Evaluation of high fidelity patient simulator in assessment of performance of anaesthetists

J. M. Weller 1*, M. Bloch 2, S. Young 3, M. Maze 2, S. Oyesola 2, J. Wyner 2, D. Dob 2, K. Haire 2, J. Durbridge 2, T. Walker 2 and D. Newble 4

1Department of Surgery, Wellington School of Medicine, Otago University, Private Bag 7343, Wellington South, New Zealand. 2Magill Department of Anaesthetics, Chelsea and Westminster Hospital, London, UK. 3Department of Clinical Engineering, Charing Cross Hospital, London, UK. 4Department of Medical Education, Northern General Hospital, Sheffield, UK

*Corresponding author. E-mail: jennifer.weller@wnmeds.ac.nz

Background. There is increasing emphasis on performance-based assessment of clinical competence. The High Fidelity Patient Simulator (HPS) may be useful for assessment of clinical practice in anaesthesia, but needs formal evaluation of validity, reliability, feasibility and effect on learning. We set out to assess the reliability of a global rating scale for scoring simulator performance in crisis management.

Methods. Using a global rating scale, three judges independently rated videotapes of anaesthetists in simulated crises in the operating theatre. Five anaesthetists then independently rated subsets of these videotapes.

Results. There was good agreement between raters for medical management, behavioural attributes and overall performance. Agreement was high for both the initial judges and the five additional raters.

Conclusions. Using a global scale to assess simulator performance, we found good inter-rater reliability for scoring performance in a crisis. We estimate that two judges should provide a reliable assessment. High fidelity simulation should be studied further for assessing clinical performance.

Br J Anaesth 2003; 90: 43–7

Keywords: anaesthetists, clinical competence; computers, computer simulation; education, educational measurement

Accepted for publication: June 25, 2002

Assessment of clinical competence is moving away from testing what a physician knows, towards assessing what a physician does in clinical practice.1 Direct observation of performance in the workplace should be the most valid method for assessing clinical performance, but has obvious limitations. Assessment of performance in a crisis is particularly difficult.

As a substitute for direct observation of real-life performance, the High Fidelity Patient Simulator (HPS) has certain advantages. The clinical conditions can be standardized, scheduled, repeated, and videotaped and there is no need to intervene for reasons of patient safety during critical events.

As with any new assessment method, the HPS requires evaluation before use in formal certification. The HPS has been tested as an assessment instrument.2 3 These studies have used extensive checklists to score performance. However, there is good evidence that global judgements concerning complex behaviours are reliable and may be more appropriate than highly structured checklists.4

This study uses a global scoring system to judge performance of anaesthetists in simulated crisis events. Aspects of validity, reliability and feasibility are reported.

Methods

Videotapes from Anaesthesia Crisis Resource Management (ACRM) courses at the Chelsea and Westminster Hospital between September 1999 and September 2000 were used for this study.

Using accepted practice guidelines and evidence from human factors research, four experienced anaesthetists agreed on the tasks involved in generic management of a
crisis in the operating theatre. Based on these tasks, a list of observable criteria was derived and grouped into two categories, medical management and behaviour (Table 1).

Development of rating scale
A global rating scale for these two categories and for overall performance was used to score performance of anaesthetists in simulated crisis events. The individual criteria served as a guide for the global score and were not independently rated. A five-point rating scale was used. The points on the scale were described as fail, borderline, acceptable, good and outstanding performance.

Pilot study
A pilot study was conducted in which 14 videotaped performances were rated by three primary judges. This established that the rating form was feasible to use in real time, that all the criteria could be observed in the selected clinical events and that the raters had common understanding of the criteria.

Tape selection
After the pilot study, 28 videotapes were selected from the available bank by a research assistant who was not involved in the rating process. Tapes were selected to give a range of performance in order to test the scoring system across the five-point scale. Selection was limited to four clinical scenarios: anaphylaxis, cardiac arrest, oxygen pipeline failure and malignant hyperthermia. The tapes showed a general view of the ‘operating theatre’, the anaesthetic machine and the monitor display. The course participant could obtain help from co-participants. Staff from the simulation centre played the parts of surgeon and operating theatre staff. Participants were identified only as specialist registrars or consultants.

Rating process
The tapes were viewed independently in real-time with no video replay. Viewing was in no predetermined order, as would be the case in an actual assessment process.

Rating by primary raters
Three primary raters, all involved in developing the rating system, independently rated 27 of the 28 tapes, two rating all 28 tapes.

Rating by additional raters
To ensure that agreement between raters was not restricted to those responsible for development of the rating form, a further five raters were recruited. As raters’ time was a limited resource, each additional rater independently scored a subset of between five and 17 of the 28 tapes. These tapes were selected by the study research assistant to include a range of performances. Different raters viewed a different subset of tapes to ensure a wide selection of performances were included. In total, 25 of the study tapes were viewed by additional raters. The minimum total number of ratings for these tapes was four, and the maximum seven (median five). All raters were specialist anaesthetists who underwent a 2 h training period with a primary rater. This included a discussion of the rating form, and a joint review of non-study tapes to establish common understanding of the criteria.

Statistical analysis
Data were analysed in several ways. The Intraclass Correlation Coefficient (ICC) was used to assess the inter-rater reliability. The ICC was estimated using a two-way random effects model. The variance components were estimated by restricted maximum likelihood using the mixed procedure in SAS statistical package (SAS Institute Inc., Cary, NC, USA). ICC for mean of two or three raters was calculated using the Spearman–Brown formula.

The Kendall Coefficient of Concordance (W) was used to assess the comparative performance of the three primary raters. The Spearman’s Rank Correlation Coefficient (p) was used to assess the comparative performance of every rater against the median of all available raters and to estimate the correlation between the scores for the three categories.

Ethics
Approval for this research was obtained from the Chelsea and Westminster Hospital Ethics Committee. Consent was obtained from all subjects.
Results
There was high reliability of assessment of overall performance, behaviour and knowledge when two or three raters scored the tapes, as indicated by the ICC values of 0.79 and 0.85 respectively (Table 2).

In addition, the Kendall Coefficient of Concordance was 0.77 for overall performance for the three primary raters ($P<0.001$). This also indicates good agreement between raters.

The median of all available ratings for each participant was used as our best estimate of the true score. Table 3 presents correlations between individual raters’ scores and the median scores. Scores assigned by the five new raters recruited to the team correlated as closely as key raters’ scores. All these correlations were statistically significant ($P<0.05$).

Relationship between qualities
There was a strong tendency for participants to score similarly in all three categories (Table 4). All correlations were highly statistically significant ($P<0.001$).

Pass/fail decisions
In only one of the 28 tapes (3.56%) was there significant disagreement between raters where a clear fail was awarded to a candidate where any other rater awarded a pass.

Discussion
There have been few studies of the HPS in performance assessment. We studied important aspects of HPS assessment including validity, inter-rater reliability and feasibility.

Issues of validity
The validity of an assessment has a number of components. Face validity of the simulator is high, in that it appears to assess what it sets out to assess. It is designed to be a highly realistic representation of an anaesthetic crisis in an operating theatre. Responses from participant questionnaires support the assumption that there is high face validity.

To ensure that the content of an assessment is valid, it must test across the range of required tasks. The criteria for this assessment were based on accepted principles of medical management and non-technical skills central to performance in a crisis. These criteria are supported by the literature on human factors and incorporated into anaesthesia training programs. To this extent, the content of this assessment is valid.

Construct validity can be evaluated by comparing the performance of groups of clinicians, where substantially better performance in one group would be expected. Although participants of varying levels of experience were included in the present study, the numbers were insufficient to allow meaningful comparison of groups with different experience. However, this has been confirmed in several studies assessing simulator performance where experienced anaesthetists performed better than less experienced anaesthetists, supporting the proposition that simulator performance has some validity as a measure of real-life performance.

Ideally, we should compare HPS performance with performance in real life. However, as there are few opportunities to do this for crisis management, HPS assessment could be compared with other assessments. Morgan’s comparison of medical undergraduate simulator performance with written exams and scores awarded by supervisors during clinical placements and found there was poor correlation between the three different assessment methods. If there is poor correlation between HPS performance and results of established assessments it will be difficult to decide which assessment is a more valid measure of performance in a crisis.

Most previous studies used extensive checklists to score performance in the simulator. This may be more objective than relying on the subjective opinion of judges. However, checklists direct examiners’ attentions to the details rather than the performance as a whole. Subjective judgements are still required to weight the individual items. The result is that the sum of the parts does not necessarily reflect the standard of the whole performance.
A global system is well established as an appropriate way of scoring complex performance. An advantage of HPS is the opportunity to observe overall performance. Global rating scales allow examiners to apply their professional judgement to assess performance as a whole, maximizing this advantage.

We found a high level of agreement between judges in all categories, and suggest that a global rating scale is a more valid measure of performance than checklists at this level of performance.

Issues of reliability
Several studies investigated agreement between raters when scoring simulator performance. Devitt and colleagues reported consistent scoring between raters observing anaesthetists instructed to work at a consistent level of performance. Gaba and colleagues found acceptable inter-rater reliability for technical aspects of performance of anaesthetists attending training courses. Morgan and Cleave-Hogg demonstrated good inter-rater reliability in assessing the performance of medical students in the simulator.

We found good inter-rater reliability for overall performance, medical management and behaviour. This agreement extended beyond the panel of three anaesthetists who developed the rating process, to five anaesthetists previously uninvolved in the study, suggesting this assessment process is widely applicable. Inter-rater reliability was high despite the more subjective scoring system.

In general, it is easy to recognize a clear pass or a clear fail. In this study, there was only one instance where a single tape was awarded a clear fail by an examiner where any other examiner awarded a pass.

A borderline assessment indicates the judge has difficulty deciding. Several strategies could be used to manage this borderline group. They could all be failed, reducing the risk to the public of certifying a poorly performing doctor, or they could all be passed, giving the candidate the benefit of the doubt. An alternative approach would be to extend testing time for borderline candidates.

Correlation between categories of performance
The high level of correlation between the three scores suggests one of several possibilities. Either participants tend to perform at a similar level across the two categories, or behaviour and medical management are interdependent (i.e. effective medical management requires good leadership, teamwork, communication and resource allocation) and effective behavioural strategies are dependent on sound medical knowledge. Although the individual criteria provide a framework for observation and feedback, for the purposes of summative assessment there may be little advantage in using more than the score for overall performance.

Feasibility
Total rating time in this study was 35 min per simulation, which is practical. The process is also feasible in terms of the number of examiners required to produce a reliable result.

Limitations
We did not assess performance in routine anaesthesia practice. Extreme circumstances may reveal deficiencies or outstanding qualities not obvious in routine practice and may make it easier for judges to agree on the level of performance. Performance in a crisis is only one aspect of anaesthesia practice, though of major importance. Many other aspects of performance, including crisis prevention, will be more appropriately assessed in other ways, including direct observation in the workplace.

This study assessed performance in a single simulated case. Performance between cases varies and extensive testing is required for a reliable result. We did not address this issue and the number of cases required for a reliable assessment for a particular candidate is unknown.

In conclusion, we have developed a global rating scale to assess performance of anaesthetists managing a simulated clinical crisis. We found a high level of agreement between raters. The method is feasible in terms of time and the number of raters required. Ongoing evaluation of HPS-based assessment will eventually provide us with a useful additional measure of performance for anaesthetists.

Acknowledgements
Elizabeth Whistance, Centre Administrator, helped coordinate the data collection. Dr Gordon Purdie, Biostatistician, Department of Public Health, Wellington School of Medicine, helped with the statistical analysis.

References
5 Bartko J. The Intraclass Correlation Coefficient as a measure of reliability. Psychol Rep 1966; 19: 3–11.
15 Byrne A, Jones J. Responses to simulated anaesthetic emergencies by anaesthetists with different durations of clinical experience. *Br J Anaesth* 1997; 78: 553–6