Anesthesiology 2006; 104:475–81 

# *Evaluation of Patient Simulator Performance as an Adjunct to the Oral Examination for Senior Anesthesia Residents*

*Georges L. Savoldelli, M.D., M.Ed.,* Viren N. Naik, M.D., M.Ed.,† Hwan S. Joo, M.D.,‡ Patricia L. Houston, M.D., M.Ed.,§ Marianne Graham, M.D.,‖ Bevan Yee, M.B., Ch.B., F.A.N.Z.C.A.,# Stanley J. Hamstra, Ph.D.***

***Background:*** **Patient simulators possess features for performance assessment. However, the concurrent validity and the "added value" of simulator-based examinations over traditional examinations have not been adequately addressed. The current study compared a simulator-based examination with an oral examination for assessing the management skills of senior anesthesia residents.**

***Methods:*** **Twenty senior anesthesia residents were assessed sequentially in resuscitation and trauma scenarios using two assessment modalities: an oral examination, followed by a simulator-based examination. Two independent examiners scored the performances with a previously validated global rating scale developed by the Anesthesia Oral Examination Board of the Royal College of Physicians and Surgeons of Canada. Different examiners were used to rate the oral and simulation performances.**

***Results:*** **Interrater reliability was good to excellent across scenarios and modalities: intraclass correlation coefficients ranged from 0.77 to 0.87. The within-scenario between-modality score correlations (concurrent validity) were moderate: $r$ = 0.52 (resuscitation) and $r$ = 0.53 (trauma) ($P < 0.05$). Forty percent of the average score variance was accounted for by the participants, and 30% was accounted for by the participant-by-modality interaction.**

***Conclusions:*** **Variance in participant scores suggests that the examination is able to perform as expected in terms of discriminating among test takers. The rather large participant-by-modality interaction, along with the pattern of correlations, suggests that an examinee's performance varies based on the testing modality and a trainee who "knows how" in an oral examination may not necessarily be able to "show how" in a simulation laboratory. Simulation may therefore be considered a useful adjunct to the oral examination.**

This article is featured in "This Month in Anesthesiology." Please see this issue of ANESTHESIOLOGY, page 5A.

Additional material related to this article can be found on the ANESTHESIOLOGY Web site. Go to http://www.anesthesiology.org, click on Enhancements Index, and then scroll down to find the appropriate article and link. Supplementary material can also be accessed on the Web by clicking on the "ArticlePlus" link either in the Table of Contents or at the top of the Abstract or HTML version of the article.

ANESTHESIOLOGISTS have pioneered the use of full-scale simulators in medicine.[1,2] In our specialty, highly realistic mannequins are increasingly regarded as invaluable educational and research tools for the study of human performance.[3–9] However, their use in the assessment of performance and competence remains controversial and is still under scrutiny.[10,11] Many questions regarding the feasibility, reliability, and validity of simulator-based examinations have been raised.[11] Recent studies have shown that well-constructed scenarios and measurement instruments can reach fair to excellent levels of reliability[9,12–20] and can demonstrate construct validity.[9,13,17,21,22] Nevertheless, compared with traditional summative assessments, simulators require substantial human resources, organization, and funds. Hence, the value added by simulation-based examinations and their place in "high-stakes" examinations deserve further investigation.

In most countries, anesthesia board examinations still rely on traditional written and oral tests to determine competence for independent clinical practice. The former is directed at the assessment of factual knowledge, whereas the latter is the usual method for evaluating clinical assessment skills and judgment. Miller[23] suggested that written and oral examinations assess clinical competence at the "knows" and "knows how" stages, respectively, whereas performance-based examinations such as objective structured clinical examinations using standardized patients[24] assess clinical competence at the "shows how" stage. The highest level of assessment is the "does" and consists of observing what practitioners actually do in clinical settings. According to Miller,[23] the higher the level of assessment is, the better the validity of the test is. Ideally, the evaluation of clinical competence should include testing modalities that assess various stages of Miller's framework. However, using standardized patients to evaluate management skills of simulated life-threatening situations and invasive procedures has obvious limitations. Observing these skills in the clinical setting is also challenging and difficult. Emerging simulation technologies can help to overcome these limitations and may represent a practical

* Fellow, † Assistant Professor, St. Michael's Anesthesia Research into Teaching (SMART) Simulation Group, Department of Anesthesia, St. Michael's Hospital, and Wilson Centre for Research in Education, University Health Network, ‡ Assistant Professor, § Associate Professor, # Fellow, St. Michael's Anesthesia Research into Teaching (SMART) Simulation Group, Department of Anesthesia, St. Michael's Hospital, ‖ Assistant Professor, Department of Anesthesia, Sunnybrook and Women's College Health Science Centre, ** Associate Professor, Wilson Centre for Research in Education, University Health Network, and Department of Surgery, University of Toronto.

Received from the Patient Simulation Centre, St. Michael's Hospital, University of Toronto, Toronto, Ontario, Canada. Submitted for publication April 21, 2005. Accepted for publication November 21, 2005. Supported by the physicians of Ontario through a grant from the Physicians' Services Incorporated Foundation, North York, Ontario, Canada. Dr. Savoldelli was also supported by the University Hospitals and the Faculty of Medicine of Geneva, Switzerland, and by the Eugenio Litta Foundation, Geneva, Switzerland.

Address correspondence to Dr. Naik: Department of Anesthesia, St. Michael's Hospital, University of Toronto, 30 Bond Street, Toronto, Ontario, Canada M5B 1W8. naikv@smh.toronto.on.ca. 

alternative.[25] Oral examination and simulator-based examination are thus considered to be complementary, but few studies in anesthesia have compared simulators with other measures of clinical performance.

The purpose of this study was to investigate the potential of simulation for assessing the performances of senior anesthesia residents and to correlate and compare simulator performance with a mock oral examination modeled on a genuine board certification examination. This study was designed to assess residents' clinical competence in similar content domains using these two modes of evaluation. We examined the interrater reliability and the sources of variation in the scores. Correlations of scores obtained with the two assessment modalities provided a measure of their concurrent validity.

## Materials and Methods

*Participants*

After institutional review board approval (St. Michael's Hospital, University of Toronto, Toronto, Ontario, Canada), all final-year anesthesia residents were invited to participate. Residents were recruited during their weekly seminars and were free to decline participation. In addition to informed consent, confidentiality agreements were obtained to ensure that the content of the clinical scenarios would not be disseminated before the end of the study.

*Scenario Development*

We developed two standardized patient management scenarios (clinical cases), testing different domains: a resuscitation scenario and a trauma scenario (additional information regarding this is available on the ANESTHESIOLOGY Web site at http://www.anesthesiology.org). These domains were chosen for two reasons. First, these two domains are always tested during the 12-station oral examination for certification in Anesthesia given by the Royal College of Physicians and Surgeons of Canada (RCPSC). Second, they represent topics that lend themselves well to simulation. To ensure face and content validity, the study cases were derived from questions that had been previously used by the Anesthesia Oral Examination Board of the RCPSC. In addition, four faculty anesthesiologists involved in postgraduate medical education and two current RCPSC examiners reviewed their content. Particular attention was given to developing scenarios that had the capacity for simulation. For each of the clinical scenarios of resuscitation and trauma, an oral script and a simulator script were developed. Oral scripts modeled the format of traditional oral RCPSC questions. The simulator scripts consisted of programming scenarios into the computer software (Laerdal Scenario Builder) that controls the Laerdal SimMan Universal Patient Simulator (Laerdal Medical Canada Ltd., Toronto, Ontario, Canada). A thorough list of critical procedures, appropriate or inappropriate, that the examinees could potentially perform was included in the programming, so that the mannequin would react accordingly and in a standardized way from one candidate to another. In addition, the script described the timing and the content of verbal responses and actions that actors could provide to the candidate during the simulation session. To allow meaningful comparison between the two assessment modalities and minimize case specificity, scenarios and critical features given in the simulator were constructed to resemble the corresponding oral scenario. However, to avoid pattern recognition, the opening scenarios differed between the two assessment modalities.

*Study Design*

Each participant's oral examination was followed immediately by his or her simulator-based examination. In each assessment modality, the resuscitation scenario was followed by the trauma scenario. No feedback was provided between the two assessment modalities. Participants were not told that oral examination scenarios and simulator scenarios would be similar. After completing the study, participants were debriefed on their oral and simulator performances. During each study day, the participants were "quarantined" from each other between assessments. A questionnaire was administered to ascertain their perceptions of the oral and simulator-based examinations.

*Oral Examination Phase*

Each participant was assessed individually during an oral examination similar to the RCPSC examinations. Two examiners evaluated all participants. Time to discuss each question was limited to 12 min, as is customary during RCPSC examinations. At the end of the oral session, the examiners independently rated the performance using a global rating system.

*Simulator-based Examination Phase*

Before their simulator session, all participants were reoriented to the simulation environment. All simulator-based examinations were performed using the Laerdal SimMan Universal Patient Simulator. Participants were told to manage the patient as they would in the clinical setting. They were also instructed to verbalize clinical findings, differential diagnoses, management, and treatment choices. Two investigators, not involved in the rating process, played standardized roles (emergency nurse and senior medical student during the resuscitation scenario, and emergency nurse and general surgeon during the trauma scenario). They were present to provide assistance as required by the candidate. The duration of the scenarios varied slightly according to the candidate performance; the average duration was 11

min. The entire session was videotaped, and a graphical display of the patient's vital signs was overlaid on the videotape footage. Two examiners, different from those who evaluated the oral performance, independently reviewed and scored these tapes.

*Scoring System and Scoring Process*

Participants' performances were scored using a nine-item scale developed and validated by the Anesthesia Oral Examination Board of the RCPSC (appendix). Five aspects of the performance were rated: evaluation of the patient, creation of differential diagnoses, problem-solving ability, application of knowledge, and communication skills. The first four aspects were rated twice: once when the examinee was dealing with the initial clinical problem and once more after the clinical situation had changed. Each item was rated on a five-point scale with anchored performance criteria. For historical reasons, the five-point scale uses the numbers 64, 67, 70, 73, and 76, rather than 1–5. In a recent study using a simpler version of the scale, Kearney *et al.*[26] confirmed that the RCPSC oral examination provided fair to good interrater and intrarater reliability. Before the scoring process, the examiners agreed that a performance score of 70 would be the minimum deemed acceptable for an independent practitioner. The nine scores of the scale were then averaged to produce a total score for the performance (minimum score 64, maximum score 76). Examiners were also asked to give each participant a pass or fail rating.

The four examiners were clinical faculty involved in education. Two of them, one in each assessment modality, were current members of the Anesthesia Oral Examination Board of the RCPSC. All were trained to use the scoring system for the assessment modality they were judging. To facilitate their task, they were provided with scenario templates developed with the clinical cases. These templates described the critical features and the major decision points that an ideal examinee's performance should include for each scenario in each modality. (Additional information regarding this is available on the ANESTHESIOLOGY Web site at http://www.anesthesiology.org.)

*Statistical Analysis*

Statistical analysis was performed using SPSS 11.5 (SPSS Incorporated, Chicago, IL). Descriptive statistics were used to analyze demographic data. Reliability and validity of the measurement instrument were studied using several approaches. The internal consistency of the nine-item measurement scale was determined using the Cronbach $\alpha$. Interrater reliability of the total scores (using the sum of all nine items) was determined using an intraclass correlation coefficient. Using an analysis of variance approach,[27] we identified the relative sources of variance in the residents' scores (*e.g.*, participants, raters, scenarios, assessment modality) within and across assessment modalities. Participants' scores were correlated between assessment modalities using the obtained Pearson correlation coefficients to assess the concurrent validity. A $P$ value of less than 0.05 was considered statistically significant.

## Results

*Demographics*

Twenty-one of 28 eligible residents participated in the study. One participant was excluded because of videotape malfunction during the two simulated scenarios. An additional simulator performance (trauma scenario) for one participant could not be scored because of technical recording difficulties. The mean participant age was 33.6 yr (± 3.8 yr). Fifteen percent of the participants were female (4 of 20). This ratio was lower than the overall proportion in our program (around 40%) as a result of fewer women at this level of training during the study period. With one exception, all participants had experienced simulation-based education in the past. The number of previous simulator experiences ranged from 0 to 6 (median 3). All participants were Advanced Cardiac Life Support certified, but only 55% were Advanced Trauma Life Support certified.

*Descriptive Statistics*

In the oral examination, the means (± SDs) of the total scores in the resuscitation and trauma scenarios were 57.1 (± 13.4) and 57.0 (± 13.0), respectively (expressed as percentages of the maximum score). The overall pass rates were 70% and 80% for the resuscitation and trauma scenarios, respectively. Corresponding scores in the simulator-based examination were 58.5 (± 14.3) and 64.8 (± 11.0), with overall pass rates of 85% and 95%, respectively. Relatively large SDs and ranges of scores indicate that despite homogeneity in the participants' level of training, raters were still able to detect individual differences in participants' performances.

*Reliability*

The internal consistency of the rating scale was high across scenarios, raters, and assessment modalities: Cronbach $\alpha = 0.93–0.98$. Interrater reliability was good to excellent across scenarios and modalities ranging from intraclass correlation coefficient (single rater) = 0.77–0.87 ($P < 0.001$). The pass/fail agreement between the raters in both modalities was good. In the oral examinations, raters agreed on 37 of the 40 performances. In the simulator-based examinations, examiners agreed on all but one performance.

*Variance Components Analysis*

The analysis of variance components was performed separately for each assessment modality (fully crossed

**Table 1. Variance Components of the Resident Scores within Each Assessment Modality**

| | Oral | | Simulator | |
|---|---|---|---|---|
| Component | Estimate | % | Estimate | % |
| Participant | 92.59 | 51.80 | 99.57 | 53.37 |
| Rater | 0.38 | 0.21 | 2.64 | 1.41 |
| Scenario | 0.0 | 0.0 | 13.48 | 7.22 |
| Participant × rater | 13.57 | 7.59 | 8.66 | 4.64 |
| Participant × scenario | 51.23 | 28.66 | 42.79 | 22.94 |
| Rater × scenario | 0.0 | 0.0 | 1.18 | 0.63 |
| Residual* | 20.99 | 11.74 | 18.24 | 9.78 |

* Residual = triple order interaction (participant × rater × scenario) plus all unexplained sources of variance.

**Table 2. Variance Components for the Average Scores across Assessment Modalities**

| Component* | Estimates | % |
|---|---|---|
| Participant | 59.75 | 40.01 |
| Scenario | 1.04 | 0.69 |
| Participant × scenario | 6.43 | 4.31 |
| Participant × modality | 44.98 | 30.12 |
| Scenario × modality | 2.06 | 1.38 |
| Residual† | 35.07 | 23.49 |

* The assessment modality was considered a fixed factor; hence it does not appear in the main effects. † Residual = participant × scenario × modality plus all unexplained sources of variance.

design: participants by rater by scenario). This provided a way of exploring the sources of variance in measuring performances. Table 1 illustrates that the sources of variance in both modalities followed similar patterns. The main source of variance was the participants, signifying that individual abilities accounted for more than half of the variance in observed scores. Because the purpose of an evaluation is to detect differences between examinees, most of the variance should be accounted for by the participants. The rater component reflects differences in rater mean scores; the low values indicate that the raters' stringency did not vary substantially. The scenario component reflects differences in scenario mean score; this provides an estimate of the variation in relative difficulty of the scenarios. The low value indicates that overall, the scenarios were of similar difficulty. The participant-by-rater interaction component provides an estimate of how the various raters differ in their ranking of the examinees. In a fair test, this component should be equal or close to zero. The participant-by-scenario interactions were the second largest source of variance, indicating that rank ordering of examinee differed for each scenario. The last variance components are the residual variances that include the triple-order interaction and all other unexplained sources of variance; this is generally considered to be an estimate of the error variance.

After averaging the scores of the two raters, a single score was obtained for each participant in each scenario and in each modality. This provided a way of exploring the role of the assessment modality as a source of variance (participant by scenario by modality). Table 2 shows a variance component analysis of the average scores. Again, the main source of variance was the participant, signifying that individual abilities accounted for 40% of the variance in observed scores. The largest interaction variance component was the participant-by-modality interaction. The latter component indicates that rank ordering of examinee during a similar scenario (resuscitation or trauma) differs substantially between each assessment modality. This finding alone strongly suggests that the addition of the simulator as a modality picks up strengths or weaknesses in some residents' performances that would otherwise be missed by the oral examination.

### *Correlation between Assessment Modalities*

Oral examination scores correlated moderately with simulator scores: Pearson $r = 0.52$ and 0.53 (resuscitation and trauma scenario respectively; $P < 0.05$). This indicates a relatively low level of concurrent validity between the two testing modalities. Finally, any participant who received a "fail" rating in the simulator-based evaluation also received a "fail" rating from at least one examiner during the corresponding oral scenario.

### *Residents' Opinions of Oral and Simulator Assessment*

Table 3 presents the percentage of participants that agree and disagree with the statements proposed in the poststudy questionnaire. Most residents agree that the simulator may be a useful adjunct but express caution before widespread adoption—local and national access to simulation-based education and evaluation must be improved. The participants also noted the need for further validation research.

## Discussion

Our results suggest that clinical judgment and management skills of senior anesthesia residents can be assessed using a simulator with a level of reliability equivalent to that of an oral examination. In addition, simulator and oral examinations demonstrate only moderate concurrent validity.

In this study, we used the same rating scale for both assessment modalities. However, the performance expectations were somewhat different for the oral and the simulator-based examination. Oral scorings rewarded logical sequencing of reasoning, verbalization of management plans, and the ability to convey opinions in a precise and concise way. Simulator scorings mostly rewarded logical sequencing of management steps, demonstration of proper technical abilities, and ability to communicate and lead a medical team (see Web En-

**Table 3. Residents' Opinions of Oral and Simulator Assessment (n = 20)**

| Statement | Strongly Agree or Agree, % | Strongly Disagree or Disagree, % |
|---|---|---|
| Simulation-based examination is useful as an assessment tool | 75 | 5 |
| Further research in the field is needed before its use in formal evaluation | 90 | 0 |
| For ACLS and ATLS management: simulator-based examination reflects clinical competency better than oral examination | 75 | 0 |
| Prior simulator experience is mandatory before its use as an evaluation tool | 90 | 5 |
| Simulator-based examination has the potential for being incorporated into the specialty certification exam in anesthesia | 50 | 10 |

Responses derived from a five point Likert scale (1 = strongly disagree; 2 = disagree; 3 = undecided; 4 = agree; 5 = strongly agree). The "undecided" category is not shown in the table.

ACLS = Advanced Cardiac Life Support; ATLS = Advanced Trauma Life Support.

hancement). The following are some examples of suboptimal performances observed during the simulated sessions that were not observable in the oral: inappropriate protection of the cervical spine during airway management, failure to perform chest compression despite recognizing the absence of pulse, failure to perform needle decompression of a tension pneumothorax, incorrect sequence of physical evaluation and/or treatment during patient management. Therefore, the differences between oral and simulated scores reflected differences between the oral and simulated performance endpoints.

Despite similarity in the domains of knowledge tested, performance scores obtained with the two assessment modalities correlated only moderately. The coefficient of approximately 0.5 indicates that scores obtained during an oral examination explained only approximately one quarter of the variance of the scores obtained in the simulator assessment. In addition, the participant-by-modality interaction was an important source of the scores variance, indicating that some candidates performed better in the oral examination than in the simulator and *vice versa*. This signifies that an examinee's performance varies based on the testing modality and suggests that a trainee who "knows how" in an oral examination may not necessarily be able to "show how" in a simulation laboratory. These results are in accordance with Miller's conceptual framework of performance assessments described in the introduction.[23]

Several other sources of variance may explain the differences in the scores. The variance component analysis clarifies why the scores differed from one evaluation to another. A legitimate concern is the rater-related variance linked to judges' subjectivity, particularly when the raters use holistic evaluation tools and know some of the candidates, as was the case in our study. There is, however, some evidence that, when expert raters assess complex performances, global scales have excellent psychometric properties and are even more likely to capture increasing levels of expertise compared with checklists.[28,29] Their main advantage is to capture the adequacy of timing and sequencing of patient management that are characteristics of complex patient care. In our study, the total variance linked to the examiners (main effect plus participant-by-rater interaction) was modest (7.8% of the oral and 6.1% of the simulator scores). Previous studies using experts and global ratings have reported similar results.[20,30]

Another well-known source of variance is the participant-by-scenario interaction. In our study, this was the second largest variance component in each modality. This indicates that participants' performance varies between scenarios, and performance in one scenario moderately predicts performance in another scenario. However, this latter component was lower in our study than in previous reports.[19,20,30] One plausible explanation is that the two domains that were tested, resuscitation and trauma care, share some generic management principles. It is likely that had we tested very different domains, the participant-by-scenario variance would have been greater. Interestingly, after averaging the scores of both raters to assess the effect of the testing modality, the participant-by-scenario interaction decreased significantly. Another source of variance, the participant-by-modality interaction, became more important than the effect of the scenario itself. This suggests that if the two modalities are combined, fewer encounters might be needed to reliably test examinees' competence in one clinical domain. Finally, the residual variance components, an estimate of the error variance, were acceptable and comparable to what has been reported.[20]

Every participant who received a "fail" rating in the simulator-based evaluation also received at least one "fail" rating during the corresponding oral scenario. This adds further support to the concurrent validity of the simulator compared with the oral examination. One may

argue that if the purpose of the test is to reach a pass/fail decision, why bother doing complex and costly simulator tests? However, our results also suggest that the simulator measures important untested abilities in trainees that would otherwise be unfavorably rated in the oral examination alone. In addition, the current study was limited to testing only two scenarios in two specific clinical domains.

Few studies have compared simulator performance with other modes of evaluation. Morgan *et al.*[15,16] found no or weak correlations between simulator assessments of medical students and their written or clinical evaluations during their anesthesia rotation. The poor correlation with the written examination can be explained by the fact that this test measures factual knowledge at Miller's stage of "knows," *i.e.*, at a lower stage compared with the oral examination. In a multicenter study, Schwid *et al.*[21] compared assessments from mock oral and simulation examinations and found results that were similar to ours ($r = 0.47$). However, this was not the primary purpose of their study, and unrelated scenarios and different domains of knowledge tested in each modality may have influenced their results. This source of variation in performance assessment is known as "case specificity" and may have led to a moderate correlation.[23] In the current study, we have used similar scenarios in the two evaluation modes in an attempt to control for this confounding variable. Interestingly, although this design may have favored a stronger correlation, our results are comparable to those reported by Schwid *et al.*[21]

Responses to the poststudy questionnaire suggest that future simulator-based assessments will likely gain acceptance among examinees. A majority of the participants thought that simulation was a useful assessment tool and that for advanced cardiac and trauma management, their simulator performance reflected their clinical competency better than oral examinations did. They also agreed that a simulator-based examination has the potential for being incorporated into the specialty certification examination in anesthesia. Nevertheless, the majority recognized that further research was needed and that previous experience and access to simulation was required before consideration could be given to its use for high-stakes evaluation. Interestingly, the only study participant who had no previous experience with simulation did not seem to be at a disadvantage compared with other candidates. In addition, although their scores tended to be higher in both modalities, the performances of residents with Advanced Trauma Life Support training did not statistically differ from nontrained residents.

The current study has limitations. As mentioned, the relatively small number of scenarios limits the generalizability of the results. The use of multiple scenarios would have permitted us to draw conclusions on a larger range of skills. Recent studies using a generalizability theory approach[27] have addressed this important issue using a variety of measurement instruments.[19,20,30] Depending on the study, the number of raters, and the scoring system, the authors found that between 8 and 15 short simulated encounters would be necessary to achieve a reliable assessment. The second limitation is that we used identical raters for two scenarios within each assessment modality. We designed our study this way for two reasons: first, to keep the simulator examiners blinded from the oral performance, and second, to be consistent with what is currently done during a Royal College examination (external validity). The constant order of presentation of the assessment modalities, oral followed by simulation, is another limitation. This design does not allow us to assess the effect of sequencing on the results. It was chosen because of a limited sample. A larger sample, divided into two groups, one starting with the oral and the other with the simulator, would have permitted to assess the effect of sequencing. The similarities between oral and simulated scenarios might have further influenced the impact of sequencing. Our attempt to control for "case specificity" (by using similar scenarios) was a rational choice in this experimental study. However, it would certainly be worthwhile to repeat the study using different case scenarios in the two modalities, randomly chosen from a universe of cases. This may in fact better reflect the way simulation might be used for future testing.

In conclusion, our results suggest that examinees differ in performance based on how the skills are assessed. Resident scores on the oral examination ("knows how") were not a good predictor of how a particular resident performed in a simulator-based assessment ("shows how"). Examinees may differentially benefit or be penalized from an assessment based on either of these two modalities. Simulation may therefore be considered a useful adjunct to the oral examination to measure clinical competence of senior anesthesia residents. To our knowledge, only one country, Israel, has incorporated simulation in its national board certification examination in anesthesiology.[31] It is probably necessary that before its wide acceptance by licensing authorities for high-stakes examination, larger studies must be conducted. Future studies should look at the relation between simulator performance and real clinical practice. However, the ultimate question of whether one assessment modality predicts best what a clinician actually does in practice will be difficult to answer. Therefore, trainees' competency, even in a single competence domain such as trauma or resuscitation management, may need to be assessed using multiple assessment modalities. Finally, issues of equal accessibility to simulators during training and familiarity with the tool before the examination will need to be addressed.

The authors thank the fifth-year anesthesiology residents for their participation in this study. They also thank Bruce Karatzoglou, B.Sc., and Roger Chow (Simulation Centre Coordinators, Patient Simulation Centre, St. Michael's Hospital, Toronto, Ontario, Canada) for their contribution. Finally, the authors thank Jodi Herold McIlroy, Ph.D. (Assistant Professor), and Glenn Regehr, Ph.D. (Professor, Wilson Centre for Research in Education, University of Toronto, Toronto, Ontario, Canada), for their thoughtful advice while designing the study.

## References

1. Abrahamson S, Denson JS, Wolf RM: Effectiveness of a simulator in training anesthesiology residents. J Med Educ 1969; 44:515–9
2. Gaba DM, DeAnda A: A comprehensive anesthesia simulation environment: Re-creating the operating room for research and training. ANESTHESIOLOGY 1988; 69:387–94
3. Gaba DM, Howard SK, Fish KJ, Smith BE, Sowb YA: Simulation-based training in anesthesia crisis resource management (ACRM): A decade of experience. Simul Gaming 2001; 32:175–93
4. Ziv A, Wolpe PR, Small SD, Glick S: Simulation-based medical education: An ethical imperative. Acad Med 2003; 78:783–8
5. Weller J, Wilson L, Robinson B: Survey of change in practice following simulation-based training in crisis management. Anaesthesia 2003; 58:471–3
6. Blum RH, Raemer DB, Carroll JS, Sunder N, Felstein DM, Cooper JB: Crisis resource management training for an anaesthesia faculty: A new approach to continuing education. Med Educ 2004; 38:45–55
7. Howard SK, Gaba DM, Smith BE, Weinger MB, Herndon C, Keshavacharya S, Rosekind MR: Simulation study of rested *versus* sleep-deprived anesthesiologists. ANESTHESIOLOGY 2003; 98:1345–55
8. Chopra V, Gesink BJ, de Jong J, Bovill JG, Spierdijk J, Brand R: Does training on an anaesthesia simulator lead to improvement in performance? Br J Anaesth 1994; 73:293–7
9. Devitt JH, Kurrek MM, Cohen MM, Cleave-Hogg D: The validity of performance assessments using simulation. ANESTHESIOLOGY 2001; 95:36–42
10. Kapur PA, Steadman RH: Patient simulator competency testing: ready for takeoff? Anesth Analg 1998; 86:1157–9
11. Byrne AJ, Greaves JD: Assessment instruments used during anaesthetic simulation: Review of published studies. Br J Anaesth 2001; 86:445–50
12. 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
13. Devitt JH, Kurrek MM, Cohen MM, Fish K, Fish P, Noel AG, Szalai JP: Testing internal consistency and construct validity during evaluation of performance in a patient simulator. Anesth Analg 1998; 86:1160–4
14. Morgan PJ, Cleave-Hogg D, Guest CB: A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. Acad Med 2001; 76:1053–5
15. Morgan PJ, Cleave-Hogg DM, Guest CB, Herold J: Validity and reliability of undergraduate performance assessments in an anesthesia simulator. Can J Anaesth 2001; 48:225–33
16. Morgan PJ, Cleave-Hogg D, DeSousa S, Tarshis J: High-fidelity patient simulation: Validation of performance checklists. Br J Anaesth 2004; 92:388–92
17. Forrest FC, Taylor MA, Postlethwaite K, Aspinall R: Use of a high-fidelity simulator to develop testing of the technical performance of novice anaesthetists. Br J Anaesth 2002; 88:338–44
18. Weller JM, Bloch M, Young S, Maze M, Oyesola S, Wyner J, Dob D, Haire K, Durbridge J, Walker T, Newble D: Evaluation of high fidelity patient simulator in assessment of performance of anaesthetists. Br J Anaesth 2003; 90:43–7
19. Boulet JR, Murray D, Kras J, Woodhouse J, McAllister J, Ziv A: Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. ANESTHESIOLOGY 2003; 99:1270–80
20. Murray DJ, Boulet JR, Kras JF, Woodhouse JA, Cox T, McAllister JD: Acute care skills in anesthesia practice: A simulation-based resident performance assessment. ANESTHESIOLOGY 2004; 101:1084–95
21. Schwid HA, Rooke GA, Carline J, Steadman RH, Murray WB, Olympio M, Tarver S, Steckner K, Wetstone S: Evaluation of anesthesia residents using mannequin-based simulation: A multiinstitutional study. ANESTHESIOLOGY 2002; 97:1434–44
22. Byrne AJ, Jones JG: Responses to simulated anaesthetic emergencies by anaesthetists with different durations of clinical experience. Br J Anaesth 1997; 78:553–6
23. Miller GE: The assessment of clinical skills/competence/performance. Acad Med 1990; 65:S63–7
24. Harden RM, Gleeson FA: Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979; 13:41–54
25. Issenberg SB, McGaghie WC, Hart IR, Mayer JW, Felner JM, Petrusa ER, Waugh RA, Brown DD, Safford RR, Gessner IH, Gordon DL, Ewy GA: Simulation technology for health care professional skills training and assessment. JAMA 1999; 282:861–6
26. Kearney RA, Puchalski SA, Yang HY, Skakun EN: The inter-rater and intra-rater reliability of a new Canadian oral examination format in anesthesia is fair to good. Can J Anaesth 2002; 49:232–6
27. Brennan RL, Johnson EG: Generalizability of performance assessments. Educ Meas Iss Pract 1995; 14:9–12
28. Regehr G, MacRae H, Reznick RK, Szalay D: Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med 1998; 73:993–7
29. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M: OSCE checklists do not capture increasing levels of expertise. Acad Med 1999; 74:1129–34
30. Weller JM, Robinson BJ, Jolly B, Watterson LM, Joseph M, Bajenov S, Haughton AJ, Larsen PD: Psychometric characteristics of simulation-based assessment in anaesthesia and accuracy of self-assessed scores. Anaesthesia 2005; 60:245–50
31. Sidi A, Ziv A, Berkenstadt H: The process of incorporating simulation to the Israeli national board examination in anesthesiology: From preliminary to obligatory program (abstract). ANESTHESIOLOGY 2004; 101:A1412

## Appendix: Global Rating Scale of Performance

Circle the appropriate rating for each aspect of the candidate's performance.

| | | | | |
|---|---|---|---|---|
| **1. Initial evaluation of the patient and gathering of information. Hx, Px, Lab** | | | | |
| 64 | 67 | 70 | 73 | 76 |
| Major omissions | | Acceptable | | Exemplary |
| **2. Creation of DDX** | | | | |
| 64 | 67 | 70 | 73 | 76 |
| Illogical and/or inappropriate | | Acceptable | | Comprehensive |
| **3. Problem solving** | | | | |
| 64 | 67 | 70 | 73 | 76 |
| Illogical and/or inadequate | | Acceptable | | Exemplary |
| **4. Application of knowledge** | | | | |
| 64 | 67 | 70 | 73 | 76 |
| Illogical and/or inadequate | | Acceptable | | Exemplary |
| **5. Ability to deal with changing situations: Evaluation of patient and gathering information. Hx, Px, Lab** | | | | |
| 64 | 67 | 70 | 73 | 76 |
| Illogical and/or inadequate | | Acceptable | | Exemplary |
| **6. Ability to deal with changing situations: Creation of DDX** | | | | |
| 64 | 67 | 70 | 73 | 76 |
| Illogical and/or inappropriate | | Acceptable | | Comprehensive |
| **7. Ability to deal with changing situations: Problem solving** | | | | |
| 64 | 67 | 70 | 73 | 76 |
| Illogical and/or inadequate | | Acceptable | | Exemplary |
| **8. Ability to deal with changing situations: Application of knowledge** | | | | |
| 64 | 67 | 70 | 73 | 76 |
| Illogical and/or inadequate | | Acceptable | | Exemplary |
| **9. Communications skills/leadership** | | | | |
| 64 | 67 | 70 | 73 | 76 |
| Illogical and/or inadequate | | Acceptable | | Exemplary |
| **Overall on this procedure, should the candidate: Pass/Fail** | | | | |

70 is the minimum score expected from an independent practitioner.

DDX = differential diagnosis; Hx = history taking; Lab = laboratory; Px = physical examination.