General Anesthesia

Validity and reliability of undergraduate performance assessments in an anesthesia simulator

Pamela J. Morgan MD CCFP FRCPC,*
Doreen M. Cleave-Hogg BA MA PhD,†
Cameron B. Guest MD FRCPC MED,*
Jodi Herold BSc (Pt) MA†

Purpose: To examine the validity and reliability of performance assessment of undergraduate students using the anesthesia simulator as an evaluation tool.

Methods: After ethics approval and informed consent, 135 final year medical students and 5 elective students participated in a videotaped simulator scenario with a Link-Med Patient Simulator (CAE-Link Corporation). Scenarios were based on published educational objectives of the undergraduate curriculum in anesthesia at the University of Toronto. During the simulator sessions, faculty followed a script guiding student interaction with the mannequin. Two faculty independently viewed and evaluated each videotaped performance with a 25-point criterion-based checklist. Means and standard deviations of simulator-based marks were determined and compared with clinical and written evaluations received during the rotation. Internal consistency of the evaluation protocol was determined using inter-item and item-total correlations and correlations of specific simulator items to existing methods of evaluation.

Results: Mean reliability estimates for single and average paired assessments were 0.77 and 0.86 respectively. Means of simulator scores were low and there was minimal correlation between the checklist and clinical marks ($r=0.13$), checklist and written marks ($r=0.19$) and clinical and written marks ($r=0.23$). Inter-item and item-total correlations varied widely and correlation between simulator items and existing evaluation tools was low.

Conclusions: Simulator checklist scoring demonstrated acceptable reliability. Low correlation between different methods of evaluation may reflect reliability problems with the written and clinical marks, or that different aspects are being tested. The performance assessment demonstrated low internal consistency and further work is required.

Objectif : Vérifier la validité et la fiabilité de l'évaluation de la performance des étudiants de médecine à l'aide d'un simulateur d'anesthésie.

Méthode : Le comité d'éthique et les participants à l'étude ayant donné leur consentement, 135 étudiants de dernière année en médecine et 5 étudiants choisis ont participé à une simulation vidéo à l'aide du Link-Med Patient Simulator (CAE-Link Corporation). Les scénarios étaient fondés sur les objectifs pédagogiques du programme d'enseignement en anesthésie de l'université de Toronto. Pendant les sessions de simulation, les examinateurs ont utilisé un scénario pour guider les interactions de l'étudiant avec le mannequin. Deux examinateurs ont visionné et évalué indépendamment la performance enregistrée sur chacune des cassettes vidéo selon une liste de 25 critères. Les écarts types et moyens des points accordés selon le simulateur ont été déterminés et comparés avec les évaluations cliniques et écrites reçues pendant la rotation. La conformité interne du protocole d'évaluation a été définie à l'aide de corrélations des éléments entre eux, et des éléments avec l'ensemble, ainsi que des corrélations d'éléments spécifiques du simulateur avec des méthodes existantes d'évaluation.

Résultats : Les estimations moyennes de fiabilité pour les évaluations individuelles et appariées moyennes ont été de 0,77 et 0,86 respectivement. Les moyennes des scores au simulateur ont été basses et il n'y a eu qu'une corrélation minimale entre la liste et les rendements cliniques ($r = 0,13$), entre la liste et les évaluations écrites ($r = 0,19$) et entre le rendement clinique et l'examen écrit ($r = 0,23$). Les corrélations des éléments entre eux et des éléments avec l'ensemble ont varié fortement et la corrélation entre les unités du simulateur et les outils d'évaluation existant a été faible.

Conclusion : La cotation selon les critères liés au simulateur a démontré une fiabilité acceptable. Une faible corrélation entre différentes méthodes d'évaluation peut traduire des problèmes de fiabilité avec les notes obtenues aux examens écrits et cliniques, ou le fait que d'autres aspects sont testés. L'évaluation de la performance a démontré une faible conformité interne et la nécessité d'expérimentation ultérieure.

From the Department of Anesthesia,* Sunnybrook and Women's College Health Sciences Centre, University of Toronto and the Centre for Research in Education,† University of Toronto, Toronto, Ontario, Canada.

Address correspondence to: Dr. P.J. Morgan, Department of Anesthesia, Sunnybrook & Women's College Health Sciences Centre, Women's College Campus, 76 Grenville Street, Toronto, Ontario M5S 1B2 Canada. Phone: 416-323-6400 Ext. 4349; Fax: 416-323-6307; E-mail: pam.morgan@utoronto.ca

This study was supported by a research award from the Canadian Anesthesiologists' Society.

Accepted for publication November 18, 2000.

ASSESSMENT of medical competence has been a focus for discussion by many educators over the years.¹⁻³ Licensing bodies have expressed dissatisfaction with traditional methods of clinical evaluation.⁴ This, in turn, has caused concern among clinicians initiating an impetus for change. Performance in the anesthesia clerkship is evaluated by methods that have not been subjected to rigorous tests of reliability or validity.

To improve reliability of assessments, other specialties have introduced standardized patients in the undergraduate curriculum by means of Objective Structured Clinical Examinations (OSCE).^{5,6} Due to the nature of the practice of anesthesia, standardized patients have limitations as assessment techniques but simulation technology offers exciting opportunities for exploration of standardized evaluation methods.⁷ The use of 'bench model simulations' has been introduced in the surgical training program at the University of Toronto to test technical skills.⁸ Other 'simulation' technologies have been used to teach and evaluate performance of different skills.⁹⁻¹¹

The Anesthesia Simulation System provides a realistic operating room experience that offers opportunities for working through a situation structured to challenge a student at the expected level of competence. The sessions can be easily videotaped which allows the provision of constructive feedback to both students and faculty instructors.

The safety of the patient is at all times the primary concern. Students cannot be allowed to manage situations in an operating room that may impact on outcome. Therefore, student assessment of certain competencies in the operating room situation becomes problematic. Advances in technology have allowed opportunities to expand medical education outside the realm of 'live patients'.¹² One of these new technologies is the Anesthesia Simulator which has the potential to offer a controlled environment where clerks' skills can be assessed. However, rigorous assessment of any new technique, particularly if it involves evaluation, must be performed.

There is very little in the literature addressing the evaluation of medical undergraduates in the simulator environment. In a study by Devitt *et al.*, clerks, residents and practicing anesthesiologists were evaluated in a simulation setting.¹³ Scenarios, however, were developed for the most expert group. There was little correlation between the clerks' simulator evaluation and their standard clinical assessment or between the clerks' simulator evaluation and the written examination. It was the authors' opinion that appropriate pre-defined educational objectives should guide the

evaluation process. Third year medical students' performances were observed in a simulated environment in a pilot study by Tome *et al.*¹⁴ These students had been given a six-hour supplemental virtual problem-based learning curriculum prior to testing in the simulator. The authors stated that further study was required to determine the validity and reliability of a simulator evaluation process.

This study was designed to improve the undergraduate evaluation process and to address issues regarding the reliability and validity of a simulator-based performance assessment.

Methods

After ethics board approval, all final year medical students (n=177) at the University of Toronto were invited to participate in this study during their anesthesia rotation. Information regarding the project was provided and written consent was obtained from participants. Eighteen groups of students attended the Simulation Centre on the eighth day of their 10-day rotation in anesthesia. Each student, therefore, had seven days of operating room experience before the simulator session.

On the day of the simulation session, students viewed a 10-min video orientation addressing the purpose of the study, the capabilities of the simulator mannequin and an orientation to the simulator 'operating room', equipment and drugs. Students received an outline of a patient scenario, complete with history, physical examination and laboratory findings and the planned procedure. (Figure) The attending faculty answered questions related to the scenario. During the sessions, medical students worked through a 15-min case scenario requiring them to achieve multiple skill and knowledge objectives. Attending faculty guided the session with a script including specific questions related to the scenario. Students were informed that they should 'manage' the case to the best of their ability and that the faculty supervisor would be a limited resource. Faculty assisted by administration of appropriate induction agents once the student had identified the preoperative considerations, checked for the appropriate drugs and airway equipment and verbalized a management plan. The supervisor responded to student requests as if he/she were a circulating nurse i.e., procurement of equipment or personnel. A written protocol outlining the dialogue that faculty was to follow with each student was presented during a workshop and employed during the student sessions. Faculty were allowed to provide 'prompts' to the student according to a fixed format thereby standardizing the faculty-student interaction. These 'prompts' were limited; they

Patient Mr. M. B.

History:

24-yr-old previously healthy man presents to the emergency department complaining of increasing abdominal pain, now most severe in the right lower quadrant. He has had mild abdominal pain for two days with exacerbation this morning. He feels generally unwell. He ate lunch about three hours ago and has been feeling nauseated since that time but has not vomited. He has no other associated symptoms. He has no other medical illnesses, takes no medications and has no allergies. There is no personal or family history of problems with general anesthetics. He has all his own teeth with no caps, dentures, loose teeth or bridge work.

Physical Examination:

T: 38.4°C	BP: 120/80 mmHg
HR: 105 bpm	RR: 20 bpm
Airway:	Normal
Respiratory System:	Normal
CVS:	Heart sounds normal
	No murmurs

Lab:

Hb	134
WBC	16.7
Electrolytes	Normal

Provisional Diagnosis: Acute appendicitis

Plan: Emergency laparotomy for appendectomy

FIGURE Clerkship Simulation Study Scenario 1

did not direct the students' management nor did they correct mistakes.

All sessions were videotaped for subsequent analysis. After the sessions, students received feedback from the faculty supervisor and were asked to complete a survey evaluating the simulator session.

Ten faculty (who were not involved as faculty in the simulator with the students), attended an instructional workshop and subsequently viewed the videotapes of student performances. Faculty were randomly assigned to 'marking pairs' but all evaluations were done independently of each other. Each faculty pair, therefore, evaluated three or four groups of 8-10 students over the course of the academic year. Evaluators had no or limited interaction with the students before assessing their video performance.

Case design

At the University of Toronto, the anesthesia rotation is scheduled in two-week blocks during a six-week rotation. A total of six scenarios were developed. This number was deemed necessary in order to optimize confidentiality of case content between groups of students. These cases were rotated based on a predetermined sequence. All students working in the simulator on any given day received the same scenario. Each case was based on the published curriculum objectives. The learning objectives of the six case scenarios are outlined in Table I. A printed handout information sheet containing the pertinent history, physical examination and laboratory findings was developed for each case. Students were expected to perform common tasks such as bag and mask ventilation, laryngoscopy and tracheal intubation, and to make medical judgments based on information available to them. Events specific to each scenario were entered into the computer controlling the anesthesia mannequin.

Performance evaluation protocols

Each student's performance was evaluated using a 25-point criterion based checklist. All six scenarios were evaluated under four headings: 1) Preoperative Considerations, 2) Preparation, 3) Induction of Anesthesia, 4) Intraoperative Problems. A sample of a performance protocol for one scenario is illustrated in Table II. All scenarios had the same headings with identical total scores for each category. Protocols were developed based on the curriculum expectations i.e., students are expected to be able to list the important preoperative considerations of the morbidly obese patient presenting for anesthesia and surgery or students are expected to know how to mask ventilate the lungs of an unconscious patient.

Statistical analysis

Statistical analysis was performed using SPSS 10.0.1 for Windows 95 (SPSS Inc., Chicago, Illinois).

Reliability

To assess reliability, intraclass correlation coefficients were calculated for 'single' and 'mean two-rater' student ratings by each of the rater pairs, using a two-way random effects model.^{15,16} The Spearman-Brown formula was used to estimate the number of raters required to achieve a reliability of 0.90.¹⁷

Reliability is an expression of the extent to which a measurement reproducibly determines its target quantity. The intraclass correlation coefficient (ICC) or reliability coefficient is the ratio of variance between subjects to the total variance, with a value of 1 implying no measurement error. A 'mean two-rater' ICC

TABLE I Learning Objectives of Case Scenarios

<i>Scenario #</i>	<i>Objective #1</i>	<i>Objective #2</i>	<i>Objective #3</i>	<i>Objective #4</i>
1	Preoperative assessment and management of aspiration risk	Rapid sequence induction	Recognition, & management of post-laryngoscopy hypertension & tachycardia	Recognition of bronchial intubation
2	Preoperative assessment of intravascular volume, administration of blood	Rapid sequence induction	Recognition & management of bronchial intubation	Recognition & management of post-induction hypotension
3	Preoperative considerations for the morbidly obese patient	Induction of anesthesia and tracheal intubation	Intubation & recognition of postintubation hypoxemia	Recognition and management of bronchospasm
4	Preoperative considerations for emergency surgery	Fluid resuscitation of hypovolemic patient	Rapid sequence induction	Recognition and management of hypotension after induction
5	Preoperative considerations for smoker	Induction of anesthesia and tracheal intubation	Determinants of adequate ventilation	Recognition and management of anaphylactic reaction
6	Advantages of regional vs general anesthesia	Induction of anesthesia and tracheal intubation	Recognition and management of hypotension and myocardial ischemia	Recognition and management of ventricular tachycardia

estimates the reliability of measurement when a student's score is reported as the mean of two raters' independent assessments. A 'single rater' ICC estimates the reliability of measurement if only one rater was used to determine a student's final score.

Content validity

In order to ensure that the evaluation process measured behaviors stated in the published learning objectives, undergraduate education committee members determined the coupling of objectives and performance tasks.

Convergent validity

Clinical evaluations of medical students during their anesthesia rotation are the mean of the numerical summation of 7-10 daily evaluation cards of six university-designated competencies: preoperative assessment, technical skills, application of knowledge and judgment, interpersonal skills, reliability and independent learning (scale 1-5, 1=unsatisfactory, 5=excellent).

Faculty members work in a 1:1 situation with the medical students in the operating room and their daily assessments are based on the interaction occurring between student and faculty during the day. Students usually work with different faculty each day, but may be assigned to the same faculty more than once during the two-week rotation.

The written examination comprises 10 short answer questions (based on the learning objectives of the rotation) with a possible total score of 100. Each examination included at least two questions pertaining to the simulator scenario content. Pearson correlation coefficients were determined for students' scores on their clinical, written, and (average) simulator assessments (n=135).

Internal consistency

ITEM ANALYSIS

Item-total correlation coefficients were determined for each of the four items or scores in each of the six scenarios.

Case analysis

Inter-item alpha (internal consistency) coefficients were determined for all six case scenarios. The mean simulator marks obtained in the preoperative considerations (knowledge) were correlated to written examination marks. Mean scores were compared between the clinical and simulator evaluations in two competencies: technical skills (induction of anesthesia) and judgement (intraoperative problem). Correlation between items and existing evaluation tools was calculated using a Pearson product moment correlation. Descriptive statistics of scenarios were calculated using means and standard deviations.

Results

Of the original student population of 177, 42 students did not participate in the study. Of these 42 students, 22

TABLE II Performance Evaluation Protocol (Scenario 1)
(Emergency surgery, full stomach)

Category	Evaluation Criteria
Preoperative considerations	Credit student if the following points are stated upon direct questioning (1 mark each): Total: 5 a) Patient at risk for aspiration, b) 2 reasons why pt is aspiration risk, c) should perform rapid sequence induction (RSI), d) describes RSI, e) one risk of RSI
Preparation	Credit student if the following equipment is checked (1 mark each): Total: 6 Suction, source of oxygen, circuit available, laryngoscope, endotracheal tube and stylet
Induction	Credit student if the following items are done (1 mark each): Total 8 a) assistant told to be ready for cricoid pressure, b) preoxygenation for at least 30 sec, c) faculty told to give drugs in rapid sequence, d) avoidance of positive pressure ventilation, e) laryngoscopy, f) trachea intubated on first attempt, g) chest auscultation, h) presence of CO ₂ checked
Intraoperative	For credit student should react/perform the following (1 mark each): Total Problems 6 a) verbally note tachycardia/hypertension in timely fashion, b) state laryngoscopy likely cause, c) increase depth of anesthesia in some manner, d) verbally note desaturation, e) auscultate lungs, f) recognize and correct bronchial intubation

opted not to attend the simulator session. Six students were absent due to interviews for postgraduate training and eight did not arrive because of a weather-related public transit shutdown. Evaluations of five students were not performed due to an audio failure related to microphone malfunction. One student was not present due to not having completed the requirements of the previous year's curriculum. This left a total of 135 students. Data from five elective students from other universities were included in the reliability results (n=140). Five pairs of faculty analyzed students' video performances. Each of these faculty pairs was assigned a group of between 25 and 34 students. (Table III).

Inter-rater reliability estimates (via intraclass correlation coefficients) for single and mean two-rater checklist scores are shown in Table III. Data from each of the rater pairs was used to generate an estimate of reliability for that data set. The mean reliability across rater pairs is the numerical average of these reliability estimates. For a single rater, the mean reliability across rater pairs was 0.77 (range 0.58-0.93). For average two-rater assessments, the mean reliability across rater pairs was 0.86 (range 0.74-0.96). To achieve a reliability of 0.9, an estimated 2.68 raters would be required.

Pearson correlation coefficients for mean checklist simulator scores and the clinical and written evaluations (n=135) were: simulator:written, $r=0.19^*$ simulator:clinical, $r=0.13$ and clinical:written, $r=0.23^*$ (*denotes statistically significant correlation, $P < 0.05$).

Inter-item correlations, means and standard deviations for the six case scenarios are summarized in Table IV. Correlation between items and existing evaluation tools was low: knowledge (simulator:written examination), $r=0.09$, technical skills (simulator:clinical), $r=0.10$ and judgement (simulator:clinical), $r=0.12$. In terms of differences between groups on the overall case score, there was an effect of group (i.e., which scenario they encountered) although this accounted for only 8% of the variance in total case score.

Corrected item-total correlation coefficients were computed for each of the four items within each case or scenario. (Table V). While a number of items demonstrated coefficients within the commonly

TABLE III Estimated reliability (and 95% confidence intervals) via intraclass correlation coefficients, for single and paired (average) ratings on simulator checklist scores. (n=140)

Reliability	examiner pair 1	examiner pair 2	examiner pair 3	examiner pair 4	examiner pair 5
n	26	26	34	25	29
single rating	.73 (.48-.87)	.58 (.26-.79)	.93 (.86-.96)	.84 (.66-.92)	.77 (.57-.89)
average rating	.84 (.65-.93)	.74 (.41-.88)	.96 (.92-.98)	.91 (.80-.96)	.87 (.73-.94)

TABLE IV Inter-item correlations, means and standard deviations of six scenarios

Scenario #	Inter-item <i>alpha</i>	Mean (Score 0-25)	SD	n
1	0.74	12.71	5.01	56
2	0.43	15.69	3.75	52
3	0.24	15.4	2.75	48
4	0.34	14.07	3.35	42
5	0.01	13.52	2.58	36
6	0.5	14.3	3.74	46
Total		14.3	3.84	280

TABLE V Item-total Correlation Coefficients for Four Items in Each of Six Scenarios

	Item Total Correlation Coefficients					
	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Preoperative Considerations	0.537	0.128	0.034	-0.112	-0.267	-0.08
Preparation	0.575	0.079	-0.04	0.039	-0.08	0.393
Induction of Anesthesia	0.5	0.367	0.41	0.126	0.143	0.452
Intraoperative Problems	0.631	0.381	0	0.215	0.327	0.074

acceptable range, a large number of items displayed fairly low correlations.

Discussion

The anesthesia simulator offers a useful environment for standardized testing of students or faculty.⁷ It is crucial that any innovative method of evaluating medical students be rigorously examined. The high cost of both the operation of a simulation centre and the labour-intensive nature of the endeavour compels careful consideration of the reliability and validity of these evaluation methods.^{18,19}

Our study demonstrated acceptable inter-rater reliability using faculty pairs to determine scores on medical students' videotaped performances. These results were similar to those reported in a pilot study with a different cohort of students (n=24) undertaken the year prior to this study.²⁰ Inter-rater reliability in this pilot project was 0.87, $P < 0.05$. Devitt *et al.* also reported high inter-rater reliability when performance of practising anesthesiologists was studied, ($r=0.96$, $P < 0.001$).²¹

In the current study, examination of each faculty pair's results revealed a range of reliability, with one pair (#2) being distinctly lower than other pairs. After the involved faculty members were interviewed, it was evident that one faculty member had been interrupted and distracted throughout much of the video viewing and did not give careful attention to the evaluation. This

experience highlights the importance of carefully addressing the commitment of evaluators to focus their full attention to the task at hand. If videotaped simulator evaluations are to be used as a part of a student's final grade, this issue becomes particularly important.

The decision as to what constitutes adequate reliability for student evaluation is somewhat arbitrary, although some authors have recommended requiring values of approximately 0.9 for tests used to make decisions about individuals.²² To use video assessments in educational practice, it would be reasonable to assess new faculty raters' reliability on a pilot series of videotapes, paired against an existing faculty rater.

The question regarding reliability estimates being based on single ratings or average ratings depends on the intended use of the test.²³ Given that our averaged scores from rating pairs achieved estimated reliabilities of approximately 0.9 despite imperfect examiner performance as described above, we feel that two trained raters would likely compose an adequate sample to provide acceptable reliability in simulator checklist assessment.

Correlations between the simulator checklist marks and clinical and written evaluations were low. These results may have multiple causes. One likely explanation is that the criteria for assessment of simulator performance are different from that used in the daily operating room assessment, or on the written examination. Performance in the simulator involves integration of clinical skills, judgment and applied knowledge. Ideally, these tasks would also be evaluated in the operating room but, in reality, patient safety issues and medico-legal considerations do not always allow for such experiences. Certainly, if a patient's condition were deteriorating for some reason, faculty would not allow students the time or latitude to assess, problem-solve, and independently manage the situation. Another possible cause is that daily evaluations are often determined by a student's ability to answer factual knowledge questions rather than evaluation of the full range of clinical performance. This fact may well account for the lack of correlation between the clinical and simulator evaluation. Similarly, the written examination is based on core knowledge and does not involve skills or hands-on medical management problems.

Validity is an important aspect in the development of any evaluation process. The validation process determines the degree of confidence we can place on inferences made about people based on their scores from that evaluation.²⁴ In this study, internal consistency of the simulator evaluations and correlations with existing methods of evaluation was used to address validity.

Inter-item correlations reflect the degree of association between the different components of a measurement. Our results demonstrated a wide variation in inter-item correlations. This is not surprising since, in some cases, one item or objective had no relationship with the others and therefore the scores on the items would not necessarily be similar (i.e., Scenario 5, objectives 1 and 4). In other cases (i.e., Scenario 1, objectives 1 and 2) the objectives were closely associated.

Item-total correlations indicate the degree to which an item contributes to the overall score. If the correlation is low, then the item should be considered for possible elimination on the basis that it may be measuring something other than the construct of interest. While the item total correlation coefficients and internal consistency coefficients may be low, this is not entirely surprising given that each of the four scores or 'items' are purporting to measure different components, namely Preoperative Considerations, Preparation, Induction of Anesthesia and Intraoperative Problems.

Means of simulator test scores were low. A number of reasons may account for these findings. The simulator experience was foreign to all of the students as was some of the equipment. Despite the orientation session, the lack of familiarity with the environment, may have negatively affected performance. As well, the knowledge that their performance was being videotaped and was being assessed by faculty members, may have caused some anxiety. Perhaps an even more important factor in the finding of low scores may be that we were expecting a multimodal performance. Students had to formulate an anesthetic plan based on knowledge of a patient case, then were required to perform the induction of anesthesia, facilitated by faculty and manage intraoperative problems. Although all of the cases and events were knowledge/clinical objectives of the undergraduate curriculum, few students would have encountered many of the problems in the operating room. Even if a similar situation had been clinically experienced, it would be unlikely that the student would have been allowed to manage the real life situation. This fact attests to the strengths of the simulator as a learning experience but may require more careful consideration when used as an evaluation tool.

The correlations of simulator specific objectives with existing methods of evaluation were also low. The low correlation of the knowledge component of the simulator assessment and the written examination was surprising. Students' final written examination was held at the end of the six-week block and it is possible that, at the time of the simulator performance assessment, they had not yet studied the material and therefore performed poorly. It is also possible that

some students had discussed similar cases with faculty during their rotation and therefore performed well.

Correlations between specific simulator and clinical items, (skills and judgement) were also low. The reason for this finding may be related to problems with the simulator assessment or problems with the clinical evaluations. It is well recognized that clinical ratings may be very subjective and somewhat arbitrary.²¹ Another cause for the differences may include the type of skills that were tested in each domain. Intravenous skills were not assessed in the simulator but were assessed in the clinical ratings. Tracheal intubation and mask ventilation was assessed in both environments, but this skill can be fairly difficult to perform in the simulator due to the inherent stiffness of the mannequin and lack of lubrication of the upper airway.

With respect to the assessment of judgement, faculty members are asked to rate students in the operating room using categories of unsatisfactory to excellent. Descriptors accompany these categories. Satisfactory is described as: common basic problems understood, and excellent as: applies advanced knowledge. It is likely that assessment of the judgment category on the clinical evaluation is based on dialogue that has occurred between faculty and student rather than the student having demonstrated judgement during the management of the case. Most students are assigned to procedures that involve healthy patients so that the student can perform hands-on skills and be able to discuss the case or other anesthesia curriculum objectives with the attending faculty. Therefore, most students wouldn't actually see any or many intraoperative problems. In the simulator, however, likely for the first time in many cases, students are faced with critical events and are expected to respond. This unique experience may be reflected in a relatively low score as compared to the clinical ratings.

A limitation of our study was the evaluation of student performance in only one scenario. It was impossible, therefore, to determine whether difficulty in case content accounted for differences in performance evaluations. We chose to have students participate in one scenario only for two reasons. The first was for purposes of case confidentiality. The second reason was based on the feasibility of having 177 medical students spend 90 min each, managing six cases. In a two-week rotation, this task would have been extremely onerous. It is however, important to appreciate differences in case difficulty and to be able to take this into account in any assessment method.

Unfortunately, there is no gold standard of undergraduate evaluation in anesthesia. This fact limits the comparison of simulator-based evaluations to existing

evaluation methods to determine validity. Although the written anesthesia examination at the University of Toronto has demonstrated excellent inter-rater reliability, the internal consistency of the examination is unknown.²⁵ Inter-rater reliability of the clinical assessments is unknown. Therefore, it is impossible to know what evaluation method is invalid when comparing the simulator assessments to existing evaluation methods. Nonetheless, our expectation of the student to initiate, manage and complete an anesthesia case may have been overzealous.

The simulator performance assessment methodology in this study demonstrated low internal consistency and requires further study before implementation. It may be necessary to limit the assessment process to the management of discrete events or testing on a particular skill, rather than attempting to evaluate a lengthier case scenario. In addition, prior exposure to the simulator environment is likely to improve student comfort level with the evaluation process. Once a valid performance assessment is developed, evaluation of videotapes of simulator performances by two raters can be considered a reliable method. Further work is planned to develop and integrate simulator evaluations into the undergraduate curriculum at the University of Toronto.

Acknowledgements

The authors would like to acknowledge the invaluable support of Mr. M. O'Donnell, Mr. B. Chong and Mr. L. Joy in the simulator operation for this study. As well, we would like to thank the faculty of the Department of Anesthesia and the students at the University of Toronto for their significant contribution.

References

- 1 Linn RL. Education assessment: expanded expectations and challenges. *Education Evaluation and Policy Analysis* 1993; 15: 1-16.
- 2 van der Vleuten CPM, Newble DI. How can we test clinical reasoning? *Lancet* 1995; 345: 1032-4.
- 3 Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: written and computer based simulations. *Assessment and Evaluation in Higher Education* 1987; 12: 220-46.
- 4 Rothman AI, Cohen R. Understanding the objective structured clinical examination: issues and options. *Annals RCPSC* 1995; 28: 283-7.
- 5 Hodges B, Regehr G, Hanson M, McNaughton N. Validation of an objective structured clinical examination in psychiatry. *Acad Med* 1998; 73: 910-2.
- 6 Stillman PL, Regan MB, Swanson DB, et al. An assessment of the clinical skills of fourth-year students at four New England medical schools. *Acad Med* 1990; 65: 320-6.
- 7 Gaba D, DeAnda A. A comprehensive anesthesia simulation environment: re-creating the operating room for research and training. *Anesthesiology* 1988; 69: 387-93.
- 8 Reznick R, Regehr G, MacRae HK, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg* 1996; 173: 226-30.
- 9 Taffinder N, Sutton C, Fishwick R, McManus IC, Darzi A. Validation of virtual reality to teach and assess psychomotor skills in laparoscopic surgery: results from randomised controlled studies using the MIST VR laparoscopic simulator. *Stud Health Technol Inform* 1998; 50: 124-30.
- 10 Rudman DT, Stredney D, Sessanna D, et al. Functional endoscopic sinus surgery training simulator. *Laryngoscope* 1998; 108: 1643-7.
- 11 Jambon AC, Dubecq-Princeteau F, Dubois P, et al. SPCI: a training simulator for initial formation in gynecologic laparoscopy. *J Gynecol Obstet Biol Reprod* 1998; 27: 536-43.
- 12 Isenberg SB, McGaghie WC, Hart IR, et al. Simulation technology for health care professional skills training and assessment. *JAMA* 1999; 282: 861-6.
- 13 Devitt JH, Kurrek M, Cohen MM. Can medical students be evaluated by a simulator based evaluation tool developed for practicing anesthesiologists? *Anesthesiology* 1997; 87: A947.
- 14 Tome JA, Fletcher J, Lydell DR. Performance assessment of medical students educated in a simulator environment. *Anesthesiology* 1997; 87: A946.
- 15 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979; 86: 420-8.
- 16 Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep* 1966; 19: 3-11.
- 17 Fleiss JL. Reliability of measurement. In: Fleiss JL (Ed.). *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons, 1986: 1-32.
- 18 Kurrek MM, Devitt JH. The cost for construction and operation of a simulation centre. *Can J Anaesth* 1997; 44: 1191-5.
- 19 Spence AA. Simulators in anesthesia. In: Ikeda K, Doi M, Kazama T (Eds.). *State-of-the-Art Technology in Anesthesia and Intensive Care*. Amsterdam: Elsevier, 1998: 171-3.
- 20 Morgan P, Cleave-Hogg D. Evaluation of medical students' performance using the anaesthesia simulator. *Med Educ* 2000; 34: 42-5.
- 21 Devitt JH, Kurrek MM, Cohen MM et al. Testing the raters: inter-rater reliability of standardized anaesthesia simulator performance. *Can J Anaesth* 1997; 44: 924-8.

- 22 *Streiner DL, Norman GR* Reliability. *In:* Streiner DL, Norman GR (Eds.). *Health Measurement Scales: A Practical Guide to their Development and Use*. Oxford: Oxford University Press, 1995: 104–27.
- 23 *Ebel RL*. Estimation of the reliability of ratings. *Psychometrika* 1951; 16: 407–24.
- 24 *Streiner DL, Norman GR* Validity. *In:* Streiner DL, Norman GR (Eds.). *Health Measurement Scales: A Practical Guide to their Development and Use*, 2nd ed. Oxford: Oxford University Press, 1995: 144–62.
- 25 *Tarshis J, Morgan PJ, Devitt JH* Marking of student written examinations: interrater reliability. *Anesthesiology* 1998; 89: A68.