Assessment instruments used during anaesthetic simulation: review of published studies

A. J. Byrne\textsuperscript{1} and J. D. Greaves\textsuperscript{2}*  

\textsuperscript{1}Morriston Hospital, Swansea SA6 6NL, UK. \textsuperscript{2}Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, UK  

*Corresponding author

This review was undertaken to discover what assessment instruments have been used as measures of performance during anaesthesia simulation and whether their validity and reliability has been established. The literature describing the assessment of performance during simulated anaesthesia amounted to 13 reports published between 1980 and 2000. Only four of these were designed to investigate the validity or reliability of the assessment systems. We conclude that the efficacy of methodologies for assessment of performance during simulation is largely undetermined. The introduction of simulator-based tests for certification or re-certification of anaesthetists would be premature.

Br J Anaesth 2001; 86: 445–50  

Keywords: education, simulation of anaesthesia; anaesthesia, simulation; publications

Accepted for publication: October 27, 2000

Complex simulation of anaesthesia is an exciting development in medical education and has strong appeal as a way of measuring the performance of anaesthetists without the risks of using real patients.\textsuperscript{1} However, 15 yr ago, Vreuls and Overmayer wrote: ‘Performance measurement in otherwise sophisticated training simulators is so poorly designed that the measures are virtually useless.’\textsuperscript{2} In order to determine how much progress has been made since then we have examined five aspects of the reported assessments to evaluate a simulator-based assessment system.

(i) The type of simulator. Chopra\textsuperscript{3} developed a classification of simulators that divides simulators into three groups depending upon whether the simulation environment is of high, intermediate or low fidelity.

(ii) The type of simulation. A simulator can be used in a full theatre environment with a full complement of staff present, for short simulations with only the anaesthetist present, or as a device for training for specific tasks.

(iii) The efficacy of the assessment. Assessment of performance during simulation must be valid and reliable.\textsuperscript{4} A valid assessment tests what it is supposed to measure (Table 1). The reliability of a test reflects the consistency and precision of the observations, e.g. between assessors or between tests.

(iv) The aspects of performance to be tested. The simulation and assessment must be suited to the aspect of performance being measured. Assessment may be used to measure: propositional knowledge (facts), procedural knowledge, psychomotor skill, decision-making skills and professional behaviour.

(v) The assessment instruments applied in conjunction with the simulation. There are a number of measures of outcome that can be used during or after a simulator session (Table 2). These can be directed towards ‘single’ aspects of performance or they can depend upon a score given for the overall ‘global’ performance. Multiple instruments may be applied to a single simulation, especially when the sessions are videotaped.

Methods

Search strategy and criteria for assessment of the papers

Articles were sought that described the assessment of performance during simulated anaesthesia and that were published between 1980 and 2000. The review was restricted to articles that had been peer reviewed. Internet search engines were used to search for anaesthesia simulation sites and identify key workers in the field. The Medline and Embase databases were then searched using the words ‘anaesthesia’ and ‘anaesthesiology’ (in all their variations), ‘simulation’, ‘performance’, ‘competence’ and ‘assessment’, as well as the names of authors previously identified. The references in each report were examined for relevance to simulation and its associated authors were
made the basis of a repeat search. This was continued through three cycles. Studies were included only if they included an attempt to measure performance.

Principal findings

The features of the assessments in these reports are shown in Table 3. Gaba and DeAnda\(^5\) studied 19 first- and second-year anaesthesia residents during simulation on the Comprehensive Anaesthesia Simulation Environment (CASE) simulator. Their purpose was to determine the response of anaesthetists to critical incidents. As part of the study they attempted to determine whether there were differences between individual trainees or between experience levels in the rapidity of detection and correction of intra-operative problems. The simulations contained six adverse events. They assessed ‘time to detect’; ‘time to solve’ and adherence to protocols based on accepted resuscitation guidelines. Acceptable actions were classified as either ‘compensation criteria’ or ‘corrective criteria’, depending on whether they merely reduced the consequences of the problem or treated the cause. Deviations from the protocols were classified as ‘major’ or ‘minor’. The investigators showed that second-year residents corrected problems significantly more quickly than first-year residents (\(P<0.05\)) but that variations in performance were wide. ‘Time to detect’ and rates of adherence to protocols did not vary between the two groups studied.

DeAnda and Gaba\(^6\) studied eight experienced anaesthetists during the same simulation that had been used with trainee anaesthetists in the study above. The outcomes for the experienced anaesthetists were compared with those previously obtained from learners. There was a trend for earlier detection of the problems and this reached statistical significance during some of the critical incidents. Again, the main finding was of marked variability in performance.

Howard and colleagues\(^7\) used simple written questions to assess the understanding of Anesthetic Crisis Resource Management (ACRM) principles before and after a training course, using a high-fidelity theatre-type simulator (TTS). They showed that residents improved their scores after the simulation, but experienced practitioners did not.

Gaba and colleagues\(^8\) reported a study with the primary purpose of evaluating the inter-rater variability of a system of scoring the behaviour and technical performance of anaesthetists during simulations. They retrospectively reviewed videotapes of 72 participants in ACRM courses. Each subject was rated by five observers for crisis management behaviour and by three observers for technical actions. The study showed that, for technical actions, agreement between observers was good, whereas that for crisis management skills was less good.

Schwid and O’Donnell\(^9\) studied 30 anaesthetists during simulations using the Anaesthesia Simulator Consultant (ASC), a screen-based simulation programme. The purpose of the study was to ‘assess the ability of anaesthetists to recognise diagnostic clues, to make the diagnosis rapidly, and to evaluate the patient’s response during simulated critical incidents’. Each subject was tested for five simulations. Performance was assessed by adherence to management protocols and rated as either ‘correct’ or ‘incorrect’. There was no consistent relationship between experience

---

**Table 1** Some aspects of the validity of an assessment

<table>
<thead>
<tr>
<th>Type of validity</th>
<th>Definition/comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face validity</td>
<td>A measure of how appropriate the test items are to the purpose of the assessment. Can be determined by the opinion of an expert group.</td>
</tr>
<tr>
<td>Content validity</td>
<td>The extent to which a test is representative of the knowledge or skills that have been learned. To have high content validity a test must sample a full range of the learning objectives identified by the curriculum.</td>
</tr>
<tr>
<td>Construct validity</td>
<td>The extent to which the assessment reflects the concept that is being tested, e.g. does an IQ test really test ‘general intelligence’?</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>Relates the candidate’s performance to established standards of practice.</td>
</tr>
<tr>
<td>Concurrent validity</td>
<td>Relates the test performance to the candidate’s performance in tests that are believed to assess the same attribute.</td>
</tr>
</tbody>
</table>

**Table 2** Measures of outcome for use in conjunction with simulation

<table>
<thead>
<tr>
<th>Type of assessment</th>
<th>Definition/comments</th>
<th>Performance measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Debriefing</td>
<td>Written enquiries or interview can be used after simulations.</td>
<td>Single aspect</td>
</tr>
<tr>
<td>Simulated ‘mortality’</td>
<td>Can be used during simulations where the ‘patient’ is severely ill and will soon die unless correct treatments are started immediately. If the anaesthetist chooses the correct treatment, the patient lives and the anaesthetist is therefore deemed to be competent.</td>
<td>Global</td>
</tr>
<tr>
<td>Time to solve or detect</td>
<td>Represents the time between the onset of the problem and the point at which the anaesthetist identifies or solves the problem.</td>
<td>Global</td>
</tr>
<tr>
<td>Observation and scoring systems</td>
<td>These break down the task of treating the patient into a series of actions, each being given weighting. Observers can then evaluate subjects using a check list or score sheet, noting each of the subject’s actions and allocating marks to each.</td>
<td>Single aspect</td>
</tr>
<tr>
<td>Psychological tests</td>
<td>Observation of the subject’s performance in psychological terms.</td>
<td>Global or single aspect</td>
</tr>
<tr>
<td>Ref. No.</td>
<td>No. of subjects</td>
<td>Type of simulator</td>
</tr>
<tr>
<td>---------</td>
<td>----------------</td>
<td>------------------</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>hi-fi TTS</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>hi-fi TTS</td>
</tr>
<tr>
<td>7</td>
<td>hi-fi TTS</td>
<td>no</td>
</tr>
<tr>
<td>8</td>
<td>72</td>
<td>hi-fi TTS</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
<td>screen based</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>intermediate fidelity TTS</td>
</tr>
<tr>
<td>11</td>
<td>36</td>
<td>intermediate fidelity TTS</td>
</tr>
<tr>
<td>12</td>
<td>28</td>
<td>hi-fi TTS</td>
</tr>
<tr>
<td>13</td>
<td>80</td>
<td>intermediate fidelity TTS</td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>hi-fi TTS</td>
</tr>
<tr>
<td>15</td>
<td>25</td>
<td>hi-fi TTS</td>
</tr>
<tr>
<td>16</td>
<td>89</td>
<td>hi-fi TTS</td>
</tr>
<tr>
<td>17</td>
<td>24</td>
<td>hi-fi TTS</td>
</tr>
</tbody>
</table>
and performance. The major finding was a clear relationship between the ability of subjects to follow Advanced Cardiac Life Support (ACLS) guidelines and the period of time since ACLS training. No attempt was made to validate the assessment protocol.

Byrne and colleagues\textsuperscript{10} published a study using the Anaesthetic Computer-Controlled Emergency Situation Simulator (ACCESS) system, an intermediate-fidelity TTS. The simulations were short scenarios containing simple problems with well-defined solutions. Performance was assessed by ‘time to solve’ and ‘mortality’. The inexperienced subjects took longer to solve the problems and mortality was higher, but this was a preliminary study and no conclusions about the efficacy of the performance assessment were drawn from the documented data.

In 1997, Byrne and Jones\textsuperscript{11} studied more anaesthetists using the same technique as in the previous study. The anaesthetists were divided into four groups depending on the length of their anaesthetic experience. Anaesthetists with <1 yr’s experience took longer to treat the problem and caused more ‘deaths’. Although the result was statistically significant, there were wide variations in performance within each group.

Chopra and colleagues\textsuperscript{12} studied the performance of 28 trainee anaesthetists over two simulations using a high-fidelity TTS, the Leiden Anaesthesia Simulator. The purpose of the study was ‘to quantitatively evaluate the efficiency of simulation as a training tool in anaesthesia’. Subjects were asked to solve a crisis 4 weeks after being challenged with either the same crisis or a control unrelated emergency. Measurement of performance was based on a scoring system developed from accepted guidelines. Numerical weightings were assigned to each appropriate action as well as a judgement as to whether the action was of major or minor importance. For each individual, a ‘treatment score’ represented the total score for positive actions, a ‘deviation score’ represented the total of actions omitted and a ‘total performance score’ was calculated from the difference between the two scores. The time taken to ‘first response’ was also noted. For all four measured variables, performance during the simulation of malignant hyperthermia was significantly better in the group exposed to a training simulation of malignant hyperthermia, when compared with their baseline performance.

Lindekaer and colleagues\textsuperscript{13} studied performance in managing ventricular fibrillation. Eighty anaesthetists worked in 40 teams of two people during a simulation session of the SOPHUS simulator, a high-fidelity TTS. The simulator sessions were videotaped and reviewed along with the subjects. The performance of each team was compared with the European Resuscitation Guidelines, modified for an anaesthetized patient. The number of teams making each intervention and the times taken from the onset of arrhythmia were noted for each team. There was poor adherence to the resuscitation protocol.

Devitt and colleagues\textsuperscript{14} studied the inter-rater reliability of two anaesthetists who independently observed and scored a total of 30 clinical problems. Five problems were generated in each hour-long scenario. For each problem, the actions of the subject were rated as: 0, no response; 1, compensatory intervention; 2, corrective treatment. The two raters were in complete agreement over the score in 29 out of 30 problems. The authors report this as demonstrating a high level of inter-rater reliability.

In 1998, Devitt and colleagues\textsuperscript{15} described a study primarily designed to test the items in a rating system developed specifically to evaluate the performance of anaesthetists in a simulated system. The principal objective of the study was to determine whether the test used was reliable as judged by the internal consistency between test items. A secondary item was to see whether the test item could discriminate between two groups of anaesthetists with different lengths of experience. This would demonstrate that the test had construct validity. They studied the performance of 25 anaesthetists during a 1 h simulator session. The inter-rater reliability of the scoring system had been evaluated previously.\textsuperscript{14} The more experienced anaesthetists produced significantly higher scores ($P<0.001$). Individuals did not score consistently across all five scenarios, i.e. there was poor internal consistency between test items. This variation was mostly in two items and when these items were removed from each scenario the consistency was improved.

In 1998, Kurrek and colleagues\textsuperscript{16} studied the performance of 89 anaesthetists during a simulation, on a high-fidelity TTS, that contained an episode of ventricular fibrillation. Performance was rated on a three-point scale. Seventy per cent of participants had taken an ACLS course and these subjects achieved higher scores than those who had not been trained ($P<0.05$). The assessment was limited to measuring adherence to resuscitation guidelines in a small part of a lengthy simulation.

Morgan and Cleave Hogg\textsuperscript{17} reported a pilot study of an assessment of medical students who undertook the anaesthetic management of simulated patients in a high-fidelity TTS. Six 15 min scenarios were developed and each student worked through one of these. Each student’s performance was videotaped and five evaluators separately assessed it using standardized performance evaluation criteria. The authors do not explain how their scoring system operated. The simulator scores showed high inter-rater reliability. The simulator marks showed poor correlation with marks allocated for the clinical attachment and short answer questions.

Discussion

All the reports reviewed described assessment of performance in the course of conducting a simulated general anaesthetic during which critical incidents were presented.
What aspect of performance was tested?

No study directly assessed propositional knowledge. All the studies made some assessment of procedural knowledge using several assessment techniques: (i) compliance with a recognized treatment algorithm, e.g. ALS; (ii) compliance with criteria determined by the researchers as representing good practice; (iii) compliance with management systems developed for dealing with anaesthetic crises; (iv) complex observations with associated scoring systems; and (v) clinical judgement and decision making were assessed in several studies by the measurement of ‘time to respond’ or ‘time to solve’. The study by Gaba and colleagues included assessment of professional behaviours in the form of crisis management scoring. None of these investigations specifically assessed psychomotor skills.

Assessment instruments used

In many of these studies scoring was by review of videotapes. A number of workers used measures of ‘time to solve’ or ‘time to detect’. Byrne and colleagues used ‘patient death’ as a measure of outcome. Psychological tests have been used in the operating room to measure the performance of anaesthetists. There are no reports of this form of assessment being used in conjunction with anaesthetic simulation.

The efficacy of assessments used in the simulator

In this review we have found only four studies that were designed primarily to investigate aspects of the validity and reliability of assessments undertaken in the simulator.

Validity

None of the studies attempted to prove the face or content validity of their assessments. These aspects of the assessments cannot be taken for granted. The construct validity of the assessment system was investigated directly by Devitt and colleagues in a small study of 25 subjects. The individual test items showed poor internal consistency such that conclusions cannot be made about the performance of individuals. Few of these papers make reference to criterion validity, i.e. the extent to which the subject’s performance is judged adequate. They record the performance of individuals but do not attempt to transfer this into a measurement of competence. The raters in the study by Gaba and colleagues considered some of the performances ‘deficient’ if two of the observers considered that the subjects had omitted two actions that the simulator crew had rated as essential. Other studies do not identify any reference criteria for marking a performance as unsatisfactory.

Morgan and Cleave-Hogg discovered poor correlation of their simulator assessment with either written tests or the opinions of supervising consultants. The authors suggest that their new test measures different aspects of their students’ performance.

A number of studies found that individual test subjects had a wide variation in response to apparently similar test items. This can be explained as lack of concurrent validity in the assessment system. It is well recognized with other assessment systems, such as multiple choice questions, that apparently well designed questions can lead to unpredicted responses. These questions are usually discarded. It must be presumed that a similar situation exists in simulator assessments.

Reliability

Devitt and colleagues investigated the inter-rater reliability of scoring. Gaba and colleagues, using three raters, showed good inter-rater reliability in scoring technical actions. There was less inter-rater agreement between five raters scoring ACRM behaviours (such as orientation to case, communication, feedback, anticipation, etc.). The authors analyse the statistics of inter-rater reliability in depth and note that their scoring system requires further development before it can be used for assessment of competence in high-stakes situations such as certification. Morgan and Cleave-Hogg found that the inter-rater reliability of their assessment in 25 subjects was good.

Thirteen reports include comparative assessment of subjects working in anaesthesia simulators. In two of these the construct validity of the assessment is addressed and in three measures are described that attempt to determine the inter-rater reliability of the assessment. (i) Simulators can generate a variety of tasks that can be used as the basis for the assessment of performance. (ii) Simulators can be used to measure adherence to protocols. (iii) There is some preliminary evidence that scoring actions in response to simulated situations appears to show inter-rater reliability. The validity of observations is not established. (iv) The reported within-subject and within-group variability calls into question some of the stimulus–response expectations of the investigators. (v) Attempts to assess the performance of anaesthetists are in their infancy and few of the studies to date have specifically designed assessment to address the questions of validity and reliability. It must be concluded that any moves towards using performance in administering simulated anaesthesia, as assessment in training or in recertification procedures are premature.

References

11 Byrne AJ, Jones JG. Responses to simulated anesthetic emergencies by anesthetists with different duration of clinical experience. Br J Anaesth 1997; 78: 553–6
19 Weigner MB, Herdon OW, Gaba DM. The effect of electronic record keeping and transesophageal echocardiography on task distribution, workload and vigilance during cardiac anesthesia. Anesthesiology 1997; 87: 144–55