Educational and Psychological Testing have since redefined construct validity to encompass the whole of validity (5, 8).

Content validity refers to whether the rating instrument covers all the relevant domains of CRM. As the traits mentioned in the Ottawa GRS follow those listed by Gaba (1), and the Ottawa GRS was reviewed by both simulation and CRM instructors from across Canada, the Ottawa GRS appears to have content validity. As the simulator cases also underwent a similar peer review process, they also appear to have content validity.

Response process refers to the integrity of data and the maximum control and/or elimination of error associated with test administration (8). Both the support staff training and rater training process were provided to standardize test administration in both case delivery and evaluation of performance. Residents received an orientation session to familiarize themselves with the simulator environment, and each resident participated in identical scenarios. Therefore, the standardization of case delivery of case appears to be optimal. Given that an extensive Delphi rater training process was undertaken to complete the Ottawa GRS and that all videotapes were digitally converted to a uniform viewing format for rater evaluation, the rating process itself also appears to be meet optimal criteria to minimize error in the scoring process itself. From a response process perspective, both the medium of high-fidelity simulation and Ottawa GRS scoring process appear to support the evidence for construct validity.

Internal structure refers to the statistical or psychometric properties of the instrument itself (8). These include the difficulty (or ease) of each test item, the discriminatory ability to separate high and low-scoring participants with each test item, and the reliability of each item. There is no system to quantify the difficulty of each scenario, but the second scenario appeared to be more difficult than the first, given the lack of knowledge and/or training for residents training outside surgical and/or emergency medicine. Despite this apparent difference, both scenarios were able to demonstrate statistically significant differences between PGY-1 and PGY-3 residents in Ottawa GRS scores with all three raters. Thus, a discriminatory ability appears to be present in both scenarios. In addition, as all categories within the Ottawa GRS demonstrated a similar statistical significance between groups, the Ottawa GRS appears to have discriminatory abilities present as well. From a reliability perspective, the Ottawa GRS demonstrated acceptable interrater reliability. The overall CRM performance, leadership, problem-solving, and situational awareness categories all demonstrated similar ICC scores. These results are encouraging, given the novelty of the Ottawa GRS. All of these results suggest that the internal structure data supports the evidence for construct validity.

The hypotheses generated in examining PGY-1 and PGY-3 performance comprise the component of relationship to other variables (in this case, level of training). The study results indicate that the Ottawa GRS could differentiate between PGY-1 and PGY-3 performance during simulator scenarios. These results support the premise that PGY-3 residents should outperform PGY-1 residents in CRM during the simulated resuscitation of a critically ill patient. Although this premise appears both plausible and likely, it certainly has never been proven.

Readers may question whether the Ottawa GRS differentiated performance simply due to a difference in medical knowledge between the PGY-3 and PGY-1 group. It should be noted that less than one third of the PGY-3 group (nine of 28) came from a surgical or emergency room residency training background, yet they outperformed the PGY-1 group during the trauma scenario as well as the cardiac scenario. A difference in case-specific knowledge alone would appear insufficient to explain this difference. In addition, the exclusion criterion of prior simulator experience prevented anesthesia PGY-3 residents from study participation. By eliminating one of the resident groups with the broadest set of knowledge and skills required to resuscitate acutely ill patients, one could have anticipated that the difference in CRM skill levels between PGY-1 and PGY-3 residents would be not as pronounced if medical skills and knowledge were solely responsible for effective CRM (Table 1). Despite this, all three raters consistently and reliably noted a significant difference in overall CRM scores between the PGY-1 and PGY-3 resident groups (Table 2). These results would suggest that the Ottawa GRS differentiated groups based not on differences in medical knowledge but on some other skill set, such as CRM performance. The fact that all CRM categories in the Ottawa GRS demonstrated this effect would suggest that from a relationship-to-other-variables perspective, construct validity is present (Table 3).

The results of this study support the presence of construct validity and interrater reliability when using the Ottawa GRS to evaluate CRM performance during high-fidelity simulated emergencies. However, there are several limitations of this study. First, although a significant difference between PGY-1 and PGY-3 residents was observed, these results only suggest that the Ottawa GRS will distinguish CRM performance at different levels of expertise. A more robust analysis would involve PGY-1 residents returning in their PGY-2 and PGY-3 years to participate in new simulator sessions. In this manner, each resident’s first PGY-1 score would serve as a true control score. This design change has already been incorporated for future studies.

Second, a more important issue is whether the differences in CRM performance are clinically relevant. As a score of 5 was designated as indicative of competent performance, the fact that the mean score of PGY-3 residents exceeds

Table 6. Ottawa Crisis Resource Management Global Rating Scale scores postgraduate year (PGY)-1 vs. PGY-3 by individual rater (overall crisis resource management performance)

<table>
<thead>
<tr>
<th>Session No./Rater</th>
<th>PGY-1 Score</th>
<th>PGY-3 Score</th>
<th>Mean Difference (95% Confidence Interval)</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 1</td>
<td>3.97 ± 1.51</td>
<td>5.42 ± 1.14</td>
<td>1.45 (0.71–2.19)</td>
<td>.0002</td>
</tr>
<tr>
<td>Session 2</td>
<td>4.00 ± 1.33</td>
<td>5.63 ± 0.97</td>
<td>1.63 (1.00–2.27)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Rater B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 1</td>
<td>4.21 ± 1.86</td>
<td>5.58 ± 1.28</td>
<td>1.37 (0.48–2.25)</td>
<td>.0032</td>
</tr>
<tr>
<td>Session 2</td>
<td>4.93 ± 1.66</td>
<td>6.11 ± 1.37</td>
<td>1.18 (0.35–2.02)</td>
<td>.0061</td>
</tr>
<tr>
<td>Rater C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 1</td>
<td>3.34 ± 1.23</td>
<td>5.25 ± 0.97</td>
<td>1.91 (1.29–2.52)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Session 2</td>
<td>4.07 ± 1.30</td>
<td>5.63 ± 1.18</td>
<td>1.56 (0.88–2.23)</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Crit Care Med 2006 Vol. 34, No. 8

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.
this figure and the mean score of the PGY-1 residents falls below this level suggests that the observed difference may have clinical relevance. However, as the study was not designed to address the issue of clinically relevant differences in scoring or the issue of summative scoring, future studies will be required to address these important issues.

Third, no significant difference in CRM scores was observed from the first to the second scenario. However, given the short time interval between the first and second simulator scenario and the lack of formal CRM instruction between sessions, this result was not surprising. In fact, this result would appear to strengthen the ability of high-fidelity simulation to formally assess CRM performance, as participation in the first session did not appear to be influenced performance in the second scenario. However, as the educational effect of study participation was not a primary end point, it is premature to state whether a participation effect is present. A more robust analysis on the effect of simulation participation and/or CRM instruction would be the repeated observation of resident performance over a longer interval of time after having received formal CRM instruction.

Finally, although an ICC score of 0.60 represents an acceptable level of interrater reliability, it is not ideal for high stakes (or “summative”) evaluation. Both the Ottawa GRS and high-fidelity simulation are unproven tools and/or mediums for evaluation. Therefore, reliability issues with the Ottawa GRS could have been due to errors in instrument design, errors in rater training, or a combination of both factors. The poor ICC scores seen in the resource utilization and communication categories indicate that at least some revisions in the scoring system and/or descriptive anchors are required. Conversely, in viewing the individual rater scores, the so-called “dove/hawk” rater effect was observed (39). Rater B clearly demonstrated overall CRM performance scores higher than rater A and rater C in both simulator scenarios, with a mean score almost a half-point greater in scenario 2. This finding suggests that the rater training process may be partially responsible for some of the variability between raters. Therefore, it is likely that revisions to the Ottawa GRS and rater training process will be required before greater interrater reliability can be demonstrated.

Although validity and reliability are key characteristics of any potential evaluation device, feasibility is another key trait. Each session takes a minimum of 20 mins to complete, and another 20 mins for evaluation when using videotape review. At least one instructor and two actors are required to be present for each session. One issue that requires further exploration is the portability of the Ottawa GRS to other sites. Another feasibility issue for the future use of high-fidelity simulation lies in the issue of content specificity (40, 41). Medical education literature indicates that with conventional examinations, problem solving is dependent on the knowledge base required to solve a particular problem in question. Therefore, this would suggest that multiple simulator sessions are required to reliably and accurately assess resident performance. The principle of content specificity certainly holds true for conventional examination—most summative Objective Structured Clinical Examination (OSCE) examinations use a minimum of 10 to 15 stations. It is also understood that an absolute lack of knowledge would be the most important weakness in a medical emergency.

Reviews of OR emergencies indicate that errors in CRM, and not medical knowledge, are most responsible for human error (1–3). Because acute resuscitation is a medical skill that conventional examination methods have been unable to explore or evaluate, content specificity may be less applicable to CRM. In terms of content, two diametrically opposed cases were chosen in this study—cardiac ischemia with arrhythmias vs. the trauma patient. Yet, in both cases, each individual rater was able to clearly separate PGY-1 and PGY-3 groups by Ottawa GRS scores (Table 2). Moreover, designing the second scenario of a trauma patient where at least half the residents (nonsurgical = 31 residents) do not participate in trauma resuscitation should have produced a lower overall CRM score for the second scenario. The fact that the overall CRM score actually increased, albeit not to a significant level, does appear counterintuitive to the notion that content specificity is a key factor in the Ottawa GRS assessment of CRM skills. This study’s findings would suggest that perhaps high-fidelity simulation does examine a skill set that differs from conventional medical evaluation. However, to assume that content specificity is not a factor based on one single study would also contradict >30 yrs of medical literature. Therefore, further evaluation of the role of content specificity in CRM and high-fidelity simulation is required before final conclusions can be drawn.

Despite the lack of evidence for the role of computerized human patient simulation as an educational or evaluation tool, it is nevertheless being integrated into multiple academic centers (42). This phenomenon is likely in recognition of the fact that simulation represents the only safe alternative to real-life practice for the education and evaluation of acute resuscitation and CRM. This fact is likely the driving force behind the mandatory incorporation of high-fidelity simulation for training and evaluation in other high-risk professions such as the airline, aerospace, and nuclear industries. Interestingly, these professions have incorporated high-fidelity simulation as a mandatory component to evaluation, despite a paucity of evidence to support its use (43–48). Given the encouraging results in terms of validity for the Ottawa GRS in measuring CRM performance and the use of high-fidelity simulation as the medium for CRM performance evaluation, future studies are warranted. Future research will provide a more robust assessment of the Ottawa GRS as a formal tool to evaluate CRM performance after revisions are completed.

CONCLUSION

High-fidelity simulation remains largely untested as an evaluation medium in medical education. Due to its unique ability to simulate acute emergencies, it provides the opportunity to evaluate a skill set previously untested by conventional examinations: crisis resource management (CRM). CRM appears universally important during an acute emergency, and errors in CRM account for the majority of critical incidents due to human error.

In this study, raters using the Ottawa GRS demonstrated the ability to differentiate between PGY-1 and PGY-3 performance to a statistically significant degree in two separate simulator scenarios. This difference was noted in all Ottawa GRS categories and with each individual rater. This finding supports the presence of construct validity and acceptable interrater reliability when using the Ottawa GRS to evaluate CRM performance during high-fidelity simulation scenarios. Since a higher level of interrater reliability is required for high-stakes ("summative") evaluation, further revision of the Ottawa GRS design and rater training process is required before it can be used as a summative evaluation tool.
ACKNOWLEDGMENTS

We acknowledge Arthur Rothman, EdD, Ontario Institute for Studies in Higher Education/University of Toronto, for his assistance by review of the original study design as well as in the design and rater training process for the Ottawa GRS. We also acknowledge Paul Hebert, MD, Clinical Epidemiology Unit, University of Ottawa, and Dean Ferguson, PhD, Clinical Epidemiology Unit, University of Ottawa, for their assistance by review of the original study design and final manuscript.

REFERENCES

APPENDIX A – OTTAWA CRISIS RESOURCE MANAGEMENT (CRM) GLOBAL RATING SCALE ("Ottawa GRS")

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 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 #: ______________________________________

Staff: ______________________________________

Date: ______________________________________

Time: ______________________________________

OVERALL PERFORMANCE

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice; all CM skills require significant improvement</td>
<td>Advanced novice; many CM skills require moderate improvement</td>
<td>Competent; most CM skills require minor improvement</td>
<td>Clearly superior; few, if any CM skills that only require minor improvement</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

I. LEADERSHIP SKILLS
Loses calm and control for most of crisis; unable to make firm decisions; cannot maintain global perspective

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loses calm/control frequently during crisis; delays in making firm decisions (or with cueing); rarely maintains global perspective</td>
<td>Remains calm and in control for most of crisis; makes firm decisions with little delay; usually maintains global perspective</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

II. PROBLEM SOLVING SKILLS
Cannot implement ABC's assessment without direct cues; uses sequential management despite cues; fails to consider any alternative in crisis

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incomplete or slow ABC assessment; mostly uses sequential management approach unless cue; gives little consideration to alternatives</td>
<td>Satisfactory ABC assessment; without cues; mostly uses concurrent management approach with only minimal cueing; considers some alternatives in crisis</td>
<td>Thorough yet quick ABC without cues; always uses concurrent management approach; considers most likely alternatives in crisis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

III. SITUATIONAL AWARENESS SKILLS
Becomes fixated easily despite repeated cues; fails to re-assess and re-evaluate situation despite repeated cues; fails to anticipate likely events

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avoids fixation error only with cueing; rarely reassesses re-evaluates situation without cue; rarely anticipates likely events</td>
<td>Usually avoids fixation error with minimal cueing; reassesses re-evaluates situation frequently with minimal cue; usually anticipates likely events</td>
<td>Avoids any fixation error without cue; constantly reassesses re-evaluates situation without cue; constantly anticipates likely events</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

IV. RESOURCE UTILIZATION SKILLS
Unable to use resources & staff effectively; does not prioritize tasks or ask for help when required despite cues

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Able to use resources with minimal effectiveness; only prioritizes tasks or asks for help when required with cue</td>
<td>Able to use resources with moderate effectiveness; able to prioritize tasks and/or ask for help with minimal cue</td>
<td>Clearly able to utilize resources to maximal effectiveness; sets clear task priority and asks for help early with no cue</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

V. COMMUNICATION SKILLS
Does not communicate with staff; does not acknowledge staff communication, never uses directed verbal/non-verbal communication

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communicates occasionally with staff, but unclear and vague; occasionally listens to but rarely interacts with staff; rarely uses directed verbal/non-verbal communication</td>
<td>Communicates with staff clearly and concisely most of time, listens to staff feedback; usually uses directed verbal/non-verbal communication</td>
<td>Communicates clearly and concisely at all times, encourages input and listens to staff feedback; consistently uses directed verbal/non-verbal communication</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>