Anesthesiology 2004; 101:1084–95

Acute Care Skills in Anesthesia Practice

A Simulation-based Resident Performance Assessment

David J. Murray, M.D., John R. Boulet, Ph.D.,† Joseph F. Kras, M.D.,‡ Julie A. Woodhouse, B.S.N.,§ Thomas Cox, M.D.,‡ John D. McAllister, M.D.‡*

Background: **A recurring initiative in graduate education is to find more effective methods to assess specialists' skills. Life-sized simulators could be used to assess the more complex skills expected in specialty practice if a curriculum of relevant exercises were developed that could be simply and reliably scored. The purpose of this study was to develop simulation exercises and associated scoring methods and determine whether these scenarios could be used to evaluate acute anesthesia care skills.**

Methods: **Twenty-eight residents (12 junior and 16 senior) managed three intraoperative and three postoperative simulation exercises. Trainees were required to make a diagnosis and intervention in a simulation encounter designed to recreate an acute perioperative complication. The videotaped performances were scored by six raters. Three raters used a checklist scoring system. Three faculty raters measured when trainees performed three key diagnostic or therapeutic actions during each 5-min scenario. These faculty also provided a global score using a 10-cm line with scores from 0 (unsatisfactory) to 10 (outstanding). The scenarios included (1) intraoperative myocardial ischemia, (2) postoperative anaphylaxis, (3) intraoperative pneumothorax, (4) postoperative cerebral hemorrhage with intracranial hypertension, (5) intraoperative ventricular tachycardia, and (6) postoperative respiratory failure.**

Results: **The high correlation among all of the scoring systems and small variance among raters' scores indicated that all of the scoring systems measured similar performance domains. Scenarios varied in their overall difficulty. Even though trainees who performed well on one exercise were likely to perform well in subsequent scenarios, the authors found that there were considerable differences in case difficulty.**

Conclusion: **This study suggests that simulation can be used to measure more complex skills expected in specialty training. Similar to other studies that assess a broad content domain, multiple encounters are needed to estimate skill effectively and accurately.**

This article is accompanied by an Editorial View. Please see: Gaba DM: What makes a "good" anesthesiologist? ANESTHESIOLOGY 2004; 101:1061–3.

* Director, Washington University Clinical Simulation Center, Professor, Department of Anesthesiology, Washington University School of Medicine, ‡ Associate Professor, Department of Anesthesiology, Washington University School of Medicine, § Administrator, Washington University Clinical Simulation Center, St. Louis, Missouri. † Vice President, Education Council for Foreign Medical Graduates, Philadelphia, Pennsylvania.

Received from the Department of Anesthesiology, Washington University School of Medicine, St. Louis, Missouri. Submitted for publication April 28, 2004. Accepted for publication July 29, 2004. Supported by a Foundation for Anesthesia Education and Research Education Grant (to Dr. Kras, Principal Investigator), Park Ridge, Illinois; and the Department of Anesthesiology, Washington University, St. Louis, Missouri.

Address reprint requests to Dr. Murray: Washington University Clinical Simulation Center, Professor, Department of Anesthesiology, Washington University School of Medicine, Box 8054, 660 South Euclid, St. Louis, Missouri 63110. Address electronic mail to: murrayd@notes.wustl.edu. Individual article reprints may be purchased through the Journal Web site, www.anesthesiology.org.

‖ Accreditation Council for Graduate Medical Education: ACGME Outcome Project, 2004. Available at: www.acgme.org. Accessed September 27, 2004.

THE advent of life-sized simulators affords an opportunity for objective assessment without accompanying endangerment to patient well-being. One of the key advantages of training and evaluating physicians with life-sized mannequins is that acute diagnosis can be integrated with the ongoing demands of managing and stabilizing a changing medical or surgical condition. Trainees develop skills and practice different management strategies in a standardized setting, without endangering the health or life of human patients.[1–4] Currently, numerous medical schools and graduate training programs provide formative educational programs that use integrated (mannequin-based) simulators for teaching and assessment.[5,6] To initiate any evaluation method relevant to anesthesia practice, a number of steps in test development process are necessary: (1) ensuring the content of the evaluation is relevant and important skills are being measured, (2) modeling settings that have high fidelity for actual clinical care, (3) effectively isolating the skill of the provider from the multiple external factors that influence patient outcome, (4) establishing a scoring method that provides reliable ability estimates, and (5) providing evidence to support the validity of the simulation scores.

The need to assure that physicians in training as well as in practice develop and maintain the skills expected of a consultant remain a high priority in graduate and continuing medical education. The Accreditation Council for Graduate Medical Education has implemented phase 2 of an initiative that requests that training programs assess a resident's competence in six separate domains of medical practice.‖ Clearly, domains that include communication, technical skills, or integrating complex diagnostic and therapeutic skills cannot be measured well using traditional paper-and-pencil examinations. As a result, performance-based assessments, including evaluations using standardized patients, have been developed and implemented to measure some of the more basic skills expected in practice. For advanced skills, especially those involving patient management, clinical case simulations and integrated simulators have been incorporated in various programs.[7] A set of acute care simulation exercises could be used to evaluate clinical competence in relevant (*e.g.*, crisis) clinical situations.[8]

Studies indicate that properly constructed simulation exercises not only have a high resemblance to the clinical environment, but also improve trainee skill and teamwork.[9–16] Historically, simulation studies have generally used a single prolonged scenario that poses multiple additive challenges.[5,10,12,17–20] In some studies,[21,22]

sets of exercises were developed that are similar in administrative design and scoring to the standardized patient assessments currently used to evaluate graduating physicians for certification and licensure decisions.[23,24] By applying a similar methodology to advanced training situations, it may be possible to produce a set of reproducible exercises that could effectively measure the clinical skills and decision-making processes expected of anesthesia consultants working in high-acuity settings.

One of the potential limitations of evaluating physicians *via* simulation is that expert raters are usually required to review performance.[17,18,25,26] The scoring task may be particularly arduous if comprehensive checklists are used to score the simulation exercises. Similarly, while global scoring has been proposed as a method to score performance-based assessments,[27,28] including mannequin-based simulations, recruiting and training qualified raters can be difficult. Overall, there is little agreement on which, if any, scoring method is most appropriate for quantifying performance on simulation exercises. Checklists, although easy to generate and score, may not capture degrees of expertise. Furthermore, as documented in the standardized patient literature, they may reward thoroughness as opposed to advanced skill.[29,30] For acute care scenarios typically modeled using mannequins, there is little doubt that some actions are more important than others. Although weighting these actions more heavily in the scoring rubrics may alleviate the problem, experts must reach consensus on the specific checklist items and weights. Holistic scoring, where experts provide global ratings of overall performance, seems apropos. Experts can capture timing (*e.g.*, diagnosing quickly) and sequencing (*e.g.*, establishing an airway before providing fluids) in their ratings. However, without extensive training and calibration of the raters, holistic ratings can be biased, resulting in error-prone estimates of ability. As a result, additional psychometric studies focusing on mannequin-based scoring systems are certainly warranted.

The purpose of this study was to develop and evaluate a set of scripted simulated scenarios that isolate and measure an individual physician's clinical skills in anesthesia acute care management. Assessment scores from a sample of residents were used to evaluate the psychometric properties of various scoring systems and to contrast the performance of more- and less-experienced practitioners.

Materials and Methods

Simulation Laboratory

This project was conducted in our simulation laboratory that contains a sensorized life-sized electromechanical patient mannequin developed by MEDSIM-EAGLE® (MEDSIM Eagle, Fort Lauderdale, FL). The physiologic and pharmacologic models that determine cardiac and respiratory responses are used to simulate acute medical conditions such as anaphylaxis, hemorrhage, pneumothorax, and cardiac tamponade. A SUN workstation (Sun Microsystems, Inc., Santa Clara, CA) serves as the command computer that drives an additional mannequin computer as well as the mechanical interfaces. The mannequin features breath sounds and heart sounds, a full set of pulses, motors to operate chest excursions during spontaneous ventilation, carbon dioxide exhalation, and standard hookups for blood pressure electrocardiography and invasive hemodynamic monitoring, such as pulmonary artery and central venous pressure monitoring. The mannequin's voice (when required) is a speaker mounted in the occipital region and controlled by personnel in a separate control room. The simulator offers simple as well as advanced programming actions to create and then save a unique scenario for repeated evaluation of performances. For example, the myocardial ischemia event was programmed using a sliding scale to determine the degree of ST-segment changes; the amount of myocardial irritability, such as ventricular arrhythmias; the speed of onset of the changes; and the severity of myocardial depression. A variety of additional features (*e.g.*, heart rate, lung compliance, vascular resistance) can be manipulated independently to create a unique but reproducible event that effectively tests the skill level of the provider.

Scenario Selection and Development

The first step in scenario development was to create a list of perioperative events that could be simulated in our laboratory and that a resident should be able to recognize and manage after completion of training. The list was developed by the simulation faculty and staff (D.J.M., J.F.K., J.A.W.). Many of the scenarios considered for development had been used in previous simulation training exercises during the simulation centers preceding years of operation (1996). This list was then cross-referenced with the topic list in the American Board of Anesthesiologists Content Outline to assure that the selections were recognized as important content for an anesthesia consultant. A goal of the simulation training assessment was to create a set of exercises that could be completed during a 1-h evaluation session. The scenarios selected for development included (1) postoperative anaphylaxis, (2) intraoperative myocardial ischemia, (3) intraoperative atelectasis, (4) intraoperative ventricular tachycardia, (5) postoperative stroke with intracranial hypertension, and (6) postoperative respiratory failure. Each simulated scenario was developed to model a situation that required a rapid diagnosis and acute intervention in a brief encounter (5 min total time). Simulation faculty and staff reviewed each exercise to assure, based on the simulated findings, patient history, and associated anesthesia management, that each encounter was suffi-

ciently realistic. The events were designed with the goal of requiring a minimum intervention by the console operator. The simulator operator's main role was to make sure that all actions of the trainee were effectively entered and to allow the predetermined simulation algorithms to respond to the participant interventions. For example, the simulated blood pressure and heart rate responses to narcotics and β blockers during the myocardial ischemia scenario were entirely based on the mannequin's software algorithms.

After the set of six simulations were designed and developed by the simulation faculty, a review of the scenarios was conducted by the research team. The purpose of the review was to evaluate the simulated findings during the 5-min simulation exercise, to assess mannequin responses to diagnostic and therapeutic interventions, to determine whether the real-time correct responses could be accomplished during the exercise, and to enumerate the expected participant actions as well as to script any verbal or console responses required by the simulator operator. After this review, modifications were made to each exercise. Before initiating the study, a senior resident and faculty member pilot tested the six scenarios. Neither of the participants was involved in scenario design and development process. For the three intraoperative scenarios, the participant entered the simulation approximately 30 min after anesthesia induction. For the three postoperative scenarios, the trainee entered the simulation 20 min after the patient's arrival in the recovery room. Before starting each assessment scenario, the participants were supplied with the patient's history and physical findings, an anesthetic record, and, when indicated, a recovery room note.

Study Participants

The protocol was approved by the institutional review board (Washington University School of Medicine, St. Louis, Missouri). Informed written consent was obtained from each resident before participation in the simulation sessions. The residents also signed a confidentiality agreement before entering the simulation training. The residents (n = 28) were recruited from a single residency training program of 33 residents. The trainees were individually evaluated during a 2-month period close to the end of their respective training year (clinical anesthesia year 1 [CA-1], CA-2, or CA-3). All of the residents had experience with the simulator and the simulation center in small group training exercises during the initial months of training. There were 19 male participants and 10 female participants, including the one resident who piloted the scenarios. This resident who was recruited for the pilot testing was not included in the analysis. Eight of the residents graduated from foreign medical schools. Four residents did not participate primarily because of conflicts with clinical assignments (n = 2) or graduated from the program before the simulation training could be scheduled (n = 2).

The residents were each assigned to one of two groups (junior, n = 12; senior, n = 16) based on their previous clinical experience. The junior residents had completed 2 yr of postgraduate training. In addition to a year of internship, these residents had completed a further year of training in anesthesia. Nearly all of this training year was comprised of providing anesthesia for general, orthopedic, urologic, ambulatory, and pediatric patients. Most of the junior resident training experiences were providing care for elective operations to patients with American Society of Anesthesiologists (ASA) physical status I, II, or III. Although these residents did provide anesthesia for some emergency surgical procedures, the majority of these participants had limited clinical exposure to trauma care and patients with ASA physical status IV and V. The senior residents had completed the 2 yr of training and had either 1 or 2 additional years in more advanced anesthesia subspecialty experiences, including at least 1-month experiences in surgical intensive care, cardiac anesthesia, obstetric anesthesia, cardiovascular intensive care, transplant anesthesia, and pain management. The senior residents also participated in on-call and emergency anesthesia services at a large tertiary care hospital. This participation led to a more extensive experience with anesthesia for trauma surgery and with patients with ASA physical status IV and V.

Simulation Exercise

The simulation exercise was conducted in a single 75- to 90-min individual training session for each resident. The trainees managed each simulated event without assistance. One of two faculty members (D.J.M. or J.F.K.) and a console operator (J.A.W.) observed the simulation session from the control room adjacent to the simulation laboratory. Each trainee was given similar instructions by the simulation faculty or staff before beginning the exercises. The participants were instructed (1) to perform all diagnostic and therapeutic actions considered appropriate by the participant and (2) to verbalize actions taken during the scenario. The participants could request information about the participant's condition, but the responses from the faculty or console operator were scripted based on the input of the faculty and staff during scenario development and pilot testing. The six exercises were presented in identical order to each participant. After every two simulation encounters, the supervising faculty member discussed the case management for the preceding exercises. The sessions were completed during a 2-month period in the final months of the trainee's clinical year of training.

Each participant's performance was videotaped and recorded on a four-quadrant screen that included two separate video views of the participant and the mannequin. Two microphones were suspended from the ceil-

ing to capture audio during the scenarios. The third screen of the four quadrant video recording was the simultaneous full display of patient vital signs (electrocardiography, pulse oximetry, and monitoring of inspired and expired gas, blood pressure, and central venous pressures). The simulator control room staff or faculty could enter text information identifying scenario details or clarify participant or simulator actions using keyboard entries. This information appeared in the fourth quadrant of the videotaped recording.

Scoring

To obtain a quantitative performance measure for the exercises, we developed and contrasted four different scoring systems, three analytic and one holistic: a traditional checklist of diagnostic and therapeutic actions, time to key action for the most important three actions (diagnosis, initiate therapy, definitive therapy), key action, and a simple global rating.

The raters scored the pilot scenarios in a group session. During this session, the raters developed criteria for successful completion of actions. The three raters who provided time-based as well as global scores met to define the time-based actions and criteria for global scores. All of the raters independently observed and scored the residents' performances from the videotaped recordings. Each resident performed the scenarios in the same sequence during the evaluation. Two faculty members and the nurse clinician scored the performances using the scenario-specific checklists; three faculty anesthesiologists used the key action scoring system. The three faculty members who used the key action scoring system also provided a global rating of the resident's performance. This ultimately led to four separate scenario scores. One of the checklist raters was blinded to resident training level.

Checklist Scoring. The first scoring system developed was a comprehensive checklist of all the expected correct actions for each scenario. This list was created during the review of the scenario design and content by the investigators. This list was revised during the pilot testing of the scenarios. The completed list of actions was then resubmitted to each of the faculty members who were asked to add additional actions as well as to select the three or four most important diagnostic and therapeutic actions for each scenario. Using the returned lists, the five most frequently selected actions were determined for each exercise. These actions were submitted once more to the faculty to develop the three key actions for each scenario. These key actions were weighted in the checklist and were also used to develop the time-based as well as key-action scoring systems (table 1). Three of the raters used a checklist (analytic) scoring system with 11–16 possible actions for each scenario. The raters were asked to indicate whether a specific action described on the checklist had been performed by the candidate. The highest cumulative score defined the best possible performance in the checklist scoring system. The maximum possible score on the scenarios ranged from 14 to 22 points and, for scoring purposes, was expressed as a percentage value.

Time to Key Action. In previous simulation studies, we found that some of the checklist items correlated highly with overall performance on each exercise.[21,22] Checklist items that used a time limit for various actions were particularly useful in discriminating between more- and less-experienced trainees. For this reason, we hypothesized that a scoring system based on the time a candidate required to correctly diagnose a condition and initiate the correct therapy would provide valid and reliable estimates of ability. The key actions for each scenario included a diagnostic action, an initial treatment, and a definitive treatment. These key actions for each scenario were derived from the checklist as described.

Three of the raters scored the exercises using the abbreviated key action scoring system. The raters recorded the elapsed time until each of the key actions were performed by the trainee. This time-based key action score was based on the premise that more experienced residents would accomplish the key actions in less time than their junior colleagues. To derive a time to key action score, the time (in seconds) until the action was taken was subtracted from 300 (total time available for the encounter). This difference was then divided by 60, and 1 was added to the total. For example, if the diagnosis was obtained in 1 min, the trainee would receive a score of 5. If a key action never occurred, the trainee would receive a score of 0. Hypothetically, a trainee could receive a score of 6 for an action, but this would entail performing an action immediately at the beginning of the scenario. The three time-based action scores were added to yield an overall scenario score. In practice, scores could range from 0 (no actions taken) to 15 (all actions completed in 1 min). A score greater than 10 generally indicated the participant had performed all three actions during the scenario. A score of 12 indicated that the participant had accomplished all three actions in less than 3 min. One of the main concerns recognized with this scoring system is that a participant who accomplished all three actions might receive the same score as a resident who performed one action. For example, an overall score of 5 might indicate that one action was performed early in the scenario or possibly that all three actions were performed later in the scenario. For this reason, a key action score was also used to assess performance.

Key Action. Although the time-based key actions rewarded speed, the faculty believed that differences in time to each action might be less important than determining whether a participant actually performed the action during the scenario. For this reason, a scoring

Table 1. Scoring Items

Scenario	Checklist Scoring Items	Time-based Scoring Items
Anaphylaxis—PACU	Establish neuromuscular recovery (1 point), examine/inquire airway/blood loss/secretions (1 point), F_{IO_2} of 100% rebreathing mask or Ambu bag and mask (1 point), auscultate chest (1 point), diagnose bilateral wheeze/coarse breath sounds (1 point), increase intravenous fluids (1 point), anaphylaxis diagnosed within 3 min (2 points), anaphylaxis diagnosis (2 points), epinephrine within 3 min (3 points), epinephrine any dose (1 points), epinephrine correct dose ($>$ 50 μg, $<$ 300 μg) (1 point)*, pharmacologic treatment of hypotension (1 point), inhaled β agonists (1 point), intravenous diphenhydramine (1 point), intravenous steroids (1 point)	(a) Time to diagnosis of anaphylaxis (b) Time to treatment regimen for suspected anaphylaxis (c) Time to dose of epinephrine
MI—intraoperative	Diagnose ischemia (2 points), confirm ischemia (rhythm strip, ST analysis, check other leads) (2 points), increase F_{IO_2} to 100% (1 point), increase anesthetic depth (1 point), maximum heart rate during scenario less than 110 beats/min (1 point)†, maximum heart rate during scenario less than 120 beats/min (1 point), nitroglycerin therapy (1 point), titrate nitroglycerin (1 point), β-blocker therapy (2 points), titrate β-blocker therapy (1 point), inform surgery team of ischemia (1 point), heart rate less than 100 beats/min at end of scenario (1 point), heart rate less than 95 beats/min at end of scenario (1 point)†, systolic blood pressure less than 150 beats/min, diastolic blood pressure less than 100 beats/min at end of scenario (1 point)	(a) Time to diagnose ischemia by ST analysis or electrocardiographic rhythm strip (b) Any treatment directed at improving ischemia (c) Time to reduce heart rate less than 100 beats/min
Atelectasis—intraoperative	F_{IO_2} to 100% (2 points), review ventilator settings (1 point), diagnose hypoventilation/atelectasis (2 points), increase tidal volume/PEEP (2 points), mechanical to hand ventilation (1 point), auscultate chest (1 point), diagnose diminished breath sounds bilaterally (1 point), effective ventilation by hand (increase oxygen saturation to 90%, increase chest excursion) (1 point), lowest oxygen saturation greater than 80% (2 points), pass suction catheter *via* endotracheal tube (2 points), oxygen saturation to 90% at anytime during scenario (1 point), oxygen saturation to 95% before 120 s (1 point), oxygen saturation to 95% at any time during scenario (2 points)	(a) Time to 100% F_{IO_2}, hand ventilation, and auscultation (b) Time to reverse decline in oxygen saturation and improve oxygen saturation to 90% or greater (c) Time to oxygen saturation greater than 95%
Ventricular tachycardia—intraoperative	Diagnose ventricular tachycardia (1 point), palpate pulse or auscultate heart sounds (1 point), indicate patient is unstable or need for immediate shock (1 point), F_{IO_2} to 100% (1 point), defibrillator to bedside (1 point), correct joule (200+) (1 point), correct procedure for shock (1 point), administer shock within 60 s (1 point)‡, administer shock within 3 min (1 point), administer shock (2 points), abort operative procedure (1 point), lidocaine bolus/infusion (2 points), laboratory tests and 12-lead electrocardiogram (1 point)	(a) Time to diagnosis of ventricular tachycardia (b) Time to initiate any correct therapy (lidocaine/shock) (c) Time to shock
Cerebral hemorrhage—PACU	Establish patient is unresponsive (1 point) or unresponsive to pain (2 points), auscultate (1 point), conduct neurologic evaluation (1 point), indicate neurologic event (1 point), indicate potential increased ICP (1 point), neurology consult/CT scan (1 point), diagnosis within 2 min (1 point), prepare for intubation (1 point), F_{IO_2} to 100% (1 point), intubate (2 points), ventilate and auscultate (1 point), does not attempt to lower blood pressure (1 point)	(a) Time to establish patient unresponsive to verbal/pain or neurologic examination (b) Time to diagnose cerebral event/CT scan (c) Time to intubation
Aspiration—PACU	Establish patient is unresponsive to verbal (1 point), auscultate chest (1 point), request arterial blood gas (1 point), diagnose respiratory failure (2 points), prepare to intubate (1 point), Ambu bag and mask oxygen before intubation (1 point), sedation/anesthesia before or after intubation (1 point), laryngoscopy and intubation technique (1 point), intubated in less than 2 min (2 points), effective ventilation after intubation (2 points), indicate ventilator/PEEP required (1 point)	(a) Time to diagnose respiratory failure (b) Time to intubation (c) Time to effective ventilation after intubation

* Anaphylaxis. † MI: If the resident received a point for maximum heart rate less than 110 beats/min, he/she also received a point for maximum heart rate less than 120 beats/min. If the resident received a point for heart rate less than 95 beats/min at the end of the scenario, he/she also received a point for heart rate less than 100 beats/min. ‡ Ventricular tachycardia: If the resident received a point for administering a shock within 60 s, he/she also received a point(s) for administering a shock with 3 min and administering a shock during scenario.

CT = computed tomography; F_{IO_2} = fraction of inspired oxygen; ICP = intracranial pressure; MI = myocardial ischemia; PACU = postanesthesia care unit; PEEP = positive end-expiratory pressure.

system that merely documented whether the participant performed the action during the scenario was also used to measure performance. The key action score was the number of actions (diagnosis, initial treatment, definitive treatment) successfully completed and could range from 0 to 3.

Global Scoring. In addition to recording the time required to accomplish the essential diagnostic or therapeutic actions, the three time-based raters also rated the overall performance of each trainee. This holistic rating was based on time to diagnosis and treatment and allowed raters to consider potentially egregious or unnecessary diagnostic or therapeutic actions made by the trainee during the scenario. The raters were instructed to make a mark on a 10-cm horizontal line based on their assessment of the trainee's performance. The global rating system was anchored by the lowest value 0 (unsatisfactory) and the highest value 10 (outstanding). Before scoring the participants, raters agreed that a performance that met a score of 7 or more would be considered a standard expected in consultant practice for each exercise. Each performance was independently rated by each of the faculty raters.

Raters

Five anesthesiologists and one nurse clinician independently rated the residents' performances. All five of the anesthesiologists were clinical faculty who spent more than 70% of their time in patient care activities, either in clinical instruction or in direct patient care, often supervising the residents who were study participants. All of the faculty anesthesiologists had been board certified for a period of greater than 5 yr. The nurse clinician was blinded to the training background and experience of the residents.

Analysis

Several analyses were performed to investigate the utility of the scores from the simulation exercises. First, analysis of variance was used to test for score differences as a function of training (junior *vs.* senior residents). It was hypothesized that senior residents (CA-2, CA-3) would perform better on the simulation exercises, regardless of the scoring method. Second, to investigate the properties of the scores as a function of scoring method, various psychometric analyses were performed. Correlation coefficients were used to quantify the strength of the relations between simulator scores. Case discrimination statistics (correlation between case score and total score) were calculated to investigate how well each scenario could identify low- and high-ability residents. Finally, generalizability theory[31] was used to determine the reliability of each of the scoring systems and to identify the facets (*e.g.*, rater, scenario) that best explained the variability in resident scores.

Results

Comparison of Junior and Senior Residents

For each of the four scoring systems, a two-way analysis of variance was conducted to test the null hypothesis that was no difference in overall performance between the junior and senior residents. For the four analyses, the independent variables were resident group (junior, senior) and case (scenarios 1–6). The dependent variables were the four different scores (*i.e.*, weighted checklist, time to key action, key action, global).

For the analysis based on the weighted checklist, the case × group interaction was not significant. This indicates that the relative performance of the individuals in each group did not vary as a function of the case. However, there was a significant main effect attributable to resident group ($F = 5.8$, $P < 0.05$). This result reveals that, averaged over the 6 cases, there was a significant difference in mean scores between the junior and senior residents. As shown in table 2, the senior residents performed approximately 6 points better than the junior residents. The analysis also revealed a significant case effect ($F = 8.7$, $P < 0.01$), indicating that, averaged over the two study groups, there were performance differences by case. Overall, average performance was worst on postoperative stroke (scenario 5) and best on postoperative respiratory failure (scenario 6).

The results for the analyses of variance based on global, time to key action, and key action scores were similar to those for the weighted checklist. For all three analyses, there was no significant case × residency group interaction. This indicated that the differential performance of junior and senior residents was not dependent on the type of scenario (table 2). There were also significant main effects attributable to residency group ($F_{global} = 11.5$, $P < 0.01$; $F_{time} = 7.9$, $P < 0.01$; $F_{action} = 15.6$, $P < 0.01$) and case ($F_{global} = 4.9$, $P < 0.01$; $F_{time} = 14.9$, $P < 0.01$; $F_{action} = 10.7$, $P < 0.01$). Overall, regardless of scoring system or scenario, the senior resident group outperformed the junior resident group, and some scenarios were more difficult than others.

Relations Among Scores

The summary (averaged over 6 cases) simulator scores for each scoring system were strongly correlated. The correlation of the key action scores with the time-based scores was 0.89, with the checklist scores was 0.84, and with the global scores was 0.88. Almost 80% of the variance in the weighted checklist scores could be explained by the global scores. Likewise, nearly 80% of the variance in time to key action scores could be explained by the simple sum of the key actions completed in 5 min. The case–total correlations (discrimination statistics) among the cases were also positive. This indicates that residents who performed well on a given case tended to perform well overall. These correlations among individ-

Table 2. Performance of Junior and Senior Residents, by Scenario

Scenario	Scoring	Junior (n = 12), Mean ± SD	Senior (n = 16), Mean ± SD
Anaphylaxis	Weighted checklist (%)	37.5 ± 13.6	45.4 ± 10.8
	Global (0–10)	5.9 ± 2.7	7.3 ± 1.5
	Time to key action (max = 15)*	6.1 ± 4.5	8.9 ± 3.2
	Key action (max = 3)	2.2 ± 1.3	2.8 ± 0.4
MI	Weighted checklist (%)	50.2 ± 18.5	51.7 ± 14.9
	Global (0–10)	6.1 ± 2.1	7.3 ± 1.5
	Time to key action (max = 15)*	9.1 ± 4.0	10.1 ± 3.1
	Key action (max = 3)	2.0 ± 0.8	2.4 ± 0.7
Atelectasis	Weighted checklist (%)	47.5 ± 17.9	60.3 ± 14.8
	Global (0–10)	6.2 ± 2.2	7.7 ± 1.3
	Time to key action (max = 15)*	8.3 ± 3.3	11.1 ± 3.7
	Key action (max = 3)	2.1 ± 0.8	2.6 ± 0.6
Ventricular tachycardia	Weighted checklist (%)	57.1 ± 21.0	61.4 ± 13.7
	Global (0–10)	7.9 ± 2.0	8.4 ± 1.0
	Time to key action (max = 15)*	13.9 ± 3.9	13.7 ± 3.6
	Key action (max = 3)	2.8 ± 0.5	2.9 ± 0.3
Cerebral hemorrhage	Weighted checklist (%)	35.9 ± 20.0	42.5 ± 24.4
	Global (0–10)	5.6 ± 2.7	5.8 ± 2.6
	Time to key action (max = 15)*	6.0 ± 3.5	7.2 ± 3.1
	Key action (max = 3)	1.3 ± 0.8	1.7 ± 0.7
Postoperative respiratory failure	Weighted checklist (%)	60.1 ± 15.3	64.3 ± 9.9
	Global (0–10)	6.2 ± 1.9	7.5 ± 0.9
	Time to key action (max = 15)*	6.8 ± 3.6	8.4 ± 2.7
	Key action (max = 3)	2.2 ± 0.7	2.6 ± 0.4
Overall	Weighted checklist (%)	48.1 ± 11.3	54.3 ± 7.3
	Global (0–10)	6.3 ± 1.7	7.3 ± 0.8
	Time to key action (max = 15)*	8.3 ± 2.4	9.9 ± 1.3
	Key action (max = 3)	2.1 ± 0.6	2.5 ± 0.2

* Maximum points possible for time to key action is 15 if all three actions are completed in 1 min.

MI = myocardial ischemia.

ual cases and total score on cases ranged from 0.43 to 0.89, depending on the case and the scoring system used to measure performance (table 3). In general, the discrimination statistics were highest for postoperative respiratory failure (scenario 6) and lowest for postoperative stroke (scenario 5). The high positive correlations among each of the cases and the overall performance suggest that the performance domain being measured by each of these scenarios must be similar, but no one case could be used to determine an overall ability.

Variance Component Analysis

The variance components were determined using the generalizability analysis. This analysis partitioned the sources of variation in the scores into various components to investigate the sources of measurement error in the simulation scores. The reproducibility and overall reliability of the evaluation could be determined by assessing the sources of variation. In this study, we expected that the sources of variability in simulation scores might vary somewhat as a function of scoring method, but the principle source of variance would be attributable to differences in individual resident's abilities. This analysis could be used to measure variance in scores due to raters and rater interactions (participant and scenario). The small rater variance by scenario indicates raters rank each scenario of equal difficulty. An estimate of rater by participant provided a measure of whether

Table 3. Case–Total Correlations

Scenario	Weighted Score	Global	Time to Key Action	Key Action
Anaphylaxis—PACU	0.61	0.71	0.71	0.79
MI—intraoperative	0.59	0.67	0.59	0.71
Atelectasis—intraoperative	0.63	0.73	0.47	0.66
Ventricular tachycardia—intraoperative	0.51	0.68	0.43	0.43
Cerebral hemorrhage—PACU	0.53	0.57	0.40	0.43
Aspiration—PACU	0.73	0.89	0.69	0.76

Each scenario correlation coefficient (Spearman) with overall participant score on six scenarios. All cases correlate with trainees' performance on the entire exercise. Aspiration is strongest correlation with overall score.

MI = myocardial ischemia; PACU = postanesthesia care unit.

Table 4. Variance Components for Simulation Scores

	Weighted Checklist		Global		Time		Action	
	Estimate	%	Estimate	%	Estimate	%	Estimate	%
Component								
Examinee (E)	54.0	13.4	1.3	22.8	2.0	10.2	0.14	18.9
Case (C)	75.4	18.6	0.5	9.1	6.0	30.3	0.17	22.2
Rater (R)	4.0	1.0	0.2	4.3	0.1	0.0	0.00	0.0
E × C	202.2	49.9	2.1	37.3	10.2	51.4	0.34	46.0
E × R	2.0	0.5	0.3	5.3	0.0	0.1	0.00	0.5
C × R	6.2	0.2	0.1	2.0	0.3	1.7	0.01	1.7
Triple (C × R × P) variance, error	60.9	15.0	1.1	19.1	1.3	6.4	0.08	11.3
Generalizability coefficient	0.59		0.72		0.53		0.69	

raters rank order participants similarly. In this study, where some of the raters knew the participants, this type of analysis was useful to assess "halo" effects. We anticipated that the scoring method most likely to be affected by rater subjectivity, the global scoring method, would also have the highest rater variance.

The estimated variance components for the resident scores are shown in table 4. The analyses were done separately for each of the scoring methods. The examinee (resident) variance component is an estimate of the variance across examinees of examinee level mean scores. Ideally, most of the variance should be here, indicating that individual abilities account for differences in observed scores. The case components are the estimated variances of case mean scores. For all four scoring methods, the estimates were greater than zero, suggesting that the cases varied somewhat in average difficulty. The rater components are the variances of the rater mean scores. The relatively small values indicate that raters did not vary appreciably in terms of average stringency (table 4). As expected, the rater variance component was greatest (as a percentage of total variance) for global scoring (4.3%; table 4), where the scores would be most likely to be influenced by subjective factors. Despite more variance as a result of the global scoring, the overall reliability of the global evaluation (generalizability coefficient = 0.72; table 4) was even more reproducible than the traditional checklist scoring method (generalizability coefficient = 0.59). The largest interaction variance component, for all four scoring methods, was examinee by case. The magnitude of these components suggests that there were considerably different rank orderings of examinee mean scores for each of the various cases. This variance indicates that to reliably assess a resident's ability, a single encounter is not adequate to effectively assess skill. The relatively small examinee by rater components (table 4) suggest that raters rank ordered residents similarly, indicating that the rating criteria were consistently applied by all raters whether blinded or unblinded to the participant's training background. Likewise, the small rater by case components indicate that the raters rank ordered the difficulty of the cases similarly and must use the rating systems in an equivalent manner, consistently defining the same endpoints for each action. The final variance components are the residual variances that include the triple-order interactions (rater, scenario, and participant) and all other unexplained sources of variation. These variances were the lowest for the key action scoring methods (time-based and key action) (table 4).

For the weighted checklist data, the generalizability coefficient, based on six cases and three independent raters, was 0.59. (For global scoring, the generalizability coefficient was 0.72.) The reproducibility of the scores was lowest for the time to key action method (0.53). Here, case specificity (examinee × case) accounted for more than 50% of the variance in observed scores. The simple key action scores were moderately reproducible (0.69). Similar to the other scoring methods, the case and examinee by case variance components were the largest, indicating that the cases were not of equal difficulty and that examinee performance can vary as a function of the scenario content.

As shown in table 4, the rater facets and associated interactions do not contribute much to the variability of observed scores, regardless of scoring method. The small rater, person by rater, and rater by case variance components indicate that the choice and number of raters has little impact on the reproducibility of the resident scores.

Discussion

Similar to results from previous studies, more experienced practitioners outscored less experienced ones in simulated scenarios designed to assess their ability to manage acute care situations.[22] We anticipated that the additional training and clinical experiences of senior residents who were more familiar with emergency situations and confident with diagnosis as well as treatment of these conditions would lead to higher scores. It is reasonable to assume that these physicians would be more prepared to effectively translate their knowledge into a logical and orderly sequence of actions that would lead to rapid diagnosis and treatment of the patient. The

residency program sampled, like every other residency program, incorporates more sophisticated subspecialty training experiences for individuals as they progress through the program. As expected, advanced training experiences in intensive care and other subspecialties, such as cardiovascular anesthesia, on-call emergency anesthesia, and transplant anesthesia, provide senior residents with the necessary experience to manage these simulated emergency situations. Overall, the fact that junior residents performed less well on the acute care exercises, although based on a small select sample of participants, provides some evidence to support the discriminant validity of the simulation scores.

The use of simulation in medicine is becoming more widespread. However, unlike the plethora of research related to standardized patient assessments, relatively little work has been done to explore scoring systems for evaluations based on integrated simulators. An important goal of this study was to evaluate scoring strategies for mannequin-based simulation exercises and propose an effective method for use in future studies. The high correlation among all of our scoring systems indicates that, regardless of the rubric used, comparable performance domains are measured.

In evaluating scoring systems for advanced training, our goal was to develop objective measures of performance. A traditional comprehensive checklist scoring system is one of the most common methods used to score performance evaluations.[27] However, depending on the quality of test construction, inclusive checklists may have drawbacks, including documentation errors[32] and the potential to reward thoroughness as opposed to skill. Checklist scoring systems also do not capture sequencing issues that are fundamental to patient management in acute care situations. Moreover, physicians who rapidly assess a patient and effectively manage a condition may be penalized for not performing certain actions demarked on the checklist. In our study, residents achieved lower checklist scores primarily because these checklists were inclusive of actions. This may explain why residents achieved mean checklist scores often less than 50% of the maximum possible points. For example, the ventricular tachycardia event was promptly recognized and effectively treated by most residents as evidenced by the mean key action scores were more than 90% of the maximum possible points (key action 2.9 out of 3) and time to key action scores (mean $>$ 13 of 15), but resident mean scores on the checklist were less than 60% of the maximum possible points. On the other end of the spectrum, a trainee could perform many actions listed on a checklist but fail to perform the most essential diagnostic or therapeutic ones. Although weighting the checklist items may partially alleviate this problem, consensus must be achieved among experts as to what these weights should be. An additional concern with checklists relates to the difficulty of finding raters to score the detailed actions. This task requires careful observation during the scenario review. Depending on the number of examinees and the number of scenarios, each rater may be required to observe numerous performances. Here, factors such as fatigue could lead to scoring inaccuracy.[33] Given that other scoring modalities (*e.g.*, global rating, key actions) adequately capture levels of performance, the use of "objective" checklists may not be necessary.

Almost all of these scoring systems require some subjective rater judgments. In the time-based scoring system, raters were required to determine when an action endpoint occurred, such as time to effective ventilation in the respiratory failure scenario or time to diagnosis of myocardial ischemia or anaphylaxis in these respective scenarios. The small variances among raters suggest that a consistent endpoint could be determined with a relatively limited amount of rater training before beginning the scoring. Although trainees were rank ordered similarly regardless of scoring method, the use of the key action scoring system offers some distinct advantages, including ease of rater use and the ability to identify scoring discrepancies quickly.

The holistic evaluation, which included a single global rating of performance, added more rater variance, accounting for 4.3% of the variance, but the overall reliability and reproducibility of the global scoring system was similar to more objective scoring methods. Unlike the other scoring systems, the global assessment allowed raters to evaluate all of the actions (correct and incorrect) and to assess the sequence of actions in assigning an overall score. The global scoring was also the only scoring method that could potentially be used to determine a competence standard. In this scoring system, raters agreed before beginning the rating that a score of 7 (using a 0–10 scale) would be expected of an anesthesia consultant. Although understanding a performance consistent with an anesthesia consultant would be a major advantage of this scoring system, the faculty who assigned global ratings might have been influenced by their knowledge of the residents training level, skills in actual clinical practice, and time to graduation. An additional concern with global scores is that rater discrepancies, especially their root causes, are more difficult to resolve when the rating criteria are not as explicit. In future studies, a method to define a standard of performance expected of a competent specialist using more objective criteria would help to determine whether differences in scores among participants and groups are relevant. In the absence of a definition of a consultant or accepted performance standard, a significant difference in a checklist or key action scenario or overall score is difficult to interpret in the context of training or potentially certification of practitioner skill.

All three analytic scoring methods (checklist, time to key action, key action) correlated with a single global

faculty assessment. This suggests that the checklist items and key actions adequately captured complex performance even though these analytic scoring systems did not include a method to document unnecessary or potentially detrimental actions or define a competence standard. A method to identify and review added actions not considered in the analytic scoring systems would be a useful addition to the evaluation.

Rather than providing an examinee with a single prolonged scenario that posed multiple additive challenges, we instead presented multiple reproducible exercises that measured a defined, targeted management issue. The selection of scenarios, goals, and actions expected for each simulation encounter were based on a number of steps involving multiple faculty members. The end result was the development of a set of brief, targeted simulation exercises relevant to anesthesia practice that could be scored efficiently and reliably. As a result, estimating a resident's skill level was more reproducible because the score was based on multiple performance samples. Before beginning this study, we were unsure of how many encounters might be required to provide a reliable evaluative experience. For this reason, we arbitrarily selected a set of six encounters that could be completed in an hour-long individual training session. The generalizability coefficients were moderate from this group of six scenarios, suggesting that if even more precise measures of ability were required, additional performance samples would be needed. We found that there was considerable variation attributable to case content. In fact, case specificity, not rater stringency/leniency/bias, was the major determinant leading to variation in simulation scenario scores. Although the positive correlation between scenario scores, regardless of method, suggests that similar aspects of clinical performance are measured, the broad range of scores and different rank ordering of trainees on each exercise suggest that a variety of simulation exercises are required to effectively assess acute care skills.

In medical assessment, there has traditionally been concern about the potential subjectivity of expert raters. Higher interrater variability is expected because of the potentially subjective nature of the assessment and halo effects from knowledge of a participant's training or background.[10,14,17–19,27,34] In this study, the same three raters used identical scoring systems to score all of the participants. Although this methodology may help to assure relatively consistent rater scores, five of our six raters knew the training level and background of participants. This could lead to halo effects, particularly when using a global rating system. The scenario design, development, and scoring procedures for the three analytic systems were implemented to provide more objective scoring methods and to achieve rater consensus on scenario scoring. These study design factors may have helped reduce the rater-related variance in scores. All of the rater-related variances were small, suggesting that at least in this simulation study, concern about subjectivity of the scoring systems and rater reliability might not have been as important a limitation of simulation assessment as reported in previous studies. The consistent finding in this study and a previous study of graduating medical students using a similar methodology[22] was that the reliability of an examinee's score is far more dependent on the number of tasks or scenarios as opposed to the number of raters per given task.

To increase the confidence of the ability measure, additional scenarios would contribute more to the reproducibility of the performance assessment than using more raters. To achieve a reliability estimate for the evaluation consistent with the generalizability coefficients reported in certification examination (> 0.75), increasing the number of encounters to 10 or 12 would achieve a predicted reliability of more than 0.8 for most of the reported scoring systems. For example, using a key action score, with double the number of scenarios ($n = 12$) and the same number of raters ($n = 3$), the estimated reliability coefficient would be 0.82 (SEM = 0.18). If only 1 randomly selected rater were used to score a 12-scenario assessment, the reliability of the scores would only decrease to 0.78 (SEM = 0.20). Results were similar for the other scoring methods. A more detailed analysis of a larger set of scenarios would aid in determining, based on the purpose of the assessment, the optimal number of scenarios and associated content mix.

Previous studies dealing with performance assessment have cast a broad net and have included skills in communication and behavior as part of the equation.[10,11,17,18] The duration of scenarios and the multiple skills that each encounter attempts to measure has generally led to arduous and complicated scoring rubrics.[10,14,17–19,35] In comparison to previous simulation studies, our methodology did not provide a mechanism to assess skills in communication or leadership. To effectively assess skills in communication and leadership, additional scenario complexity would be required that might include "standardized" nurses and peers as well as scoring systems that effectively measure these skills. In this study, a directed encounter that isolated and measured participant skill in a single acute care setting was used to measure ability. Our results indicate that many of the psychometric concerns associated with mannequin-based assessment can be overcome by this test battery approach, which is similar in structure to the standardized patient assessments used for certification and licensure decisions.[36,37]

Although the battery approach typically used for standarized patient assessments could be an effective method to measure anesthesia trainee and consultant skills, special attention must be paid to the particular choice of scenarios. In general, the exercises must be

chosen to reflect the actions expected of a trainee or experienced practitioner in practice settings.[38] The scores on individual scenarios positively correlated with overall scores, indicating that similar performance skills were assessed by these exercises. A more detailed analysis of each scenario would be helpful to determine which skills and scenarios require additional emphasis during training. For example, we found that one of the three postoperative scenarios (respiratory failure) was most effectively managed by anesthesia residents and was also highly correlated with overall performance (case discrimination ranged from 0.69–0.89). In contrast, participants' scores were the lowest on the other two postoperative scenarios (anaphylaxis and stroke). This suggests that a range of events may be required for a comprehensive assessment, including preoperative, intraoperative, and postoperative problems. In addition, from a content perspective, deciding the percentage of scenarios to be sampled from various categories of conditions (*e.g.*, cardiovascular, respiratory, metabolic, equipment) is essential for developing a valid, broad-based, assessment. For standardized patient assessments, this is often accomplished by matching assessment content to patient visit statistics available from national medical surveys.[39] A similar strategy for anesthesia could be used. Here, one could identify clinical practice situations that lead to perioperative patient morbidity and mortality. Relevant scenarios could be identified through the analysis of Medicare databases, concentrating on postoperative patients in whom certain complications and conditions result in a higher mortality.[40] Overall, although a multiple-scenario assessment is required to reliably measure acute care skills in anesthesia, the validity of assessment is highly dependent on the type, breadth, and content of the cases that are chosen to be modeled.

The broader significance of this, and related, investigations is the introduction of methods to measure and access competence in clinical practice. In a recent review of studies that define and assess competence, Epstein and Hundert[41] suggest that current assessment formats test knowledge but may underemphasize a number of important domains of competence, including integration of knowledge into clinical practice. This review highlights the need for evaluations that target additional important abilities, including situations that require higher-order clinical reasoning skills, pattern recognition, and directed actions. Unfortunately, because of a shortage of resources and expertise, few anesthesia training programs are capable of developing evaluations that meet the high psychometric standards required of examinations used by licensure boards. Nevertheless, with faculty support and an appropriate simulation environment, locally developed assessments can be used to measure the advanced skills of healthcare professionals. Additional psychometric studies involving larger numbers of residents and a broader sampling of the universe of simulated conditions are required to support the generalization of our conclusions. More important, validation studies concentrating on the relation of the simulation scores to other measures of clinical performance as well as provider skill with "real" patients are necessary.

References

1. Gaba DM: Improving anesthesiologists' performance by simulating reality. ANESTHESIOLOGY 1992; 76:491–4
2. 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
3. Kapur PA, Steadman RH: Patient simulator competency testing: Ready for takeoff? Anesth Analg 1998; 86:1157–9
4. Murray DJ: Clinical simulation: Technical novelty or innovation in education. ANESTHESIOLOGY 1998; 89:1–2
5. Good ML: Patient simulation for training basic and advanced clinical skills. Med Educ 2003; 37(suppl 1):14–21
6. Gordon JA, Tancredi DN, Binder WD, Wilkerson WM, Shaffer DW: Assessment of a clinical performance evaluation tool for use in a simulator-based testing environment: A pilot study. Acad Med 2003; 78:S45–7
7. Federation of State Medical Boards, Inc. and National Board of Medical Examiners: 2004 USMLE Step 3 Content Description and Sample Test Materials. Philadelphia, FSMB and NBME, 2003
8. Schwid HA: Anesthesia simulators: Technology and applications. Isr Med Assoc J 2000; 2:949–53
9. 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
10. Holzman RS, Cooper JB, Gaba DM, Philip JH, Small SD, Feinstein D: Anesthesia crisis resource management: Real-life simulation training in operating room crises. J Clin Anesth 1995; 7:675–87
11. Howard SK, Gaba DM, Fish KJ, Yang G, Sarnquist FH: Anesthesia crisis resource management training: Teaching anesthesiologists to handle critical incidents. Aviat Space Environ Med 1992; 63:763–70
12. Jacobsen J, Lindekaer AL, Ostergaard HT, Nielsen K, Ostergaard D, Laub M, Jensen PF, Johannessen N: Management of anaphylactic shock evaluated using a full-scale anaesthesia simulator. Acta Anaesthesiol Scand 2001; 45:315–9
13. Lindekaer AL, Jacobsen J, Andersen G, Laub M, Jensen PF: Treatment of ventricular fibrillation during anaesthesia in an anaesthesia simulator. Acta Anaesthesiol Scand 1997; 41:1280–4
14. 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
15. O'Donnell J, Fletcher J, Dixon B, Palmer L: Planning and implementing an anesthesia crisis resource management course for student nurse anesthetists. CRNA 1998; 9:50–8
16. McGaghie WC, Issenberg SB, Petrusa ER: Simulation: Savior or Satan? A rebuttal. Adv Health Sci Educ Theory Pract 2003; 8:97–103
17. 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
18. 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
19. 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
20. Schwid HA, O'Donnell D: Anesthesiologists' management of simulated critical incidents. ANESTHESIOLOGY 1992; 76:495–501
21. Murray D, Boulet J, Ziv A, Woodhouse J, Kras J, McAllister J: An acute care skills evaluation for graduating medical students: A pilot study using clinical simulation. Med. Educ 2002; 36:833–41
22. 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
23. Whelan G: High-stakes medical performance testing: The Clinical Skills Assessment program (letter). JAMA 2000; 283:1748
24. Federation of State Medical Boards, Inc. and National Board of Medical Examiners: 2004 USMLE Step 2 CS Content Description and General Information Booklet. Philadelphia, FSMB and NBME, 2003
25. Devitt JH, Kurrek MM, Cohen MM, Cleave-Hogg D: The validity of performance assessments using simulation. ANESTHESIOLOGY 2001; 95:36–42
26. Schwid HA, Rooke GA, Michalowski P, Ross BK: Screen-based anesthesia simulation with debriefing improves performance in a mannequin-based anesthesia simulator. Teach Learn Med 2001; 13:92–6

27. 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

28. Boulet JR, Rebbecchi TA, Denton EC, McKinley DW, Whelan GP: Assessing the written communication skills of medical school graduates. Adv Health Sci Educ Theory Pract 2004; 9:47–60

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. Hodges B, McIlroy JH: Analytic global OSCE ratings are sensitive to level of training. Med Educ 2003; 37:1012–6

31. Brennan RL: Generalizability Theory 1. New York, Springer-Verlag, 2001, pp 1–538

32. Boulet JR, McKinley DW, Whelan GP, Hambleton RK: Quality assurance methods for performance-based assessments. Adv Health Sci Educ Theory Pract 2003; 8:27–47

33. Humphris GM, Kaney S: Examiner fatigue in communication skills objective structured clinical examinations. Med Educ 2001; 35:444–9

34. Gaba DM: Two examples of how to evaluate the impact of new approaches to teaching. ANESTHESIOLOGY 2002; 96:1–2

35. Morgan PJ, Cleave-Hogg D, DeSousa S, Tarshis J: High-fidelity patient simulation: Validation of performance checklists. Br J Anaesth 2004; 92:388–92

36. Norcini J, Boulet J: Methodological issues in the use of standardized patients for assessment. Teach Learn Med 2003; 15:293–7

37. Gimpel JR, Boulet DO, Errichetti AM: Evaluating the clinical skills of osteopathic medical students. J Am Osteopath Assoc 2003; 103:267–79

38. Boulet JR, Gimpel JR, Errichetti AM, Meoli FG: Using National Medical Care Survey data to validate examination content on a performance-based clinical skills assessment for osteopathic physicians. J Am Osteopath Assoc 2003; 103:225–31

39. Ziv A, Boulet JR, Burdick WP, Friedman Ben-David M, Gary NE: The use of national medical care surveys to develop and validate test content for standardized patient examinations, Proceedings of the Eighth Ottawa Conference on Medical Education and Assessment (1). By Melnick D. Philadelphia, National Board of Medical Examiners, 2000, pp 99–105

40. Silber JH: Anesthesiologist direction and patient outcomes. LDI Issue Brief 2000; 6:1–4

41. Epstein RM, Hundert EM: Defining and assessing professional competence. JAMA 2002; 287:226–35