High-stakes assessment of the non-technical skills of critical care trainees using simulation: feasibility, acceptability and reliability

The use of simulation in assessment is an evolving area. Advocates have suggested that it will revolutionise health care, with credentialing mechanisms of the future based on experiential processes such as simulation. Simulation has been advocated for assessment of procedural competency and for its ability to recreate a complete patient encounter, including the need for communication, teamwork, decision making and technical skills. It has been suggested that simulation should be integrated into a comprehensive curriculum that includes specialist certification. It has been stated that traditional methods of assessment of clinical competence are “incomplete and archaic at best” and that summative assessment through simulation offers consistency and objectivity.

The psychometric characteristics of an assessment item include validity, reliability, feasibility and acceptability. Psychometric evaluation of assessment using simulation lags behind its implementation. A simulation station has been incorporated into the Australasian intensive care specialist certification examination.

Simulation can be used for assessment of the performance of critical care trainees in two domains: technical and non-technical skills (NTS). There have been several studies into psychometric characteristics in technical skills assessment. These studies were performed by simulation enthusiasts in a simulation suite environment, rather than as part of a broad-based multimodality high-stakes examination.

NTS assessment has been less well studied. NTS include effective communication, team leadership, task management, situational awareness and decision making. A deficit in NTS has been associated with poor patient outcomes in critical care. Formal training in NTS is increasing over time.

Although there is an increased interest in teaching NTS, there remains uncertainty about the best way to assess these skills. Tools have been created for the purpose of assessing a trainee’s non-technical performance in a simulated or actual patient management scenario, the most relevant being the Anaesthesia Non-Technical Skills (ANTS) scale and the Ottawa Global Rating Scale (GRS).

Objective: To evaluate the use of high-fidelity simulation for summative high-stakes assessment of intensive care trainees, focusing on non-technical skills (NTS), testing feasibility and acceptability of simulation assessment, and the reliability of two NTS rating scales.

Design, setting and participants: Prospective observational study of senior intensive care trainees in a simulated specialist examination.

Methods: Participants undertook a simulated patient-management scenario and were assessed using two rating scales: the Anaesthesia Non-technical Skills (ANTS) scale and the Ottawa Global Rating Scale (GRS). Assessors were trained, currently active, high-stakes examiners. Participants also completed a survey on simulation-based summative assessment.

Outcome measures: The inter-rater reliability of two rating scales for NTS assessment. We evaluated the feasibility of simulation-based assessment, and used survey results to assess acceptability to participants.

Results: Simulation assessment was feasible. Participants considered simulation-based high-stakes assessment to be acceptable and felt their scenario performance was reflective of real-world performance. Participants identified a need for debriefing following scenario-based assessment. Inter-rater reliability was fair for the ANTS and Ottawa GRS scores (intra-class correlation coefficient, 0.39 and 0.42, respectively). There was only fair agreement between raters for an NTS pass or fail (weighted kappa, 0.32) and for a technical skills pass or fail (weighted kappa, 0.36).

Conclusions: Summative high-stakes assessment using a single simulated scenario was feasible and acceptable to senior intensive care trainees. The low inter-rater reliability for the ANTS and Ottawa GRS rating scales and for pass or fail discrimination may limit its incorporation into an existing examination format.
The ANTS scale was designed to assess the workplace performance of anaesthesia trainees but is easily adaptable to the critical care environment; the Ottawa GRS is specific for intensive care practice. The ANTS system was devised after detailed task analysis by a team including industrial psychologists, and the Ottawa GRS was principally devised by practising clinicians. Construct validity has been established.

Neither of these tools was designed for summative assessment, but some would consider it desirable to specifically assess NTS in a high-stakes process. There is an established practice of summative assessment of NTS in other industries, particularly aviation. Validated rating systems exist, and a pilot’s failure in the non-technical component of their mandatory simulator assessment may result in a flying ban until non-technical performance has been remediated.

In this study we set out to evaluate summative assessment of senior intensive care medicine trainees in a simulated scenario. We incorporated a high-fidelity simulated scenario inserted into the current College of Intensive Care Medicine (CICM) of Australia and New Zealand examination format, that is, as one of a series of 10-minute, viva-voce-style stations. Although we gathered data about technical and NTS assessment, our focus was on NTS, given the larger existing body of research on technical skills.

Our primary aim was to evaluate the feasibility and acceptability of NTS assessment in a high-stakes format. We tested the reliability of the psychometric characteristics of the ANTS scale and the Ottawa GRS. We also sought to identify if there was a correlation between assessed technical performance and non-technical performance. We aimed to perform assessments using a real-world format that could be included in the existing specialist exam. We used existing CICM high-stakes examiners and ensured realistic requirements for infrastructure, time and examiner training.

### Methods

#### Participants
Study participants were 24 senior trainees in intensive care who were attending a pre-examination course structured in a mock-exam format. The course is a recognised last point of preparation for the exit examination. Most participants were completing their Australasian specialist exit examinations. Participants were a mean of 10.7 years from graduation and had completed a mean of 4 years of intensive care practice. Seven of the 24 participants described themselves as familiar with managing simulated patients.

#### Study design
Consent was obtained from the study participants. The study protocol was approved by the Princess Alexandra Hospital Human Research Ethics Committee.

The scenario station was one of a series of 10-minute viva-voce-style examination stations. Study participants were asked to manage a simulated patient in a 10-minute scenario. The scenario involved a trauma patient with a deteriorating level of consciousness using a mannequin patient simulator (Laerdal SimMan). Key participant actions included intubation, initial management of intracranial hypertension, liaison with neurosurgery and radiology staff, and management of junior staff. Confederate actors performed the roles of intensive care nurse and junior doctor. Before study commencement, the scenario was reviewed by five practising specialists to establish content validity. A standardised script was created and the scenario was rehearsed and road-tested to ensure reproducible delivery. Twenty-four participants completed the scenario and, despite a lack of prior participant familiarisation, all were completed without major technical issues.

The video recordings of participant performance were provided to six raters. All raters were current examiners.
for the CICM specialist examination. College requirements mandate that examiners regularly attend calibration and standardisation exercises. Before performing the NTS ratings, examiners underwent an NTS training package. The package included presentations and reading material on the subject of NTS and NTS assessment, as well as a number of test scenarios. The examiners were asked to rate the test scenarios, with the results calibrated against consensus marks. Completion time for the examiner training package was about 1 day.

After completing the training and calibration exercise, examiners were asked to rate each participant’s performance using the ANTS scale and the Ottawa GRS. We used a slight modification of the Ottawa GRS, as suggested by the original authors (Dr John Kim, Assistant Professor of Medicine, University of Ottawa, personal communication, 2009). Rating scales are shown in Appendix 1 and Appendix 2. Examiners were asked to provide a pass or fail assessment of each participant’s technical performance, non-technical performance, and an overall pass or fail mark. No criteria were provided for determining overall pass or fail, as we wished to identify factors leading an examiner to attribute an overall passing grade. Participants were asked to complete a survey questionnaire regarding the acceptability of the NTS assessment. Response options were on a five-point Likert scale, ranging from “strongly disagree” to “strongly agree”.

### Statistical analysis

SAS 9.2 for Windows (SAS Institute) was used for analysis. Intraclass correlation was used to estimate reliability coefficients for the ANTS and Ottawa scales. Simple kappa values were used for estimating agreement of all categorical variables.

The ANTS scale comprises 15 elements in four categories and statistical analysis of ANTS scores is reported using category scores only. Analysis using the sum of individual element scores within each category produced comparable results. The first element of the Ottawa GRS is overall performance and, to avoid redundancy, we did not include it in the summed overall score.

### Results

Six assessors completed the ANTS and Ottawa GRS scores for all participants. Due to technical factors, four assessors completed the pass or fail determinations. Inter-rater agreement for technical skills pass or fail was 0.36 (weighted average kappa) and for overall pass or fail, 0.28. For inter-rater reliability of the NTS scores, ANTS and Ottawa scores intraclass correlation coefficients (ICC) were 0.39 and 0.42, respectively (Table 1). Inter-rater agreement for NTS pass or fail discrimination was similar to inter-rater agreement for the ANTS and Ottawa scores; the weighted average kappa for all four examiners was 0.32 for NTS pass or fail.

Each rater scored each candidate for pass or fail in each of three domains (NTS, technical skills and overall). We examined the correlation between the pass or fail results given to an individual candidate by an individual rater (Table 2). The strongest agreement was between technical skills pass or fail and overall pass or fail ($\kappa$, 0.89), ie, if an individual rater gave a candidate a pass for technical skills, they were very likely to attribute the same candidate an overall pass, and were less likely to attribute an overall pass to a candidate who had achieved a non-technical pass ($\kappa$, 0.62).
The final component of our study was a survey of participants’ attitudes towards formal assessment of NTS using a simulated scenario. Participants were surveyed immediately after completing the scenario, with a response rate of 100%. Results are shown in Table 3. The survey results indicate strong support for the acceptability of formal assessment of NTS using a simulated scenario. Participants generally agreed that their behaviour in the scenario was representative of their capabilities in the workplace. An unexpected finding was a perceived need for debriefing, with 84% agreeing that “candidates must have an opportunity to debrief after a scenario-based assessment”.

Discussion

Our results show that summative assessment of technical and NTS in an exam-format simulated scenario is feasible, and highly acceptable to candidates. However, our data show limited inter-rater reliability for the currently available NTS assessment tools that are relevant to intensive care. Both tools performed similarly. Additionally, there was only fair inter-rater agreement for technical skills assessment and a simple pass or fail designation.

Our high-fidelity simulation scenario was run externally to a simulation centre, using a portable mannequin and equipment. In terms of feasibility, candidates perceived the environment, scenario and actors to be realistic. Those without prior simulation exposure were able to complete a scenario without familiarisation, and candidates felt that their simulated scenario performance reflected their real-life performance in a similar situation.

In terms of acceptability, candidates felt that it was reasonable to be assessed using a simulated scenario, and agreed that NTS assessment in a simulated scenario was reasonable.

However, the reliability of assessment using a simulated scenario was limited. Previously published studies evaluating performance using simulation in a low-stakes context have shown generally higher inter-rater reliability. The statistics used in different studies vary somewhat (eg, ICC vs Pearson correlation coefficient $r$), but they are roughly comparable. Reported figures range from $\text{ICC} = 0.36$ to $r = 0.97$. These studies generally used single-dimension scales and checklists to assess technical skills performance.

Assessing technical skills was not our primary aim, and we only assessed technical skills at the level of pass or fail. We found limited inter-rater reliability for technical skills pass or fail. In some countries (such as Israel), assessment of technical performance in simulated scenarios has already been incorporated into high-stakes specialist certification and recertification. In those jurisdictions, assessment of technical performance has been implemented using trained simulation educators in a simulation facility.

Fewer studies have examined inter-rater correlation of NTS assessment. Reported figures have ranged from $r = 0.57$ to $r = 0.88$ using one to eight scenarios, also performed in a simulation suite. NTS scores have been shown to be more stable in multiple scenarios than technical skills scores, thus should lend themselves better to single-scenario assessment.

Our finding was that summative assessment of technical skills and NTS using performance in a single simulated scenario was subjective to a worrying degree when performed by high-stakes examiners who were not all simulation educators. This leads us to caution against the use of this modality currently for high-stakes assessment in intensive care.

Our data demonstrated lower inter-rater correlation than previous reports using a single scenario. Our study is unique in using currently active high-stakes examiners rather than simulation educators to rate our participants. Our examiners had all undertaken examiner-calibration training exercises and all underwent NTS training, but were not chosen due to a particular interest in simulation. Our results may therefore be similar to those generated by incorporation of a simulation component into an actual high-stakes examination.

By asking the raters to make a determination of pass or fail in two domains as well as overall pass or fail, we have identified the skills considered by examiners to be most important. The closest correlation was between a technical skills pass or fail and an overall pass or fail, indicating that examiners gave substantially more weight to technical skills performance than to NTS performance when determining an overall pass or fail mark. This may indicate examiner bias towards technical skills performance over NTS.

Our study is limited by the small sample size. Another limitation is the use of video for assessment. Video recording was essential for the purposes of our study, but may not be representative of real-world assessment, in which the examiner must make a determination on the spot without the luxury of being able to review a candidate’s performance.

Shortcomings of our study include the short duration of examiner training, the use of a single 10-minute examination episode, and the fact that our high-stakes format was simulated rather than a real, high-stakes examination, but these limitations are to some extent intentional. The mode of implementation represents a real-world scenario for adoption of simulation-based technical and
NTS assessment into a current exam format. It would be unrealistic to expect examiners, who are practising senior clinicians, to spend longer than a day on NTS training in addition to their existing examination writing, assessment and calibration duties.

We used one scenario incorporated into the existing exam format, which reflects recent practice in the CICM exit examination. Having multiple scenarios would significantly lengthen the exam.

The use of a simulated (rather than real) high-stakes examination permitted the use of multiple raters and the calculation of inter-rater correlation. The trainees were very close to completing their actual exit exam, and trainee survey data indicate that the mock exam was taken as seriously as a real exam.

Conclusions
Our data indicate that summative assessment of technical skills and NTS, using simulation, is acceptable to participants and can be feasibly integrated into an existing examination format, but the inter-rater reliability of a single assessment episode was limited. Correlation was limited for the ANTS and Ottawa GRS scores, and simple pass or fail determinations. We caution against adoption of high-stakes, summative assessment of NTS using a simulated scenario incorporated into an existing examination format. If it were seen to be desirable to summatively assess NTS in intensive care trainees using simulation, this would require further psychometric validation before implementation. It may be preferable to establish a dedicated testing session in a simulation facility rather than incorporation into an existing examination format.

Acknowledgements
We thank Jeremy Cohen, Amod Karnik and Mary Pinder for assessing the performance of trainees, and Kim Vidhani for trainee assistance.

Competing interests
None declared.

Author details
Leo Nunnink, Senior Lecturer
Carole Foot, Clinical Associate Professor
Bala Venkatesh, Professor
Charlie Corke, Associate Professor
Manoj Saxena, Senior Lecturer
Mark Lucey, Senior Lecturer
Mark Jones, Senior Lecturer

References
1 University of Queensland, Brisbane, QLD, Australia.
2 University of Sydney, Sydney, NSW, Australia.
3 Geelong Hospital, Geelong, VIC, Australia.
4 University of New South Wales, Sydney, NSW, Australia.
Correspondence: l.nunnink@uq.edu.au

Appendix 1. Anaesthesia Non-Technical Skills (ANTS) scale (enter rating for each category and element)

<table>
<thead>
<tr>
<th>Category</th>
<th>Element</th>
<th>Element rating (out of 4)</th>
<th>Category rating (out of 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task management</td>
<td>Planning and preparing</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prioritising</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Providing and maintaining standards</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Identifying and utilising resources</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team working</td>
<td>Co-ordinating activities with team</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Exchanging information</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Using authority and assertiveness</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assessing capabilities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Support</td>
<td>Supporting others</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Situation</td>
<td>Gathering information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>awareness</td>
<td>Recognising and understanding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision making</td>
<td>Identifying options</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Balancing risks and selecting options</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Re-evaluating</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Rating descriptions: 1 Poor (endangered or potentially endangered patient safety, serious remediation required); 2 Marginal (cause for concern, considerable improvement needed); 3 Acceptable (satisfactory standard but could be improved); 4 Good (consistently high standard, enhancing patient safety; could be used as positive example for others); Not observed: skill could not be observed in this scenario.
### Appendix 2. Ottawa global rating scale

<table>
<thead>
<tr>
<th>Skill assessed</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall performance</td>
<td>Novice; all CRM skills require significant improvement</td>
</tr>
<tr>
<td>I. Problem solving</td>
<td>Cannot implement ABC assessment without direct cues; uses sequential management despite cues</td>
</tr>
<tr>
<td>II. Situational awareness</td>
<td>Becomes fixated easily despite repeated cues; fails to re-assess and re-evaluate despite repeated cues</td>
</tr>
<tr>
<td>III. Anticipation and planning</td>
<td>Fails to anticipate likely events; lacks any organisation in plan of action; completely reactive in plan of action</td>
</tr>
<tr>
<td>IV. Leadership</td>
<td>Loses calm and control for most of crisis; unable to make firm decisions; cannot maintain global perspective</td>
</tr>
<tr>
<td>V. Clinical management</td>
<td>Fails to diagnose major events; fails to implement definitive management, and is unable to implement appropriate doses of agents, despite repeated cues</td>
</tr>
<tr>
<td>VI. Resource utilisation</td>
<td>Unable to allocate resources appropriately; fails to prioritise tasks or call for help when required, despite cues</td>
</tr>
<tr>
<td>VIIa. Communication (output)</td>
<td>Rarely initiates communication with staff; fails to communicate clearly and concisely; fails to use directed communication with staff</td>
</tr>
<tr>
<td>VIIb. Communication (input)</td>
<td>Fails to acknowledge team input despite cues; fails to seek team input when appropriate</td>
</tr>
</tbody>
</table>

ABC = airway, breathing and circulation. CRM = crisis resource management.