Anesthesiology 2007; 107:705–13

Performance of Residents and Anesthesiologists in a Simulation-based Skill Assessment

David J. Murray, M.D., John R. Boulet, Ph.D.,† Michael Avidan, M.D.,‡ Joseph F. Kras, M.D.,‡ Bernadette Henrichs, C.R.N.A.,§ Julie Woodhouse, R.N.,‖ Alex S. Evers, M.D.#*

Background: **Anesthesiologists and anesthesia residents are expected to acquire and maintain skills to manage a wide range of acute intraoperative anesthetic events. The purpose of this study was to determine whether an inventory of simulated intraoperative scenarios provided a reliable and valid measure of anesthesia residents' and anesthesiologists' skill.**

Methods: **Twelve simulated acute intraoperative scenarios were designed to assess the performance of 64 residents and 35 anesthesiologists. The participants were divided into four groups based on their training and experience. There were 31 new CA-1, 12 advanced CA-1, and 22 CA-2/CA-3 residents as well as a group of 35 experienced anesthesiologists who participated in the assessment. Each participant managed a set of simulated events. The advanced CA-1 residents, CA-2/CA-3 residents, and 35 anesthesiologists managed 8 of 12 intraoperative simulation exercises. The 31 CA-1 residents each managed 3 intraoperative scenarios.**

Results: **The new CA-1 residents received lower scores on the simulated intraoperative events than the other groups of participants. The advanced CA-1 residents, CA-2/CA-3 residents, and anesthesiologists performed similarly on the overall assessment. There was a wide range of scores obtained by individuals in each group. A number of the exercises were difficult for the majority of participants to recognize and treat, but most events effectively discriminated among participants who achieved higher and lower overall scores.**

Conclusion: **This simulation-based assessment provided a valid method to distinguish the skills of more experienced anesthesia residents and anesthesiologists from residents in early training. The overall score provided a reliable measure of a participant's ability to recognize and manage simulated acute intraoperative events. Additional studies are needed to determine whether these simulation-based assessments are valid measures of clinical performance.**

This article is accompanied by an Editorial View. Please see: Weinger MB: Experience ≠ expertise: Can simulation be used to tell the difference? ANESTHESIOLOGY 2007; 107:691–4.

PERIOPERATIVE critical events remain a leading cause of adverse patient outcomes.[1,2] Prompt diagnosis and rapid treatment are considered essential to reduce the morbidity and mortality associated with these acute conditions.[3] However, the near random incidence of these infrequent, life-threatening events challenges a physician's ability to acquire and maintain the skills needed to manage them. For this reason, a resident will often have few, if any, opportunities to develop the skills needed to manage these critical events. The introduction of full-scale electromechanical mannequins provides an educational setting where medical students, residents, and anesthesiologists can acquire skills and develop expertise necessary to recognize and manage many of these high-risk conditions.[3–18]

In previous studies, we found that an anesthesia resident's ability to recognize and manage one event did not generalize very well to the management of other critical conditions.[5,11] While participants who effectively managed one event were more likely to receive higher scores on other events, an adequately reliable overall performance measure would require each trainee to manage numerous simulated events.[5,11,19,20] For this reason, an inventory of simulated events was needed for future studies. Similar to the previously designed simulated exercises, these scenarios and associated scoring actions require analysis to assure that participants' scores are consistent and reliable. In addition, studies are needed to determine whether the resulting scores are valid and predict an individual's performance in an operating room. Most studies that evaluate simulation-based training and assessment are designed and targeted at the skills of medical students and residents.[3–20] The effectiveness of many of these interventions is difficult to assess because skill is rapidly evolving as a result of the comprehensive learning environment associated with training. In contrast, a specialist's skill is likely more stable and primarily influenced by their experiences in patient care settings. There are few, if any, studies that evaluate the skills of anesthesiologists managing simulated intraoperative events. An assessment of anesthesiologists could provide valuable data to evaluate the reliability and validity of the training exercises and to determine how practice experience affects the acquisition and maintenance of skills in a simulated environment.

The purpose of this study was to determine whether a multiple-scenario simulation-based assessment could be used to provide a reliable and valid measure of anesthesia residents' and anesthesiologists' performances in a simulated intraoperative environment. More specifi-

* Carol B. and Jerome T. Loeb Professor in Medicine and Director, ‖ Administrator, Washington University Clinical Simulation Center, ‡ Associate Professor, # Henry E. Mallinkrodt Professor and Chairman, Department of Anesthesiology, § Co-Director, Nurse Anesthesia Services, Department of Anesthesiology, Washington University School of Medicine. † Associate Vice President, Research and Data Resources, Foundation for Advancement of International Medical Education and Research, Philadelphia, Pennsylvania.

Received from the Department of Anesthesiology, Washington University School of Medicine, St. Louis, Missouri. Submitted for publication May 5, 2006. Accepted for publication June 28, 2007. Supported by the Foundation for Anesthesia Education and Research, Rochester, Minnesota (to Dr. Murray, Principal Investigator [PI]); AHRQ 1 U18 HS016652-01 from the National Institutes of Health, Bethesda, Maryland (to Dr. Murray, PI); and the Clinical Research Division, Department of Anesthesiology, Washington University, St. Louis, Missouri (to Dr. Avidan, PI).

Address correspondence to Dr. Murray: Washington University Clinical Simulation Center, Washington University School of Medicine, Box 8054, 660 South Euclid, St. Louis, Missouri, 63110. murrayd@wustl.edu.

cally, the variability in participant scores, as a function of both choice of scenario and choice of rater, was used to estimate the reliability and reproducibility of this type of assessment. For validity, the scores from individual simulation exercises were examined to determine whether factors such as training and experience resulted in performance differences and also whether an individual scenario score correlated with overall scores.

Materials and Methods

For this project, 12 scripted simulated scenarios were developed to measure essential skills in intraoperative acute care management (table 1). These events were selected based on input from five faculty educators. These faculty members were asked to list critical events that an anesthesia consultant should be able to recognize and manage. The lists were collated and then resubmit-

Table 1. Description of Events and Scoring Items for 12 Scenarios

Scenarios	Scoring Items
Bronchospasm One minute after beginning the simulation, oxygen saturation level decreases to 85%, and heart rate increases to 120 beats/min. Blood pressure remains at 105/60 mmHg.	Listen to chest, increase inspired oxygen, state diagnosis, administer β agonist/epinephrine
Anaphylaxis At 1 min, blood pressure decreases to 80/50 mmHg, heart rate increases from 64 to 115 beats/min, and oxygen saturation decreases to 88%. At 3 min, a skin rash clue is given.	Increase inspired oxygen, auscultate, check blood pressure, state diagnosis, stop antibiotic infusion, administer epinephrine
Unstable ventricular tachycardia Thirty seconds after beginning the simulation, wide-complex tachycardia occurs with decreased blood pressure. Heart rate increases from 70 to 170 beats/min, and blood pressure decreases from 120/70 mmHg to 80/60 mmHg.	State diagnosis, increase inspired oxygen, deliver shock, deliver synchronized cardioversion, give/request antiarrhythmic
Myocardial ischemia Electrocardiogram indicates ST segment depression is at 3 mm. Heart rate increases to 125 beats/min, and blood pressure is 170/90 mmHg.	State diagnosis, check or order 12-lead electrocardiogram, titrate narcotic or β-blocker therapy to decrease heart rate, request nitroglycerin drip or apply paste
Right bronchial intubation At the beginning of the simulation, vital signs are stable, except oxygen saturation is 91%. There is no chest wall movement on the left side of the chest.	Auscultation or inspection of chest, increase inspired oxygen, state diagnosis, reposition endotracheal tube
Tension pneumothorax Pneumothorax is present from the beginning of the simulation, blood pressure decreases to 85/55 mmHg, heart rate increases to 120 beats/min, and oxygen saturation continues to decrease to 85%.	Auscultation of chest, increase inspired oxygen, state diagnosis, relieve with needle or place chest tube
Malignant hyperthermia Within 1 min of beginning the simulation, heart rate and blood pressure increase to 115 beats/min and 180/90 mmHg, respectively. Increased end-tidal carbon dioxide level.	State diagnosis, turn off agent, call for dantrolene or malignant hyperthermia cart
Blocked endotracheal tube At the beginning of the simulation, vital signs are stable, except oxygen saturation is 90%. Breath sounds are distant. Elevated airway pressures during controlled ventilation.	Auscultate, increase inspired oxygen, recognize increased airway pressures, pass suction catheter, state diagnosis, remove blocked endotracheal tube
Total spinal Within 1 min after beginning the simulation, blood pressure starts decreasing to 60/40 mmHg, and heart rate decreases to 40 beats/min.	Increase inspired oxygen, check blood pressure, turn off agent, increase fluids, state diagnosis, give epinephrine
Loss of pipeline oxygen Fifteen seconds after beginning the simulation, the pipeline oxygen is turned off. At 30 s, the alarm sounds. Vital signs remain stable.	State diagnosis, open oxygen tank 1, open oxygen tank 2
Hyperkalemia At the beginning of the simulation, blood pressure is 170/90 mmHg, and heart rate is 75 beats/min. Heart rate increases, ventricular irritability increases, and peaked T waves are evident on electrocardiogram.	Order or check electrolytes or arterial blood gas or potassium, state diagnosis, institute appropriate treatment
Acute hemorrhage One minute after beginning the simulation, blood pressure begins to decrease to 85/50 mmHg, and heart rate increases to 115 beats/min. One liter of "blood" is in the suction canister.	Ask about blood loss or evaluate for excessive blood loss (check suction canister), increase intravenous fluids, state diagnosis, request hemoglobin or hematocrit or blood product

Table 2. Anesthesiologists' and Residents' Demographic Profiles

	CA-1 (New) (n = 31)	CA-1 (Adv) (n = 12)	CA-2/CA-3 (n = 21)	Anesthesiologists (n = 35)
Mean age, yr	28 ± 2	31 ± 3	32 ± 4	43 ± 6
Female/male, n	12/19	4/8	6/21	6/29
Practice years				11 ± 4
Practice setting, community/academics				26/9
Training, months				
In anesthesia	13 ± 2	21 ± 2	34 ± 4	
Overall	13 ± 2	27 ± 13	37 ± 10	
Anesthesia certification	NA	NA	NA	
ABA certified/board eligible				32/2
Foreign				1
Subspecialty	NA	NA	NA	17
Other specialty	0	3	2	3

Results are mean ± SD.

ABA = American Board of Anesthesiology; Adv = advanced; NA = not applicable.

ted to the same faculty who rank-ordered them based on their perceived importance in practice. The 12 events that were most frequently selected by faculty included (in alphabetical order) (1) acute hemorrhage, (2) anaphylaxis, (3) blocked endotracheal tube, (4) bronchospasm, (5) hyperkalemia, (6) loss of pipeline oxygen, (7) malignant hyperthermia (MH), (8) myocardial ischemia, (9) pneumothorax, (10) right main stem intubation, (11) total spinal, and (12) ventricular tachycardia. All of the selected events were topics included in the American Society of Anesthesiologists–American Board of Anesthesiology Content Outline.

The 12 intraoperative events were then developed into a set of scripted simulated scenarios. Each event was designed to require a rapid diagnosis and treatment. The goal of scenario development was to design an exercise that evaluated diagnostic and therapeutic skills required to manage a critical intraoperative event. The events were programmed in a manner that would compel a practitioner to rapidly diagnose and intervene in a short time period. The clinical practice domain included skills involving pattern recognition (*e.g.*, for main stem intubation [absent chest movement on left, decreased breath sounds, decreased oxygen saturation, increased airway pressure and endotracheal tube at 25 cm]), diagnostic skills such as inspection for chest movement, auscultation of simulated breath and heart sounds and palpation of pulses, algorithmic responses, psychomotor skills (*e.g.*, airway management, tracheal intubation, defibrillation), and situational awareness (progressive "clinical" deterioration in myocardial ischemia, MH, total spinal, anaphylaxis).

After receiving institutional review board (Washington University School of Medicine, St. Louis, Missouri) approval for the protocol, we obtained informed written consent from 64 residents. The residents, who were all from a single training program, were divided into three groups based on their training. The new CA-1 residents (n = 31, 13.1 ± 1.1 months of anesthesia residency) had completed an internship and were in their first month of anesthesia training. The advanced CA-1 residents (n = 12, 20.5 ± 2.1 months of anesthesia residency) had training experiences in intraoperative anesthesia care but minimal, if any, experiences in subspecialty anesthesia practice. The CA-2/CA-3 residents (n = 21, 34.3 ± 4.4 months of anesthesia residency) had additional anesthesia training in all subspecialties, including obstetric, pediatric, pain, intensive care, and cardiovascular anesthesia (table 2).

The 35 anesthesiologists responded to a solicitation letter sent by the investigators. Two hundred fifty letters were mailed to anesthesiologists in selected postal zip codes in the St. Louis metropolitan area. The anesthesiologists received an honorarium to compensate them for their time and travel during the study period.

This project was conducted in our laboratory that contains a MEDSIM-EAGLE® (Fort Lauderdale, FL) simulator. A SUN® workstation (Sun Microsystems, Santa Clara, CA) served as the command computer that drove an additional mannequin computer as well as the mechanical interfaces. The anesthesiologists who participated had limited experience working in a simulated environment. For this reason, a standardized orientation was developed to ensure that all participants received similar orientation to the mannequin, the anesthesia machine, and monitoring equipment. This orientation included reviewing a 17-min videotape that detailed the features of the mannequin, the function of the anesthesia machine, the monitors and equipment, the location of supplies, and how procedures are performed on a simulator. After this orientation, all of the participants conducted an anesthesia induction to review the anesthesia machine, monitors, and equipment and to practice airway management for the mannequin. During this orientation, participants had an opportunity to ask questions regarding the mannequin and equipment before they actually participated in the scenarios. This orientation period continued until the participants indicated that they were ready to begin the performance assessment.

To collect and analyze psychometric data on a wider range of events, 12 exercises were designed. The ad-

vanced CA-1 and CA-2/CA-3 and the experienced anesthesiologists each managed 8 scenarios. Each set of 8 was randomly selected and then presented in a random order. The result was an intensive participant assessment period of 75–90 min. The same distribution of scenario assignments and order were used in each group so as to balance the evaluation across groups. Thus, proportionate numbers of individuals in each group were exposed to the same assessment content in identical order. The 31 new CA-1 residents also managed 8 cases, but each individual encountered only 3 of 6 intraoperative scenarios (anaphylaxis, myocardial ischemia, right main stem intubation, blocked endotracheal tube, pneumothorax, and hemorrhage). These 6 intraoperative events were selected by two of the investigators (D.J.M., J.F.K.). These selected scenarios had either been used in previous studies or extensively tested in our simulation laboratory.[5,11] The CA-1 residents managed an additional 5 preoperative and postoperative scenarios, but only scores for the overlapping intraoperative scenarios, noted above, were used in the comparative analysis and reliability estimation.

The participants were instructed to (1) perform all diagnostic or therapeutic actions; (2) expect that all actions (medication administration, *etc.*) occur in "real" time; and (3) verbalize diagnosis and treatment during the scenario so that scores for some elements could be accurately coded and, if indicated, simulation staff could respond to any requests by the participant (*e.g.*, obtaining laboratory studies or x-ray results). The participants were allowed to ask questions about the patient's condition but were instructed to obtain necessary information from the record or an evaluation of the mannequin. Many of the responses to participants' questions were predetermined by a scripted response developed by the research team. The participants managed each 5-min scenario alone. After each exercise, the supervising faculty member discussed the diagnosis and management with the residents. The practicing anesthesiologists did not receive feedback during the simulation session.

Each participant was expected to diagnose and manage the simulated patient within the 5-min period. Between scenarios, the resident or anesthesiologist reviewed the next preoperative assessment and accompanying anesthesia record in an adjacent conference room. During this time, the research team prepared the mannequin and simulation laboratory for the next scenario.

The anesthetic record provided information about the previous and current anesthetic management as well as the mannequin's cardiovascular and respiratory signs. At the start of each scenario, the life-sized mannequin's simulated clinical signs matched the vital signs recorded on the anesthetic record. In some scenarios, the intraoperative crisis developed over a period of 1–2 min after the participant entered the simulation laboratory (bronchospasm, anaphylaxis, ventricular tachycardia, total spinal, acute hemorrhage, and loss of pipeline oxygen). For the myocardial ischemia, blocked tube, MH, hyperkalemia, and right main stem intubation scenarios, cardiorespiratory changes existed at the start of the scenario (table 1).

An audio–video recording of each simulation exercise integrated the video from two cameras and audio from ceiling microphones. An additional quadrant of the four quadrant audio–video record included the mannequin's monitor screen that provided the hemodynamic variables and anesthetic agent concentrations. The final quadrant included participant and scenario identifying information.

The faculty investigators developed scoring measures for each scenario; these essential items included three to six key diagnostic or therapeutic actions (table 1). For the advanced CA-1 residents, CA-2/CA-3 residents, and experienced anesthesiologists, two raters blinded to the identity and prior training of the participants scored the scenarios. Before beginning the scoring, the raters were trained by one of the investigators (J.F.K.). Raters reviewed the goals and criteria for each key action of the scenarios and scored a number of pilot performances to assure consistency with predetermined operational definitions. These two raters then independently scored the time taken for participants to perform each key action. For key action items where the two raters disagreed or had differences of more than 30 s regarding when the action occurred, a third rater reviewed the entire scenario and provided independent key action scores. This third rater's score was used to reach a majority decision about whether or not a participant had completed a key action during the scenario. The scenarios used for the new CA-1 residents were scored by only one of the primary raters because previous studies indicated that these selected scenarios could be reliably scored using a single rating.[5,11]

The key action score was used rather than the time to key action because our previous research suggested that, for brief scenarios such as the ones studied, simply performing the task within the short time period provided a valid and reliable score.[11] For the analysis, the time to key action scores were simply converted to key action scores; if the task was completed in the allotted time, the participant received credit, otherwise not. For each scenario, the number of key actions accomplished could range from three to six. In addition, a percent score, based on the number of key actions credited out of the possible key actions, was computed. A total score for each participant was computed as the average of percentage encounter scores across the total number of scored scenarios (n = 8 exercises for advanced CA-1 residents, CA-2/CA-3 residents, and experienced anesthesiologists; n = 3 exercises for new CA-1 residents).

Several analyses were performed to investigate participant scores from individual scenarios as well as overall

scores. To investigate potential performance differences, by group, a two-way analysis of variance was used. The dependent variable was the key action score. The independent variables were scenario and group (new CA-1, advanced CA-1, CA-2/CA-3, and experienced anesthesiologists). For any significant interactions and main effects, one-way analysis of variance and Scheffé multiple-comparison procedure were used in follow-up analyses. Scenario discrimination statistics (correlation between a participant's mean scenario score and his or her overall score on the allotted scenarios) were used to investigate whether the score obtained on each individual scenario correlated with the overall score. Generalizability theory was used to determine the reliability of the participant scores and to identify the facets (*e.g.*, rater, scenario) that best explained the variability in individual performance.[21,22] The estimated variance components can be used to derive the commonly reported interrater and interscenario reliability statistics. The interrater reliability in this study was the correlation between the two primary rater scores.

Results

The 99 participants included 31 new CA-1 residents, 12 advanced CA-1 residents, 21 CA-2/CA-3 residents, and 35 experienced anesthesiologists (table 2). Three of the advanced CA-1 residents and 2 of the CA-2/CA-3 residents had completed training in other specialties, but these residents' scores were analyzed with their respective anesthesia training year. The 35 anesthesiologists were from a variety of practice locations around the St. Louis area. Two primary raters scored a total of 544 simulation exercises (68 participants $\times$ 8 scenarios) for the PGY-2 and CA-2/CA-3 residents and practicing anesthesiologists. One of these raters scored the 93 simulation exercises for the PGY-1 residents (31 $\times$ 3 scenarios).

For the 68 participants who each managed 8 of the 12 scenarios, there were no statistically significant differences in the frequency of each scenario assigned by group (chi-square$_{22}$ = 1.22, P = 1.00), suggesting that the random number assignment to determine scenario mix resulted in an equally difficult set of tasks, on average, for each participant (table 3).

The two-way analysis of variance yielded both a significant scenario by group interaction (F_{27} = 1.71, $P <$ 0.02) and a significant scenario main effect (F_{11} = 36.7, $P < 0.01$). The significant scenario by group interaction indicates that the physician cohorts did not perform equally well on all scenarios. There was also a significant group main effect (F_3 = 30.1, $P < 0.01$), indicating that, averaged over all the scenarios, there were significant differences in performance between groups. Based on the follow-up analyses, there were no significant differences in average scores obtained by the advanced CA-1 resident, CA-2/CA-3 resident, and anesthesiologist groups. The new CA-1 residents had significantly lower scores than any of the more experienced anesthesiology groups. More specifically, on the anaphylaxis, myocardial ischemia, main stem intubation, tension pneumothorax, blocked endotracheal tube, and acute hemorrhage scenarios, one or more of the advanced CA-1 resident, CA-2/CA-3 resident, and anesthesiologist groups scored better than the new CA-1 group (table 3 and fig. 1).

The advanced CA-1 resident, CA-2/CA-3 resident, and anesthesiologist groups were consistently able to perform all of the actions to manage many of the scenarios, including bronchospasm, right main stem intubation, tension pneumothorax, and pipeline oxygen failure. However, across all groups, both residents and anesthesiologists often did not recognize the MH, total spinal, and hyperkalemia scenarios and often performed few, if any, of the expected actions in the allotted time. Despite the similarities in overall group performance, the variability among individuals in the groups was large for several scenarios.

The scenario discriminations (the correlation between individual scenario scores and total scenario scores) were positive, confirming that a participant's score in one scenario usually generalized to their cumulative performance on all scenarios (table 3). As expected, participants who received a higher score on a scenario were more likely to receive a greater overall score, and conversely, failure to recognize any scenario correlated with lower cumulative scores. However, the correlations for individual scenarios were at best only moderately positive, indicating that individual scenario scores could not be used to predict the overall score. This result implies that the overall assessment covered a broad range of content and multiple scenarios were needed to evaluate skill. Some scenarios were more likely to predict a participant's overall score; (*e.g.*, hyperkalemia, myocardial ischemia, blocked endotracheal tube, ventricular tachycardia, acute hemorrhage), whereas others (*e.g.*, total spinal, discrimination = 0.07) did not correlate at all with the overall score.

A detailed investigation of participants' completed actions in each scenario provided information about whether changes were needed in the design or content of the exercise as well as to indicate which practice domains might require more attention and emphasis during and after training. Most participants did not recognize the MH event, perhaps because temperature remained unchanged during the 5-min period (table 3). This scenario was designed to represent an early onset of MH with increasing tachycardia (heart rate 140 beats/min at onset, 180 beats/min at 5 min) associated with increasing premature ventricular contractions, hypertension, and increasing expired carbon dioxide levels. The myocardial ischemia exercise differentiated training and experience more effectively than the other scenarios.

Table 3. Mean Performance (% Key Actions) on Simulation Exercises and Scenario Discrimination

Scenario	No. of Total Participants in Each Group	% Key Actions	Minimum Score	Maximum Score	Discrimination
Bronchospasm					
Total	44	96 ± 11	50	100	0.26
CA-1 (new)	0				
CA-1 (adv)	8	100 ± 0	100	100	
CA-2/CA-3	13	96 ± 14	50	100	
Anesthesiologists	23	95 ± 11	75	100	
Anaphylaxis					
Total	55	67 ± 23	20	100	0.33
CA-1 (new)	14	47 ± 10	40	60	
CA-1 (adv)	7	60 ± 16	40	80	
CA-2/CA-3	13	75 ± 19*	40	100	
Anesthesiologists	21	76 ± 24*	20	100	
Ventricular tachycardia					
Total	43	71 ± 18	40	100	0.51
CA-1 (new)	0				
CA-1 (adv)	8	66 ± 15	40	80	
CA-2/CA-3	13	69 ± 13	60	100	
Anesthesiologists	22	75 ± 21	40	100	
Myocardial ischemia					
Total	67	51 ± 24	17	100	0.59
CA-1 (new)	17	33 ± 18	17	67	
CA-1 (adv)	8	38 ± 23	17	67	
CA-2/CA-3	16	56 ± 19*	17	83	
Anesthesiologists	26	62 ± 22*	17	100	
Right main stem intubation					
Total	62	85.5 ± 22.4	0	100	0.53
CA-1 (new)	14	66.1 ± 33.4	0	100	
CA-1 (adv)	8	96.9 ± 8.8*	75	100	
CA-2/CA-3	16	89.1 ± 15.7*	50	100	
Anesthesiologists	24	90.6 ± 14.4*	50	100	
Tension pneumothorax					
Total	59	84.4 ± 19.7	40	100	0.38
CA-1 (new)	14	74.3 ± 18.3	60	100	
CA-1 (adv)	8	100 ± 0*	100	100	
CA-2/CA-3	14	87.1 ± 20.2	40	100	
Anesthesiologists	23	83.5 ± 20.6	40	100	
Malignant hyperthermia					
Total	45	22.6 ± 34.0	0	100	0.36
CA-1 (new)	0				
CA-1 (adv)	9	13.3 ± 29.8	0	66.7	
CA-2/CA-3	14	38.9 ± 38.9	0	100	
Anesthesiologists	22	19.6 ± 33.5	0	100	
Blocked endotracheal tube					
Total	66	58.5 ± 26.8	20	100	0.55
CA-1 (new)	17	36.5 ± 16.2	20	80	
CA-1 (adv)	9	80.0 ± 28.3*	40	100	
CA-2/CA-3	14	68.6 ± 25.7*	20	100	
Anesthesiologists	26	60.8 ± 23.7*	20	100	
Total spinal					
Total	47	49.6 ± 20.0	16.7	100	0.07
CA-1 (new)	0				
CA-1 (adv)	9	40.7 ± 18.8	16.7	66.7	
CA-2/CA-3	14	46.4 ± 19.8	16.7	83.3	
Anesthesiologists	24	55.1 ± 19.7	16.7	100	
Loss of pipeline oxygen					
Total	42	90.5 ± 18.5	33.3	100	0.31
CA-1 (new)	0				
CA-1 (adv)	7	90.5 ± 16.3	66.7	100	
CA-2/CA-3	12	91.7 ± 15.1	66.7	100	
Anesthesiologists	23	89.9 ± 21.1	33.3	100	
Hyperkalemia					
Total	43	29.5 ± 40.0	0	100	0.61
CA-1 (new)	0				
CA-1 (adv)	7	42.9 ± 46.0	0	100	
CA-2/CA-3	13	43.6 ± 49.8	0	100	
Anesthesiologists	23	17.4 ± 28.2	0	100	

(continued)

Table 3. Continued

Scenario	No. of Total Participants in Each Group	% Key Actions	Minimum Score	Maximum Score	Discrimination
Acute hemorrhage					
Total	47	63.3 ± 38.3	0	100	0.67
CA-1 (new)	17	39.7 ± 36.5	0	100	
CA-1 (adv)	8	78.1 ± 41.1	0	100	
CA-2/CA-3	16	79.7 ± 31.9*	25	100	
Anesthesiologists	23	64.1 ± 36.0	0	100	

Results are mean ± SD.

* Group performance that is significantly different from that of CA-1 (new) residents ($P < 0.025$).

Adv = advanced.

Senior residents and anesthesiologists had significantly higher scores than CA-1 and advanced CA-1 residents. The main difference was that the more experienced residents and anesthesiologists were more likely to confirm the diagnosis of myocardial ischemia using multiple leads or ST-T wave segment analysis (50% of CA-2/CA-3 residents and practicing anesthesiologists compared with 12% of CA-1 and 13% of the advanced CA-1 residents). In addition, 25% of the senior residents and 31% of the experienced anesthesiologists (31%) effectively titrated β-blockers to reduce heart rate, whereas none of the CA-1 or advanced CA-1 residents did so during this scenario.

The variance components were estimated using generalizability theory. This analysis partitions the sources of score variability into various components and provides a means to establish the reliability of the assessment scores. Even though each of the participants only encountered 8 of the 12 simulation exercises (or 3 of 6 exercises for the CA-1 resident group), an analysis of reproducibility of the scores for a randomly selected 8-scenario evaluation could still be performed. Based on the final scenario scores, the generalizability coefficient was 0.56, indicating that an 8-scenario assessment affords a modestly reliable overall measure of a participant's skill in the management of acute care scenarios. For our assessment, 41% of the variance in individual encounter scores could be attributed to the choice of scenarios. The scenarios vary in terms of average difficulty (table 3). An additional major cause of scoring variance could be ascribed to the scenarios content (51%). The magnitude of this variance component indicates that participant performance can vary considerably from scenario to scenario. For this reason, increasing the number of scenarios as well as selecting those scenarios that more effectively discriminate performance could augment the reliability of a participant's score. For example, if each participant were evaluated across all of the 12 scenarios, instead of 8, the estimated generalizability coefficient would be 0.66.

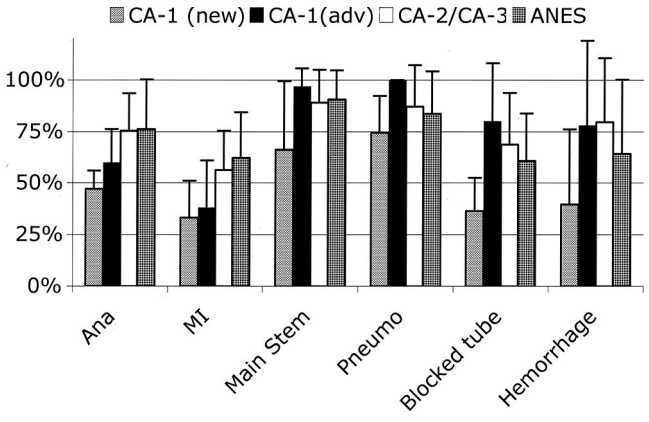

Fig. 1. Mean percentage and SDs of key actions achieved on the six exercises by the CA-1 (new), CA-1 (adv), and CA-2/CA-3 resident and anesthesiologist (ANES) groups. The six scenarios include anaphylaxis (Ana), main stem intubation (Main Stem), myocardial ischemia (MI), pneumothorax (Pneumo), blocked endotracheal tube (Blocked Tube), and acute hemorrhage (Hemorrhage). The CA-1 (adv) residents, CA-2/CA-3 residents, and anesthesiologists achieved higher scores than the CA-1 residents ($P < 0.05$). Adv = advanced.

Based on the two primary raters' scores (percentage of key actions credited) in the advanced CA-1 resident, CA-2/CA-3 resident, and anesthesiologist cohorts, the interrater reliability for the 544 encounters was $r = 0.91$. By scenario, interrater reliability ranged from a low of $r = 0.59$ on right main stem intubation to a high of $r = 0.97$ on hyperkalemia. An analysis of rater differences in scoring each exercise provides information about how different evaluators interpret case scoring rubrics and whether they score the endpoints of each action similarly. For each simulation encounter, the two primary raters scored three to six key actions. In 75 of the 544 encounters, the two primary raters who provided initial scores disagreed about when (or whether) a participant had performed one of the key actions. The number of discrepancies for each scenario ranged from 1 (0.8%) in the 132 ratings for the MH scenario (44 participants with three scoring actions) to 11 (7%) in the 141 ratings for loss of pipeline oxygen (47 participants with three scoring actions).

Discussion

There are no previous studies available that provide a measure of intraoperative management skills of practicing anesthesiologists and compare them with skills of

anesthesia residents. In this simulation-based performance assessment of residents and practicing anesthesiologists, advanced CA-1 residents, CA-2/CA-3 residents, and anesthesiologists received higher scores than CA-1 residents. This suggests that the simulation-based scenarios are a valid assessment, at least of the more basic skills acquired early in anesthesia training. The 12 intraoperative events covered a broad range of acute intraoperative skills and included a number of events that the majority of participants rapidly recognized and treated, as well as scenarios that were unrecognized by many participants. Using a set of brief exercises, the overall score provided a moderately reliable measure of a participant's ability in managing simulated intraoperative events.

The scenarios provided a valid method to discriminate between anesthesia trainees with limited training (CA-1) and those with greater anesthesia practice experience and training. The absence of differences among the more experienced anesthesiologists and the other resident cohorts (CA-2/CA-3 and advanced CA-1) may indicate that the scenarios may have been either too easy or too difficult to effectively discern skill differences among these three groups. Alternatively, these results could suggest that the more experienced resident groups and anesthesiologists did not differ in their ability to recognize and manage simulated critical events. This result would agree with many studies that have assessed the skills of experienced physicians and residents that have found experience in practice does not result in improved performance.[23] Although this study does not directly measure practice skills, the infrequent nature of these events may be a reason that experienced anesthesiologists performed similarly to the residents. There may be a variety of additional explanations for these results. First, the anesthesiologist and resident samples were from a limited geographic area and may not be representative in terms of ability. The anesthesiologists were selected based on their response to a mail solicitation and a financial incentive. The anesthesiologists' motivation to perform optimally during the simulation session may have differed from the resident cohort. In addition, the anesthesiologists' familiarity with the electromechanical mannequin, despite the orientation provided before the study, may have been less than residents, who may have benefited from their previous exposure to simulation-based training. Finally, unlike the resident cohorts, the anesthesiologists did not receive any feedback between scenarios. For the advanced CA-1 residents, the similar training environment may be a reason that these residents with less training were able to achieve scores that were similar to the CA-2/CA-3 residents. A larger study that included a more diverse group of participants (*i.e.*, different training programs) would help to establish validity by determining whether performance differs among residents at different levels of training and experienced anesthesiologists in different geographic areas.

One of the goals of the study was to collect data about the reliability and validity of a broad sample of simulated conditions. Twelve scenarios were developed and tested to obtain psychometric data from a range of simulated conditions. The raters scored the scenarios consistently and agreed on most scenario key action items (good interrater reliability). Despite this high level of agreement overall, we asked an independent expert to rescore performances when the two primary raters disagreed. The source of rating differences often yields information about flaws in the structure of the assessment, suggests modifications in scoring criteria, and indicates whether additional rater training might further enhance the reliability of the assessment for future participants. The scenario associated with the most frequent need for expert arbitration was loss of pipeline oxygen with an interrater reliability of 0.46. Raters often disagreed on whether the participant made the diagnosis. Additional rater training or modifications in the scenario would be helpful to improve interrater reliability in future performance assessments.

The 12 scenarios, as selected by the faculty, covered a broad spectrum of acute events. The majority of more experienced participants readily managed many of the easier scenarios. These scenarios were designed by faculty to reflect valid content, but these exercises were less able to provide discriminant validity that might be useful to differentiate the more advanced skills expected in specialty practice. These simpler, straightforward scenarios such as bronchospasm, main stem intubation, and loss of pipeline oxygen were managed effectively by most participants and did not contribute appreciably to the reliability of the overall score. For this reason, these exercises may be more useful to determine minimum performance expectations rather than to rank-order advanced skills. Even though the hyperkalemia and MH scenarios were difficult, these exercises more effectively discriminated among participants. In contrast, scenarios that did not discriminate along the ability continuum, such as the total spinal scenario, may be flawed as a performance measure and should be revised or excluded from future use in such assessments.

In general, participants' scores fluctuated as a function of scenario content and associated difficulty. In addition, there was a wide range of performances by individuals in each group. This variation in scores obtained by individuals in each group merits further investigation to determine how the skills to manage simulated intraoperative events are acquired and maintained. Some scenarios were much more challenging than others, regardless of participant experience and training. More important, and similar to the results reported from other multistation performance assessments, an individual's score was dependent, at least to some extent, on the content of the

simulated exercise.[5,11,20] A participant's performance on one scenario may not be a consistent predictor of performance on other exercises.[5,11] For this reason, the reliability of the overall assessment (0.56) was less than what might be considered adequate for a licensure or specialty certification examination. While modifications to the scenarios, associated scenario scoring, or even rater training, could improve the confidence of the measurement, more performance samples (scenarios) would be the most effective method to improve the reliability of the performance assessment.[5,11,12,20,22] Once a participant's score can be objectively and reproducibly measured and the validity of the exercises are established, follow-up investigations could be designed to set performance standards and provide a competency-based assessment.[24,25]

In summary, a multiple-scenario assessment provided a reliable measure of anesthesia residents and practicing anesthesiologist performance in managing simulated intraoperative events. The differences between CA-1 residents and more experienced residents and anesthesiologists support the validity of the assessment as a measure of basic skills in managing intraoperative events. Additional research is needed to determine how simulation-based assessments predict performance in clinical settings and how best these evaluations could be used to measure progress of training during residency.

References

1. Silber JH, Kennedy SK, Even-Shoshan O, Chen W, Koziol LF, Showan AM, Longnecker DE: Anesthesiologist direction and patient outcomes. ANESTHESIOLOGY 2000; 93:152–63

2. Silber JH, Kennedy SK, Even-Shoshan O, Chen W, Koziol LF, Showan AM, Longnecker DE: Anesthesiologist board certification and patient outcomes. ANESTHESIOLOGY 2002; 96:1044–52

3. Gaba DM: What makes a good anesthesiologist. ANESTHESIOLOGY 2004; 101: 1061–3

4. Barsuk D, Ziv A, Lin G, Blumenfeld A, Rubin O, Keidan I, Munz Y, Berkenstadt H: Using advanced simulation for recognition and correction of gaps in airway and breathing management skills in prehospital trauma care. Anesth Analg 2005; 100:803–9

5. Boulet JR, Murray DJ, Kras J, Woodhouse J, McAllister JD, Ziv A: Assessing the acute-care skills of medical students and recent graduates: Reliability and validity of patient simulator scores. ANESTHESIOLOGY 2003; 99:1270–80

6. 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

7. Issenberg SB, McGaghie WC, Petrusa ER, Gordon DL, Scalese RJ: Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. Med Teach 2005; 27:10–28

8. Jha AK, Duncan BW, Bates DW: Simulator-based Training and Patient Safety, Making Health Care Safer: A Critical Analysis of Patient Safety Practices (Evidence Report/Technology Assessment No. 43, AHRQ Publication 01-E058). Edited by Shojania KG, Duncan BW, McDonald KM, Wachter RM. Rockville, Maryland, Agency for Healthcare Research and Quality, 2001, pp 510–7

9. Morgan PJ, Cleave-Hogg D, McIlroy J, Devitt JM: Simulation technology: A comparison of experiential and visual learning for undergraduate medical students. ANESTHESIOLOGY 2002; 96:10–6

10. Murray DJ, Boulet J, Ziv A, Kras J, McAllister JD, Woodhouse J: An acute care skills evaluation for graduating medical students: a pilot study using clinical simulation. Med Educ 2002; 36:833–41

11. Murray DJ, Boulet J, Kras JF, Woodhouse JA, Cox T, McAllister JD: Acute care skills in anesthesia practice: A simulation-based resident performance assessment. ANESTHESIOLOGY 2004; 101:1085–94

12. Murray DJ, Boulet JR, Kras JD, Cox T, McAllister JD: A simulation-based acute skills performance assessment for anesthesia training. Anesth Analg 2005; 101:1127–34

13. Murray DJ: Clinical simulation: Measuring the efficacy of training. Curr Opin Anesth 2005; 18:645–8

14. Murray DJ: Clinical skills in acute care: A role for simulation training (editorial). Crit Care Med 2006; 34:252–3

15. Olympio MA, Whelan R, Ford RPA, Saunders ICM: Failure of simulation training to change residents' management of oesophageal intubation. Br J Anaesth 2003; 91:312–8

16. Schwid HA, O'Donnell D: Anesthesiologists' management of simulated critical incidents. ANESTHESIOLOGY 1992; 76:495–501

17. 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 multi-institutional study. ANESTHESIOLOGY 2002; 97:1434–44

18. Steadman RH, Coates WC, Huang Y-M, Matevosian R, Larmon BR, McCullough L, Ariel D: Simulation-Based Training is Superior to Problem-Based Learning for the Acquisition of Acute Care Management Skills. Crit Care Med 2006; 34:289–93

19. Margolis MJ, De Champlain AF, Klass DJ: Setting examination-level standards for a performance-based assessment of physicians' clinical skills. Acad Med 1998; 73 (suppl):S114–6

20. Rothman AI, Blackmore D, Dauphinee WD, Reznick R: The use of global ratings in OSCE station scores. Adv Health Sci Educ 1997; 1:215–9

21. Brennan RL, Gao X: Variability of estimated variance components and related statistics in a performance assessment . Appl Meas Educ 2001; 14:9–12

22. Boulet J: Generalizability theory: Basics, Encyclopedia of Statistics in Behavioral Science. Edited by Everitt BS, Howell DC. Chichester, England, John Wiley and Sons, 2005, pp 704–11

23. Choudhry NK, Fletcher RH, Soumerai SB: Systemic review: The relationship between clinical experience and quality of health care. Ann Intern Med 2005; 142:260–73

24. Boulet JR, De Champlain AF, McKinley DW: Setting defensible performance standards on OSCEs and standardized patient examinations. Med Teach 2003; 25:245–9

25. McKinley DW, Boulet JR, Hambleton RK: A work centered approach for setting passing scores on performance-based assessments. Eval Health Prof 2005; 28:349–69