Anesthesiology 2003; 99:1270–80

Reliability and Validity of a Simulation-based Acute Care Skills Assessment for Medical Students and Residents

John R. Boulet, Ph.D., David Murray, M.D.,† Joe Kras, M.D.,‡ Julie Woodhouse, Bs.N.,§ John McAllister, M.D.,‡ Amitai Ziv, M.D.‖*

Background: **Medical students and residents are expected to be able to manage a variety of critical events after training, but many of these individuals have limited clinical experiences in the diagnosis and treatment of these conditions. Life-sized mannequins that model critical events can be used to evaluate the skills required to manage and treat acute medical conditions. The purpose of this study was to develop and test simulation exercises and associated scoring methods that could be used to evaluate the acute care skills of final-year medical students and first-year residents.**

Methods: **The authors developed and tested 10 simulated acute care situations that clinical faculty at a major medical school expects graduating physicians to be able to recognize and treat at the conclusion of training. Forty medical students and residents participated in the evaluation of the exercises. Four faculty members scored the students/residents.**

Results: **The reliability of the simulation scores was moderate and was most strongly influenced by the choice and number of simulated encounters. The validity of the simulation scores was supported through comparisons of students'/residents' performances in relation to their clinical backgrounds and experience.**

Conclusion: **Acute care skills can be validly and reliably measured using a simulation technology. However, multiple simulated encounters, covering a broad domain, are needed to effectively and accurately estimate student/resident abilities in acute care settings.**

AN important goal in healthcare education is to develop teaching and evaluation techniques that measure a provider's performance in settings that reflect clinical practice.[1–5] Historically, standardized patients have often been used to assess the history and physical examination skills of medical students and graduate physicians.[1–4] Likewise, the knowledge required to manage acute care scenarios has traditionally been assessed using paper-and-pencil examinations. Unfortunately, critical care events are not easily modeled with standardized patients, and knowledge of acute care management strategies does not ensure that the physician can actually provide treatment. Nevertheless, medical students and graduate physicians are expected to be able to manage a variety of acute medical situations after medical school and internship. The importance of these acute care management skills is reflected in the course objectives of virtually all Liaison Committee for Medical Education–accredited medical schools and on step 3 of the U.S. Medical Licensing Examination. Furthermore, almost one fourth of the U.S. Medical Licensing Examination content is devoted to the history, physical examination, diagnosis, and clinical interventions and therapy for urgent care situations. As a result, the ability to develop a high-fidelity simulated environment to assess skills in managing these life-threatening conditions would be an important contribution to medical student and graduate physician education and evaluation.[2,5]

Previous studies indicate that simulation exercises can be used to recreate complex crisis situations in the operating room.[6–13] This technology provides a realistic setting in which to train anesthesia providers.[6–12] The ability to create scenarios that are reproducible over time, highly realistic, and, for the purpose of evaluating physicians, eliminate threats to patient safety are key advantages to this type of training.[6–8] From an evaluation perspective, one additional advantage of life-sized mannequins is that a trainee's skill in acute diagnosis can be integrated with the need to manage and stabilize a variety of medical and surgical conditions.[2,5–17] Furthermore, the simulation exercises currently used in anesthesia could be modified to teach and evaluate graduate physicians in a variety of disciplines.[2,5,16–21]

Currently, crisis management exercises used in anesthesia training often use a complicated, prolonged encounter that is used to evaluate a trainee's technical and behavioral skills.[8,13,14] This type of exercise is an effective training technique for the advanced practitioner but would be more difficult to use to evaluate medical students and graduate physicians. For the latter purpose, a brief, simulated acute care condition could be used to assess a discreet set of diagnostic and therapeutic skills. Depending on content and reliability considerations, a number of these simulations would be required to assess the range of acute care skills expected of a graduate physician.

The purpose of this study was to develop and test simulation exercises that could be used to evaluate the

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

* Assistant Vice President, Research and Evaluation, Educational Commission for Foreign Medical Graduates, Philadelphia, Pennsylvania. † Professor, ‡ Assistant Professor, § Nurse Clinician, Department of Anesthesiology, Washington University School of Medicine. ‖ Director, Israel Center for Medical Simulation, Tel-Hashomer, Israel.

Received from the Department of Anesthesiology, Washington University School of Medicine, St. Louis, Missouri. Submitted for publication January 6, 2003. Accepted for publication June 12, 2003. Supported by grant No. 24-9899 from the National Board of Medical Examiners Medical Education Research Fund, Philadelphia, Pennsylvania (to Dr. Murray), and an Education Grant from the Foundation for Anesthesia Education and Research, Mayo Clinic, Rochester, Minnesota (to Dr. Kras).

Address reprint requests to Dr. Boulet: Research and Evaluation, Educational Commission for Foreign Medical Graduates, 3624 Market Street, Philadelphia, Pennsylvania, 19104-2685. Address electronic mail to: jboulet@ecfmg.org.

clinical skills of final-year medical students and recent graduates. More specifically, efforts were directed at providing evidence to support the reliability and validity of scores procured from these simulation-based exercises. For reliability, both the consistency of individual performances over a number of simulation exercises and the uniformity of multiple expert raters' scoring of specific exercises were studied. For validity, the ability of the scores from individual simulation exercises to discriminate between low- and high-ability examinees was assessed. In addition, the performances of individuals with varying degrees of experience and training were contrasted. It was hypothesized that relevant, content-related experience would be positively associated with performance outcomes.

Materials and Methods

Simulator

This study was conducted in our laboratory using a life-size simulator developed by MEDSIM-EAGLE® (MedSim USA, Inc., Ft. Lauderdale, FL). The simulator uses a software program that determines cardiac and respiratory physiologic and pharmacologic responses that imitate acute medical conditions such as anaphylaxis, hemorrhage, pneumothorax, cardiac tamponade, and others. 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, noninvasive blood pressure, electrocardiogram, eyelid and pupillary responses, and invasive hemodynamic measures such as arterial, 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 computer software provides 25 predefined patients (conditions include healthy, morbid obesity, mitral stenosis, coronary artery disease, and others). More than 30 critical events can be programmed to occur with each of the patients. The combination of predefined patients and events that can be varied using features such as speed of onset, severity of symptoms, and response to therapy offers a method to tailor the simulation to the educational level of the trainee.

Scenario Development

Simulation exercises (cases) were developed through a standardized, multistage process. In the initial stage, a list of clinical cases was developed on the basis of objectives of the current medical school curriculum at Washington University (St. Louis, Missouri) and the capabilities of the simulator to effectively recreate the specific clinical environment. To ensure content representativeness, this list was cross-referenced with the U.S. Medical Licensing Examination step 3 evaluative framework. Scenarios were selected from a wide range of clinical contexts within the critical care domain. Only cases that could be realistically simulated on the patient simulator were considered for development.

Four faculty members reviewed the list of acute care scenarios. All four of the faculty members were on the curriculum committee and directed the medical student rotation for their respective departments or divisions (Emergency Medicine, Surgery, Anesthesia, and Trauma Surgery). In addition, they were all actively involved in the teaching as well as the evaluation of medical students during the student rotations on their respective clinical services. The faculty members were given the task of prioritizing individual simulation exercises based on their expert opinions regarding the potential utility of the task for measuring acute care skills.

From this initial list, 10 scenarios were selected and modeled using the simulator. Appropriate patient histories and physical examination findings were developed (table 1). The goal of each of the scenarios was to define both the skill and the content domains that were relevant and unique to critical and emergency care. The clinical findings and the responses of the simulated patient were modeled to achieve an acute situation that required both diagnosis and initiation of treatment in a 5-min period. The scenario development also included scripting regarding how often and what verbal responses should be provided either independently or in reply to a candidate's questions (*e.g.*, the nature and severity of pain, shortness of breath or lightheadedness, responses to physical examination, or therapeutic treatments). This script also defined the procedures that candidates must perform and actions that would be provided based on a verbal order from the trainee. For example, a request for laboratory studies such as a type and cross-match or blood gas were considered verbal orders that an accompanying nurse or technician would obtain for the trainee. The request results were not available during the simulation, but certain key items were available for interpretation by the trainee when requested (*e.g.*, electrocardiogram suggestive of anterior wall ischemia in myocardial ischemia; chest radiograph in pneumothorax; radiograph showing fractured femur in simulated patient with hemorrhagic hypotension). The mannequin was connected to a monitor, and a continuous electrocardiographic tracing was provided. Blood pressure and oxygen saturation were continuously available *via* the monitor if requested by the trainee. Clinical examinations such as auscultation, palpation, and neurologic evaluation and therapeutic actions such as administration of oxygen, drugs, needle or chest tube insertion, bag and mask ventilation, or tracheal intubation were actions expected of the trainee. The faculty defined an initial list

Table 1. Simulation Scenarios

Scenario (Vital Signs on Request, Features of Physical Examination)	Set 1	Set 2
Trauma–hemorrhagic hypotension–femur fracture: A previously healthy 25-yr-old man arrives in the ED after a construction accident (a steel girder fell on his left thigh). He has left thigh pain and arrives in the ED 20 min after the accident. Electrocardiogram-HR = 130 beats/min, RR = 18 breaths/min (BP = 80/60 mmHg, pulse = 130, oxygen saturation = 98%).	X	X
Myocardial infarction: A 61-yr-old man is admitted with "crushing" chest pain of increasing severity for the past 4 h. He is 5′8″ tall and weighs 100 kg. Patient is diaphoretic. Electrocardiogram-HR = 110 beats/min—approximately 8 PVCs/min, RR = 18 breaths/min (BP = 180/100 mmHg, pulse = 110, oxygen saturation = 97%).	X	X
Pneumothorax in closed-chest trauma: A 25-yr-old, 70-kg man fell off his bicycle in Forest Park. He walked into the ED with dyspnea. Electrocardiogram-HR = 145 beats/min. He is alert but dyspneic, with HR = 145 beats/min, RR = 30 breaths/min BP = 100/60 mmHg, oxygen saturation = 86%).	X	
Ectopic pregnancy–hypovolemia and abdominal pain: A 22-yr-old woman arrives in the ED with a 12-h history of abdominal pain. She reports pain and lightheadedness. Electrocardiogram-HR = 145 beats/min, RR = 22 breaths/min, (BP = 75/50 mmHg, oxygen saturation = 100%, temperature = 37.9°C).	X	
Cerebral hemorrhage–herniation: A 66-yr-old woman is brought to the ED unresponsive. She could not be awakened this morning. She has had no major medical problems in the past. RR = 8 breaths/min electrocardiogram-HR = 49 beats/min (BP = 200/80 mmHg, oxygen saturation = 96%, patient examination reveals unresponsive dilated left pupil).	X	
Ventricular tachycardia: An 87-yr-old woman is brought from a nursing home with chest pain and shortness of breath. She is receiving nitroglycerin ointment daily and 50 mg oral metoprolol tartrate daily. Electrocardiogram-HR = 170 beats/min (BP = 70/40 mmHg).	X	
Respiratory failure–intubation required: You are called to see a 68-yr-old woman who was admitted with dyspnea as a result of an exacerbation of chronic bronchitis. She is unresponsive, with rapid shallow breathing and diaphoresis. HR = 126 beats/min, RR = 26 breaths/min (BP 130/95 mmHg, temperature = 37.8°C, oxygen saturation = 78%, ABG, if requested, P_{O_2} = 48 mmHg, P_{CO_2} = 56 mmHg, pH = 7.26).		X
Asthma exacerbation: A 15-yr-old girl with asthma is brought to the ED with severe dyspnea. She reports increasing dyspnea despite use of inhalers. Dyspnea prevents full-sentence responses. Electrocardiogram-HR = 135 beats/min, RR = 30 breaths/min, temperature = 36°C (oxygen saturation = 88%).		X
Pulsatile abdominal mass: A 68-yr-old man with severe abdominal pain is brought to the ED by family members. He reports excruciating abdominal and back pain. Electrocardiogram-HR = 130 beats/min, RR = 20 breaths/min, (BP = 90/60 mmHg, oxygen saturation = 95%, patient examination—pulsatile abdominal mass).		X
Syncope with heart block: A 76-yr-old woman collapsed in the hospital while visiting her friend. Unresponsive, electrocardiogram-HR = 44 beats/min, (BP = 75/50 mmHg, oxygen saturation = 95%).		X

ABG = arterial blood gas; BP = blood pressure; ED = emergency department; HR = heart rate; P_{CO_2} = partial pressure of carbon dioxide; P_{O_2} = partial pressure of oxygen; PVC = premature ventricular contraction; RR = respiratory rate.

outlining the student/resident priorities in care during this phase of the study (see Appendix). Each of the scenarios was reviewed independently by each of the four faculty members. Where appropriate, modifications to the presenting history, physical findings, and vital signs were made.

After the scenarios were developed, the faculty formulated a detailed checklist of expected actions to be performed by the students (see Appendix). The checklist items were limited to less than 20 actions. A scoring weight ranging from 1 to 4 was also provided for each checklist item. The magnitude of the weight reflected the importance of the particular action in terms of patient care. For some scenarios, additional checklist items were added that not only reflected necessary actions but also whether the procedures were completed in a set period of time. For example, a student or intern who completed the entire primary survey for the hypotensive trauma patient scenario in less than 1 min was given an additional checklist credit. Each scenario checklist was revised after a pilot test session involving a senior medical student and an intern. These individuals were not study participants. The pilot phase was used to determine how well the scenario matched the clinical environment and whether the checklist items reflected expected actions.

The instructions to trainees were identical for each clinical encounter. The trainees were told to diagnose and treat a patient with the presenting condition in a 5-min period. A nurse would be available to assist in the evaluation. Trainees could request consultations, but no assistance would arrive during the 5-min encounter.

Study Participants

There were a total of 40 participants recruited for the study. Of these, 37 were attending or had attended Liaison Committee for Medical Education–accredited medical schools. Because of variations in medical training worldwide and the questionable representativeness of the sample of international medical graduates, data for these individuals (n = 3) were not included in this study. There were 13 first-year residents (2 emergency medicine, 10 anesthesia, 1 surgery) and 24 fourth-year U.S. medical students.

The fourth-year medical students were chosen to participate on the basis of their selection of anesthesia during their fourth year of medical school. The intern

participants were recruited from emergency medicine/surgery residencies. They were all on their required anesthesia rotation during internship or, if anesthesia residents, their initial month of anesthesia training.

The participants were required to provide written consent to videotape their performance and also agreed not to disclose the scenarios to peers. After the assessment, debriefing sessions were conducted to discuss performance, review the educational content of each scenario, and gather data regarding the fidelity, realism, and content of the simulation exercises.

Evaluation of Student/Resident Performance

Two faculty members and two nurse clinicians scored the simulation encounters independently. None of the faculty involved in scenario development served as raters. The two faculty members had met many of the study participants before the evaluation. Both nurse clinicians were unaware of the previous training and the clinical performance of the students and residents. Before the initiation of scoring, the raters met to discuss the objectives of the evaluation and the specific application of the defined scoring rubric. All scoring was done from videotapes. A four-quadrant video screen that included two separate camera views of the participant and mannequin was used for evaluation. In addition, one of the quadrant's video recordings was the simultaneous full display of patient vital signs (electrocardiogram, pulse oximetry, blood pressure, central venous pressures). Although this videotaped replay scoring strategy may introduce some imprecision in the measures, previous research suggests that performance, both technical and behavioral, can be assessed without live observation.[7]

The raters were required to indicate whether a specific action described on the checklist had been performed by the student/resident. A score for each scenario was derived by multiplying the credited actions (scored 0 for no and 1 for yes) by their respective weights and summing. To generate student/resident scores, by scenario, the average of the checklist ratings for the four raters was used. These scores were converted to percentages based on the resultant mean performance and the maximum score that could be obtained for a given simulation scenario.

Study Design and Analysis

Although 10 simulation scenarios were available for use, each participant was only required to work through six. Two forms (set 1, set 2) were developed, each containing six simulation scenarios (table 1). Two scenarios were common to both sets. There were 15 fourth-year students and 4 residents who performed the cases in set 1. Nine fourth-year students and 9 residents completed the cases in set 2.

Various analysis strategies were used to examine student/resident performance. Descriptive statistics were used to summarize performance by case. Case–total score correlations (discrimination statistics) were calculated to discern whether certain scenarios yielded scores that were better able to distinguish between more- and less-abled participants. To investigate the reproducibility of scores, Generalizability theory provides a mechanism for disentangling the error term into multiple sources.[22] For the current investigation, each student/resident performed six simulation exercises. Each of the exercises (cases) was individually scored by four trained raters. This is a fully crossed candidate $\times$ rater $\times$ case generalizability study design. Generalizability analyses were conducted to estimate variance components. These variance components were then used to derive generalizability and dependability coefficients. The generalizability coefficient, similar to coefficient α for a model in which there is a single (mean) score for each scenario, provides information on the consistency of trainees' performances across cases. The dependability coefficient, which also considers the choice of rater in estimation of the reproducibility of trainee scores, is appropriate when scores are given absolute interpretations. For example, rater (or rater by case) stringency effects would be important if pass/fail decisions were based on the assessment scores but not if one simply wanted to rank order the students/residents. In addition to generalizability and dependability coefficients, the standard error of measurement (SEM) was also calculated. The SEM, which is on the same metric as the original measures, provides information on the precision of individual student/resident scores. Analyses were performed separately for set 1 and set 2 data. Finally, to support the validity of the assessment, various score comparisons were made between students/residents with more and less clinical experience and training. The significance of any mean score differences was assessed using analysis of variance.

Results

Descriptive Statistics

Descriptive statistics are displayed in table 2. Overall, according to the defined scoring rubric, the most difficult case was cerebral hemorrhage–herniation (case 1). The easiest case, judging from the average total score, was myocardial infarction (case 2). As evidenced by the relatively high SDs, there was considerable variation by case in student/resident performance.

Case–Total Correlations (Discrimination)

Correlation coefficients were calculated to investigate how well the individual scenario scores discriminated between low- and high-ability students/residents. For each case set ($n = 6$), individual scenario scores were correlated with the sum of the other five scenario scores. Provided that the scenarios are measuring the same sets

Table 2. Descriptive Statistics

Case No.	n (Ratings)	Scenario (Case)	Mean, %	SD
1	148*	Trauma–hemorrhagic hypotension–femur fracture	52.2	16.4
2	148	Myocardial infarction	71.5	13.8
3	76	Pneumothorax in closed-chest trauma	61.9	10.2
4	76	Ectopic pregnancy–hypovolemia and abdominal pain	57.9	16.6
5	76	Cerebral hemorrhage–herniation	48.8	19.1
6	76	Ventricular tachycardia	61.2	21.4
7	72	Respiratory failure—intubation required	62.6	19.7
8	72	Asthma exacerbation	59.2	11.8
9	72	Pulsatile abdominal mass	67.0	16.6
10	72	Syncope with heart block	52.3	10.9

* 4 raters × 37 participants.

of skills, this sum can be considered to be a criterion performance measure. The discrimination values for each scenario set are presented in table 3. These values can range from −1 to 1. In general, the case discriminations were relatively high and positive, indicating that students/residents who performed well on one particular case tended to perform well overall.

Variance Component Analysis

The estimated variance components for the student/resident scores are shown in table 4. The analyses were performed separately for the two case sets. The person (student/resident) variance component is an estimate of the variance across persons of person 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 both analyses, the estimates were greater than zero (σ^2_c = 43.40 and 68.35), suggesting that the cases vary somewhat in average difficulty. The rater components are the variances of the rater mean scores. The relatively small values indicate that raters do not vary appreciably in terms of average stringency. The largest interaction variance component, for both data sets, was person × case. The magnitude of these components suggests that there are considerably different rank orderings of person mean scores for each of the various cases. The relatively small person × rater components suggest that the various raters rank order persons similarly. Likewise, the small rater × case components indicate that the raters rank order the difficulty of the cases similarly. The final variance components are the residual variances that include the triple-order interactions and all other unexplained sources of variation.

For set 1 data, the generalizability coefficient (ρ^2), based on six cases and four raters, was 0.74 (SEM = 5.6). The dependability coefficient (Φ) was 0.69 (SEM = 6.2). For set 2 data, the generalizability and dependability coefficients were 0.53 (SEM = 5.1) and 0.44 (SEM = 6.2), respectively. As noted previously, however, the rater facet and associated interactions do not contribute much to the variability of observed scores. This becomes evident when generalizability coefficients for a design involving six cases but only one randomly selected rater are estimated. For set 1, the generalizability coefficient is only reduced to ρ^2 = 0.69 (SEM = 6.3). For set 2, the generalizability coefficient is reduced to ρ^2 = 0.46 (SEM = 5.9).

The small rater, person by rater, and rater by case variance components indicate that the choice of rater has little impact on the reproducibility of the student/resident scores. Interrater reliability for set 1 was $\rho_{xx'}$ = 0.97. Interrater reliability for set 2 was $\rho_{xx'}$ = 0.95.

Validity Coefficients

The residents and fourth-year students were asked to indicate how much elective time they had spent in various rotations. It was hypothesized that those individuals who spent more time in critical care electives would perform better on the simulation exercises. The correlation between elective time (weeks spent in rotation) and

Table 3. Case–Total Correlations (Discriminations)

Case No.	Scenario (Case)	Set 1	Set 2
1	Trauma–hemorrhagic hypotension–femur fracture	0.64	0.20
2	Myocardial infarction	0.63	0.32
3	Pneumothorax in closed-chest trauma	0.23	
4	Ectopic pregnancy–hypovolemia and abdominal pain	0.63	
5	Cerebral hemorrhage–herniation	0.21	
6	Ventricular tachycardia	0.39	
7	Respiratory failure–intubation required		0.32
8	Asthma exacerbation		0.34
9	Pulsatile abdominal mass		0.38
10	Syncope with heart block		0.19

Table 4. Simulation Study Variance Components

Source of Variability	Set 1 Estimate*	Set 1 $n_c = 6, n_r = 4$	Set 2 Estimate*	Set 2 $n_c = 6, n_r = 4$
Person	87.56	87.56	29.58	29.58
Rater	0.82	0.20	0.00	0.00
Case	43.40	7.23	68.35	11.39
Person × rater	5.87	1.47	3.65	0.89
Person × case	167.23	27.87	140.93	23.49
Rater × case	4.07	0.17	1.81	0.08
Residual	37.75	1.57	49.19	2.05

* Provides a decomposition of the total observed score variance for a single scoring of a single case.

the total simulation score (averaged over six cases), by elective, is presented in table 5. Participants who did not perform a rotation in a certain discipline were assigned a value of 0 for that elective. The total number of weeks in critical care electives (anesthesiology, cardiology, critical care, emergency medicine, pulmonary medicine, surgery, and trauma) was also calculated and correlated with the total simulation score. In general, the more time an individual spent in an elective rotation, the better his/her simulation performance was. The only exception was anesthesiology, where there was a negative, albeit nonsignificant, association between weeks spent in the rotation and total simulation score. Individuals who had done an elective in critical care (n = 18, mean = 13.8, SD = 1.6), regardless of number of weeks spent, scored significantly higher ($F = 5.61$, $P < 0.05$) on the simulation exercises than those who had not (n = 19, mean = 12.4, SD = 2.1). The total number of weeks spent in critical care electives was positively associated with simulator performance ($r = 0.24$, $P < 0.05$).

The correlation between total weeks spent in critical care electives and the simulator scores was also calculated for each case separately. These data are presented in table 6. For most cases, there was a positive association between total elective time and case scores. For the myocardial infarction and cerebral hemorrhage–herniation cases, these associations were moderate and statistically significant.

Table 5. Correlation of Time (Weeks) Spent in Elective Rotations and Simulation Score

Elective Rotation	Correlation
Anesthesiology	−0.05
Critical care medicine	0.37*
Cardiology	0.18
Emergency medicine	0.28
Obstetrics/gynecology	0.21
Pediatrics	0.06
Pulmonary medicine	0.07
Surgery	0.18
Trauma	0.20
Sum, weeks	0.34*

* $P < 0.05$.

Comparison of Student/Resident Cohorts

There were 13 first-year residents and 24 fourth-year U.S. medical students tested as part of this study. Given the additional medical experience of the residents, one would expect this group to perform better on the simulation exercises. The mean performance of the resident group, averaged over cases, was 64.9 (SD = 5.9). The mean performance of the fourth-year students was 57.1 (SD = 9.0). The difference in performance was significant ($F = 7.8$, $P < 0.01$) and in the hypothesized direction. This difference represents a reasonably large effect (effect size = 0.89).

Relevant Training

The students/residents also indicated whether they were currently certified in Pediatric Advanced Life Support and Advanced Cardiac Life Support. Those individuals who had current Advanced Life Support certification (n = 2) outperformed those who did not (n = 35) (mean PALS = 71.7, mean no PALS = 59.2, $F = 4.2$, $P < 0.05$, effect size = 1.4). Likewise, individuals who had current Advanced Cardiac Life Support certification (n = 16, mean = 63.2, SD = 7.0) outperformed those who did not (n = 21, mean = 57.3, SD = 9.3). This difference

Table 6. Correlation of Simulation Score with Total Elective Time, by Case

Case No.	Scenario (Case)	Elective Time*
1	Trauma–hemorrhagic hypotension–femur fracture	0.20
2	Myocardial infarction	0.41†
3	Pneumothorax in closed-chest trauma	−0.08
4	Ectopic pregnancy–hypovolemia and abdominal pain	−0.01
5	Cerebral hemorrhage–herniation	0.48†
6	Ventricular tachycardia	0.06
7	Respiratory failure–intubation required	0.25
8	Asthma exacerbation	0.07
9	Pulsatile abdominal mass	0.37
10	Syncope with heart block	0.23

* Weeks spent in anesthesiology, cardiology, critical care, emergency medicine, pulmonary medicine, surgery, and trauma electives. † $P < 0.05$.

was also significant ($F = 4.4$, $P < 0.05$) and represents a relative large effect (effect size = 0.67).

Discussion

The results of this study suggest that simulation can be used as a method to evaluate clinical performance of medical students and residents. We established the content validity of the assessment by selecting and modeling a variety of potentially life-threatening acute care conditions that, at the conclusion of medical school, a generalist physician should be able to diagnose and treat. In our study, most of the scenarios were developed to simulate acute conditions that had a well-defined set of diagnostic and treatment actions. For example, the primary survey and initial therapy of a hypotensive trauma patient, the emergency department management of a patient with chest pain, and the diagnosis and treatment of an unresponsive patient are all conditions that have a set of diagnostic and therapeutic guidelines developed by groups such as the American College of Surgeons or the American Heart Association. Unfortunately, in most previous simulation studies, the direct relations among the content of the exercise, the requisite treatment and management options, and the required skill level of the examinee had not been established.[6–9,11,12,21]

We found that the choice of rater was not very important in terms of determining the overall reproducibility of student/resident scores. This was probably, in part, a result of the careful selection of assessment content and the development of detailed, well-defined scoring rubrics. When compared to a standardized patient assessment, we anticipated that the simulated scenarios might be more difficult to reliably score because of the increased complexity of the patient interactions. The combination of diagnostic actions, judgments, and therapeutic interventions expected of the trainee could create an assessment situation that could only be validly scored by an expert rater. Gaba *et al.*[8] found that, because of the number of cognitive, psychomotor, inferential, and deductive skills that must be assessed, using a number of expert raters may be the only legitimate way to score the performance. However, our data suggest that only minimal gains in reliability could be achieved by having multiple raters per scenario. In a previous pilot study,[15] we determined that variance among raters could be reduced by assuring agreement among raters about key actions during the development of each scenario. Because most of the simulations modeled have specific guidelines for evaluation, diagnosis, and treatment, the scoring could be performed more objectively, resulting in high levels of agreement between expert raters. Similar to the results from previous studies involving multitask performance assessments,[23] the person by case variance component was large, indicating that choice and number of cases are the most important factors affecting the reliability of the performance-based scores.

The validity of the assessment scores was investigated through a detailed analysis of case performance and a comparison of participants with various levels of training and experience. Medical students are able to select from a variety of rotations in their final year of training, often resulting in considerable variability in patient contact. Although most students and residents are familiar with diagnosis as well as treatment of the acute care situations, students who have an increased exposure to acute management situations would be more likely to effectively translate their experience into a logical and orderly sequence of actions that would lead to rapid diagnosis and treatment of the condition. One argument for the validity of the assessment was based on a determination of whether the content of the student's final year helped prepare him or her for this type of exercise. As expected, the students who had certain types of clinical rotations (critical care, emergency) performed better on the simulation encounters. The positive associations between experience (*e.g.*, number and type of clinical rotations) and the simulator scores suggest that meaningful interpretations of individual scores can be made. However, some elective experiences were not positively related to performance. For example, clinical rotations on anesthesiology did not correlate with performance on the simulation exercises. This is likely because all of the students and residents had spent at least 4 weeks on an anesthesia rotation, effectively equalizing training experiences and minimizing the variability in performance, thereby attenuating the true association. In terms of the case scores, we found that all of the simulation exercises discriminated between low- and high-ability examinees. This finding suggests that the test development process, including the development of scenario-specific analytic scoring rubrics, is likely to yield appropriate assessment contents, resulting in valid simulator scores.

The relation between clinical experience and responses to simulated anesthetic emergencies has been studied previously.[11,21] Our results indicated that interns performed significantly better than medical students on the simulation exercises. This performance difference provides further evidence to support the validity of the simulation scores. In addition, on average, students and interns who had participated in Advanced Cardiac Life Support training or Pediatric Advanced Life Support scored significantly higher on the simulation exercises than those who did not. This result might be attributed in part to the two scenarios that are included in the content of Advanced Cardiac Life Support exercises (ventricular tachycardia and heart block) but could also be an indication of the interest and background of the trainee. That is, students or residents who obtained this certification were often interested in specialties such as emergency medicine. Nevertheless, the finding that pre-

vious training on a simulator is associated with better performance on our assessment is noteworthy. Overall, as expected, trainees may perform better, at least for simulated emergencies, when they have had some experience with the actual training device or a related device.

For this study, each trainee participated in only six simulations. The generalizability and dependability coefficients were only moderate, suggesting that if more precise measures of ability are required, additional performance samples are needed. Even though trainees who performed well on one exercise were likely to perform well in subsequent scenarios, we found that there was considerable variation in student/resident scores attributable to case content. In fact, case specificity was the major determinant leading to variation in simulation scenario scores. This finding is consistent with previous research in the performance assessment domain where task and person by task variability have been identified as key sources of variability in examinee scores.[24] That is, performances do not generalize extremely well from one patient problem to another.

The limited correlation among the scenario scores supports a decision to use multiple scenarios to measure the range of skills expected in acute care. It was not surprising that a student or intern who was able to diagnose and treat a pneumothorax might not be able to diagnose and treat a patient with myocardial ischemia. If the correlations among scenario scores were higher, this would suggest that a student or intern's acute care skills could be measured on the basis of performance across fewer scenarios. However, the fact that all scenarios discriminated along the ability continuum indicates that the scenarios measure similar aspects of clinical performance. For future studies, it would be informative to explore the relations among specific treatment and management options, both within and across scenarios. This would aid in case development and provide some insight as to which acute care skills are generalizable across patient conditions and which are not.

A central question in developing an acute care skills evaluation is to determine how many encounters might be required to accurately assess a trainee's ability. The generalizability and dependability coefficients were only moderate, suggesting that if more precise measures of ability are required, additional performance samples are needed. A more detailed analysis of a larger set of scenarios might indicate how many and what mix would be required to effectively sample the range of acute care skills expected of a physician. Based on our results, and depending on the purpose of the assessment, it is likely that a relatively large number of performance scenarios will be required to obtain sufficiently accurate ability estimates.

Although we have provided some evidence to support the validity and reliability of the simulator scores, there are several additional investigations that warrant attention. First, because of the relatively small sample of students and residents studied, it would be worthwhile to replicate the assessment with a larger, more representative cohort. This would enhance the generalizability of the findings. Second, from a scoring perspective, alternate strategies should be tried. For example, a holistic or global approach in which experts provided a summative judgment of overall performance has been used for other performance-based examinations.[25] This scoring strategy, although potentially more subjective, capitalizes on expert opinion and could yield more valid scores. Third, from a validity perspective, it would be extremely valuable to gather additional data on the students and residents. The performance on the simulation exercises is directed not only to knowledge about acute care, but also to an organized, orderly, and rapid diagnosis and treatment of a patient problem. The skills in acute care are primarily tested in a subset of U.S. Medical Licensing Examination step 3. One would expect that there would be a positive relation between step 3 scores and performance on the simulation exercises. Unfortunately, the students' and interns' performances on step 3 were not available at the time of the study. Finally, although the fidelity of the modeled exercises is high, there is no guarantee that an examinee's performance in a simulated environment will translate to real-life situations. To our knowledge, there have been no comprehensive studies comparing patient care outcomes between candidates who met success criteria on the simulator and those who did not. These investigations are surely needed, especially if simulators are going to be applied for formal certification of credentialing decisions.

The results of this study indicate that reasonably reliable and valid measures of clinical performance can be obtained from simulation exercises, provided that care is taken in the development and scoring of the scenarios. Students and residents are unlikely to encounter many of the life-threatening conditions that may occur in patients during the course of training in medical school or during internship or residency training, but these skills are expected of a generalist physician. A simulator, which reacts in a physiologically and pharmacologically appropriate manner, allows examinees to demonstrate clinical skills in a controlled and reproducible environment. This type of training and assessment, which provides immediate feedback about patient care decisions, should lead to better performance. With this in mind, the development and testing of additional simulation scenarios and associated scoring systems are warranted.

References

1. Dupras DM, Li J: Use of an objective structured clinical examination to determine clinical competence. Acad Med 1995; 70:1029–34

2. Issenberg SB, McGaghie WS, 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 assessment. JAMA 1999; 282:861–6

3. The Medical School Objectives Writing Group: Learning objectives for medical student education: Guidelines for medical schools. Acad Med 1999; 74:13–8
4. 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
5. Murray DJ: Clinical simulation: Technical novelty or innovation in education. ANESTHESIOLOGY 1998; 89:1–2
6. Chopra V, Gesink BJ, DeJong 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
7. 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
8. 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
9. Jacobsen J, Lindekaer AL, Ostergaard HT, Nielsen K, Ostergaard D, Laub M, Jensen PF, Johannessen N: Management of anaphylactic shock using a full scale anaesthesia simulator. Acta Anaesth Scan 2001; 45:315–9
10. Kapur PA, Steadman RH: Patient simulator competency testing: Ready for takeoff? Anesth Analg 1998; 867:1157–9
11. Devitt JH, Kurrek MM, Cohen MM, Cleave-Hogg D: The validity of performance assessments using simulation. ANESTHESIOLOGY 2001; 95:36–42
12. Schwid HA, Rooke GA, Carline J, Steadman RH, Murray BA: Evaluation of anesthesia residents using mannequin-based simulation. ANESTHESIOLOGY 2002; 97:1434–44
13. Byrne AJ, Greaves JD: Assessment instruments used during anaesthetic simulation: Review of published studies. Br J Anaesth 2001; 86:445–50
14. McGaghie WC, Issenberg SB, Gordon DL, Petrusa ER: Assessment instruments used during anaesthetic simulation. Br J Anaesth 2001; 87:647–8
15. 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
16. Wong SH, Ng KF, Chen PP: The application of clinical simulation in crisis management training. Hong Kong Med J 2002; 8:131–5
17. Innovative Simulations for Assessing Professional Competence: From Paper-and-pencil to Virtual Reality. Edited by Tekian A, McGuire CH, McGaghie WC. Chicago, Department of Medical Education, University of Illinois at Chicago, 1999
18. Sica GT, Barron DM, Blum R, Frenna TH, Raemer DB: Computerized realistic simulation: A teaching module for crisis management in radiology. AJR Am J Roentgenol 1999; 172:301–4
19. Garden A, Robinson B, Weller J, Wilson L, Crone D: Education to address medical error: A role for high fidelity patient simulation. N Z Med J 2002; 115:133–4
20. Schwid HA: Anesthesia simulators: Technology and applications. Isr Med Assoc J 2000; 2:949–53
21. 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
22. Brennan RL: Generalizability Theory. New York, Springer-Verlag, 2001
23. Shavelson RJ, Baxter GP, Gao X: Sampling variability of performance assessments. J Educ Meas 1993; 30:215–32
24. Brennan RL: Performance assessments from the perspective of generalizability theory. Appl Psychol Meas 2000; 24:339–53
25. Rothman AI, Blackmore D, Dauphinee WD, Reznick R: The use of global ratings in OSCE station scores. Adv Health Sci Educ Theory Pract 1997; 1:215–9

Appendix

The faculty defined an initial list outlining the student/resident priorities in care during this phase of the study (table 7). Each scenario was reviewed independently by each of the four faculty members. Appropriate modifications to the presenting history, physical findings, and vital signs were made. After developing the scenarios, the faculty formulated a detailed checklist of expected actions to be performed by the students, limited to less than 20 actions. A scoring weight ranging from 1 to 4 was also provided for each checklist item. The magnitude of the weight reflected the importance of the specific action in terms of patient care. For some scenarios, additional checklist items were added that reflected necessary actions as well as whether the procedures were completed in a set period of time (tables 8 and 9).

Table 7. Common Scenarios

Hemorrhagic Hypotension	Myocardial Infarction
Establish patient is responsive to verbal	Establish patient is responsive to verbal
Auscultate chest long enough to determine RR	Auscultate chest
Request oxygen saturation	Auscultate chest long enough to determine RR
Request BP	Request BP
Expose the patient	Expose the patient
Determine the presence of deformity (left femur)	Oxygen administration
Primary survey in < 1 min	Request 12-lead electrocardiogram
IV access and initiate IVF	Interpret 12-lead electrocardiogram
Provide appropriate IVF	325 mg oral ASA or antiplatelet agent
Determine need for type and cross	Nitroglycerine SL or spray
Proper sequencing of survey	Establish IV
All of the above in <3 min	Morphine IV
Premorbid history; penicillin anaphylaxis	β blockade (metoprolol or atenolol) IV
Circulation examination of lower extremity	Establish absence of previous medical problems
Neuroexamination of lower extremity	Cardiology consultation
Determine need for extremity radiograph	Steps to determine thrombolytic therapy
Determine and implement immobilization of left leg	Heparin IV
Provide analgesia	Laboratory studies: CK with or without isoenzymes

ASA = acetylsalicylic acid; BP = blood pressure; CK = creatine kinase; IV = intravenous; IVF = intravenous fluid; RR = respiratory rate; SL = sublingual.

Table 8. Additional Scenarios—Set 1

Pneumothorax in Closed-chest Trauma	Ectopic Pregnancy—Hypovolemia and Abdominal Pain	Cerebral Hemorrhage–Herniation	Unstable Ventricular Tachycardia
Establish patient is responsive to verbal	Establish patient is responsive to verbal	Establish patient is not responsive to verbal	Establish patient is not responsive to verbal
Request oxygen saturation	Auscultate chest long enough to determine RR	Establish patient is unresponsive to pain	Recognize arrhythmia as ventricular tachycardia
Request BP	Request oxygen saturation	Auscultate chest long enough to assess RR	Palpate pulse
Expose chest	Request BP	Request oxygen saturation	Request BP
Expose patient	Expose/inspect entire patient	Request BP	Apply oxygen
Establish asymmetric chest wall movement	Establish IV access	Expose the patient	Auscultate/inspect chest
Auscultate chest	Begin appropriate fluid replacement	Presence of dilated left pupil	Request oxygen saturation
Absence of right-sided breath sounds	Resuscitation begins is < 2 min	Provide supplemental oxygen	Defibrillator to bedside
Palpate chest—determine right chest wall pain	Request type and cross	Establish IV and appropriate fluid	Defibrillator paddles applied to chest in < 60 s
Supplemental oxygen	Supplemental oxygen	Prepare for intubation (ETT, laryngoscope, suction)	Cardioversion
Establish IV and appropriate fluid	Suggest O-negative blood	Ambu bag and mask For > 30 s before intubation	Synchronized
Above issues in <2 min	Directed gynecologic history (e.g., LMP, vaginal bleeding, pregnancy)	Appropriate laryngoscopy technique	100+ J
Request chest radiograph	Abdominal examination—Palpate	Intubation attempt <30 s	Assess rhythm/pulse/BP after cardioversion
Interpret chest radiograph	Order beta-HCG	If successful, inflate cuff	Assess responsiveness (minimal response)
Suggest needle decomp/ CT placement	Gynecology consult	If successful, auscultate chest	Lidocaine bolus/infusion
Define location for needle decomp or CT	Suggest utrasound/culdocentesis/ gynecologic evaluation	Does not attempt to lower BP	Request 12-lead electrocardiogram
Provide analgesia	Consider immediate operative intervention	Recognize need for immediate CT scan	Essential diagnostic and therapy steps in <3 min
Definitive treatment in < 3 min		Suggest mannitol	

BP = blood pressure; CT = computed tomography; Decomp = decompression; ETT = endotracheal tube; HCG = human chorionic gonadotropin; IV = intravenous; RR = respiratory rate

Table 9. Additional Scenarios—Set 2

Respiratory Failure—Intubation Required	Asthma Exacerbation	Pulsatile Abdominal Mass	Syncope with Heart Block
Establish patient is unresponsive to verbal	Establish patient is responsive to verbal	Establish patient is responsive to verbal	Establish patient is minimally responsive to verbal
Auscultate chest	Request oxygen saturation	Auscultate chest long enough to assess RR	Assess airway
Auscultate/inspect chest >5 s; assess RR	Expose chest	Request oxygen saturation	Assess breathing—RR and oxygen saturation
Request oxygen saturation	Determine RR	Request BP	Assess circulation—BP and HR
Request BP	Auscultate chest, diagnose wheezing	Expose/inspect entire patient	Administer oxygen
Provide oxygen	Provide oxygen	Establish IV access and begin appropriate fluid replacement in <60 s	Request transcutaneous pacemaker
Establish IV access and appropriate fluid	Non-rebreathing mask with oxygen	Determine need for type and cross	Establish IV access
Initiate Ambu bag and mask oxygen	Request BP	Focus abdominal examination	Above issues in <1 min
Above steps in <1 min	Above issues in <1 min	Determine the presence of deformity (distended abdomen)	Administer atropine
Prepare for intubation (ETT, Laryngoscope, Sx)	IV access and IV fluid at increased maintenance rate	Determine presence of pulsatile abdominal mass	Assess response following atropine (HR, BP, LOC)
Appropriate laryngoscopy technique	Peak air flow assessment	Suggest immediate surgical consultation	Begin dopamine or epinephrine infusion
Intubation attempt >30 s	Begin nebulizer therapy (any β-agonist or combined atrovent)	Above issues in < 3 min	Administer epinephrine bolus
If successful, inflate cuff	History—inhaler use and frequency	Dx study—cross table lateral, echocardiography, or CT	Transfer to medical ICU
If unsuccessful, resume Ambu bag ventilation	Asthma history-multiple serious ED/ICU admits	Limit fluid resuscitation when BP>110 mmHg systolic	Laboratory for CK-MB or troponin
Attach Ambu bag and ventilate	History—previous intubation/ICU admission		Request 12-lead electrocardiogram
Confirm ETT placement	Corticosteroid IV βAgonist IV Suggest ABG Order chest x-ray Indicate potential for intubation		Request cardiology consultation

ABG = arterial blood gas; BP = blood pressure; CK = creatine kinase; CT = computed tomography; Dx = diagnosis; ED = emergency department; ETT = endotracheal tube; HR = heart rate; ICU = intensive care unit; LOC = level of consciousness; RR = respiratory rate; Sx = suction.