Anesthesiology 2001; 95:36–42

The Validity of Performance Assessments Using Simulation

J. Hugh Devitt, M.D., M.Sc., F.R.C.P.C.,* Matt M. Kurrek, M.D.,† Marsha M. Cohen, M.D., M.H.Sc., F.R.C.P.C.,‡ Doreen Cleave-Hogg, B.A., M.A., Ph.D.§

Background: **The authors wished to determine whether a simulator-based evaluation technique assessing clinical performance could demonstrate construct validity and determine the subjects' perception of realism of the evaluation process.**

Methods: **Research ethics board approval and informed consent were obtained. Subjects were 33 university-based anesthesiologists, 46 community-based anesthesiologists, 23 final-year anesthesiology residents, and 37 final-year medical students. The simulation involved patient evaluation, induction, and maintenance of anesthesia. Each problem was scored as follows: no response to the problem, score = 0; compensating intervention, score = 1; and corrective treatment, score = 2. Examples of problems included atelectasis, coronary ischemia, and hypothermia. After the simulation, participants rated the realism of their experience on a 10-point visual analog scale (VAS).**

Results: **After testing for internal consistency, a seven-item scenario remained. The mean proportion scoring correct answers (out of 7) for each group was as follows: university-based anesthesiologists = 0.53, community-based anesthesiologists = 0.38, residents = 0.54, and medical students = 0.15. The overall group differences were significant ($P < 0.0001$). The overall realism VAS score was 7.8. There was no relation between the simulator score and the realism VAS ($R = -0.07$, $P = 0.41$).**

Conclusions: **The simulation-based evaluation method was able to discriminate between practice categories, demonstrating construct validity. Subjects rated the realism of the test scenario highly, suggesting that familiarity or comfort with the simulation environment had little or no effect on performance.**

Additional material related to this article can be found on the ANESTHESIOLOGY Web site. Go to the following address, click on Enhancements Index, and then scroll down to find the appropriate article and link. http://www.anesthesiology.org

* Associate Professor, † Assistant Professor, Department of Anaesthesia, ‡ Professor, Departments of Health Science Administration and Anaesthesia, and Senior Research Scientist, The Centre for Research in Women's Health, § Assistant Professor and Associate Director, Medical Education Department of Anesthesia, Centre for Research in Education, University of Toronto, Sunnybrook and Women's College Health Sciences Centre.

Received from the Department of Anaesthesia, Sunnybrook and Women's College Health Sciences Centre, The Centre for Research in Women's Health, the Department of Health Science Administration, and the Clinical Epidemiology and Health Care Research Program, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada. Submitted for publication November 22, 1999. Accepted for publication January 4, 2001. Supported by grant No. 94-36 from the physicians of Ontario through the Physician's Services Incorporation Foundation, Toronto, Ontario, Canada, and a Senior Scientist Award from the Medical Research Council of Canada, Ottawa, Ontario, Canada (to Dr. Cohen). Presented in part at the annual meetings of the American Society of Anesthesiologists, Orlando, Florida, October 20, 1998, and Dallas, Texas, October 11, 1999. Drs. Kurrek and Devitt have participated in educational activities that have been sponsored in part by Eagle Simulation, Inc., Binghamton, New York (formerly CAE).

Address reprint requests to Dr. Devitt: Queen Elizabeth II Health Sciences Centre, Department of Anaesthesia, 1796 Summer Street, Halifax, Nova Scotia, Canada B3H 3A7. Address electronic mail to: exhd@qe2-hsc.ns.ca.

THE assessment of clinical competence is an imperfect art.[1,2] The best estimation of competence at present are measures of clinical performance where "competence" describes a physician's capabilities and performance reflects that physician's actual practice.[3]

Although an evaluation tool must be practical and reliable, it must also be valid before being adopted into widespread use.[1,4] A test is said to be valid if it can actually measure what it is intended to measure.[1,4] Validity can be assessed by a number of different methods. A test is said to have construct validity if the test results are in keeping with expectations.[1,4,5] We previously demonstrated construct validity using a simple evaluation scheme in a small number of subjects.[6] This study focused on construct validity of a scenario using a large number of subjects with a wide range of expertise.

Over the last 25 years, the interest in simulation-based training has grown and expanded rapidly.[7] More recently, medical simulation has been used to develop evaluation methods for anesthesiologists.[6,8,9] Advantages of the simulation environment include no risk to patients, scenarios that can allow exploration of uncommon but serious clinical problems, scenarios that can be standardized for comparisons across practitioners, errors that could be allowed, which, in a clinical setting, would require intervention by a supervisor, and performance can be measured objectively.[10,11]

Previous studies have demonstrated that simulator-based evaluation processes contain acceptable reliablity.[6,8,9] However, simulator-based evaluation processes still require validation. The objective of the current study was to determine whether a simulator-based evaluation technique assessing clinical performance could discriminate the level of training of a large and diverse group of anesthesiologists and thus demonstrate construct validity. In addition, we asked participants to rate the realism of their experience during the simulation-based evaluation process.

Methods

The study was approved by the Sunnybrook Health Sciences Centre research ethics board (Toronto, Ontario, Canada), and written informed consent was obtained from all subjects. In addition, all participating subjects were asked to sign a statement guaranteeing confidentiality of the content of the evaluation scenario and problems.

Our simulation center consists of a mock operating room containing an anesthesia gas machine, patient physiologic monitors, anesthesia drug cart, operating table, instrument table, and electrocautery machine. Drapes, intravenous infusions, and surgical instruments are used to enhance the realism of the simulation. The details of our simulation center have been described elsewhere.[8,12] The patient mannequin (Eagle Patient Simulator, version 2.4.1; Eagle Simulation, Inc., Binghamton, NY) was positioned on the operating table. The role of the circulating nurse was scripted and acted by a knowledgeable research assistant. The circulating nurse was instructed to provide the appropriate responses during the simulation and was prompted as necessary by means of a radio frequency communication system. The surgeon was a mannequin with a built-in speaker operated by the simulation director. Except where scripted, the "surgeon" only responded to direct questions or, on occasion, asked questions to clarify ambiguous responses or statements of the participants.

None of the subjects had actual previous simulator experience, although residents and teaching staff may have heard anecdotal reports of activities in the simulation center. Subjects were drawn from five different groups consisting of (1) final-year anesthesiology residents, (2) medical students, (3) community-based anesthesiologists, (4) university-based (academic) anesthesiologists, and (5) anesthesiologists referred for practice assessment. The first group consisted of final-year anesthesia residents (5 years of postgraduate training) who were within 6 months of finishing their residency. The second group was medical students in their final year of training. (All medical students take a mandatory 2-week course in anesthesiology at the University of Toronto and participated in the study during the second week of their anesthesia rotation). The third group consisted of community-based anesthesiologists from the greater Toronto area. All community-based anesthesiologists were engaged in active practice. The fourth group was drawn from the teaching faculty of the Department of Anesthesia at the University of Toronto. All members of this group were engaged in independent clinical practice. The fifth and final group consisted of anesthesiologists identified as having practice deficiencies and had been referred by their practice hospitals or provincial licensing authorities. Subjects in all five groups were volunteers, and the latter three groups were paid a stipend for their participation. We had hypothesized that scores from our evaluation process would be highest among university-based anesthesiologists, followed by anesthesiology residents, community-based anesthesiologists, practice-referred anesthesiologists, and medical students.

Demographic data, including age, training (residency and clinical fellowship), and location of practice, were collected from all participants. All participants received a 30-min familiarization of the mannequin, gas machine physiologic monitor, and simulation facility. All participants were given the same scenario and patient information in the form of a preoperative assessment form, electrocardiogram, and chest radiograph results. All subjects were asked to verbalize their thoughts and actions during the simulation as if there were a medical student present.

The simulated patient in the scenario was a 66-yr-old man weighing 79 kg who presented with a diagnosis of carcinoma of the colon for an elective left hemicolectomy. The anticipated duration of surgery was 2.5 h. The patient had no allergies, and his current medications included isosorbide dinitrate and diltiazem. His medical history was remarkable for an uneventful myocardial infarction 5 years previously with residual stable class I postinfarction angina, a 5-year history of mild hypertension, and a 30 pack-year history of smoking. The preoperative physical examination was unremarkable, and preoperative hematology and biochemistry were normal. The preoperative electrocardiogram documented a normal sinus rhythm, normal axis, QS complex in leads II, III, and AVF, and the preoperative chest radiograph was interpreted as normal with mild hyperinflation, consistent with chronic obstructive lung disease.

A 1.5-h clinically realistic scenario was developed, containing nine anesthetic problems (items). The problems were developed by a panel of four clinical anesthesiologists who were actively engaged in clinical practice at large university-based residency training programs (in the United States and Canada) and certified in anesthesiology by the American Board of Anesthesiology. All members of the panel were knowledgeable about the capabilities of simulator technology. The test items were chosen after panel discussion to reflect a variety of clinical problems, taking into consideration the capabilities and realism of the simulator hardware and software. The items were designed to evaluate problem recognition, formulation of a medical diagnosis, and the institution of treatment. The development of each clinical problem consisted of defining the problem, determining the appropriate computer settings, and developing a script for the roles of the surgeon and circulating nurse. Each of the items was reproducible so that there was standardization of the scenario and problems. The clinical problem description and identification are listed in table 1 (a detailed description of each item is presented in an appendix on the ANESTHESIOLOGY Web site). Problems were presented in a sequential manner over the 1.5-h period of the simulated anesthetic. There was a specified time interval (5 min) between each problem where the patient's physiological parameters were returned to normal, signifying the end of the problem and resulting in a period of relative inactivity before introduction of the next item.

Table 1. Scenarios and Scoring System

Problem Number	Criteria for Compensation and Management of Problems			
	Problem	Problem Manifestation*	Compensating Intervention Score = 1	Definitive Management Score = 2
1	CO_2 canister leak	3 l/min circuit leak, duration 5 min	Increase fresh gas flow	Correction of leak
2	Missing inspiratory valve	$ETCO_2$ >60 mmHg, inspired CO_2 > 3 mmHg, duration 5 min	Increase fresh gas flow or use of bag valve ventilation device after induction	Replacement of valve before induction
3	Hypotension, mesenteric traction	Systolic BP < 80 mmHg, duration 5 min	Administration of vasopressor or fluid	Request relief of surgical stimulus
4	Atelectasis	SpO_2 < 89%, duration 5 min	Increase FIO_2	Vital capacity breath or addition of PEEP
5	Coronary ischemia	ST depression > 4 mm, ventricular ectopy, duration 5 min	Increase FIO_2 or administration of fluid or vasopressors	β Blockers or nitrate administration
6	Pneumothorax	SpO_2 < 70%, unilateral breath sounds, decreased pulmonary compliance, duration 5 min	Increase FIO_2	Needle or tube thoracostomy
7	Anaphylaxis	Pulmonary wheezing, systolic BP < 70 mmHg, duration 5 min	Any of administration of fluid, antihistamines or steroids or increase FIO_2	Administration of epinephrine
8	Hypothermia	Temperature decreases to 33°C at 1 h into the case	Warming blankets, intravenous fluid warmer or heating of respiratory gases	Use of radiant heater or convective heater or increase room temperature
9	Anuria, obstructed catheter	Absence of accumulating urine in catheter bag from beginning of case	Administration of fluid, diuretic or dopamine	Relief of catheter obstruction

* Details can found in an appendix published on the ANESTHESIOLOGY Web site.

CO_2 = carbon dioxide; $ETCO_2$ = end-tidal carbon dioxide; BP = blood pressure; SpO_2 = peripheral saturation; FIO_2 = inspired fraction of oxygen; PEEP = positive end-expiratory pressure.

For each item, a rating scale defined the appropriate score based on preset criteria (table 1). No response to the situation by the participant resulted in a score of 0; undertaking a compensating intervention resulted in a score of 1; and corrective treatment resulted in a score of 2 ("correct" score) recorded by the observer. A compensating intervention was defined as a maneuver undertaken to correct perceived abnormal physiologic values. A corrective treatment was defined as definitive management of the presenting medical problem. Appropriateness of compensating intervention and corrective treatment were defined by consensus after referencing with standard anesthesiology textbooks.

All participants were asked to anesthetize the "patient" for an elective surgical procedure as the first case at the beginning of the day. All subjects were expected to assess the patient, check anesthetic drugs and equipment, and induce and provide maintenance anesthesia for the scenario. All external cues were standardized, rehearsed, and presented in a similar manner to all study participants.

Each subject was evaluated by one of two trained raters certified in anesthesiology by The Royal College of Physicians and Surgeons of Canada and the American Board of Anesthesiology. A rating sheet that detailed possible responses and scores was given to the evaluators for each participant. The observers did not know in advance the background of each candidate as all appointments and preliminary data collection were made by a research assistant. As a result, subjects presented on a first-come, first-served basis over the course of the study, and it was not possible to predict based on scheduling the category of the subject. All participants wore operating room headgear and masks, but it was not possible to completely blind the raters as to the group in which the participant fell. All performances by participants were recorded on videotape for subsequent review and assessment.

After completion of the simulator scenarios, participants were asked to rate the realism of their experience on a 10-point visual analog scale (VAS), where a rating of 0 indicated an unrealistic experience and a rating of 10 indicated a completely realistic experience. A discussion of the experience was undertaken with each subject at the end of the simulation after each subject had rated the realism of the experience.

Statistical Analysis

Age, years of training, and years in practice were compared for each group using one-way analysis of variance. Internal consistency of the test items was estimated using the Cronbach coefficient α. A Cronbach coeffi-

Table 2. Description of Study Participants

Group	Number	Age (yr)* (Mean ± SD)	Years of Training† (Mean ± SD)	Years in Practice‡ (Mean ± SD)
University anesthesiologists	33	40 ± 8	5 ± 2	8 ± 8
Community anesthesiologists	46	46 ± 10	4 ± 1	15 ± 10
Residents (final year)	23	32 ± 4	4 ± 1	NA
Medical students	37	27 ± 3	NA	NA
Practice assessment	3	67 ± 7	4	35 ± 11

* $P < 0.0001$. † P = not significant (NS). ‡ $P < 0.0001$.
NA = not applicable.

cient $\alpha > 0.6$ was considered adequate for internal consistency.[6] An item analysis was performed by recalculating the Cronbach coefficient α with each item deleted to determine if any of the items in the scenario contributed to poor internal consistency. Those items, which reduced the overall Chronbach α on the item analysis, were eliminated, and the Cronbach coefficient α was reestimated for the remaining items.

We determined the difference in simulator scores between the four groups as follows. For each subject, the proportion of items scored as 2 was calculated for each of the items. The sum of the proportions for participants in each group was determined and divided by the number of subjects in the group to calculate the mean proportion. The difference between the mean proportion of correct answers across the four groups was determined using a one-way analysis of variance followed by pairwise comparisons (Tukey).

Mean realism VAS scores for all groups were compared using a one-way analysis of variance followed by pairwise comparisons, with $P < 0.05$ considered significant. Years in practice for those individuals engaged in active clinical practice (community- and university-based anesthesiologists, and those referred for practice assessment) were compared with the simulation score using a Pearson correlation coefficient. The simulator score was compared with the realism VAS using the Pearson correlation coefficient, with $P < 0.05$ considered significant.

Results

A total of 142 subjects participated in the study. There were significant differences in mean age between the various groups (table 2), with medical students and residents being younger than all other groups. The anesthesiologists referred for practice assessment were the oldest, while university-based anesthesiologists were younger than those in the community. There was no significant difference in length of training for those engaged in clinical practice, but anesthesiologists referred for assessment of practice were in practice longer than those in community- or university-based practices. All subjects signed and agreed to the terms of the confidentiality statement.

The Cronbach coefficient α for the nine-item evaluation process was 0.62. The item analysis is presented in table 3. Problems 1 (carbon dioxide canister leak) and 2 (missing inspiratory valve) reduced overall internal consistency. Removing these problems increased the Cronbach coefficient α to 0.69. Because these items demonstrated a lack of internal consistency (*i.e.*, lack of reliability), they were dropped from further consideration, leaving seven items in the analysis.

The practice assessment group was not considered in further group comparisons because of the small numbers (n = 3). There was a significant difference in the mean proportion of correct scores across the remaining four groups ($P < 0.0001$; fig. 1). University anesthesiologists scored significantly higher than community anesthesiologists ($P < 0.003$) and medical students ($P < 0.0001$). Community anesthesiologists scored significantly higher than medical students ($P < 0.0001$) Residents scored significantly higher than community anesthesiologists ($P < 0.005$) and medical students ($P < 0.0001$).

The distribution of scores by test item and group is presented in figure 2. There was a weak but significant correlation between years in clinical practice and simulator score ($R = -0.49$, $P = 0.0001$).

The overall mean realism VAS score was 7.8. No relation was found overall between the simulator score and the participants' perception of realism as assessed by the

Table 3. Item Analysis

	Problem Number								
	1	2	3	4	5	6	7	8	9
Median score	0	0	0	0	1	1	2	1	1
Cronbach α with item deleted	0.64*	0.64*	0.58	0.54	0.57	0.54	0.52	0.60	0.59

* Cronbach coefficient α increases with these items deleted.

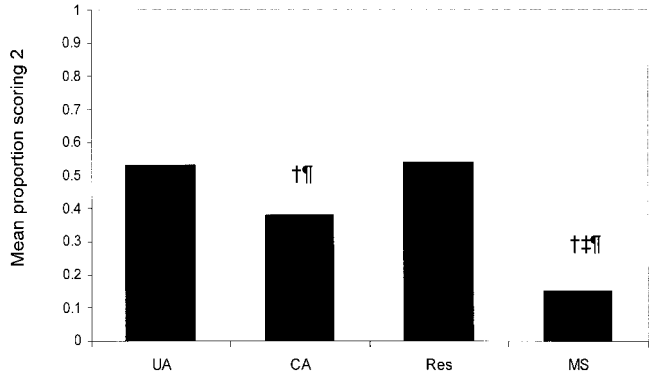

Fig. 1. Mean proportion of simulator scores by practice group. UA = university-based (academic) anesthesiologists; CA = community-based anesthesiologists; Res = final-year anesthesiology residents; MS = medical students. †P < 0.05 when compared with university anesthesiologists. ‡P < 0.05 when compared with community anesthesiologists. ¶P < 0.05 when compared with residents.

VAS rating (R = −0.07, P = 0.41; fig. 3). Each group rated the realism of their evaluation experiences as follows: university anesthesiologists, 7.3 ± 1.2 (mean ± SD); community anesthesiologists, 7.7 ± 1.3; residents, 8.1 ± 1.2; and medical students, 8.2 ± 1.0. Group differences with respect to realism were significant (P < 0.01), with medical students rating the evaluation as more realistic than university-based anesthesiologists.

Discussion

Scores across groups differed significantly, with university-based anesthesiologists and residents scoring significantly higher than all other groups. This was not unexpected because the anesthesiology residents were in their final year of training and deemed eligible by the program director to take the national specialty examination in anesthesiology. In addition, we noted that community-based anesthesiologists had significantly different scores than medical students. There were significant differences in age and the number of years in practice for subjects engaged in clinical practice.

All groups rated the simulation evaluation environment as realistic. The lack of correlation with realism in the simulator environment and the score achieved in the simulator-based evaluation process would indicate that familiarity or comfort with the simulation environment had little or no effect on performance.

A test is said to have construct validity if the test results are in keeping with expectations. We came near to demonstrating our construct in that university-based anesthesiologists and residents scored higher than all groups, and community-based anesthesiologists scored higher than medical students. We were able to show construct validity in that an evaluation system using the simulator was able to differentiate a large group of individuals based on clinical experience or training.[1,4,5] In our practice setting, most university-based anesthesiologists are actively engaged in independent clinical practice with, on average, 90% of their time devoted to clinical activities and with a minimum clinical activity of at least 50%. Teachers of residents reported that 10% of their clinical activity is conducted with residents, resulting in the remaining 90% of their clinical activity being conducted on an independent basis, as there are no

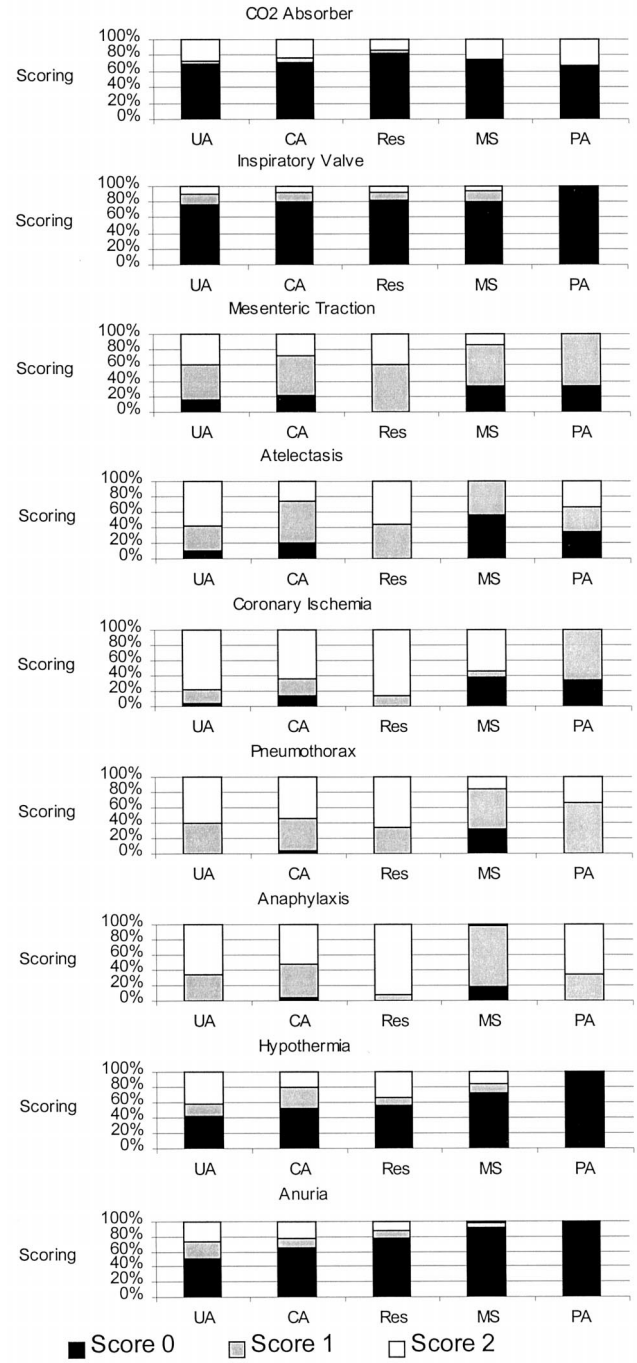

Fig. 2. Distribution of score on each item by practice group. UA = university-based (academic) anesthesiologists; CA = community-based anesthesiologists; Res = final-year anesthesiology residents; MS = medical students; PA = anesthesiologists referred for practice assessments.

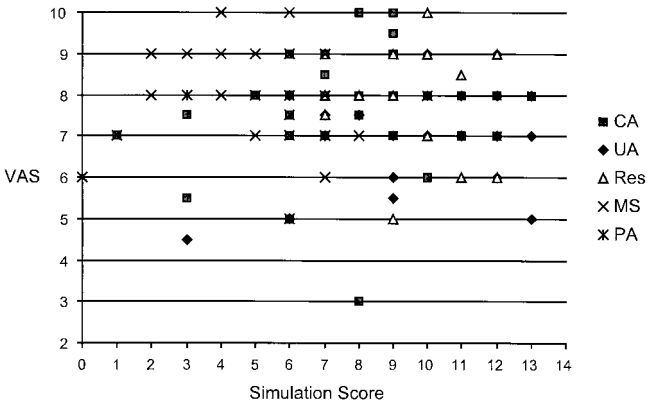

Fig. 3. Correlation of simulator score with realism visual analog scale (VAS) score for all participants (R = −0.07, *P* = 0.41). UA = university-based (academic) anesthesiologists; CA = community-based anesthesiologists; Res = final-year anesthesiology residents; MS = medical students; PA = anesthesiologists referred for practice assessments.

certified nurse anesthetists in the Canadian system. In addition, university-based anesthesiologists are exposed to more frequent rounds and educational activities than their community-based counterparts, and university-based anesthesiologists are closer to research and new developments. While university-based anesthesiologists may have been closer to simulator development and use, the university cohort of subjects was drawn from teaching hospitals that were physically remote from the simulator location.

The internal consistency on all but two items was acceptable in our study. One of our items (missing inspiratory valve) demonstrated poor internal consistency in an earlier report, while the other item (carbon dioxide canister leak) had an acceptable internal consistency in the same previous study.[6] There are several explanations for this discrepancy. The number of subjects participating in this study was increased by fivefold, thus increasing our power. Second, the breadth of experience and training of the subjects in this study encompassed a wider range than that of the previous study.[6,13] These differences in findings over the two studies suggest that internal consistency must be reviewed for each new evaluation process or study population.

A possible criticism of our study is that the simulation director (computer operator) also scored the subjects' performance. To avoid bias, the scoring system (only 3 possible scores or points) was simple and clear. The criteria for scoring were not dependent on subjective assessments but on clear action by the participants. As a result, the assessment of interrater reliability of our evaluation process was excellent.[8] A multipoint scoring system for each item would likely require greater interpretation on the part of the evaluator, resulting in possible bias and poorer interrater reliability.

It has been suggested that an evaluator independent of the simulation director might reduce any potential for bias. However, the additional person required to perform the evaluation adds other complexities. The review of videotapes by an independent observer requires additional personnel, and the videotape presents a limited window for the observer. The videotape may obscure actions and conversations that were obvious to the simulator director in the control room. We have studied interrater reliability between scores assigned by the simulation director during observation of the live scenario and the score generated by observing the videotape of the same event. Although there was good to excellent interrater agreement on all of our test items, several of the items requiring gas machine manipulation tended to have lower but acceptable agreement when compared with other items in the scenario. Our review and interpretation of these discrepancies suggested that the window presented by the videotape limited the information available to the evaluator.[13,14] The use of multiple cameras and high-definition video technology could overcome many of the aforementioned disadvantages.

Several investigators have documented that simulator-based evaluation methods can differentiate subjects on the basis of training and experience. In a previous report, our group was able to document significant scoring differences between trainees and faculty using a similar simulator-based evaluation tool.[6] Gaba and DeAnda[11] were able to demonstrate differences in time to correct critical incidents but not for time to detection of critical incidents between first- and second-year anesthesiology residents. Byrne and Jones[15] were also able to demonstrate that anesthesiologists with less than 1-year of experience were significantly slower in dealing with anesthetic emergencies than those with greater experience. Both of the latter two studies noted a wide variation in responses by all groups of subjects.[11,15] These previous studies support the construct that increased clinical experience should improve performance on simulator-based evaluation processes.

Clinical practice assessment by direct observation has been used as a method of assessing performance and competence when a practicing anesthesiologist's competence has been questioned.[16] This method of evaluation has not been subject to rigorous testing for reliability or validity.[17] The nature of the clinical anesthesia practice of those individuals referred for practice assessment, the elective nature of the scheduled practice assessment period, and the time constraints placed on the period of clinical observation results in the practice assessment being conducted on healthy elective patients.[16] Anesthesiology has advanced to the point that major adverse events are rare, so that individual practitioners are unlikely to have an actual clinical experience with such events during the assessment period.[18,19] There are a number of situations and emergencies that all clinically active anesthesiologists are expected to handle regardless of their practice situation, and yet these

situations are unlikely to occur during the time-limited period of clinical observation. A simulator-based assessment process allows the creation of relevant standardized emergency and critical incidents for use as test situations. In the interest of patient safety, critical events cannot be left untreated in real life to see if there will be an appropriate response by the anesthesiologist undergoing practice assessment. A simulator-based assessment process allows for the observation of performance during critical incidents without putting the patient at risk.

We have documented construct validity of a simulator-based evaluation process. Nonetheless, the findings of this study will require comparison with other established and validated evaluation methods of performance for practicing anesthesiologists to document criterion validity. Agreement on what constitutes an established evaluation process (gold standard) for practicing anesthesiologists may be hard to obtain. Finally, we caution that the findings of our study can only be applied to our scenario and are not necessarily able to be generalized to other simulation-based evaluation processes. We believe that our simulator-based evaluation shows promise as an adjunct to existing evaluation processes for practicing anesthesiologists.

The authors thank Melissa Shaw, R.R.T. (Department of Respiratory Therapy), and Chris Lewczuk, R.R.T. (Department of Respiratory Therapy), for help with data collection during the study; and John Paul Szalai, Ph.D. (Director, Department of Research Design and Biostatistics), and Donna Ansara, B.A. (Centre for Research in Women's Health), for assistance in performing the statistical analysis (all at Sunnybrook and Women's College Health Sciences Centre, Toronto, Ontario, Canada).

References

1. Eagle CJ, Martineau R, Hamilton K: The oral examination in anaesthesic resident evaluation. Can J Anaesth 1993; 40:947–53

2. Sivarajan M, Miller E, Hardy C, Herr G, Liu P, Willenkin R, Cullen B: Objective evaluation of clinical performance and correlation with knowledge. Anesth Analg 1984; 63:603–7

3. Norman GR: Defining competence: A methodological review, Assessing Clinical Competence. Edited by Neufeld VR, Norman GR. New York, Springer Publishing Co., 1985, pp 15–35

4. Neufeld VR: An introduction to measurement properties, Assessing Clinical Competence. Edited by Neufeld VR, Norman GR. New York, Springer Publishing Co., 1985, pp 39–50

5. Nunnally JC: Validity, Psychometric Theory, 2nd edition. New York, McGraw Hill, 1978, pp 86–116

6. Devitt JH, Kurrek MM, Cohen MM, Fish K, Fish P, Noel AG, Szalai J-P: Testing internal consistency and construct validity during evaluation of performance in a patient simulator. Anesth Analg 1998; 86:1160–4

7. Gaba DM: Human work environment and simulators, Anesthesia, 5th edition. Edited by Miller RD. Philadelphia, Churchill Livingston, 2000, pp 2613–68

8. Devitt JH, Kurrek MM, Cohen MM, Fish K, Fish P, Murphy PM, Szalai J-P: Testing the raters: Inter-rater reliability of standardized anaesthesia simulator performance. Can J Anaesth 1997; 44:924–8

9. Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R: Assessment of clinical performance during simulated crises using both technical and behavioral ratings. ANESTHESIOLOGY 1998; 89:8–18

10. Schwid HA, O'Donnell D: The anesthesia simulator-recorder: A device to train and evaluate anesthesiologists' reposes to critical incidents. ANESTHESIOLOGY 1990; 72:191–7

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

12. Kurrek MM, Devitt JH: The cost for construction and operation of a simulation centre. Can J Anaesth 1997; 44:1191–5

13. Kapur PA, Steadman RH: Patient simulator competency testing: Ready for takeoff? Anesth Analg 1998; 86:1157–9

14. Kurrek MM, Devitt JH, Cohen M, Szalai J-P: Inter-rater reliability between live-scenarios and video recordings in a realistic simulator (abstract). J Clin Monit 1999; 15:253

15. Byrne AJ, Jones JG: Responses to simulated anaesthetic emergencies by anaesthetists with different durations of clinical experience. Br J Anaesth 1997; 78:553–6

16. Devitt JH, Yee DA, deLacy JL, Oxorn DC: Evaluation of anaesthetic practitioner clinical performance (abstract). Can J Anaesth 1998; 45:A56-B

17. Wakefield J: Direct observation, Assessing Clinical Competence. Edited by Neufeld VR, Norman GR. New York, Springer Publishing Co., 1985, pp 51–70

18. Duncan PG, Cohen MM, Yip R: Clinical experiences associated with anaesthesia training. Ann RCPSC 1993; 26:363–7

19. Cohen MM, Duncan PG, Pope WDB, Biehl D, Tweed WA, MacWilliam L, Merchant RN: The Canadian four-centre study of anaesthetic outcomes: II. Can outcomes be used to assess the quality of anaesthesia care? Can J Anaesth 1992; 39:430–9