Testing the raters: inter-rater reliability of standardized anaesthesia simulator performance

J. Hugh Devitt MD MSc FRCPC,*
Matt M. Kurrek MD,*
Marsha M. Cohen MD BSc MSc FRCPC,*†
Kevin Fish MD MSc FRCPC,*
Pamela Fish MD,*
Patricia M. Murphy MD FRCPC,*
John-Paul Szalai PhD‡

Purpose: Assessment of physician performance has been a subjective process. An anaesthesia simulator could be used for a more structured and standardized evaluation but its reliability for this purpose is not known. We sought to determine if observers witnessing the same event in an anaesthesia simulator would agree on their rating of anaesthetist performance.

Methods: The study had the approval of the research ethics board. Two one-hour clinical scenarios were developed, each containing five anaesthetic problems. For each problem, a rating scale defined the appropriate score (no response to the situation: score=0; compensating intervention defined as physiological correction: score=1; corrective treatment: defined as definitive therapy score=2). Video tape recordings, for assessment of inter-rater reliability, were generated through role-playing with recording of the two scenarios three times each resulting in a total of 30 events to be evaluated. Two clinical anaesthetists, uninvolved in the development of the study and the clinical scenarios, reviewed and scored each of the 30 problems independently. The scores produced by the two observers were compared using the kappa statistic of agreement.

Results: The raters were in complete agreement on 29 of the 30 items. There was excellent inter-rater reliability (=0.96, $P < 0.001$).

Conclusion: The use of videotapes allowed the scenarios to be scored by reproducing the same event for each observer. There was excellent inter-rater agreement within the confines of the study. Rating of video recordings of anaesthetist performance in a simulation setting can be used for scoring of performance. The validity of the scenarios and the scoring system for assessing clinician performance have yet to be determined.

Objectif : En médecine, l'évaluation de la performance demeure subjective. En anesthésie, un simulateur peut être utilisé pour fournir une évaluation mieux structurée et standardisée mais on n'en connaît pas la fiabilité. Nous avons cherché à déterminer si, en anesthésie, les observateurs d'un phénomène simulé pouvaient s'entendre sur leur appréciation de la performance de l'anesthesiste.

Méthodes : Le comité d'éthique avait approuvé cette étude. Deux scénarios cliniques d'une durée d'une heure comportant cinq problèmes anesthésiques ont été élaborés. Une échelle de cotation accordait un score à chacun (aucune réponse à la situation =0, une intervention définie comme une correction physiologique =1 ; une intervention thérapeutique considérée comme le traitement définitif=2). Des enregistrements sur vidéocassettes ont servi à évaluer la concordance entre les évaluateurs. Ces enregistrements témoignaient du rôle joué pendant les deux scénarios exécutés trois fois pour un total de 30 événements. Deux anesthésistes, ignorant le déroulement de l'étude et le contenu des scénarios, ont révisé et coté indépendamment les 30 problèmes. Les deux observateurs ont comparé les scores obtenus à l'aide de la méthode statistique d'accord kappa.

Résultats : Les évaluateurs s'accordaient complètement sur 29 des 30 sujets. La fiabilité entre évaluateurs était excellente (=0,96, $P < 0,001$).

Conclusion : L'utilisation des vidéocassettes a permis de coter les scénarios en reproduisant le même événement devant chacun des observateurs. Dans le cadre de l'étude, l'accord entre les évaluateurs était excellent. On peut utiliser l'évaluation de la performance d'un anesthésiste à l'aide d'enregistrements sur vidéocassette au cours d'une simulation. La validité des scénarios et du système de cotation reste à déterminer.

From the *Department of Anaesthesia, the †Clinical Epidemiology Unit and the ‡Biostatistical Consulting Unit, Sunnybrook Health Science Centre, the †Department of Health Science Administration, Faculty of Medicine, University of Toronto, 2075 Bayview Ave, Toronto, Ontario, M4N 3M5. Supported with a grant from the physicians of Ontario through the Physician's Services Incorporation Foundation. Dr. Cohen is the recipient of a National Health Scholar Award from Health Canada.

Address correspondence to: Dr. J.H. Devitt, Phone: 416-480-4864; Fax 416-480-6039; E-mail j.hugh_devitt@mail.magic.ca

Accepted for publication May 8, 1997.

THE anaesthesia simulator has been widely acclaimed as an exciting new development in the fields of anaesthesia and medical education. It is being developed as a major educational tool in the United States, particularly for the training anaesthesia residents.[1-3] An anaesthetic simulator in a mock operating room environment has been used to train practising clinicians in "hands-on" crisis management and a number of crisis management courses are now offered on a regular basis. [4,5] A more controversial proposed role for an anaesthesia simulator is that of assessing clinical competence, both for new trainees and for practising clinicians who may have been referred for evaluation due to some question about their ability to practice clinical skills.[1-3]

The ability to assess physician performance is imperfect and has largely relied upon written examinations and oral case presentations for residents and "close observation" by academic physicians for the practising anaesthetist. These methods encompass a large subjective component. The ability to measure actual performance, defined as vigilance, interpretation of data, and formulation and implementation of a management plan, is not readily demonstrable by traditional methods. Byrick and Cohen (1995) suggested that "simulation-based learning may help us to understand how clinicians respond to warning signals and change treatment strategies when confronted with additional information."[6] Gaba *et al.* have suggested that the anaesthesia simulator could be used as a testing tool.[1] The proposed use of a simulator to assess physician performance would provide a more structured and standardized measure of performance than traditional methods. In addition, it would test both knowledge of anaesthesia practice, as well as evaluate actual performance. Thus the simulator would represent a major advance in this area. However, before there is wide-spread adoption of the technology for this sensitive purpose, the reliability and validity of any evaluation method using the anaesthesia simulator must be determined.

The anaesthesia simulator consists of an anatomically correct mannequin which is controlled by computers. The simulator mimics various human responses to drugs and anaesthetic procedures. When the mannequin is placed on an operating room table and attached to a fully functioning anaesthesia machine and patient physiological monitors, the situation of the patient in the operating room is realistically recreated. Anaesthesia simulators successfully reproduce most aspects of physiology, pharmacology and patho-physiology of the patient in the peri-operative period. This technology allows the development of structured and standardized peri-operative clinical scenarios without having to use real patients.[1-4]

In this study, we assessed inter-rater reliability of the anaesthesia simulator by determining if different observers witnessing the same clinical scenario in an anaesthesia simulator would agree on their rating of anaesthetist performance.

Methods

The Sunnybrook Anaesthesia Simulation Centre consists of a mock operating room containing an anaesthesia gas machine, patient physiological monitors, anaesthesia drug cart, operating table, instrument table, and electro-cautery machine. Drapes, intravenous infusions, and surgical instruments are used to enhance the realism of the simulation. The patient mannequin is positioned on the operating table, and the role of members of the operating room team such as the surgeon, circulating and scrub nurses were acted by the investigators. The events can be viewed through a one-way mirror and video cameras were used to provide a permanent record of the simulation. The simulated anaesthesia workspace is shown in Figure 1.

After receiving approval for the study from the research ethics board at Sunnybrook Health Science Centre, two one-hour clinical scenarios were developed each containing five anaesthetic problems. The anaesthetic problems were designed to evaluate problem recognition, formulation of medical diagnosis and institution of treatment. Each clinical problem was scripted in the following manner: the problem was defined, the appropriate computer settings were recorded (CAE patient simulator), and the actor (investigator) playing the part of the anaesthetist was briefed to provide the

FIGURE 1 Simulated Anaesthesia Workspace

Photograph of the anaesthesia workspace, simulator mannequin and simulated operating room.

appropriate responses during the simulation. The scenarios were clinically reasonable and each contained five anaesthetic problems (items) to diagnose and manage. The problems in both scenarios can be classified into the following areas; a gas machine fault, a problem induced by mesenteric traction, a respiratory problem, a major cardiovascular problem and a long term monitoring problem. Each of the items was reproducible by the computer programme so that there was standardization of the scenarios. The clinical scenarios with problem description and identification is listed in the Table.

For each individual item in a scenario, a rating scale defined the appropriate score. No response to the situation by the "anaesthetist" gave a score of 0; undertaking a compensating intervention gave a score of 1; and if corrective treatment was undertaken, a score of 2 was recorded. A compensating intervention was defined as a manoeuvre undertaken to correct perceived abnormal physiological values. Corrective treatment was defined as definitive management of the presenting medical problem. The scoring system for both scenarios is outlined in the Table.

The video output signal of the physiological monitor was processed by a scan converter and recorded simultaneously with the anaesthesia workspace camera using a video-mixer. As a result the rater reviewing the tapes was able to observe concurrently the events in the operating room and the physiological patient data presented on the monitor. In addition, a sound recording was made of all events during the scenario. The sound was played back during the review of the videotape by the rater. Figure 2 presents an example of an isolated frame from the visual data available to the raters.

Videotape recordings of the two anaesthetic scenarios for assessment of inter-rater reliability were generated through role-playing. The investigators took turns role-playing the anaesthetist and were instructed on how to respond to each of the five problems in each of the scenarios (e.g., in some situations, the actor would purposefully perform an incorrect action). A random number table was used to determine which response was presented for videotaping for each problem at each simulation. In all, three videotape recordings were made for each scenario so that all three responses to each of the five problems were possible. Thus 15 data points were created for each of the two scenarios, for a total of 30 data points.

We tested inter-rater reliability by having two clinical anaesthetists who were not involved in the development of the study review the same videotape for

TABLE Scenarios and Scoring System

Criteria for Compensation and Management of Problems in scenario 1

Event	*Compensating Intervention Score = 1*	*Definitive Management Score = 2*
CO_2 canister leak	Increase fresh gas flow	Correction of leak
Sinus bradycardia during peritoneal traction	Atropine or vasopressor administration	Request relief of surgical stimulus
Atelectasis	Increase F_iO_2	Vital capacity breath or addition of PEEP
Coronary ischaemia	Increase F_iO_2 and/or administration of fluid or vasopressors	ß-blockers or nitrate administration
Hypothermia	Warming blankets, *iv* fluid warmer or heating of respiratory gases	Use of radiant heater or convective and/ heater or increase room temperature

Criteria for Compensation and Management of Problems in scenario 2

Event	*Compensating Intervention Score = 1*	*Definitive Management Score = 2*
Missing inspiratory valve	Increase fresh gas flow or use of bag value ventilation device after induction	Replacement of valve prior to valve induction
Hypotension during peritoneal traction	Administration of vasopressor or fluid	Request relief of surgical stimulus
Pneumothorax	Increase F_IO_2	Needle or tube thoracostomy
Anaphylaxis	Any of increase F_iO_2 administration of fluid, antihistamines or steroids	Administration of epinephrine
Anuria from obstructed catheter	Administration of fluid, diuretic or dopamine	Relief of catheter obstruction

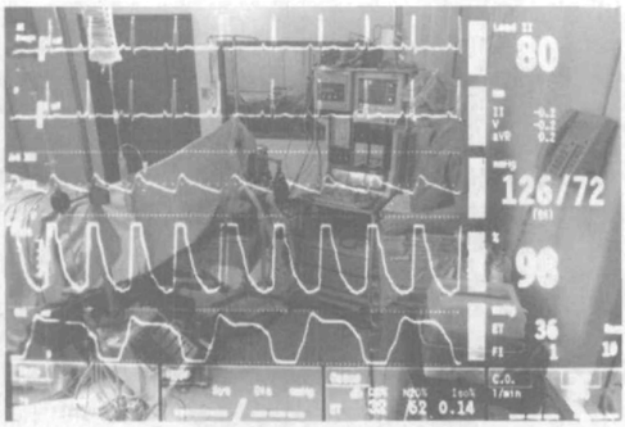

FIGURE 2 Picture of video-screen viewed by observers

A reproduction of the videotape recording with the output of the physiological monitor superimposed on the picture of the anaesthesia workspace camera.

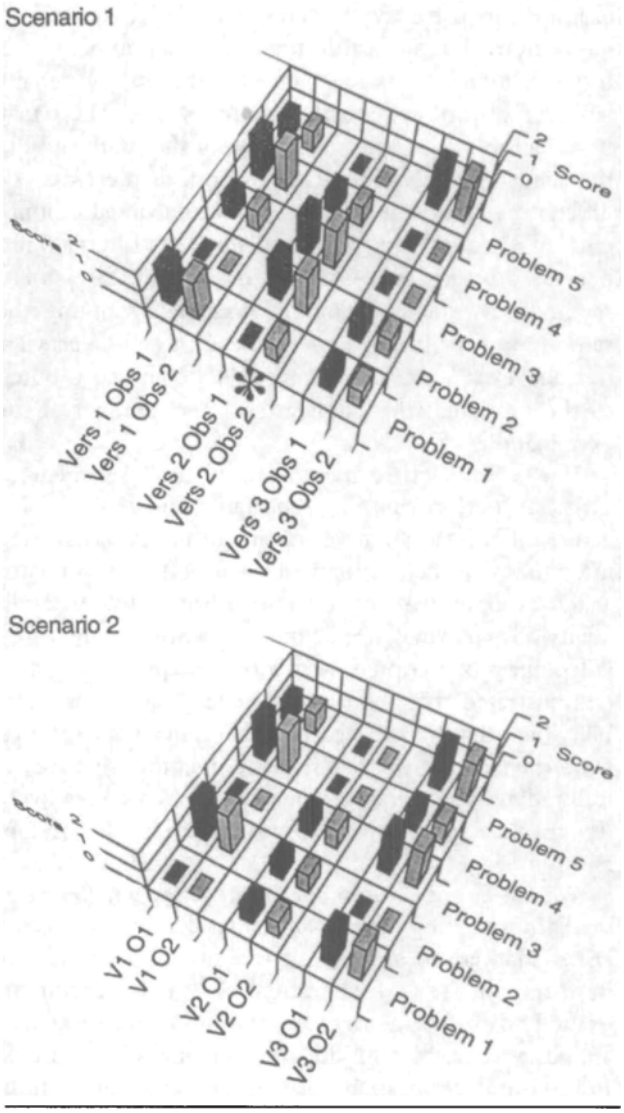

FIGURE 3 Graphical Representation of Rater Agreement

Rating of 2 scenarios by observers where V indicates the scenario version number and O the observer.

*indicates the only item on which there was observer disagreement.

each scenario and score each of the 30 problems. The two anaesthetists were certified in anaesthesia by the Royal College of Physicians and Surgeons of Canada and had been in active clinical practice for more than five years. The raters had participated in a three day course on the general principals of scenario design and construction. In addition, both anaesthetists received training in scenario content for this study and the scoring system. A special form for scoring was developed which described the rating system for each tape. Both anaesthetists reviewed the 30 problems and did not communicate with each other about the study.

The scores produced by the two raters were compared using the kappa (κ) statistic of agreement with a $\kappa > 0.75$ being considered excellent inter-rater reliability.[7]

Results

The 30 items were scored by each evaluator and the final scores recorded by the observers of the two scenarios are presented in Figure 3. For example, for the first scenario, problem 1, observer 1 scored 2 on the first version of the scenario and observer 2 scored 2. The raters were in complete agreement on 29 of the 30 items. There was only a single discrepancy between them and the inter-rater reliability was excellent ($\kappa = 0.96$, $P < 0.001$).

Discussion

Few examination tools in clinical medicine have undergone rigorous evaluation before being adopted for widespread use. Such evaluation or examination instruments should be reliable, valid, and practical. The reliability of an examination denotes that the results are repeatable and reproducible across different examiners and situations. The validity of an examination means that the test actually measures what it is intended to measure. Finally, the examination must be able to be applied in an efficient and cost effective manner. In a delicate area such as physician performance where so much is at stake, one must have a high degree of confidence in the chosen instrument.[8]

Traditional methods for evaluating clinical performance of anaesthetists are somewhat arbitrary, and no single method has proven to be best in assessing clinical competence. The standardized performance based clinical

examination (objective structured clinical examination) has been used in medical education, but a number of factors can influence its reliability.[9–11] Examiner variability has been previously reported as a problem.[12] There may be variation in the knowledge base of the candidates, in the actual conduct of the examination, or there may be differences in the rating methods of individual examiners.[12–14] The use of an anaesthesia simulator offers a number of advantages over the traditional assessment methods. By standardizing the scenarios, scripting the responses to the problems and having the observers view the same events, we have eliminated differences attributed to the "patient," the candidates or the conduct of the examination.

It is essential that any tool used for assessment of physician performance be repeatable and reproducible across different examiners and situations. As a first step in establishing the reliability of the anaesthesia simulator as a tool for assessment, we assessed the inter-rater reliability of two experienced clinicians viewing and scoring videotapes of scripted scenarios independently, and demonstrated that there was excellent inter-rater reliability between the two observers. This probably resulted from the fact that the criteria for evaluation were specifically determined, and the grading categories were designed to be very simple and relatively objective in nature.

We are aware of only one other study which examined the reliability of the anaesthesia simulator. Gaba *et al.* adapted an assessment tool used by the airline industry for use with anaesthetists.[14] This instrument graded crisis management behaviour during simulated anaesthetic crises and during real anaesthetic cases. Behavioural items such as assertion, communication, leadership and workload distribution were subjectively evaluated. Their study found poor to fair inter-rater reliability and concluded that further refinement of the scoring system and better training of the observers was necessary to improve reliability.[14]

In this study, rating of video recordings of anaesthetists in a simulator environment handling defined problems proved to be feasible and demonstrated excellent inter-rater reliability. As yet the validity or the cost-effectiveness of this approach remains to be demonstrated. Further work will be necessary before an anaesthesia simulator can be used as a technique for evaluating anaesthetist's performance.

Acknowledgments

The authors wish to thank Dr. D. Fung for viewing a similar scenario six times while grading the videotapes. In addition we would like to thank Mr. A. Noel for his assistance during the videotaping of the scenarios and Dr. B. Orser for reviewing the manuscript.

References

1 *Gaba DM, DeAnda A.* A comprehensive anaesthesia simulation environment: re-creating the operating room for research and training. Anesthesiology 1988; 69: 387–94.

2 *Gaba DM, DeAnda A.* The response of anesthesia trainees to simulated critical incidents. Anesth Analg 1989; 68: 444–51.

3 *Kurrek MM, Fish KJ.* Anaesthesia crisis resource management training: an intimidating concept, a rewarding experience. Can J Anaesth 1996; 43: 430–4.

4 *Holzman RS, Cooper JB, Gaba DM, Philip JH, Small SD, Feinstein D.* Anesthesia crisis resource management: real-life simulation training in operating room crises. J Clin Anesth 1995; 7: 675–87.

5 *Howard SK, Gaba DM, Fish KJ, Yang G, Sarnquist FH.* Anesthesia crisis resource management training: teaching anesthesiologists to handle critical incidents. Aviat Space Environ Med 1992; 63: 763–70.

6 *Byrick RJ, Cohen MM.* Technology assessment of anaesthesia monitors: problems and future directions. Can J Anaesth 1995; 42: 234–9

7 *Fliess JL.* Statistical Methods for Rates and Proportions, 2nd ed. New York: John Wiley and Sons, 1981; 218.

8 *Hart IR.* The OSCE-objective yes, but how useful? *In*: Hart IR, Harden RM, Walton HJ (Eds.). Newer Developments in Assessing Clinical Competence. Montreal, Quebec: Heal Publications Ltd, 1986: 22–8.

9 *Cohen R, Rothman AI, Ross J, Poldre P.* Validating an objective structured clinical examination (OSCE) as a method for selecting foreign medical graduates for a pre-internship program. Acad Med 1991; 66: S67–9.

10 *Newble DI, Hoare J, Sheldrake PF.* The selection and training of examiners for clinical examinations. Med Ed 1980; 14: 345–9.

11 *Robb KV, Rothman AI.* The assessment of clinical skills in general medical residents-comparison of the objective structured clinical examination to a conventional oral. *In*: Hart IR, Harden RM, Walton HJ (Eds.). Newer Developments in Assessing Clinical Competence. Montreal, Quebec: Heal Publications Ltd, 1986: 87–94.

12 *Swanson R, Swanson S, Spooner J, Haight K, Ramsden V, Tan L.* Inter-rater variability in an advanced cardiac life support course: a case study. Medical Teacher 1987; 9: 447–9.

13 *Wilson GM, Lever R, Harden RM, Robertson JIS.* Examination of clinical examiners. Lancet 1969; I: 37–40.

14 *Gaba DM, Botney R, Howard SK, Fish KJ, Flanagen B.* Interrater reliability of performance assessment tools for the management of simulated anesthetic crises. Anesthesiology 1994; 81: A1277.