Anesthesiology 2002; 97:1434–44

Evaluation of Anesthesia Residents Using Mannequin-based Simulation

A Multiinstitutional Study

Howard A. Schwid, M.D., G. Alec Rooke, M.D., Ph.D.,* Jan Carline, Ph.D.,† Randolph H. Steadman, M.D.,‡ W. Bosseau Murray, M.D.,§ Michael Olympio, M.D.,‖ Stephen Tarver, M.D.,# Karen Steckner, M.D.,** Susan Wetstone,†† the Anesthesia Simulator Research Consortium‡‡*

Background: **Anesthesia simulators can generate reproducible, standardized clinical scenarios for instruction and evaluation purposes. Valid and reliable simulated scenarios and grading systems must be developed to use simulation for evaluation of anesthesia residents.**

Methods: **After obtaining Human Subjects approval at each of the 10 participating institutions, 99 anesthesia residents consented to be videotaped during their management of four simulated scenarios on MedSim or METI mannequin-based anesthesia simulators. Using two different grading forms, two evaluators at each department independently reviewed the videotapes of the subjects from their institution to score the residents' performance. A third evaluator, at an outside institution, reviewed the videotape again. Statistical analysis was performed for construct- and criterion-related validity, internal consistency, interrater reliability, and intersimulator reliability. A single evaluator reviewed all videotapes a fourth time to determine the frequency of certain management errors.**

Results: **Even advanced anesthesia residents nearing completion of their training made numerous management errors; however, construct-related validity of mannequin-based simulator assessment was supported by an overall improvement in simulator scores from CB and CA-1 to CA-2 and CA-3 levels of training. Subjects rated the simulator scenarios as realistic (3.47 out of possible 4), further supporting construct-related validity. Criterion-related validity was supported by moderate correlation of simulator scores with departmental faculty evaluations (0.37–0.41, $P < 0.01$), ABA written in-training scores (0.44–0.49, $P < 0.01$), and departmental mock oral board scores (0.44–0.47, $P < 0.01$). Reliability of the simulator assessment was demonstrated by very good internal consistency ($\alpha = 0.71$–0.76) and excellent interrater reliability (correlation = 0.94–0.96; $P < 0.01$; $\kappa = 0.81$–0.90). There was no significant difference in METI *versus* MedSim scores for residents in the same year of training.**

Conclusions: **Numerous management errors were identified in this study of anesthesia residents from 10 institutions. Further attention to these problems may benefit residency training since advanced residents continued to make these errors. Evaluation of anesthesia residents using mannequin-based simulators shows promise, adding a new dimension to current assessment methods. Further improvements are necessary in the simulation scenarios and grading criteria before mannequin-based simulation is used for accreditation purposes.**

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

CURRENT assessment of anesthesia residents consists of departmental faculty evaluations, written in-training examination, and departmental mock oral board examinations. Each of these methods has advantages and shortcomings. While departmental faculty evaluations are based on actual clinical interactions, they can be subjective, highly influenced by a single negative experience, and based on minimal contact. Written examination is an objective measure of factual knowledge but may not evaluate clinical application of that knowledge. The oral examination is designed to test patient management based on scientific principles, but the mock examination may be administered by inexperienced, inadequately trained faculty and therefore subject to high interrater variability.

Simulation technology for anesthesia offers the possibility of a new assessment modality. Standardized clinical scenarios can be generated and trainee response can be measured without the need to intervene for patient safety. For several years, the Educational Commission for Foreign Medical Graduates clinical skills assessment has utilized standardized patients, actors presenting with characteristic signs and symptoms which the candidate diagnoses and prescribes treatment.[1] For many purposes in anesthesia, a computerized, breathing mannequin with pulses, blood pressure, electrocardiogram, and other physiologic responses could serve as the standardized patient to produce high-fidelity, reproducible, life-

* Professor of Anesthesiology, Department of Anesthesiology, University of Washington, and Staff Anesthesiologist, VA Puget Sound Health Care Service. † Professor of Medical Education, Department of Medical Education, University of Washington. ‡ Associate Clinical Professor of Anesthesiology and Vice Chair, Department of Anesthesiology, University of California–Los Angeles, Los Angeles, California. § Professor of Anesthesiology, Department of Anesthesiology, Pennsylvania State University, Hershey, Pennsylvania. ‖ Associate Professor of Anesthesiology, Department of Anesthesiology, Wake Forest University, Winston-Salem, North Carolina. # Associate Professor of Anesthesiology, Department of Anesthesiology, University of Kansas, Kansas City, Kansas. ** Staff Anesthesiologist, Department of General Anesthesia, Cleveland Clinic, Cleveland, Ohio. †† Medical Student, University of Washington School of Medicine. ‡‡ Members of the Anesthesia Simulator Research Consortium are listed in Appendix A.

Received from the Department of Anesthesiology, University of Washington, Seattle, Washington, and Surgical and Perioperative Care, VA Puget Sound Health Care Service, Seattle, Washington. Submitted for publication July 30, 2001. Accepted for publication July 16, 2002. At each participating institution, the Department of Anesthesiology provided financial support for the work conducted at that institution. There was no other outside funding for this project. Dr. Steadman has received honoraria from METI, Medical Education Technologies Inc., Sarasota, Florida, for briefing new users, but the most recent time was 2 to 3 years ago, before starting this study. The design of this study was presented at the annual meeting of the Society of Academic Anesthesiology Chairs, Chicago, Illinois, October 28, 2000. Preliminary results were presented at the annual spring meeting of the Society for Education in Anesthesia, Cleveland, Ohio, June 3, 2001.

Address reprint requests to Dr. Schwid: Surgical and Perioperative Care (112A), VA Puget Sound Health Care System, 1660 South Columbian Way, Seattle, Washington 98108. Address electronic mail to: hschwid@u.washington.edu.

threatening acute situations that are not possible in other evaluation settings.

To use the anesthesia simulator to evaluate anesthesia residents, realistic scenarios must be programmed and evaluated, and the validity and reliability of the grading system must be determined. Construct-related validity describes whether the simulator evaluation is a legitimate indicator of performance, while criterion-related validity compares the results of the simulator evaluation to other measures of resident performance.[2] Internal consistency is a measure of the quality of the items in the score. We tested four scenarios involving anesthetic critical incidents and two grading systems which assign a score to performance. Evidence of construct-related and criterion-related validity plus internal consistency, interrater reliability, and intersimulator reliability of the simulator evaluation are presented. In addition, the frequency of certain management errors was measured to increase attention in these areas for training purposes and the design of effective interventions and guidelines to improve patient safety.

Methods

Anesthesiology departments at 10 institutions with METI (Medical Education Technologies Inc., Sarasota, FL) or MedSim (MedSim Inc., formerly of Ft. Lauderdale, FL) mannequin-based simulators participated in this study. Institutional Review Board approval was obtained at each institution. Anesthesiology residents were invited to participate. At some institutions, this meant a sign was posted and residents were called to volunteer. At other institutions, the faculty directly contacted residents and asked whether they would be willing to participate. Each institution was instructed to try to obtain a mixture of weak and strong residents. Resident selection was largely determined by which residents were available during an available simulation time slot. The residents were free to decline to take part, but none refused. Ninety-nine residents signed written informed consent forms to be videotaped during their management of four scenarios and for information to be obtained from their academic records. Subject level of training ranged from 7 CB residents (clinical base year), 52 CA-1 (first year of clinical anesthesia training), 25 CA-2, and 15 CA-3. The number of residents tested at each institution varied from 4 to 19. All were tested in the simulator within 2 months of the end of their indicated level of training, and none were tested on the day postcall. All residents were familiar with the mannequin-based simulator and had previous simulator training, although none had managed or observed the particular simulator scenarios presented in this study. Prior to starting the simulator session, residents were instructed to manage the patient as they would in the operating room and verbalize all observations, possible problems, and treatments administered.

The simulation scenarios and grading forms used in this study were developed and used in a prior study involving over 30 simulator sessions.[3] The scenarios and grading forms were circulated to the 32 anesthesiologists who contributed to this study for comments and suggestions. Modifications of both the scenarios and grading forms were made, and the scenarios were programmed for both MedSim and METI simulators. In addition, the simulation actors (paramedic trainee, surgeon, and circulating nurse) at each institution were given identical scripts for the scenarios. The primary author reviewed a videotaped simulation session from each of the participating institutions to ensure scenario consistency across all centers.

The first simulated scenario involved esophageal intubation. The subject performed an anesthetic induction but allowed a "paramedic trainee" to intubate. The "paramedic" performed laryngoscopy, reported visualization of the vocal cords, and then proceeded to intubate the esophagus while maintaining that the endotracheal tube was placed correctly. Physiologic signs of esophageal intubation, including lack of breath sounds, increased airway pressure, absent exhaled carbon dioxide, and eventually decreased arterial oxygen saturation, were produced by the anesthesia simulator. The grading criteria (Appendices B and C) for this scenario included the diagnostic observations announced by the subject and time to reestablish ventilation.

A few minutes after correction of the esophageal intubation, the surgeon requested administration of antibiotic and muscle relaxant. An anaphylactic reaction was triggered with an increase in heart rate to 120 beats/min and a fall in systolic blood pressure to 50–60 mmHg refractory to treatment with ephedrine and phenylephrine. The simulator did not physically produce a rash, but the subject was informed that a rash was present if he or she inquired. The subject was given 15 min to diagnose and treat the problem. Since the airway was previously secured, grading criteria for anaphylaxis were largely geared toward making the correct diagnosis, and appropriate and timely administration of fluids and epinephrine.

Following the anaphylaxis scenario, the simulator was reset, and the subject induced anesthesia for a second patient. Shortly after intubation, the patient's preexisting COPD was exacerbated, resulting in bronchospasm with high airway pressure, decreased tidal volume, carbon dioxide retention, and decreased arterial oxygen saturation. Grading criteria were consideration of differential diagnosis for the difficulty ventilating, appropriate administration of bronchodilators, and increasing the concentration of the inhalation agent. The bronchospasm lasted 15 min or until the subject administered bronchodilators twice.

A few minutes after resolution of the bronchospasm, the fourth scenario occurred, with ST depression, tachycardia, and hypotension. Grading criteria were administration of pressors and fluids to increase blood pressure, decreasing the inhalation agent and administration of narcotics, β blockade to decrease heart rate, and appropriate titration of nitroglycerin. The subject was graded on selection of therapeutic agents, amounts administered, and time to administration.

The simulation sessions were videotaped using two camera angles to capture the subject's clinical management and the anesthesia machine and monitors. Quality of the audiovisual recording was ensured prior to subject testing at each participating institution. Two evaluators at each institution independently reviewed the videotapes of the subjects from their institution to score subject performance in the simulator. In most cases, these two internal evaluators knew the subjects' level of training and may have had previous clinical experience with the resident. A third evaluator from a different institution graded the videotape again. The outside reviewers had no knowledge of the subjects.

Two scoring systems were utilized. Both were checklists of responses to the critical incidents. The first grading system (Long Form, Appendix B) had 108 possible points, with many points determined by subject verbalization of observations. Most items on the long form were worth one point, but several items had heavily weighted items for performing essential treatment within certain time limits, and negative points for dangerous actions or omissions. The second grading system (Short Form, Appendix C) had 40 single-point items, with no weighted or negative points. Points on the short form were awarded for therapeutic actions that would directly benefit the patient with no points for observations or differential diagnosis. It should be noted that it was possible to earn multiple points for a single action with the short form. For example, administration of epinephrine in less than 5 min from the time the heart rate reached 100 beats/min earned three points: one point for less than 5 min, one point for less than 8 min, and one point for less than 12 min.

Following the simulation session, subjects rated the realism of the scenarios on a graded scale (4 = very realistic, 3 = somewhat realistic, 2 = somewhat unrealistic, 1 = very unrealistic). In addition, routine assessments of residents by departmental faculty evaluations, written in-training examination, and mock oral board examination were obtained from the resident's academic records. The departmental evaluations were based on ratings from at least six faculty members for clinical rotations within a few months of the simulator assessment. All participating departments used the same faculty evaluation form for this study with graded scores (5 = outstanding, 4 = good, 3 = satisfactory, 2 = doubtful, 1 = unsatisfactory). The written score used in this study was the percent correct from the most recent annual American Board of Anesthesiologists (ABA) in-training examination. Mock oral board scores were based on the ratings of at least two faculty members for the most recent departmental practice examination using the standard ABA grading scale (80 = definite pass, 77 = probable pass, 73 = probable fail, 70 = definite fail).

Interrater reliability was measured using Pearson correlation and κ statistics for the two evaluators from the examining institution (raters 1 and 2) and for the evaluator from the outside institution (raters 1 and 3, 2 and 3). After the interrater reliability statistics were analyzed, the average score for the three raters was used for the simulation score for the remainder of the statistics. Construct-related validity was supported by progression of simulator scores with level of training and the subjects' rating of the realism of the simulator scenarios. Criterion-related validity was determined by comparing the simulator scores with departmental faculty evaluations, written in-training examination, and mock oral board examination using the Pearson correlation statistic. Reliability of the simulator evaluation was assessed by the Cronbach α statistic for internal consistency. Intersimulator reliability, referring to the effect of simulator type, was assessed using the *t* test.

A single investigator reviewed all the videotaped simulation sessions a fourth time to document the frequency of certain errors. It was necessary to complete another review of the videotapes since some of the observed errors were not anticipated prior to the start of the study and were not included in the original grading forms. Significance of error rate differences between beginning residents (CB and CA-1) *versus* advanced residents (CA-2 and CA-3) were evaluated using the Fisher exact test.

Results

Esophageal Intubation

All anesthesia residents diagnosed and treated the esophageal intubation adequately, reestablishing ventilation in less than 5 min. However, CB residents did take significantly longer to reestablish ventilation ($P = 0.02$) than more experienced residents, with 43% (3/7) taking more than 2 min, while only 8% (7/92) of CA-1, CA-2, and CA-3 residents took more than 2 min.

Anaphylaxis

Several problems were noted in diagnosing the anaphylactic reaction (table 1). Overall, residents did not make the diagnosis 34% (34 of 99) of the time. Several residents (15%) stated that they thought the lack of measurable blood pressure was due to a faulty noninvasive blood pressure monitor during the hypotensive period. Checking the simulated patient's pulse would have revealed a weak or absent pulse. Eleven residents (11%)

Table 1. Error Frequency in Diagnosis of Anaphylaxis

Error	CB	CA-1	CA-2	CA-3	*P* Value
Failure to diagnose anaphylaxis					
Committed error	4	18	8	4	.30
Did not commit error	3	34	17	11	
Thought equipment problem					
Committed error	1	10	2	2	.19
Did not commit error	6	42	23	13	
Attempted electrical cardioversion					
Committed error	0	0	1	0	.40
Did not commit error	7	52	24	15	

The *P* value represents the significance of the Fisher exact test for error rate differences between beginning residents (CB and CA-1) *versus* advanced residents (CA-2 and CA-3).

misinterpreted sinus tachycardia (HR 120) as supraventricular tachycardia. One resident attempted to electrically cardiovert the sinus tachycardia. The most frequent incorrect diagnosis was hypovolemia from preoperative fluid deficit or surgical blood loss, as judged by subject verbalizations. In these cases, the anesthesia resident insisted there must be excessive blood loss despite disagreement by the surgeon, which delayed or prevented correct diagnosis.

In the simulated anaphylaxis scenario, hypotension was severe, and treatment of hypotension was essential regardless of whether the diagnosis of anaphylaxis was made (table 2). Although ninety residents (90%) did increase fluid administration, most only opened the intravenous fluid line, delivering at most a few hundred milliliters during the period of hypotension. The majority of residents did not use a pressurized infusion device (93%) and did not start a second intravenous line (64%). The frequency of these errors did not decrease with increasing level of training.

In addition to appropriate administration of intravenous fluids during anaphylaxis-induced hypotension, it is important to decrease the concentration of the inhalation agent and administer epinephrine. Forty percent of residents did not decrease the inhalation agent, 29% did not administer epinephrine, and 16% administered a 1-mg intravenous epinephrine bolus, a potentially arrhythmogenic dose. Of the 14 residents who verbalized

Table 2. Error Frequency in Treatment of Anaphylaxis

Error	CB	CA-1	CA-2	CA-3	*P* Value
Did not increase fluid administration rate					
Committed error	1	6	0	3	.36
Did not commit error	6	46	25	12	
Did not use pressurized infusion device					
Committed error	6	49	24	14	.54
Did not commit error	1	3	1	1	
Did not start second intravenous line					
Committed error	5	29	19	11	.06
Did not commit error	2	23	6	4	
Did not use Trendelenburg position					
Committed error	5	43	19	13	.53
Did not commit error	2	9	6	2	
Did not decrease inhalation agent concentration					
Committed error	3	18	7	12	.16
Did not commit error	4	34	18	3	
Did not administer epinephrine					
Committed error	1	20	5	3	.07
Did not commit error	6	32	20	12	
Administered excessive epinephrine					
Committed error	3	9	2	2	.14
Did not commit error	4	43	23	13	
Administered drug to slow heart rate despite hypotension					
Committed error	0	7	2	1	.36
Did not commit error	7	45	23	14	
Did not administer H1 blocker					
Committed error	6	37	17	10	.36
Did not commit error	1	15	8	5	

The *P* value represents the significance of the Fisher exact test for error rate differences between beginning residents (CB and CA-1) *versus* advanced residents (CA-2 and CA-3).

Table 3. Error Frequency in Management of Bronchospasm

Error	CB	CA-1	CA-2	CA-3	*P* Value
Did not consider differential diagnosis					
Committed error	2	27	11	10	.45
Did not commit error	5	25	14	5	
Did not increase concentration of inhalation agent					
Committed error	6	30	7	5	.002
Did not commit error	1	22	18	10	
Did not administer inhaled bronchodilator					
Committed error	1	3	4	0	.41
Did not commit error	6	49	21	15	

The *P* value represents the significance of the Fisher exact test for error rate differences between beginning residents (CB and CA-1) *versus* advanced residents (CA-2 and CA-3).

their reason for giving 1 mg, 13 gave this dose for "PEA" (pulseless electrical activity) or "severe hypotension," and one gave it for "anaphylaxis."

Ten residents (10%) administered esmolol, labetalol, or adenosine to slow the heart rate despite the presence of severe hypotension at the time. Often the drug was given before the blood pressure cuff was recycled, so the blood pressure reading was 3–5 min old. The incidence of this error (0% of CBs, 14% of CA-1s, 8% of CA-2s, 7% of CA-3s) did not improve significantly with level of training.

Bronchospasm

Numerous problems were observed in the diagnosis of the simulated bronchospasm scenario (table 3). Some residents never arrived at the correct diagnosis. One CA1 and one CA2 performed cricothyroidotomy, while two CA-1 residents diagnosed malignant hyperthermia due to high end-expired carbon dioxide and ordered dantrolene and ice to cool the patient. Fifty percent of residents did not consider a mechanical cause for increased difficulty ventilating, such as endotracheal tube kink, incorrect position, or mucus plugging.

Additional errors were observed in the treatment of bronchospasm. Eight percent did not administer a bronchodilator, and this error did not decrease significantly with training. Forty-eight percent of the residents did not increase the concentration of the inhalation agent to promote bronchodilation. The incidence of this error did improve with training ($P = 0.002$), but 33% of CA-3 residents still made this error.

Myocardial Ischemia

In the case of myocardial ischemia (table 4), errors included 22% omission of pressors, 45% omission of a β1 blocker to slow the heart rate, and 31% omission of nitroglycerin. Of those residents who did administer nitroglycerin, many made significant errors in dosing. Thirty-four percent started with an excessive dose (greater than 0.5 $\mu g \cdot kg^{-1} \cdot min^{-1}$), and 85% did not titrate to the target dose of 1–2 $\mu g \cdot kg^{-1} \cdot min^{-1}$. The frequency of most of these errors did improve with level of training.

Validity and Reliability

Subjects scored from 16 to 81 out of 108 possible points by the long form, with mean 52.1 and SD 14.0 (fig. 1). Short-form scores ranged from 9 to 31 out of 40 possible points, with mean 21.2 and SD 4.6 (fig. 2). Table 5 summarizes the observed progression of scores with

Table 4. Error Frequency in Management of Myocardial Ischemia

Error	CB	CA-1	CA-2	CA-3	*P* Value
Did not treat hypotension in 12 min					
Committed error	5	13	2	2	.013
Did not commit error	2	39	23	13	
Did not treat tachycardia in 12 min					
Committed error	5	27	7	6	.03
Did not commit error	2	25	18	9	
Did not administer nitroglycerin					
Committed error	5	18	6	2	.04
Did not commit error	2	34	19	13	
If did administer nitroglycerin, started with excessive dose:					
Committed error	1	10	4	2	.18
Did not commit error	1	23	15	11	
If did administer nitroglycerin, failed to titrate to 1 to 2 $\mu g \cdot kg^{-1} \cdot min^{-1}$					
Committed error	2	32	16	7	.004
Did not commit error	0	1	3	6	

The *P* value represents the significance of the Fisher Exact test for error rate differences between beginning residents (CB and CA-1) *versus* advanced residents (CA-2 and CA-3).

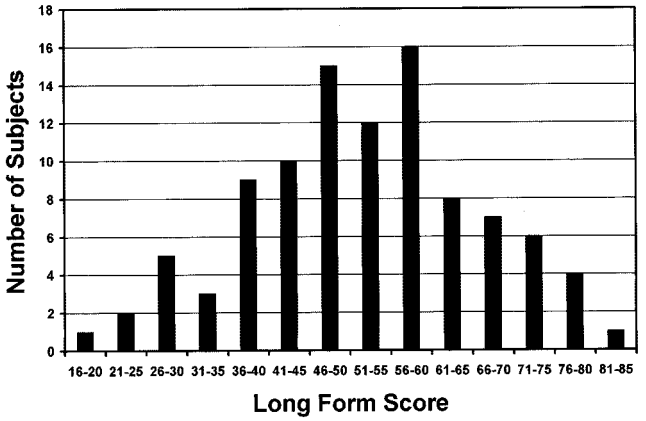

Fig. 1. Distribution of the long-form scores.

level of training. The long-form scores showed significant improvement from the CB to CA-1 yr, while the short-form change in scores from CB to CA-1 yr was not statistically significant. Long- and short-form scores both significantly improved from CA-1 to CA-2 yr. There was no significant further improvement from the CA-2 to CA-3 yr for either form.

Construct-related validity is supported by progression of scores with level of training and subjects' rating the realism of the simulation as very good (3.47 out of possible 4). Criterion-related validity is supported by moderate correlation of simulator scores with departmental faculty evaluations, ABA written in-training scores, and departmental mock oral board scores (table 6).

Reliability of the simulator assessment is demonstrated by very good internal consistency as measured by the Cronbach α statistic (0.71–0.76, table 7). A Cronbach α greater than 0.8 is considered excellent consistency.[4] In addition, excellent interrater reliability was measured (correlation = 0.94–0.96; $P < 0.01$; $\kappa = 0.81$–0.90, table 8). The third independent evaluator from an outside institution had nearly identical ratings as the two evaluators from the testing institution. No significant differences were found in long- or short-form scores between subjects evaluated using METI *versus* MedSim simulators for any level of training.

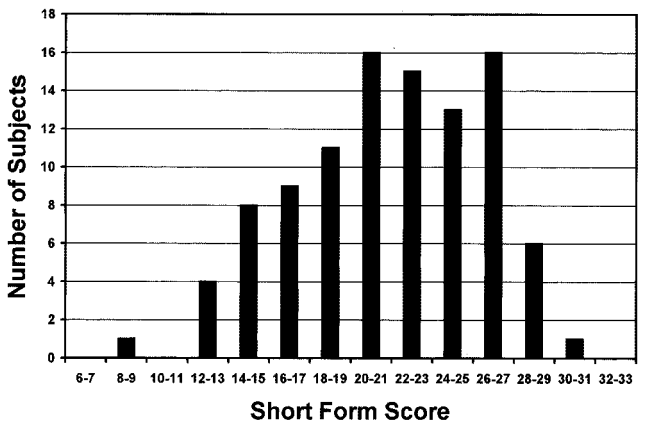

Fig. 2. Distribution of the short-form scores.

Discussion

The simulator evaluation of anesthesia residents was constructed to assess their ability to manage four relatively commonly discussed and taught anesthetic critical incidents. It is reasonable to expect competent residents to be able to manage esophageal intubation, anaphylaxis, bronchospasm, and myocardial ischemia. We found that all 99 residents adequately managed the esophageal intubation. Correct diagnosis was based on lack of exhaled carbon dioxide and absent breath sounds. While use of the "paramedic trainee" to place the endotracheal tube may have artificially raised suspicion, from this study it appears that residency training for recognition of esophageal intubation is adequate.

Anaphylaxis occurs infrequently, and it is understandable that there are more management errors associated with it. However, errors in the management of hypotension, the occurrence of which is not limited to anaphylaxis, were common and in general did not significantly improve with training. This included insufficient administration of intravenous fluids and failure to reduce the inhalation agent concentration. Further resident drill in this area or dissemination of management guidelines for hypotension, such as those developed by the VA National Center for Patient Safety,[5] may reduce these error rates.

We also observed that 15% of residents assumed that recycling of NIBP cuff, low NIBP readings, or absent pulse oximeter wave forms were due to equipment failure and did not check the patient's pulse to confirm monitor readings. This may be an artifact of a simulation environment in which subjects are more skeptical of equipment, or it may reflect an overreliance on automation. However, when monitors give unexpected results, management may be aided by clear troubleshooting guidelines, which could include checking the patient's pulse and color. A previous study by Schwid and O'Donnell[6] found that almost half of the subjects (14 of 30) made significant management errors when they noted tachycardia but did not remeasure the blood pressure. The authors in that study suggested that the result may have been an artifact of testing with the screen-based simulator. The current study confirms the problem in mannequin-based simulation with 10 residents (10%) administering a drug to slow the heart rate before rechecking the blood pressure to rule out hypotension. The simulated scenario is remarkably similar to a recent negative anesthetic outcome resulting in charges of negligent homicide.[7] In that case, the resident administered esmolol during cardiovascular collapse due to a reaction to clindamycin. Our study shows that this may be a fairly

Table 5. Progression of Simulator Scores with Level of Training

Level of Training	Number of Subjects	Long-form Score Mean (SD)	Short-form Score Mean (SD)
CB	7	36.0 (12.0)	17.7 (5.0)
CA-1	52	49.6 (13.4)	20.2 (4.4)
		$P = 0.01$ compared to CB score	$P = 0.15$ compared to CB score
CA-2	25	58.3 (13.2)	22.8 (4.4)
		$P < 0.01$ compared to CA-1 score	$P = 0.02$ compared to CA-1 score
CA-3	15	57.9 (9.8)	23.8 (3.4)

frequent inappropriate response, a system failure, and is not limited to the failure of a single individual. Additional drill or attention to published guidelines in the management of tachycardia and hypotension may benefit anesthesia residency training for this problem.

The administration of a large bolus of intravenous epinephrine during the anaphylactic reaction was another potential error. In our scenario, the simulated patient exhibited sinus tachycardia of 120 beats/min, blood pressure of 50–60 systolic, and correspondingly weak or absent pulse. Several residents diagnosed pulseless electrical activity that by ACLS guidelines should be treated with 1 mg intravenous epinephrine.[8] Although anaphylaxis is one of the causes of pulseless electrical activity, we believe that 1 mg epinephrine in this case may be excessive and lead to a ventricular arrhythmia.

While mechanical obstruction is the most common cause for difficulty ventilating, failure to consider at least one mechanical cause during simulated bronchospasm was common and did not improve through training. While it is hard to fault residents for jumping to the correct answer, we are concerned that mechanical obstruction would have been missed by some of these residents. Although most residents did make the correct diagnosis of bronchospasm, treatment errors were common. Eight percent did not administer an inhaled β agonist, and the frequency of this error did not decrease significantly with training. Failing to increase the inhalation agent decreased significantly through training but was still common.

Management of myocardial ischemia required support of blood pressure to maintain coronary perfusion, followed by reduction of heart rate to improve myocardial oxygen balance, and dilation of coronary arteries with nitroglycerin. The most common errors were omission of a β blocker and inappropriate use or omission of nitroglycerin. Reasons residents did not give a β blocker included concern about the patient's chronic obstructive pulmonary disease and concern about further decreasing blood pressure. The group of 32 anesthesiologists involved in this study reached a consensus that it was reasonable to expect administration of a β1-specific β blocker to not unduly exacerbate COPD nor cause a major exacerbation of hypotension if blood pressure was supported with pressors. In this simulation, the patient responded well to administration of fluid and pressors. The main reason verbalized for not giving nitroglycerin included concern for further lowering the blood pressure, which also could have been addressed by pressor support and careful titration of the nitroglycerin. It has been argued that there are many ways of successfully managing this scenario. The above errors are not absolute. However, it is generally desirable to achieve all three treatment goals: maintenance of coronary perfusion, slowing the heart rate, and dilation of coronaries. In this simulation, these goals were achievable. Advanced residents demonstrated better performance than beginning residents in the management of myocardial ischemia, indicating an area of success in residency training.

The progression of simulator scores from beginning residents to senior residents supports construct-related validity. The measured scores agreed with our expectations that management of these four events improves during the first 2 yr of anesthesia training since training for these events typically occurs early in the anesthesia residency. Similarly, Devitt *et al.*[9] were able to demonstrate the validity of simulator assessment by discriminating between performance of university-based anesthesiologists, community-based anesthesiologists, residents, and medical students.

In addition, construct-related validity of the simulator as an evaluation tool is supported by the survey following the simulator session since the subjects rated the simulator scenarios as very realistic. However, several shortcomings of the mannequin-based simulators were noted by subjects and evaluators, including poor quality of breath sounds, ambiguous electrocardiogram wave forms, and inaccuracy of the capnogram during slow exhalation.

Table 6. Correlation of Simulator Scores with Other Evaluation Methods

Grading Form	Departmental Faculty Evaluations	Written ABA In-Training Exam	Departmental Mock Oral Board Exam
Long form	0.37, $P < 0.01$	0.44, $P < 0.01$	0.47, $P < 0.01$
Short form	0.41, $P < 0.01$	0.49, $P < 0.01$	0.44, $P < 0.01$

Table 7. Internal Consistency of Simulator Scores

Grading Form	Rater 1	Rater 2	Rater 3
Cronbach α statistic long form	0.76	0.71	0.72
Cronbach α statistic short form	0.75	0.71	0.71

Criterion-related validity is supported by the moderate correlation between the simulator scores and departmental evaluations, written ABA in-training examination, and mock oral board examination. We expected only moderate correlations since each of these modalities measures different aspects of anesthesia knowledge and skills. Our results were similar to correlations between faculty evaluation and written examination (0.38), mock oral examination and faculty evaluation (0.43), and mock oral and written examinations (0.47) found by Schubert *et al.*[10] Morgan and Cleave-Hogg[11] found much lower correlation between their simulator evaluation and written test (0.19) or faculty ratings (0.04) for medical students on an anesthesia rotation. The reason for the lower correlations for the medical students is not clear, but we believe that the four scenarios used in our assessment are targeted to a higher level of training and are not appropriate for evaluation of medical students.

Reliability is the ability of a test to yield reproducible results. Internal consistency examines subject performance on different components of a test, while interrater reliability determines the extent that the examiner influences the score. We also evaluated the extent that the type of simulator (MedSim *vs.* METI) affected the score. Both the long- and short-form grading checklists had very good internal consistency (Cronbach α statistic 0.71–0.76). This falls slightly below internal consistency of 0.80, considered adequate for high-stakes examinations, such as board certification. The Cronbach α statistic often identifies items that if removed from the test would improve internal consistency. Removal of the esophageal intubation scenario from the long form increased Cronbach α by 0.01–0.02, but elimination of other items did not increase internal consistency. In comparison, Devitt *et al.*[12] measured internal consistency of 0.27 for their 10-scenario simulator evaluation with improvement to 0.66 with elimination of four of the scenarios.

We measured excellent interrater reliability for the simulator assessment of diagnostic and therapeutic response to anesthetic emergencies. Our results agree with those of Devitt[13] and Morgan[11] and Gaba's[14] observations of technical ratings. We also compared reliability of raters 1 and 2, who may know the subject, *versus* rater 3, who did not know the subject and found no differences. Since interrater reliability was so high, we suggest that when a detailed checklist is used, only one rater is necessary to assign a score to the simulator assessment. We further examined scores for subjects tested on MedSim simulators *versus* METI simulators and found no difference for level of training. However, both simulators have unrealistic aspects that may negatively impact the quality of the assessment. These have been mentioned above and must be corrected before the simulator is used for high-stakes evaluation. If simulator realism were improved, good residents would likely score well consistently on all scenarios (which would mathematically increase measures of internal consistency) because they would be less likely to be mislead by unrealistic and inaccurate simulator cues.

Two different grading forms were used in this study. The long form was quite complex, with weighted and negative points, and the short form was simple, with only single-point checklist items. Both forms produced scores with an adequate spread of simulator scores to distinguish superior, average, and unacceptable performance. The validity and reliability statistics for these two forms were almost identical. For simplicity, we recommend single-point checklist grading forms for future evaluation.

Potential limitations of this study include selection bias and the simulation environment. The selection process may be skewed for either strong or weak residents, and this may have affected the overall error rates, but it is unlikely that it significantly altered the patterns of errors identified. Hypervigilance or cavalier behavior in response to the simulator environment may explain the rare or unexpected behavior seen in this study. However, it most likely does not account for common patterns of errors seen in a group of residents who overall performed reasonably.

Patterns of errors were identified not only in the management of rare critical incidents, such as anaphylaxis, but also in the management of more common situations, such as severe hypotension, bronchospasm, and myocardial ischemia. Although the overall management of these four scenarios did improve with level of training, common patterns of errors persisted throughout training. The results of this study and previous studies[15–17] should call attention to these particular common errors so that

Table 8. Inter-Rater Reliability of Simulator Scores

Grading Form	Rater 1–Rater 2	Rater 1–Rater 3	Rater 2–Rater 3
Pearson correlation long form	0.94, $P < 0.01$	0.95, $P < 0.01$	0.96, $P < 0.01$
Pearson correlation short form	0.96, $P < 0.01$	0.96, $P < 0.01$	0.96, $P < 0.01$
κ Statistic long form	0.81	0.87	0.90
κ Statistic short form	0.83	0.87	0.90

they can be more effectively addressed in residency training. They also point to the need for continued development and dissemination of clinical guidelines to serve as a starting point for treating life-threatening anesthesia complications. A follow-up study after residents have been exposed to NCPS or comparable guidelines would be helpful to evaluate the effectiveness of this strategy to prepare anesthesiologists to manage critical incidents.

It is also important to note that this simulator assessment evaluates only one aspect of anesthetic care and does not address other components of clinical competence, such as preoperative evaluation, formulation of the anesthetic plan, or communication with patients and other healthcare providers. In addition, based on prior education research on OSCE (objective structured clinical examination), more than four scenarios would provide better generalizability of the results.[1] We conclude that mannequin-based anesthesia simulators show promise as a valid and reliable method to evaluate anesthesia residents, but further improvement is necessary in the realism of the simulators, design of the testing scenarios, and grading forms prior to their use for high-stakes examination purposes.

References

1. Ziv A, Ben-David MF, Sutnick AI, Gary NE: Lessons learned from six years of international administrations of the ECFMG's SP-based clinical skills assessment. Acad Med 1998; 73:S83–90
2. Oosterhof A: Developing and Using Classroom Assessments, 2nd edition. Upper Saddle River, NJ, Prentice Hall, 1999, pp 35–44
3. Schwid HA, Rooke GA, Ross BK, Michalowski P: Screen-based anesthesia simulation with debriefing improves performance in a mannequin-based anesthesia simulator. Teaching Learning Med 2001; 13:92–96
4. Nunnally JC: Psychometric Theory, 2nd edition. New York, McGraw Hill, 1978, pp 245–6
5. Department of Veterans Affairs, Veteran Health Administration: Cognitive Aids for Anesthesiology. VA National Center for Patient Safety, January 2002
6. Schwid HA, O'Donnell D: Anesthesiologists' management of simulated critical incidents. ANESTHESIOLOGY 1992; 76:495–501
7. Werth B: A marine's private war. The New Yorker 2000; Dec 18:64–77
8. International Consensus on Science: A guide to the international ACLS algorithms. Circulation 2000; 102(suppl I):I142–57
9. Devitt JH, Kurrek MM, Cohen MM, Cleave-Hogg D: The validity of performance assessments using simulation. ANESTHESIOLOGY 2001; 95:36–42
10. Schubert A, Tetzlaff JE, Tan M, Ryckman JV, Mascha E: Consistency, inter-rater reliability, and validity of 441 consecutive mock oral examinations in anesthesiology. ANESTHESIOLOGY 1999; 91:288–98
11. Morgan PJ, Cleave-Hogg D: Evaluation of medical students' performance using the anesthesia simulator. Med Educ 2000; 34:42–45
12. 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
13. Devitt JH, Kurrek MM, Cohen MM, Fish K, Fish P, Murphy PM, Szalai JP: Testing the raters: Inter-rater reliability of standardized anaesthesia simulator performance. Can J Anaesth 1997; 44:924–8
14. 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
15. Gaba DM, DeAnda A: The response of anesthesia trainees to simulated critical incidents. Anesth Analg 1989; 68:444–51
16. DeAnda A, Gaba DM: Role of experience in the response to simulated critical incidents. Anesth Analg 1991; 72:308–15
17. Byrne AJ, Jones JG: Responses to simulated anaesthetic emergencies by anaesthetists with different durations of clinical experience. Br J Anaesth 1997; 78:553–6

Appendix A: Anesthesia Simulator Research Consortium

Cleveland Clinic Foundation, Cleveland, Ohio: Karen Steckner, M.D., Staff Anesthesiologist, Department of General Anesthesia, Cleveland Clinic; Sawan AlHaddad, M.D., Staff Anesthesiologist, Department of General Anesthesia, Cleveland Clinic; Judith Haas, M.D., Staff Anesthesiologist, Metrohealth Medical Center, Cleveland, Ohio; J. Victor Ryckman, M.D., Staff Anesthesiologist, Department of General Anesthesia, Cleveland Clinic; John Tetzlaff, M.D., Program Director, Division of Anesthesiology and Critical Care Medicine, Cleveland Clinic, and Associate Professor, Ohio State University, Cleveland, Ohio; Julie Tome, M.D., Staff Anesthesiologist, Department of Cardiothoracic Anesthesia, Cleveland Clinic. *Indiana University, Indianapolis, Indiana:* Jeffrey L. Lane, M.D., Assistant Professor of Clinical Anesthesia, Department of Anesthesiology; Andrew Stasic, M.D., Associate Professor of Clinical Anesthesia, Department of Anesthesiology; Susan Baldwin, Lab Technician, Anesthesia Simulator Laboratory. *Pennsylvania State University, Hershey, Pennsylvania:* W. Bosseau Murray, M.D., Professor of Anesthesiology, Department of Anesthesiology; Arthur J. L. Schneider, M.D., Professor of Anesthesiology, Department of Anesthesiology; Clark Venable, M.D., Staff Anesthesiologist, Department of Anesthesiology; Jody Henry, Simulator Research Support Assistant, Department of Anesthesiology. *University of California–Los Angeles, Los Angeles, California:* Randolph H. Steadman, M.D., Associate Clinical Professor and Vice Chair, Department of Anesthesiology; Philip R. Levin, M.D., Assistant Clinical Professor, Department of Anesthesiology; Yue Ming Huang, M.H.S., Research Associate, Department of Anesthesiology. *University of Kansas, Kansas City, Kansas:* Stephen Tarver, M.D., Associate Professor of Anesthesiology, Department of Anesthesiology; Gregory Unruh, M.D., Associate Professor of Anesthesiology, Department of Anesthesiology. *University of Pittsburgh, Pittsburgh, Pennsylvania:* Rita M. Patel, M.D., Associate Professor of Anesthesiology, Department of Anesthesiology; William McIvor, M.D., Assistant Professor of Anesthesiology, Department of Anesthesiology; Helene Finegold, M.D. Assistant Professor of Anesthesiology, Department of Anesthesiology; Carole Cox, B.S., B.A., Simulation Technologist, Department of Anesthesiology. *University of Rochester, Rochester, New York:* David H. Stern, M.D., Clinical Associate Professor of Anesthesiology, Department of Anesthesiology; Lindsey C. Henson, M.D., Ph.D., Associate Professor of Anesthesiology, Department of Anesthesiology, and Senior Associate Dean for Medical Education; Ilya Shekhter, M.S., Senior Engineer, Department of Anesthesiology. *University of Washington, Seattle, Washington:* Howard A. Schwid, M.D., Professor of Anesthesiology, Department of Anesthesiology; Brian K. Ross, M.D., Ph.D., Associate Professor of Anesthesiology, Department of Anesthesiology; G. Alec Rooke, M.D., Ph.D. Professor of Anesthesiology, Department of Anesthesiology; Piotr Michalowski, M.D., Clinical Assistant Professor of Anesthesiology, Department of Anesthesiology; Andrew Naluai-Cecchini, M.S., Research Engineer, Department of Anesthesiology. *Wake Forest University, Winston-Salem, North Carolina:* Michael Olympio, M.D., Associate Professor of Anesthesiology, Department of Anesthesiology; Sylvia Y. Dolinski, M.D., F.C.C.P., Assistant Professor of Anesthesiology, Department of Anesthesiology; Margaret F. Brock, M.D., Assistant Professor of Anesthesiology, Department of Anesthesiology; John A. Thomas, M.D., Assistant Professor of Anesthesiology, Department of Anesthesiology; Ian Saunders, Cert. A.T., Department of Anesthesiology. *West Virginia University, Morgantown, West Virginia:* Kathleen Rosen, M.D., Associate Professor of Anesthesiology, Department of Anesthesiology; Elizabeth Sinz, M.D., Assistant Professor of Anesthesiology, Department of Anesthesiology; John Barbaccia, M.D., Associate Professor of Anesthesiology, Department of Anesthesiology; William A. Kofke, M.D., Professor of Anesthesiology, Department of Anesthesiology.

Appendix B: Long-form Scoring System

Case 1: Esophageal Intubation Followed by Anaphylaxis

Paramedic student performs esophageal intubation.
Time laryngoscope is removed from mouth: ___
___ Lack of CO_2 communicated
___ O_2 saturation communicated
___ Breath sounds auscultated
___ Stomach auscultated
___ Laryngoscopy for diagnosis
___ Notify team that must reintubate
Time resident reestablishes ventilation: ___
Elapsed time to reestablish ventilation: ___
< 2 min, 8 pts
2–5 min, 5 pts
5–8 min, 3 pts
> 8 min, 0 pts
Airway protection:
___ Esophageal tube left in place until endotracheal tube placed, or tube pulled and cricoid pressure used
___ Gastric tube placed to evacuate stomach

A few minutes after esophageal intubation is corrected, surgeon asks for antibiotic and complains about relaxation. Anaphylaxis without bronchospasm is triggered, HR increases to 120, and BP falls to 50–60 systolic.

Time HR hits 100: ___
___ Resident communicates tachycardia
___ BP is rechecked before treating HR
___ Checks breath sounds
___ Airway pressure is communicated
___ Checks skin color—tell that flushed
___ O_2 saturation communicated
___ Trendelenburg
___ Pressor other than epi
___ Notifies surgeon that there is a problem
Time agent turned off: ___
Elapsed time to turn off agent: ___
< 5 min, 3 pts
5–8 min, 2 pts
8–12 min, 0 pts
> 12 min or not done, −2 pts
___ 100% O_2
Time fluids increased: ___
Elapsed time to increase fluids: ___
< 5 min, 5 pts
5–8 min, 3 pts
8–12 min, 0 pts
> 12 min or not done, −4 pts
___ Pressure bag (2 pts)
___ Asks for second intravenous line (2 pts)
Time epi administered: ___
Elapsed time to administer epi: ___
< 5 min, 5 pts
5–8 min, 3 pts
8–12 min, 0 pts
> 12 min, −4 pts
No epi administered, −5 pts
< 20 μg, 1 pt
20–200 μg, 3 pts
201–500 μg, 2 pts
501–999 μg or none, 0 pt
1,000 or more μg, −5 pts
___ Second bolus of epi
___ Calls for help
___ Informs surgeon that possible anaphylaxis
___ Blood gas ordered
___ H1 blocker
___ H2 blocker

Case 2: Bronchospasm Followed by Myocardial Ischemia

Two minutes after intubation, bronchospasm develops.
___ Change in airway pressure communicated
___ Change in tidal volume communicated
___ Change in capnogram communicated
___ Listen to breath sounds
___ O_2 saturation is communicated
___ Try bag ventilation
___ Notify team of problem
___ Check depth of ETT
___ Look for ETT kink
___ Pass suction catheter or fiberoptic bronchoscopy (3 pts)
___ None of the diagnostic maneuvers (−3 pts)
___ Increase inhalation agent (2 pts)
___ Decrease inhalation agent (−2 pts)
___ Inhaler—appropriate agent (2 pts)
___ Inhaler—appropriate dose (2 pts)
___ Inhaler— used spacer correctly
___ Administer NMB
___ Ketamine
___ Lidocaine
___ Inhaler repeated appropriately

A few minutes after bronchospasm is corrected, the patient develops ST-segment depression, frequent PVCs, hypotension, and tachycardia.

Time ST-segment changes begin: ___
___ ST changes communicated
___ PVCs communicated
___ Tachycardia communicated
___ Hypotension communicated
___ Checks breath sounds
___ Checks airway pressure
___ Checks skin color
___ O_2 saturation is communicated
Time administer pressor: ___
Elapsed time to administer pressor: ___
< 8 min, 5 pts
8–12 min, 3 pts
> 12 min, 0 pts
No pressor, −4 pts
___ Pressor—phenylephrine (3 pts) or ephedrine (2 pts) or dopamine (2 pts)
___ Increases fluids
___ Decrease inhalation agent
Time nitroglycerin started: ___
Elapsed time to administer NTG: ___
< 8 min, 5 pts
8–12 min, 3 pts
> 12 min, 0 pts
NTG not used, −4 pts
Start at 0.25–0.5 $\mu g \cdot kg^{-1} \cdot min^{-1}$ (+2 pts)
Titrate up to 1–2 $\mu g \cdot kg^{-1} \cdot min^{-1}$ (+2 pts)
Start at 1 or more $\mu g \cdot kg^{-1} \cdot min^{-1}$ (−2 pts)
Provide adequate analgesia: morphine–fentanyl (3 pts)
Slow HR: esmolol (2 pts) or other β blocker (1 pt)
___ Treat PVCs: lidocaine

Appendix C: Short-form Scoring System

Case 1: Esophageal Intubation Followed by Anaphylaxis

Time laryngoscope is removed from mouth: ___

1. ___ Time to reestablish ventilation < 8 min
2. ___ Time to reestablish ventilation < 5 min
3. ___ Time to reestablish ventilation < 2 min
4. ___ Airway protected during reintubation
5. ___ Stomach emptied after reintubation

Anaphylaxis: time HR reaches 100: ___

6. ___ Trendelenburg
7. ___ Time to increase fluids < 12 min
8. ___ Time to increase fluids < 8 min
9. ___ Time to increase fluids < 5 min
10. ___ Use pressure bag for fluids
11. ___ Asks for second intravenous line
12. ___ Time to administer epi < 12 min
13. ___ Time to administer epi < 8 min
14. ___ Time to administer epi < 5 min
15. ___ Initial epi administered ≤ 500 μg
16. ___ 20 μg $\leq$ (initial epi administered) ≤ 200 μg
17. ___ 50 μg $\leq$ (initial epi administered) ≤ 200 μg
18. ___ Second dose of epi administered
19. ___ Calls for help
20. ___ Informs surgeon that possible anaphylaxis
21. ___ H1 blocker
22. ___ H2 blocker

Case 2: Bronchospasm Followed by Myocardial Ischemia

Two minutes after intubation, bronchospasm develops.

23. ___ Use appropriate inhaler
24. ___ Use inhaler circuit adaptor correctly
25. ___ Inhaler administration repeated
26. ___ Deepen inhalation agent for bronchospasm
27. ___ Administer ketamine
28. ___ Administer lidocaine

Time ST-segment changes begin: ___

29. ___ Increase fluid administration rate
30. ___ Decrease inhalation agent
31. ___ Administer pressor < 12 min
32. ___ Administer pressor < 8 min
33. ___ Pressor administered was phenylephrine
34. ___ Administer NTG < 12 min
35. ___ Administer NTG < 8 min
36. ___ Start NTG at 0.25–0.5 $\mu g \cdot kg^{-1} \cdot min^{-1}$
37. ___ Titrate NTG up to 1–2 $\mu g \cdot kg^{-1} \cdot min^{-1}$
38. ___ Provide adequate analgesia: morphine–fentanyl
39. ___ Slow HR with esmolol or metoprolol
40. ___ Administer lidocaine for PVCs