Simulation-based Assessment of the Management of Critical Events by Board-certified Anesthesiologists

Matthew B. Weinger, M.D., M.S., Arna Banerjee, M.B.B.S., Amanda R. Burden, M.D., William R. McIvor, M.D., John Boulet, Ph.D., Jeffrey B. Cooper, Ph.D., Randolph Steadman, M.D., M.S., Matthew S. Shotwell, Ph.D., Jason M. Slagle, Ph.D., Samuel DeMaria, Jr., M.D., Laurence Torsher, M.D., Elizabeth Sinz, M.D., M.Ed., Adam I. Levine, M.D., John Rask, M.D., Fred Davis, M.D., Christine Park, M.D., David M. Gaba, M.D.

ABSTRACT

Background: We sought to determine whether mannequin-based simulation can reliably characterize how board-certified anesthesiologists manage simulated medical emergencies. Our primary focus was to identify gaps in performance and to establish psychometric properties of the assessment methods.

Methods: A total of 263 consenting board-certified anesthesiologists participating in existing simulation-based maintenance of certification courses at one of eight simulation centers were video recorded performing simulated emergency scenarios. Each participated in two 20-min, standardized, high-fidelity simulated medical crisis scenarios, once each as primary anesthesiologist and first responder. *Via* a Delphi technique, an independent panel of expert anesthesiologists identified critical performance elements for each scenario. Trained, blinded anesthesiologists rated video recordings using standardized rating tools. Measures included the percentage of critical performance elements observed and holistic (one to nine ordinal scale) ratings of participant's technical and nontechnical performance. Raters also judged whether the performance was at a level expected of a board-certified anesthesiologist.

Results: Rater reliability for most measures was good. In 284 simulated emergencies, participants were rated as successfully completing 81% (interquartile range, 75 to 90%) of the critical performance elements. The median rating of both technical and nontechnical holistic performance was five, distributed across the nine-point scale. Approximately one-quarter of participants received low holistic ratings (*i.e.*, three or less). Higher-rated performances were associated with younger age but not with previous simulation experience or other individual characteristics. Calling for help was associated with better individual and team performance.

Conclusions: Standardized simulation-based assessment identified performance gaps informing opportunities for improvement. If a substantial proportion of experienced anesthesiologists struggle with managing medical emergencies, continuing medical education activities should be reevaluated. **(Anesthesiology 2017; 127:475-89)**

HUMAN performance is imperfect and, without dedicated periodic practice, typically degrades over time.[1–3] To this end, Maintenance of Certification (MOC) programs are intended to facilitate lifelong learning and practice improvement.[4–7] Maintenance of Certification in Anesthesiology (MOCA) and other fields has recently been revised in response to concerns about cost, relevance to practice, and inconsistent evidence of effectiveness.[5,7] Many physicians believe that their practice is safe and that they are performing optimally.[8] The ability of practicing anesthesia professionals to manage perioperative emergencies, like cardiorespiratory arrest, anaphylactic shock, or massive hemorrhage, where deficiencies may have life-or-death consequences, is largely unknown. Identifying the performance gaps of practicing clinicians could lead to more effective graduate medical education, continuing medical education, and practice improvement activities.

What We Already Know about This Topic

- Written or oral examination performances can be unreliable indicators of the real-world performance of physicians as they practice throughout a long career
- Mannequin-based simulation is used to evaluate the performance of anesthesia trainees in crisis event management

What This Article Tells Us That Is New

- To assess the technical and behavioral performance of board-certified anesthesiologists, those who were already attending simulation courses for American Board of Anesthesiology Maintenance of Certification participated in standardized study simulation scenarios that were video recorded for later scoring by blinded trained raters
- In simulated emergencies, participants successfully completed approximately 80% of critical performance elements, while approximately 25% received low holistic rating
- Higher-rated performances were not associated with previous simulation experience

This article is featured in "This Month in Anesthesiology," page 1A. Corresponding article on page 410. Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are available in both the HTML and PDF versions of this article. Links to the digital files are provided in the HTML text of this article on the Journal's Web site (www.anesthesiology.org). This article has an audio podcast.

 Anesthesiology 2017; 127:475-89

<zdoi;10.1097/ALN.0000000000001739>

Assessing the quality of perioperative event management is difficult. Critical events are uncommon and unpredictable in practice, making prospective studies of their management extremely difficult. *Post hoc* adverse event reports are typically incomplete, and their analysis has inherent selection and hindsight biases.[9] Written or oral examination performances may be unreliable indicators of real-world performance.[10,11] Mannequin-based simulation, however, provides a unique window on performance: standardized critical events (of varying levels of urgency) can be simulated with reasonable levels of realism,[12–15] and participant performance can be evaluated.[16–19]

Success in managing medical emergencies depends on both technical (*e.g.*, correct diagnosis and therapy) and behavioral (*e.g.*, leadership, communication, and resource management) skills.[20,21] Although medical education has recently incorporated behavioral skills training, it was not explicitly taught at many institutions when a preponderance of currently practicing anesthesiologists underwent their primary training.[22] In this study, we sought to quantify the distribution of technical and behavioral performance of board-certified anesthesiologists (BCAs) managing realistic perioperative simulated crises, with the following goals: (1) identifying performance gaps that could be addressed in future educational interventions; (2) investigating the feasibility of conducting simulation-based assessment at multiple sites; and (3) providing evidence to support the psychometric adequacy of the scores.

Materials and Methods

Study Design and Context

We conducted a prospective, nonrandomized, observational study at eight American Society of Anesthesiologists–endorsed simulation network programs.[1] The study sites were selected based on their research infrastructure and regular conduct of MOCA courses. Study participants were recruited from BCAs who were already attending scheduled simulation courses that satisfied their MOCA simulation training requirement.[4,23] The 6- to 8-h MOCA courses use realistic simulated encounters to foster the reflection of attendees on their care and decision-making during perioperative crises. All of the MOCA course scenarios deal with less common, unexpected clinical events of significant severity (*e.g.*, episodes of severe hypoxia and/or hemodynamic instability) requiring recognition and complex management. Course participants are not informed of or given specific training about the clinical scenarios. Each course attendee is the primary anesthesiologist (referred to colloquially as the hot seat [HS] participant), in at least one 20- to 30-min simulated clinical crisis scenario. Because teamwork is emphasized, a second anesthesiologist (the first responder [FR]), naïve to the transpiring crisis, is sequestered until he/she is called to help. Experienced simulation educators facilitate debriefings after each scenario.

We designed four standardized MOCA-compliant study scenarios that were offered in study site MOCA courses between November 2012 and June 2014. After receiving institutional review board approval, each site enrolled consenting participants and collected demographic information. Each participant performed in at least two standardized study simulation scenarios (once in the HS role and once as FR) that were video recorded for later scoring by trained raters.

Designing Four Standardized Scenarios

Four perioperative crisis scenarios were designed and iteratively piloted to do the following: (1) comply with the course requirements[4]; (2) elicit relevant technical and behavioral skills; and (3) contain critical performance elements (CPEs) that could be observed and scored. A panel of 10 independent subject matter experts (SMEs) advised the study team in creating the simulation scenarios and rating rubrics (Supplemental Digital Content 1, http://links.lww.com/ALN/B480). SMEs were selected based on their clinical and educational expertise; all participated in the American Board of Anesthesiologists examination process either as oral examiners or written examination content developers. Some were simulation instructors, but none were simulation researchers or had leadership involvement in simulation. SMEs reviewed, contributed to, and approved the scenario content and assessment metrics. They also affirmed that the scenario content and management expectations were within a BCA's scope of practice.

The four scenarios were iteratively developed, with the SMEs and research team reviewing and modifying their content as necessary, pilot-testing new iterations, and further refining the scenarios and corresponding checklists. Scenarios were approved for use by consensus of the research team and the SME panel. The resulting scenarios were as follows: (1) local anesthetic systemic toxicity (LAST) with hemodynamic collapse; (2) hemorrhagic shock from occult

Submitted for publication September 12, 2016. Accepted for publication May 8, 2017. From the Center for Experiential Learning and Assessment (M.B.W., A.B.), Vanderbilt University School of Medicine (M.B.W., M.S.S., J.M.S.), Nashville, Tennessee; Center for Research and Innovation in Systems Safety, Vanderbilt University Medical Center, Nashville, Tennessee (M.B.W., A.B., J.M.S.); Geriatric Research Education and Clinical Center, VA Tennessee Valley Healthcare System, Nashville, Tennessee (M.B.W.); Cooper Medical School, Rowan University, Cooper University Hospital, Camden, New Jersey (A.R.B.); University of Pittsburgh Medical Center and Winter Institute for Simulation Education and Research, Pittsburgh, Pennsylvania (W.R.M.); Foundation for Advancement of International Medical Education and Research, Philadelphia, Pennsylvania (J.B.); Harvard Medical School, Massachusetts General Hospital, Boston, Massachusetts (J.B.C.); Center for Medical Simulation, Boston, Massachusetts (J.B.C., F.D.); Department of Anesthesiology, University of California Los Angeles, Los Angeles, California (R.S.); Icahn School of Medicine at Mt. Sinai, New York, New York (S.D., A.I.L.); Mayo Clinic, Rochester, Minnesota (L.T.); Pennsylvania State University College of Medicine, Hershey, Pennsylvania (E.S.); Department of Anesthesiology and Critical Care Medicine and University of New Mexico Basic and Advanced Trauma Computer Assisted Virtual Experience Simulation Center, University of New Mexico School of Medicine, Albuquerque, New Mexico (J.R.); Department of Anesthesiology, Feinberg School of Medicine, Northwestern University, Chicago, Illinois (C.P.); Center for Immersive and Simulation-based Learning, Stanford University School of Medicine, Stanford, California (D.M.G.); VA Palo Alto Health Care System, Palo Alto, California (D.M.G.).

retroperitoneal bleeding (hemorrhage); (3) malignant hyperthermia (MH) presenting in the postanesthesia care unit; and (4) acute onset of atrial fibrillation with hemodynamic instability followed by ST elevation myocardial infarction (Afib/MI) (Supplemental Digital Content 2, http://links.lww.com/ALN/B481).

Standardization of Scenario Delivery

To standardize the delivery of the scenarios, detailed scripts and a guidebook of rules for scenario delivery were created. The scenario scripts delineated the contents of the simulated clinical environment (*e.g.*, the equipment and medications available), evolution of the patient's condition throughout the crises and their responses to interventions, standardized answers to anticipated participant questions, and criteria that defined successful completion of CPEs. Each script also contained the timing and content of key phrases or comments to be made by trained confederates, acting in the roles of anesthesiologists, surgeons, nurses, or the patient during the scenarios. These key phrases provided information or clinical context that otherwise would not be available from the mannequins (*e.g.*, "the patient feels warm to me" in the MH scenario). Scripted verbal prompts from confederates were used when necessary to assure timely progression of the scenarios. Key scripted events and standardized content within the scenarios have been published previously, including the rules for standardized delivery of these scenarios.[24] Before enrolling participants, investigators confirmed a site's ability to deliver the standardized scenarios by reviewing video of its pilot-trial encounters. A central database and custom video review software facilitated data collection and analysis (Supplemental Digital Content 3, http://links.lww.com/ALN/B482).

Rating Rubrics, Metrics, and Procedures

Drawing on the existing literature,[15,25–30] the project team and SMEs collaboratively developed rating rubrics and tools. Separate scoring rubrics were created for technical and behavioral performance. Because there are advantages and disadvantages of itemized *versus* global ratings,[29,31–33] we developed both types of rubrics to quantify those skills. Technical performance was measured with the percentage of the scenario's CPEs completed and holistic ordinal scores of overall technical performance. Behavioral performance was measured with numerical ratings made using behaviorally anchored rating scales (BARS) of four categories of skills: vigilance, communication, decision-making, and teamwork, as well as holistic ordinal scores of overall behavioral performance.[26] The BARS and holistic behavioral rating scales have been found to be easier to use and yield scores that are just as reliable as the Anesthetists' Non-technical Skills system,[34] a widely used but complex means of rating anesthesia providers' behavioral skills.[26] Finally, based on all of these ratings and their overall evaluation of the performance, the rater made a summative binary assessment (*i.e.*, yes or no) as to whether the participants' overall performance was at the level expected of a BCA. The raters were instructed to base their binary decision on the holistic scores for technical and nontechnical ratings. If a participant scored in the "poor" bin (see "Video Rater Training and Rating Procedures" section), the rater was instructed to rate the performance "no." If the scores were on the cusp of poor and medium performance, the rater was instructed to reconsider the technical and behavioral performance to reach the decision. Details of these metrics and scales are provided in Supplemental Digital Content 4 (http://links.lww.com/ALN/B483).

Through a Delphi Process,[35] SMEs reached consensus on 72 CPEs (16 to 20 CPEs per scenario) that represented the essential patient management steps deemed necessary in each scenario. CPEs were defined so that they could be rated as either present (*observed*) or absent (*not observed*). The CPEs were not weighted as to their importance.

Participants and Study Procedures

Figure 1 illustrates the study enrollment process. After obtaining informed consent, MOCA course attendees who volunteered for the study were allocated to study scenarios. Allocation was made by chance, although many sites assigned participants to all MOCA course (including study) scenarios that were relevant to their practice (*e.g.*, having a pain specialist perform the LAST scenario). Sites were also free to choose the study scenarios that they wished to conduct.

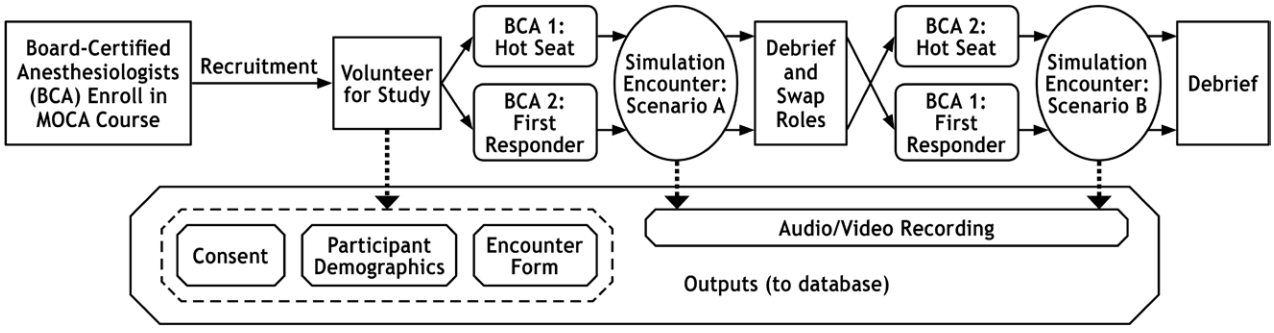

Fig. 1. Enrollment and data collection procedures. The figure shows the algorithm for enrolling participants and collecting data in the study. BCA = board-certified anesthesiologist; MOCA = Maintenance of Certification in Anesthesiology.

Participants completed a demographic survey (table 1, see also Supplemental Digital Content 3, http://links.lww.com/ALN/B482) and then participated in a standardized orientation to simulation where they were briefed on relevant mannequin characteristics, ground rules for participating in simulation encounters, and location and uses of medications, clinical equipment, and other resources (Supplemental Digital Content 5, http://links.lww.com/ALN/B484). Participants observed or took part in at least one course scenario before performing their first study encounter.

Generally, participants were studied in pairs, once each as the HS or FR in successive scenarios. To facilitate assessment of teamwork and communication skills, the FR was sequestered alone, unable to observe the evolving emergency, thereby mimicking the typical conditions for a real-world emergency response by an attending anesthesiologist. If the HS requested anesthesiologist assistance, the FR joined the simulation encounter, but not earlier than 9 min after the encounter started. If the HS did not request assistance, the FR entered the encounter 12 min after it commenced.

Table 1. Participant Demographics and Comparison with Other Cohorts of Anesthesiologists

Individual Attribute	Attribute Category	Study Participants (N = 263)*†	Comparator Cohorts		
			All Board-certified Anesthesiologists in the MOCA Process†‡	All Board-certified Anesthesiologists†‡	Physicians Billing Medicare Identified as Anesthesiologists†§
Sex	Women	37.2% [256]	33.9% [18,916]‖	29.5% [39,336]‖	24.8% [43,830]‖
Age	Yr	42 ± 7 (30, 64) [257]	43 ± 8 [18,919]‖	50 ± 10 [39,939]‖	48 ± 12 [43,544]‖
Clinical experience	Yr	9 ± 5 (0, 38) [257]	8 yr (IQR = 8) [18,730]‖	17 yr (IQR = 15) [36,716]‖	
Graduated from medical school after 1998?	Yes	63.8% [257]	54.3% [18,906]‖	25.7% [38,966]‖	39.8% [43,689]‖
Fellowship trained	Yes	46.7% [257]	24.6% [18,919]‖	12.0% [39,939]‖	
ACLS certified	Yes	90.3% [257]			
Previous simulation experience	Yes	62.6% [257]			
Clinical practice setting	Academic	47.1% [257]			
	Community	49.8% [257]			
	Other	3.1% (8) [257]			
Type of practice	Practice in a group	80.5% [256]			
	Practice primarily in a hospital setting	93.4% [257]			
Anesthetic cases performed per month	Individually performed cases	32.0 ± 40.2 (0, 250) [257]			
	Supervise others performing cases	71.1 ± 76.7 (0, 255) [257]			
Participants reporting that performing these types of cases represent a substantial component of their practice (all 257)	Ambulatory	66.4%			
	Burn or trauma	21.6%			
	Cardiac	25.4%			
	Critical care	16.4%			8.1 [43,823]‖
	General OR	79.5%			
	Geriatric patients	59.0%			
	Hepatic or transplant	9.0%			
	Neurosurgical	48.5%			
	Pain, acute	38.3%			
	Pain, chronic	10.1%			14.0 [43,823]‖
	Pediatric	41.0% [110]			
	Regional	56.7% [152]			
	Vascular	60.1% [161]			

*Data include self-reported results. The denominator (N) included all of the study participants. Some participants failed to provide demographic data. The denominator for each field is listed in brackets. †Data are presented as either mean ± SD (minimum, maximum), percentage (count) [N], or median and interquartile range (IQR). ‡Data were provided by the American Board of Anesthesiologists. Sample excludes those who were 70 yr or older as of January 2013 or known to be retired or deceased and those who were certified after January 2013. §Data were provided by the American Society of Anesthesiologists' Analytics and Research Services Department based on the Physician Compare National Downloadable Files dated 12/18/2014, 7/2/2015, 11/6/2015, 6/2/2016, and 12/19/2016. Note that individuals who graduated medical school in 2012 or later were excluded from this dataset. The subspecialty practice column shows individuals who have self-reported having additional board certifications in chronic/interventional pain or critical care; it does not necessarily mean they are actively practicing in that subspecialty. ‖Study participant population was significantly different from this national comparator group, at least $P < 0.05$ and mostly $P < 0.001$. Fisher's exact test, chi-square test, and the two-sample t test were used to compare binary, multicategoric, and quantitative demographic factors, respectively.

ACLS = advanced cardiac life support; IQR = interquartile range; MOCA = Maintenance of Certification in Anesthesiology; NA = not available; OR = operating room.

Digital audio/video recordings of each study encounter were made and, along with participant demographics and other study data, saved to the project's central database (see Supplemental Digital Content 6, http://links.lww.com/ALN/B485, for details about how the encounters were captured for later rating).

Video Rater Training and Rating Procedures

Nine academic anesthesiologists, previously unaffiliated with the study, with at least 3 yr of clinical practice after board certification and experience as educators and/or raters of clinical performance were selected as potential raters. A panel of project team members established consensus ratings on 24 exemplar study videos to be used as gold standards for rater training and assessment; these videos demonstrated a range of performances in each of the scenarios. Raters participated in a 2-day in-person training session. They were instructed on the use of the online rating software and practiced viewing and rating the exemplar videos. Project team members mentored the raters, providing one-on-one guidance, first in person and then *via* videoconference. Rater calibration was assessed regularly during training until the rater CPE ratings matched the consensus ratings exactly, their BARS scores were no more than one point from the consensus rating, and their performance ratings were within the same preliminary bin for the holistic HS and team ratings (see descriptions in the following paragraph and Supplemental Digital Content 4, http://links.lww.com/ALN/B483). Seven raters successfully completed the training and were able to rate performances in all four scenarios consistently. Raters were compensated.

After training, raters rated the randomly assigned videos of each recorded encounter an average of 1 yr after they were performed *via* a Web-based, secure application that allowed for as much review as needed to apply the scoring metrics (Supplemental Digital Content 7, http://links.lww.com/ALN/B486). The software allowed the reviewer to mark each CPE as it was observed. A CPE was counted as having been performed if the HS, FR, or a confederate under their direction completed it at any time during the encounter. Raters then scored the holistic technical and behavioral performance of the HS and the physician team (*i.e.*, HS and FR working together) by assigning the performance to one of three bins (poor, medium, or excellent) and then choosing one of three levels (low, medium, or high) within that bin (fig. 2). Thus, scores one to three were in the poor bin; four to six in the medium bin; and seven to nine in the excellent bin. This scoring system was chosen over a simple ordinal scale because it simplifies the rating process and improves rater reliability.[36] For behavioral ratings, the raters scored participants using the BARS, which is composed of detailed anchoring statements describing expected performance of those falling in the poor and excellent bins for each scale. Raters made a summative, binary (yes/no) assessment of overall performance based on the SME-chosen query: "Did this person [or team] perform at the level of a board-certified anesthesiologist?" The primary (HS) anesthesiologist was rated first, followed by the physician team. The raters also assessed whether the degree of standardization of scenario delivery was sufficient for study inclusion (*e.g.*, were there any scenario deviations serious enough to render the encounter manifestly different than intended).

Raters received batches of videos in a predetermined, counterbalanced order. The same rater was not assigned multiple encounters conducted at a single site on the same day. The raters were instructed not to score a performance if they recognized a participant.

Data Management

Data were collected directly into the study database portal *via* preconfigured data entry forms (Supplemental Digital Content 7, http://links.lww.com/ALN/B486). For logistical reasons (*e.g.*, number of courses, number of

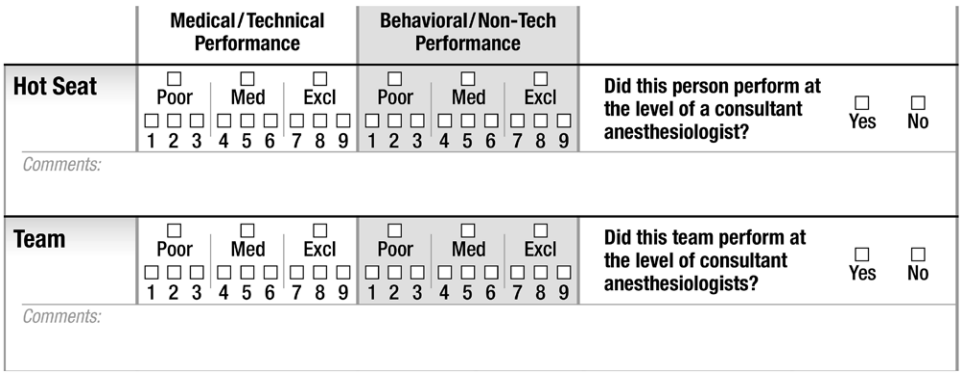

Fig. 2. Scoring rubric used for holistic performance ratings. The tool used by the trained video raters to score the participant's overall technical (*i.e.*, medical or clinical) and behavioral (*i.e.*, nontechnical or teamwork) performance. The raters first ascertained whether the technical performance of the hot-seat participant was either poor or excellent (Excl); if neither, it was determined to be in-between (Med). They then rated, within the selected performance bin, whether the performance was closest to lowest within that bin or highest; again, if neither, it was medium. Thus, a performance rated as a "7" was so categorized because it was overall excellent but low within that category.

participants per course, and efficiency of recruitment), the distribution of participant enrollment was uneven across the sites (Supplemental Digital Content 8, http://links.lww.com/ALN/B487).

Of the 342 BCAs entered into the database as participants, 24 were not an HS participant; these were all in scenarios where an extra FR was needed for an HS doing a second study scenario. For the 318 remaining study encounters, 26 (8.2%) were excluded from the final dataset due to obvious scenario standardization issues (*e.g.*, outright mannequin failure in the middle of the scenario) or inadequate audio/video capture. The raters flagged an additional eight videos as unratable, and these were excluded from the final dataset of 284 encounters (net yield of 89%).

Statistical Analysis

Reliability of Scores. Fifty encounters were scored by more than one rater. To estimate interrater reliability, 39 randomly selected encounters were scored independently by at least two raters. Variance components were calculated by scenario to estimate interrater reliability based on a model where two (of the seven) randomly selected raters provided scores. For the summative binary score, κ was calculated.

Association between Participant Characteristics and Performance. CPE data were summarized as the number and percentage of encounters in which each CPE was observed as present or absent. When an encounter was rated more than once, a CPE was scored as not performed only when all of the raters agreed. Binomial logistic regression and the associated likelihood ratio (LR) tests quantified the associations between the odds of CPE completion and participant demographics, accounting for scenario (table 1).

To derive the HS and team technical and behavioral scores in the 39 double-rated encounters, we averaged the ratings, rounding to the nearest integer. Proportional odds logistic regression and the associated LRs tested the associations between technical and behavioral performance and participant demographics, adjusting for scenario. Although the repeated ratings may be correlated among the 24 participants who performed in the HS in two different scenarios, there was insufficient information in these data to model the correlation directly (*e.g.*, using a mixed-effects regression method). Thus, these ratings were treated as independent encounters.

For the binary score in double-rated encounters, a participant's performance was only rated as not meeting the board-certified anesthesiologist criteria when all of the raters agreed (*i.e.*, all rated it "no"). Binary logistic regression and the associated LRs tested the associations between the odds of being rated a board-certified anesthesiologist and participant demographics, adjusting for scenario. The effects of each covariate were summarized using odds ratios with Wald-type 95% CI.

Because the HS and team scores were paired, a McNemar test[37] was used when assessing the fraction of technical and behavioral scores that fell in the lowest bin, as well as the fraction of performances that were rated as performing at the BCA level.

As an exploratory analysis, our assessments of hot seat and team performance were additionally adjusted by whether the provider requested assistance (*i.e.*, "call for help").

Results

A total of 263 unique HS participants performed in 284 encounters. Table 1 shows demographic information for study sample participants and several sources of data characterizing comparative population-based cohorts. When compared to all BCAs (data provided by the American Board of Anesthesiologists) and all physicians billing Medicare who self-identified as anesthesiologists (data provided by the American Society of Anesthesiologists), our study cohort was younger, had proportionately more females, and were more likely to be fellowship trained (all $P < 0.001$). These differences were less pronounced when the study cohort was compared to all BCAs in the MOCA process. The proportion of the study cohort who self-identified as being board-certified in chronic pain (10.1%) was similar to that of the Medicare billing sample (14.0%). Compared with all BCAs in the MOCA process, our cohort was twice as likely to be board-certified in critical care medicine (16.4 *vs.* 8.1%, $P < 0.001$).

Compared with the 3,461 MOCA simulation course participants in calendar years 2013–2014, the study cohort was significantly more likely to report practicing in an academic setting (47.1 *vs.* 28.0%, $P < 0.01$). Similarly, the study cohort was significantly less likely to report working in a community practice setting (49.8 *vs.* 66.0%, $P < 0.01$).

Interrater Reliability

Interrater reliability for the CPEs (percent of checklist items attained) ranged from 0.77 (myocardial infarction) to 0.93 (malignant hyperthermia) across the four scenarios (mean = 0.85). The average interrater reliability across scenarios for HS technical and behavior ratings were 0.72 and 0.83, and for team ratings they were 0.64 and 0.72, respectively. The interrater reliability for the BARS was 0.66. For the HS summative binary score, $\kappa = 0.48$; raters disagreed in 11 of 39 (28.2%) encounters with multiple ratings. For the team summative score, $\kappa = 0.27$, with disagreement in 14 (30.4%) of the encounters.

CPE Ratings

Across all of the encounters, 81% (interquartile range [IQR], 75 to 90%; table 3) of the CPEs were observed, with a range of 42 to 100%. The highest frequency of observed CPEs was in the LAST (85% [IQR, 75 to 85%]) and lowest in the hemorrhage scenario (77% [IQR, 71 to 88%]). In 46% of encounters, at least four CPEs were missed. Across all of the scenarios, 93% of participants called for help before the time when the first responder would have been sent into the scenario anyway. The likelihood of CPE

Table 2. Number (Percentage) of Board-certified Physicians Rated as Performing Representative CPEs for the Four Scenarios

Scenario	Critical Performance Element*	Specific Category	High-level Category	No. (%) CPE Rated as Performed (by Any Rater)
Laparoscopic surgery with retroperitoneal hemorrhage	Administers IV fluids (open wide infusion or deliberate bolus)	Administers IV fluids	Initial therapy (action)	64 (95.5)
	Administers vasopressor (phenylephrine: first dose 50–200 μg or ephedrine: 5–10 mg)	Administers initial treatment	Initial therapy (action)	67 (100)
	Requests delivery of blood to the OR for possible transfusion	Prepares for further treatment	Ongoing therapy (action)	58 (86.8)
	Starts administering a unit of type-specific or trauma blood	Administers treatment	Ongoing therapy (action)	28 (41.8)
	Requests that the surgeon open the abdomen	Initiates definitive treatment	Advanced communication	49 (73.1)
Sedation for gynecologic procedure with local anesthetic toxicity	Requests that the surgeon stop (or pause) the procedure	Requests diagnostic studies	Advanced communication	65 (80.3)
	Manages airway with oxygen and assisted (or controlled) ventilation and places an advanced airway device	Administers initial treatment	Initial therapy (action)	81 (100)
	Administers vasopressor (phenylephrine: first dose 50–200 μg, ephedrine or epinephrine in small doses)	Administers initial treatment	Initial therapy (action)	64 (79.0)
	Discusses clinical concerns with surgeon	Discuss concerns with proceduralist and/or team	Initial communication	78 (96.3)
	Administers initial dose of lipid emulsion of 100 ml (1.5 ml/kg) either *via* syringe or bolus infusion	Administers definitive treatment	Ongoing therapy (action)	76 (93.8)
	After diagnosis of LAST, adjusts ACLS management (reduced dose of epinephrine, avoid use of vasopressin, calcium channel and β-blockers, and local anesthetics)	Administers ongoing treatment	Ongoing therapy (action)	6 (7.4)
Endoscopic retrograde cholangiopancreatography with postoperative malignant hyperthermia	Requests that lab tests be drawn (minimum of an arterial blood gas and a potassium level)	Requests diagnostic studies	Advanced communication	56 (93.3)
	Manages airway with oxygen and assisted (or controlled) ventilation and places an advanced airway device	Manages the airway	Initial therapy (action)	55 (91.7)
	Requests MH cart (or box) containing dantrolene	Calls for (further) therapy	Advanced communication	55 (91.7)
	Gives at least one vial of dantrolene	Administers definitive treatment	Ongoing therapy (action)	46 (76.7)
Small-bowel obstruction with unstable atrial fibrillation followed by a myocardial infarction	Announces that the rhythm is or could be atrial fibrillation	Announces diagnosis	Initial communication	36 (47.4)
	Announces that the rhythm is unstable or that there is hypotension	Announces situation	Initial communication	76 (100)
	Administers initial dose(s) of suitable vasoconstrictor	Administers initial treatment	Initial therapy (action)	69 (90.1)
	Administers reasonable dose(s) of drug(s) to slow heart rate	Administers initial treatment	Initial therapy (action)	56 (73.7)
	Calls for crash cart and/or defibrillator	Calls for (further) therapy	Advanced communication	75 (98.7)
	Performs synchronized cardioversion with ≥120 J	Administers definitive treatment	Ongoing therapy (action)	62 (81.6)
	Notifies surgeon/team about ST elevation	Announces situation	Initial communication	65 (85.5)
	Discusses treatment options with cardiologist and/or surgeon including at least two of the following: (1) transfer to cardiac catheterization laboratory; (2) heparin infusion in the OR; (3) additional hemodynamic support; (4) intra-aortic balloon counter-pulsation; (5) amiodarone infusion in the OR; and/or (6) use of transesophageal or transthoracic echocardiography	Administers additional therapy	Ongoing therapy (action)	73 (96.1)

Each scenario had 16 to 20 CPEs. For brevity, only a representative subset is shown here. For the complete set of CPEs, see Supplemental Digital Content 4 (http://links.lww.com/ALN/B483).

* For CPEs deemed essential to the progression of a scenario, if the participant did not initiate the action independently, a scripted prompt by a confederate was provided at a specified time point to try to elicit the expected behavior. For example, in the atrial fibrillation/myocardial infarction scenario, if the participant(s) had not cardioverted the patient by 12 min, the confederate surgeon would prompt the behavior by stating, "Why don't we shock this patient?"

ACLS = advanced cardiac life support; CPE = critical performance element; LAST = local anesthetic systemic toxicity; MH = malignant hyperthermia; OR = operating room.

Table 3. Overall Performance Ratings by Scenario

Performance Metric		Overall Ratings (n = 284)	Ratings by Scenario: Hemorrhage (n = 67)	LA Toxicity (n = 81)	MH (n = 60)	Afib/STEMI (n = 76)
% of CPEs rated as completed*	—	81.3 (75.0–89.5) [42.1, 100.0]	76.5 (70.6–88.2) [58.8, 100.0]	85.0 (75.0–85.0) [55.0, 95.0]	84.2 (73.7, 89.5) [42.1, 100.0]	81.3 (75.0–93.8) [50.0, 100.0]
Technical score*†‡	HS	5 (3–7) [1, 9]	6 (5–7) [2, 9]	4 (3–6) [1, 8]	6 (4–7) [2, 9]	5 (3–6) [1, 9]
	Team	6 (4–7) [1, 9]	7 (5–8) [3, 9]	5 (4–6) [1, 8]	6 (4–8) [2, 9]	5 (3–7) [2, 9]
Behavioral score*†‡	HS	5 (4–7) [1, 9]	6 (5–7) [2, 9]	4 (3–7) [2, 8]	6 (4–8) [1, 9]	5 (3–7) [2, 9]
	Team	5 (4–7) [1, 9]	7 (5–8) [3, 9]	5 (4–7) [2, 8]	6 (4–8) [1, 9]	5 (4–7) [2, 9]
% Satisfactory global binary score (*i.e.*, yes rating)†‡§	HS	69.7 (198/284)	85.1 (57/67)	56.8 (46/81)	76.7 (46/60)	64.5 (49/76)
	Team	79.6 (226/284)	88.1 (59/67)	79.0 (64/81)	81.7 (49/60)	71.1 (54/76)
BARS: overall‡	—	5.4 (3.5–7.1) [1.8, 9.0]	6.5 (5.0–7.3) [2.5, 9.0]	4.3 (3.0–6.8) [2.0, 8.8]	6.1 (4.0, 7.6) [2.0, 8.8]	4.8 (3.5–6.8) [1.8, 8.3]
BARS: communication‡	—	5 (4–7) [1, 9]	7 (5–8) [2, 9]	4 (3–7) [2, 9]	6 (5–8) [1, 9]	5 (3–7) [2, 9]
BARS: decision‡	—	5 (3–7) [1, 9]	6 (5–7) [2, 9]	4 (3–6) [1, 8]	6 (4–8) [2, 9]	5 (3–6) [1, 8]
BARS: teamwork‡	—	6 (4–7) [1, 9]	7 (5–8) [2, 9]	4 (3–7) [2, 8]	6 (4–8) [1, 9]	5 (3–7) [2, 9]
BARS: vigilance‡	—	5 (3–7) [2, 9]	7 (5–7) [2, 9]	4 (3–7) [2, 8]	6 (4–8) [2, 9]	5 (3–7) [2, 9]

*Technical and behavioral ratings (1 to 9 range) and CPE (%) completion are presented as median (IQR) [minimum, maximum].

†Significant ($P < 0.001$) differences by team *versus* HS using McNemar test are shown.

‡Significant ($P < 0.05$) differences by scenario using likelihood ratio test are shown.

§Global binary (yes/no) ratings are presented as % yes (fraction).

Afib = atrial fibrillation; BARS = behaviorally anchored rating scales; CPE = critical performance element; HS = hot shot; IQR = interquartile range; LA = local anesthetic; MH = malignant hyperthermia; STEMI = ST segment elevation myocardial infarction.

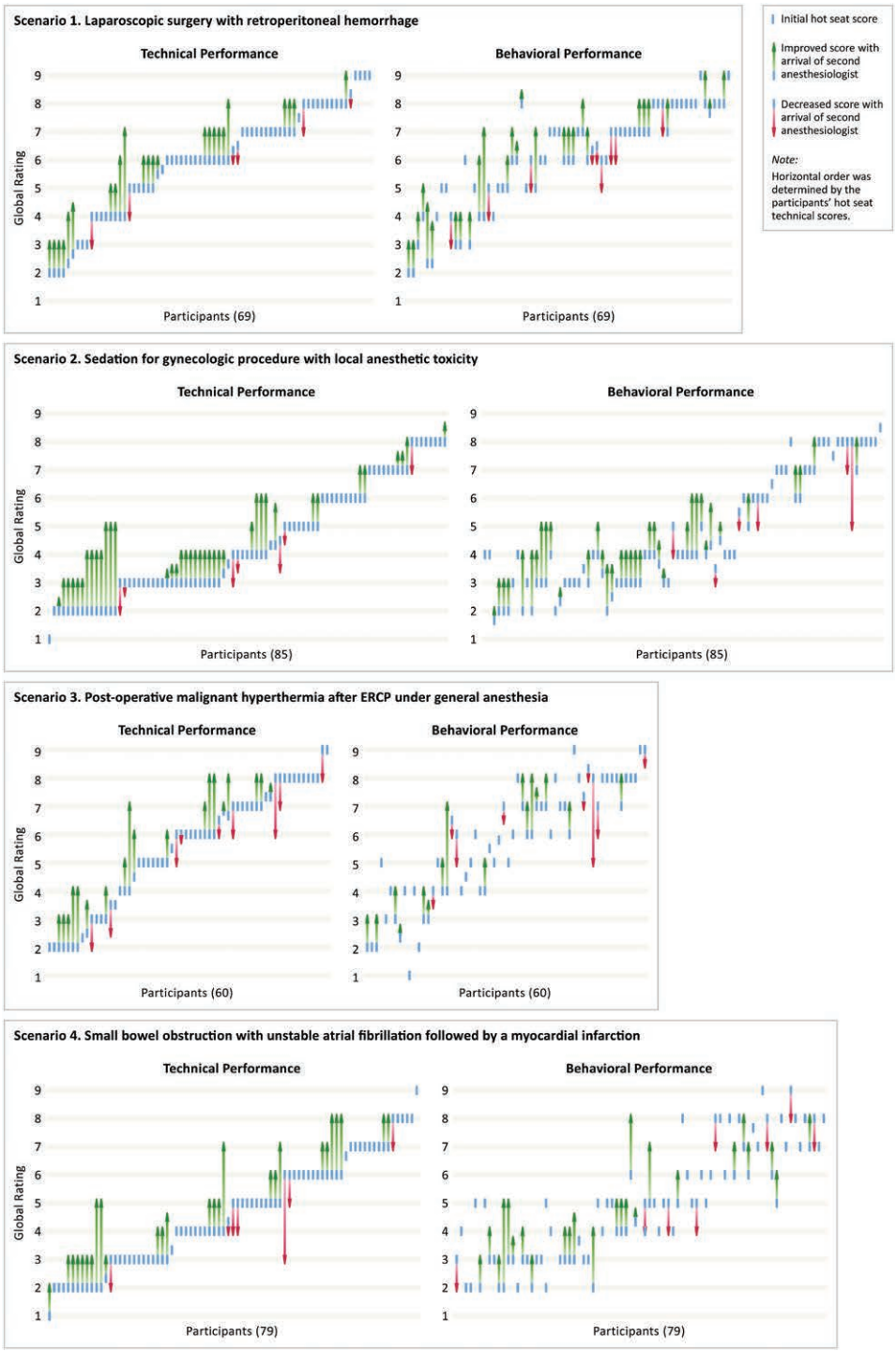

Fig. 3. Incidence of primary provider (hot seat) and team technical and behavioral ratings for each scenario. The holistic hot-seat technical (on *left*) and behavioral (on *right*) holistic (1 to 9) ratings are shown for each of the four scenarios as a *blue vertical rectangle*. If the associated team rating was different than the hot-seat rating for that encounter, it is shown as an *arrowhead*: *green and pointing upward* when the team rating was better than the hot-seat rating and *red and pointing downward* when the team rating was worse than the hot-seat rating. The participants for each scenario are ordered by their hot-seat technical performance from lowest to highest. The same hot-seat participant order is used for the behavioral rating. The data show that it is much more common for the arrival of the first responder to result in improved technical and behavioral (team) ratings. Note that, in cases where more than one rater rated the encounter, the average of all ratings were used. *Boxes* containing scenario-specific performance scores are different widths because there was a different number of participants in each scenario (*shown in parenthesis as part of the x-axis labels*).

completion differed by scenario but not by site. Table 2 provides a representative listing of CPEs by scenario and their incidence of observed performance; for a full list of CPEs, see Supplemental Digital Content 4 (http://links.lww.com/ALN/B483).

Technical and Behavioral Scores

The median technical performance rating of HS participants was five; ratings spanned the full one to nine scale. Performance varied significantly only by scenario (LR test $P < 0.001$), after adjusting for HS demographic and practice characteristics (table 3). Across all of the scenarios, team technical ratings were higher than HS ratings because the arrival of the FR often improved performance (fig. 3). Overall, 30% of the HS and 21% of team technical scores fell within the lowest performance bin (McNemar test $P < 0.001$).

Overall BARS performance was 5.4 (IQR, 3.5 to 7.1), spanning the metric range from one to nine (table 3). BARS performance varied significantly by scenario (LR test $P < 0.001$) and participant age ($P = 0.037$), after adjusting for HS demographic and practice characteristics. Similarly, the median global behavioral rating was five, spanning the full scoring range, and varied significantly by scenario (LR test $P = 0.001$). Higher participant age ($P = 0.004$), but not previous simulation experience (yes or no) or other individual factors, was associated with lower behavioral ratings. Overall, in 25% of encounters, HS behavioral scores fell in the lowest bin. Only 14% of team behavioral scores were in this bin (McNemar test $P < 0.001$ when compared with the HS scores). As seen in figure 3, the arrival of the first responder more often improved than degraded the behavioral score.

Binary Ratings

In 70% of encounters, the HS participant was rated as "having performed at the level of a board-certified anesthesiologist." Performance varied significantly by scenario (LR test $P = 0.002$), with the worst scores in the LAST scenario (43% unsatisfactory). The arrival of FRs frequently improved low HS performances; 34% of unsatisfactory HS scores were followed by satisfactory team ratings, whereas only one (<1%) satisfactory HS score was associated with an unsatisfactory team score (McNemar $P < 0.001$). HS participants in the under 40-yr age group were more likely to receive a satisfactory binary rating relative to the 40- to 50-yr (odds ratio = 1.86 [95% CI, 1.17–3.10]) and over 50-yr (odds ratio = 2.70 [95% CI, 1.36–5.35]) age groups. HS binary ratings were not associated with any other participant characteristic.

Discussion

We created a simulation-based assessment process that was reproducible across testing centers, yielded reasonably reliable assessment scores, and measured the performance of important crisis management skills of board-certified anesthesiologists. Based on multiple metrics, there was appreciable variability in the performance of board-certified anesthesiologists. CPEs were commonly omitted. Approximately 30% of encounters were rated as "poor" for overall individual technical or behavioral performance or as "unsatisfactory" for the binary rating. Arrival of the second physician commonly improved performance ratings.

The gaps in performance documented in this simulation study included four broad areas of crisis management: (1) escalation of therapy where first-line therapy is not working (*e.g.*, using epinephrine or vasopressin when phenylephrine or ephedrine and fluids are not appreciably affecting hypotension); (2) using available resources (*e.g.*, calling for help when conditions have deteriorated appreciably); (3) speaking up or engaging other team members, especially when action by them is required (*e.g.*, asking the surgeon to change the surgical approach when it is essential to effective treatment); and (4) following evidence-based guidelines (*e.g.*, giving dantrolene to a patient with obvious MH).

Age was the only statistically significant predictor of performance. Younger participants received higher ratings than older ones, although few participants were more than 60 yr of age. Our 35 participants who were 50 yr of age or older were demographically similar to the 135 participants who were 40 yr of age or younger (other than years in practice), except that they were less likely to be enrolled in MOCA (91 *vs.* 99%; $P = 0.026$) and more likely to practice in an anesthesia team model (97 *vs.* 75%; $P = 0.014$). Younger and older physicians may differ in many other ways, including the existence or nature of previous crisis management training, comfort with simulation, or simply time since completion of residency training. Degradation of skills from lack of practice or physiologic aging may explain our finding.[38]

Compared with all anesthesiologists who bill Medicare, with all board-certified anesthesiologists, and even with all BCAs in the MOCA process, our study cohort was younger and more likely to be female, be fellowship trained, and work in an academic practice. If anything, these factors may be more likely to bias our study sample toward those who were more confident about their abilities, more familiar with crisis management, and/or more comfortable with simulation and/or being assessed. We believe that such individuals would be more likely to perform better than those without these attributes. Thus, these results may well be biased toward better performances (in simulation) than might be seen in a fully representative population of all practicing anesthesiologists.

Relationship of This Study's Results to Those of Previous Studies

Our study validates and expands on results from other studies[17,19,39,40] that have assessed performance of anesthesia professionals (often residents) using simulation. We chose to study experienced anesthesiologists (BCAs) because they are the least-studied population yet provide the most patient care. Our sample of 268 BCAs was more than three times larger than that of Devitt *et al.*[40] (79 anesthesiologists) and

eight times larger than that of Henrichs *et al.*[17] (35 anesthesiologists plus 26 certified registered nurse anesthetists). Similar to previous investigations, we assessed the technical (*i.e.*, clinical) responses to simulated uncommon events and found a wide variability in the performance of fully trained anesthesia professionals. Like others, we also documented performance deficits, with a substantial rate (20% or higher) of performances rated as "poor," including many with omissions, errors, or delays in actions deemed by clinical experts *a priori* to be critical to successful patient care.

Our study methods and results go well beyond those of previous research. Previous studies concentrated on developing tools to reliably and validly measure the ability of individual clinicians. To generate reproducible scores, participants typically performed a number of short, focused scenarios (*e.g.*, 300 s) with quickly observable and unambiguous signs and symptoms.[19] Participants often worked completely alone (*i.e.*, no surgeon, nurse, or help to be called). These types of scenarios are less representative of real clinical situations and, at least from a content perspective, may yield less valid performance metrics. Finally, many previous studies assessed only technical performance, ignoring the important contribution of communication and teamwork in patient care. Our goal was to measure the performance of a large sample of experienced anesthesiologists in single scenarios. Although this strategy cannot yield reliable individual ability estimates, it allowed us to investigate group performance in simulations of higher complexity and ecologic validity.

To achieve our study aims, we designed moderate-length scenarios that had multiple credible diagnoses and treatments, thus replicating typical challenges of real events. Our participants worked in a team with trained confederate clinicians and with a second BCA in the latter half of each scenario. This design provided an environment where we could measure both technical and behavioral performance.

Relevance to Real-world Practice

Some may dismiss the variable and sometimes suboptimal performances observed in our study as the result of the artificiality of a simulated setting and contend that such deficiencies do not occur during patient care. However, we observed a variety of performance deficiencies that have been reported previously in both real and simulated events.[41] For example, almost one fifth of participants in the atrial fibrillation/myocardial infarction scenario failed to cardiovert unstable atrial fibrillation, and a similar proportion failed to request that the surgeon open the abdomen in the face of exsanguination in the hemorrhage scenario. Performance gaps observed in these simulations are known to occur during patient care, including deficiencies or delays in the following: (1) transfusing during catastrophic hemorrhage[42]; (2) cardioversion of unstable arrhythmias[43]; (3) applying appropriate pharmacologic treatment of significant hypotension[44]; and (4) effective communication between surgical and anesthesia personnel. Failure to engage the surgeon in a timely and effective fashion, including reluctance to suggest that the surgeon obtain help or use an alternate surgical approach,[45] is a well-documented pitfall during both real and simulated cases.[42,46,47] That performance gaps identified in this study occur and have been associated with poor outcomes in real cases[43,48–50] provides evidence to support the construct validity of our results.

Using comparable high-acuity scenarios, one would expect similar findings among other types of anesthesia professionals, emergency physicians, intensivists, interventional cardiologists, or surgeons. Although many other types of clinicians may only rarely face high-acuity critical events, some type of crisis management is required in nearly every clinical domain. Furthermore, issues of interprofessional communication and teamwork, effectively measured in our simulation scenarios, are important across all areas of health care.

Study Limitations

The simulated clinical environment, although realistic, was not identical to the participants' own practice environments. If faced with similar real emergencies in their familiar clinical setting with an established team of colleagues, these participants would probably perform better. Furthermore, since this study was grafted onto a learning experience, participants may not have been as motivated to perform as well as if it had been a test or a real-world crisis. Yet, many BCAs routinely find themselves in suboptimal, unstandardized, or unfamiliar environments where adaptability is essential to effective performance.

Simulating human pathophysiology is challenging, and imperfect portrayal of clinical signs and symptoms of real patients could have induced omission of correct actions or commission of incorrect ones. To mitigate this, participants were familiarized thoroughly with the mannequin and simulated care environment and were studied after having participated in or seen at least one encounter. Notably, two thirds of participants had previous simulation experience. The scenarios were designed to be realistic and appropriate to assess performance.[4] Each one contained multiple reinforcing cues to present unambiguous depictions of key events and to produce a realistic progression. Thousands of board-certified anesthesiologists have judged simulation-based MOCA courses to be effective, realistic, and relevant to their practices.[4,51] Furthermore, anesthesiologists have indicated that simulation-based training facilitated meaningful practice improvements that often had impact beyond their own individual practices.[51] Nevertheless, it is possible that some participants might have performed better with more practice in the simulation environment. Some participants may not have clinical practices that expose them to the types of cases presented during the course. However, the SMEs felt that these four scenarios typified events that all BCAs should be expected to manage.

All four scenarios were designed to depend on management according to established guidelines (*e.g.*, advanced

cardiac life support, MH, LAST). The SMEs established the CPEs for each scenario. We subsequently trained the raters based on these criteria. Many of the actions for which performance gaps were seen are indeed widely accepted as appropriate crisis management practices (see table 2 for examples).

Grafting the study onto the MOCA simulation courses constrained our study design. Course logistics mostly restricted participants from being studied more than twice–once in the HS and once as an FR. Each MOCA encounter was followed by a facilitated peer debriefing, which could have influenced subsequent performances. Querying study participants systematically about why they did what they did might have yielded greater understanding of their performance,[52] but it would have adversely affected debriefing quality and course flow for all of the course attendees.

Although raters were well trained, used sophisticated video review software, and provided reasonably reliable ratings, they could have missed subtle aspects of participant performance. Notwithstanding, we sought to measure performance fairly, within the constraints of the study design, to determine an upper bound of participant performance. For example, when more than one rater scored an encounter for the CPEs and binary ratings, we used the most favorable score. Interrater reliability was lowest for the HS and team binary ratings, where raters disagreed in approximately 30% of the encounters. There could be several explanations for why reliability was lower than for global technical and behavioral ratings of the same performances: (1) the raters agreed on the level of performance observed but had different opinions about how to rate it, possibly in part because the binary score was not explicitly defined or anchored; (2) the binary rating was the only metric that combined both technical and behavioral elements, and raters may have disagreed about the relative importance of these two aspects of performance; or (3) the raters weighted different attributes of the performance differently over time. In future research, investigators might use our archive of video recordings to test different approaches to address these limitations of holistic performance ratings.

The absence of previous simulation experience was not an independent predictor of rated performance. Because this was a yes-or-no question, we do not know how much previous simulation experience each participant had, when it might have occurred, or the type of any such experience (*e.g.*, if it targeted acute event management as did our scenarios). Furthermore, many of our demographic variables are not fully independent, so, for example, more recently trained BCAs are by definition younger and could be expected to have had more (and perhaps different types of) previous simulation-based training.

Significance and Future Directions

Practicing anesthesiologists are expected to be competent, to identify gaps in their knowledge and performance, and to participate in continuing medical education and practice improvement programs to address these gaps.[53] In particular, they must be able to detect and manage time-sensitive, potentially lethal events. Yet, the literature suggests that suboptimal individual clinician performance still contributes to adverse events during perioperative care.[54–56] The ability of an individual clinician involves a myriad of skills that cannot be captured by any single method of assessment, whether it is written or oral examinations, prospective or retrospective performance reporting, or with simulations. Nonetheless, although performance during simulated crisis events may not exactly reflect actual care, the results of this study indicate that simulation can play a key role as one important component of clinician assessment.

We measured population performance, not individual competence. Performance in a single scenario is an inadequate basis to judge the competence of any individual provider. If simulation were to be considered for use in summative performance assessment of any kind, it is clear that many scenarios would be needed to yield a reliable and valid estimate of ability. However, the data of this study, derived from a large sample of practicing anesthesiologists, provide useful feedback for training programs at all levels, from residency through MOC.

Continuing medical education and professional development currently relies largely on physicians' self-assessment of their learning.[57] Yet, it is well established that physicians have a limited ability to correctly ascertain their learning needs.[58] Furthermore, less competent physicians may be more likely to overestimate their current knowledge and abilities.[58] To improve performance, humans require accurate information about specific deficiencies (or gaps) and directed feedback from experts or a peer group to be able to inculcate and then strive, through deliberate practice, to achieve these learning goals.[1] Simulation-based training with debriefing, such as that offered as part of MOCA, provides such a structure.

Mannequin-based simulation is well suited for assessing the management of high-acuity rare events and for crisis-resource management.[59] Consequential, even potentially lethal, clinical performance gaps identified across our study cohort could be targeted for recurrent interprofessional training of both trainees and experienced personnel. Although dire events are rare, the skills needed in crises (anticipation, prevention, identification, and management of challenging occurrences) are universally important attributes of clinician expertise. Simulation allows for recurrent standardized assessment of individuals and teams, with appropriate retraining as indicated. Simulation-based training, often as part of a multimodal intervention, has been shown to improve patient care.[33,60,61]

Our findings suggest that the responses of some experienced practicing anesthesiologists during life-threatening, real-world events are suboptimal. Although we cannot say with certainty whether anesthesiologists who perform well

or poorly in simulation will respond similarly during actual events, collective experience and the literature suggest that clinician performance during real-world crises is also variable[62,63] and imperfect.[64]

Implications for Real-world Crisis Management. If performance in emergencies is suboptimal, why does harm to patients seem rare? First, although serious adverse events are relatively uncommon, when they do occur, failure to rescue may be attributed to patient illness or may go unreported.[65,66] Second, individual clinicians may self-select their practice to be specialized or even circumscribed in complexity. Clinicians thought to be lower performing than others may be protected by scheduling simpler cases or other support mechanisms. Third, clinicians uncommonly work in isolation; they are part of care systems designed in part to reduce the risk of and enhance the recovery from untoward events.[67] In some settings, many supporting clinicians can be called in to assist in an emergency, whereas in this study only one responding BCA was provided. The arrival of the second BCA usually improved performance and perhaps more so with lower-performing HS participants. The availability of experienced help in real crises depends on practice setting and time of day; many private-practice MOCA course participants comment that help from other BCAs is rarely available to them. Nevertheless, a cornerstone of safe and effective care systems remains high-performing individual clinicians, working alone and together in teams, during routine, nonroutine, and crisis situations.[68,69]

Implications of the Performance Gaps Observed. How might the performance gaps that we observed be addressed? Many parallel strategies are possible; most are commonplace in other industries of high dynamism and high intrinsic hazard, such as aerospace, nuclear power, the military, or the maritime industry. These include, for example, recurrent high-fidelity simulation training of both trainees and experienced physicians, sometimes including other team members, on the recognition and management of specific events and the use of crisis resource management techniques, as well as practice working in clinical teams to manage unfolding adverse events. Another strategy is the regular and uniform use of protocol guidance optimized for real-time use *via* emergency manuals and other cognitive aids. Other industries conduct regular formative performance assessment of individuals and teams and provide appropriate practice improvement activities, as indicated.

We need to understand more deeply why individual physicians and other clinicians do not always execute the kind of decision-making and action that are expected. We also need to investigate in greater detail the decision-making, event management, and team leadership of experienced physicians in many different simulated situations. This might require a full day of simulation training for each participant, making such programs costly, but necessary, along the path of better understanding of how to continue to improve physician performance in the pursuit of patient safety.

Acknowledgments

The authors acknowledge (in alphabetical order) the substantive contributions (*e.g.,* served as a subject matter expert [SME] or video rater, assisted in scenario development, assisted in manuscript preparation) of: Russ Beebe, B.A. (Center for Research and Innovation in Systems Safety, Vanderbilt University Medical Center, Nashville, Tennessee), Thomas Belda, B.S. (Mayo Clinic Multidisciplinary Simulation Center, Rochester, Minnesota), Edwin A. Bowe, M.D. (University of Kentucky College of Medicine, Lexington, Kentucky), Richard H. Blum, M.D., M.S.E. (Children's Hospital of Boston, Boston, Massachusetts), Brian Cammarata, M.D. (Old Pueblo Anesthesia, Tucson, Arizona), Douglas B. Coursin, M.D. (University of Wisconsin–Madison School of Medicine and Public Health, Madison, Wisconsin), Gregory J. Crosby, M.D. (Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts), Deborah J. Culley, M.D. (Brigham and Women's Hospital, Harvard Medical School), Anthony Dancel, B.S. (Massachusetts General Hospital, Center for Medical Simulation, Boston, Massachusetts), Andrew Kline, B.A. (Vanderbilt Comprehensive Care Clinic, Vanderbilt University Medical Center), Jordan Halasz, B.S. (Center for Experiential Learning and Assessment, Vanderbilt University Medical Center), Steven C. Hall, M.D. (Northwestern University Feinberg School of Medicine, Chicago, Illinois), Hans J. Hinssen, Dipl. Ing. (Penn State Hershey Clinical Simulation Center, Hershey, Pennsylvania), Joy Hawkins, M.D. (University of Colorado School of Medicine, Aurora, Colorado), Alan Johnstone, B.S. (Vanderbilt University Medical Center), Stephen J. Kimatian, M.D. (The Cleveland Clinic, Cleveland, Ohio), Jerome Klafta, M.D. (University of Chicago Pritzker School of Medicine, Chicago, Illinois), John Lutz, B.S. (Winter Institute for Simulation Education and Research, Pittsburgh, Pennsylvania), Christie Mulvey, B.S. (Penn State Hershey Clinical Simulation Center), Robert Nadelberg, M.D. (Massachusetts General Hospital, Center for Medical Simulation), Viren Naik, M.D., Med., M.B.A. (University of Ottawa Skills and Simulation Centre, Ottawa, Canada), Edward Nemergut, M.D. (University of Virginia School of Medicine, Charlottesville, Virginia), Eric Porterfield, M.S., M.S.N., R.N., F.N.P.-B.C. (Vanderbilt University Medical Center), Niraja Rajan, M.D. (Penn State Hershey Medical Center, Hershey, Pennsylvania), Lauryn Rochlen, M.D. (University of Michigan School of Medicine, Ann Arbor, Michigan), Ryan Romeo, M.D. (University of Pittsburgh School of Medicine and Winter Institute for Simulation Education and Research, Pittsburgh, Pennsylvania), Michael Seropian, M.D. (Oregon Health and Science University, Portland, Oregon), Ljuba Stojiljkovic, M.D. (Northwestern University Feinberg School of Medicine, Chicago, Illinois), Huaping Sun, Ph.D. (The American Board of Anesthesiology, Raleigh, North Carolina), Jeff Taekman, M.D. (Duke University School of Medicine, Durham, North Carolina), Christina Valle (Center for Medical Simulation), William B. Waldrop, M.D. (Baylor College of Medicine, Houston, Texas), and Cynthia Wong, M.D. (University of Iowa Carver College of Medicine, Iowa City, Iowa).

Research Support

Supported in part by grants from the Agency for Healthcare Research and Quality (No. R18 HS020415), Rockville, Maryland, and the Anesthesia Patient Safety Foundation, Rochester, Minnesota (to Dr. Weinger), and by a grant from the Foundation for Anesthesia Education and Research, Schaumburg, Illinois (to Dr. Banerjee). The American Society of Anesthesiologists, Schaumburg, Illinois, allowed the project team to use their GoToMeeting teleconferencing account.

Competing Interests

The authors declare no competing interests.

Correspondence

Address correspondence to Dr. Weinger: Center for Research and Innovation in Systems Safety, 1211 21st Avenue South, Medical Arts Bldg. Suite 732, Nashville, Tennessee 37212. matt.weinger@vanderbilt.edu. This article may be accessed for personal use at no charge through the Journal Web site, www.anesthesiology.org.

References

1. Ericsson KA: Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 2004; 79(10 suppl):S70–81
2. Smith KK, Gilcreast D, Pierce K: Evaluation of staff's retention of ACLS and BLS skills. Resuscitation 2008; 78:59–65
3. Weaver SJ, Lubomksi LH, Wilson RF, Pfoh ER, Martinez KA, Dy SM: Promoting a culture of safety as a patient safety strategy: A systematic review. Ann Intern Med 2013; 158(5 pt 2):369–74
4. McIvor W, Burden A, Weinger MB, Steadman R: Simulation for maintenance of certification in anesthesiology: The first two years. J Contin Educ Health Prof 2012; 32:236–42
5. Buscemi D, Wang H, Phy M, Nugent K: Maintenance of certification in Internal Medicine: Participation rates and patient outcomes. J Community Hosp Intern Med Perspect 2012; 2
6. Counselman FL, Carius ML, Kowalenko T, Battaglioli N, Hobgood C, Jagoda AS, Lovell E, Oshva L, Patel A, Shayne P, Tabas JA, Reisdorff EJ: The American Board of Emergency Medicine Maintenance of Certification Summit. J Emerg Med 2015; 49:722–8
7. Holmboe ES, Wang Y, Meehan TP, Tate JP, Ho SY, Starkey KS, Lipner RS: Association between maintenance of certification examination scores and quality of care for medicare beneficiaries. Arch Intern Med 2008; 168:1396–403
8. Kempen PM: Maintenance of certification and licensure: Regulatory capture of medicine. Anesth Analg 2014; 118:1378–86
9. Henriksen K, Kaplan H: Hindsight bias, outcome knowledge and adaptive learning. Qual Saf Health Care 2003; 12(suppl 2):ii46–50
10. Scott IA, Phelps G, Brand C: Assessing individual clinical performance: A primer for physicians. Intern Med J 2011; 41:144–55
11. van der Vleuten CP, Schuwirth LW: Assessing professional competence: From methods to programmes. Med Educ 2005; 39:309–17
12. Gaba DM: The future vision of simulation in health care. Qual Saf Health Care 2004; 13(suppl 1):i2–10
13. Goldberg A, Silverman E, Samuelson S, Katz D, Lin HM, Levine A, DeMaria S: Learning through simulated independent practice leads to better future performance in a simulated crisis than learning through simulated supervised practice. Br J Anaesth 2015; 114:794–800
14. Stevens LM, Cooper JB, Raemer DB, Schneider RC, Frankel AS, Berry WR, Agnihotri AK: Educational program in crisis management for cardiac surgery teams including high realism simulation. J Thorac Cardiovasc Surg 2012; 144:17–24
15. Weller J, Henderson R, Webster CS, Shulruf B, Torrie J, Davies E, Henderson K, Frampton C, Merry AF: Building the evidence on simulation validity: Comparison of anesthesiologists' communication patterns in real and simulated cases. Anesthesiology 2014; 120:142–8
16. DeMaria S Jr, Samuelson ST, Schwartz AD, Sim AJ, Levine AI: Simulation-based assessment and retraining for the anesthesiologist seeking reentry to clinical practice: A case series. Anesthesiology 2013; 119:206–17
17. Henrichs BM, Avidan MS, Murray DJ, Boulet JR, Kras J, Krause B, Snider R, Evers AS: Performance of certified registered nurse anesthetists and anesthesiologists in a simulation-based skills assessment. Anesth Analg 2009; 108:255–62
18. Khanduja PK, Bould MD, Naik VN, Hladkowicz E, Boet S: The role of simulation in continuing medical education for acute care physicians: A systematic review. Crit Care Med 2015; 43:186–93
19. Murray DJ, Boulet JR, Avidan M, Kras JF, Henrichs B, Woodhouse J, Evers AS: Performance of residents and anesthesiologists in a simulation-based skill assessment. Anesthesiology 2007; 107:705–13
20. Fletcher GC, McGeorge P, Flin RH, Glavin RJ, Maran NJ: The role of non-technical skills in anaesthesia: A review of current literature. Br J Anaesth 2002; 88:418–29
21. Kohn LT, Corrigan J, Donaldson MS, Institute of Medicine (U.S.). Committee on Quality of Health Care in America: To err is human: Building a safer health system. Washington, D.C., National Academy Press, 1999
22. Weaver SJ, Dy SM, Rosen MA: Team-training in healthcare: A narrative synthesis of the literature. BMJ Qual Saf 2014; 23:359–72
23. Steadman RH: The American Society of Anesthesiologists' national endorsement program for simulation centers. J Crit Care 2008; 23:203–6
24. McIvor WR, Banerjee A, Boulet JR, Bekhuis T, Tseytlin E, Torsher L, DeMaria S Jr, Rask JP, Shotwell MS, Burden A, Cooper JB, Gaba DM, Levine A, Park C, Sinz E, Steadman RH, Weinger MB: A taxonomy of delivery and documentation deviations during delivery of high-fidelity simulations. Simul Healthc 2017; 12:1–8
25. 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
26. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R: Anaesthetists' Non-technical Skills (ANTS): Evaluation of a behavioural marker system. Br J Anaesth 2003; 90:580–8
27. Boulet JR, Murray DJ: Simulation-based assessment in anesthesiology: Requirements for practical implementation. Anesthesiology 2010; 112:1041–52
28. Boulet JR, Murray D, Kras J, Woodhouse J: Setting performance standards for mannequin-based acute-care scenarios: An examinee-centered approach. Simul Healthc 2008; 3:72–81
29. Kim J, Neilipovitz D, Cardinal P, Chiu M: A comparison of global rating scale and checklist scores in the validation of an evaluation tool to assess performance in the resuscitation of critically ill patients during simulated emergencies (abbreviated as "CRM simulator study IB"). Simul Healthc 2009; 4:6–16
30. Flin R, Patey R, Glavin R, Maran N: Anaesthetists' non-technical skills. Br J Anaesth 2010; 105:38–44
31. Weller JM, Bloch M, Young S, Maze M, Oyesola S, Wyner J, Dob D, Haire K, Durbridge J, Walker T, Newble D: Evaluation of high fidelity patient simulator in assessment of performance of anaesthetists. Br J Anaesth 2003; 90:43–7
32. Regehr G, MacRae H, Reznick RK, Szalay D: Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med 1998; 73:993–7
33. Andreatta P, Saxton E, Thompson M, Annich G: Simulation-based mock codes significantly correlate with improved pediatric patient cardiopulmonary arrest survival rates. Pediatr Crit Care Med 2011; 12:33–8
34. Watkins SC, Roberts DA, Boulet JR, McEvoy MD, Weinger MB: Evaluation of a simpler tool to assess nontechnical skills during simulated critical events. Simulat Healthc 2017; 12:69–75
35. Stewart J, O'Halloran C, Harrigan P, Spencer JA, Barton JR, Singleton SJ: Identifying appropriate tasks for the preregistration year: Modified Delphi technique. BMJ 1999; 319:224–9

36. Norcini JJ, Blank LL, Duffy FD, Fortna GS: The mini-CEX: A method for assessing clinical skills. Ann Intern Med 2003; 138:476–81
37. McNemar Q: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947; 12:153–7
38. Baxter AD, Boet S, Reid D, Skidmore G: The aging anesthesiologist: a narrative review and suggested strategies. Can J Anaesth 2014; 61:865–75
39. Schwid HA, Rooke GA, Carline J, Steadman RH, Murray WB, Olympio M, Tarver S, Steckner K, Wetstone S; Anesthesia Simulator Research Consortium: Evaluation of anesthesia residents using mannequin-based simulation: A multiinstitutional study. Anesthesiology 2002; 97:1434–44
40. Devitt JH, Kurrek MM, Cohen MM, Cleave-Hogg D: The validity of performance assessments using simulation. Anesthesiology 2001; 95:36–42
41. Stiegler MP, Gaba DM: Decision-making and cognitive strategies. Simul Healthc 2015; 10:133–8
42. Dutton RP, Lee LA, Stephens LS, Posner KL, Davies JM, Domino KB: Massive hemorrhage: A report from the anesthesia closed claims project. Anesthesiology 2014; 121:450–8
43. Mhyre JM, Ramachandran SK, Kheterpal S, Morris M, Chan PS; American Heart Association National Registry for Cardiopulmonary Resuscitation Investigators: Delayed time to defibrillation after intraoperative and periprocedural cardiac arrest. Anesthesiology 2010; 113:782–93
44. Nanji KC, Patel A, Shaikh S, Seger DL, Bates DW: Evaluation of perioperative medication errors and adverse drug events. Anesthesiology 2016; 124:25–34
45. Novick RJ, Lingard L, Cristancho SM: The call, the save, and the threat: Understanding expert help-seeking behavior during nonroutine operative scenarios. J Surg Ed 2015; 72: 302–9
46. Scherrer V, Compere V, Loisel C, Dureuil B: Cardiac arrest from local anesthetic toxicity after a field block and transversus abdominis plane block: A consequence of miscommunication between the anesthesiologist and surgeon. A A Case Rep 2013; 1:75–6
47. Arriaga AF, Gawande AA, Raemer DB, Jones DB, Smink DS, Weinstock P, Dwyer K, Lipsitz SR, Peyre S, Pawlowski JB, Muret-Wagstaff S, Gee D, Gordon JA, Cooper JB, Berry WR; Harvard Surgical Safety Collaborative: Pilot testing of a model for insurer-driven, large-scale multicenter simulation training for operating room teams. Ann Surg 2014; 259:403–10
48. Walsh M, Devereaux PJ, Garg AX, Kurz A, Turan A, Rodseth RN, Cywinski J, Thabane L, Sessler DI: Relationship between intraoperative mean arterial pressure and clinical outcomes after noncardiac surgery: Toward an empirical definition of hypotension. Anesthesiology 2013; 119:507–15
49. Ramachandran SK, Mhyre J, Kheterpal S, Christensen R, Tallman K, Morris M, Chan PS: Predictors of survival from perioperative cardiopulmonary arrests: A retrospective analysis of 2,524 events from the National Registry of Cardiopulmonary Resuscitation. Anesthesiology 2013; 119: 1322–39
50. Monk TG, Bronsert MR, Henderson WG, Mangione MP, Sum-Ping ST, Bentt DR, Nguyen JD, Richman JS, Meguid RA, Hammermeister KE: Association between intraoperative hypotension and hypertension and 30-day postoperative mortality in noncardiac surgery. Anesthesiology 2015; 123:307–19
51. Steadman RH, Burden AR, Huang YM, Gaba DM, Cooper JB: Practice improvements based on participation in simulation for the maintenance of certification in anesthesiology program. Anesthesiology 2015; 122:1154–69
52. Blum RH, Boulet JR, Cooper JB, Muret-Wagstaff SL; Harvard Assessment of Anesthesia Resident Performance Research Group: Simulation-based assessment to identify critical gaps in safe anesthesia resident performance. Anesthesiology 2014; 120:129–41
53. Blendon RJ, DesRoches CM, Brodie M, Benson JM, Rosen AB, Schneider E, Altman DE, Zapert K, Herrmann MJ, Steffenson AE: Views of practicing physicians and the public on medical errors. N Engl J Med 2002; 347:1933–40
54. Saber Tehrani AS, Lee H, Mathews SC, Shore A, Makary MA, Pronovost PJ, Newman-Toker DE: 25-Year summary of US malpractice claims for diagnostic errors 1986-2010: An analysis from the National Practitioner Data Bank. BMJ Qual Saf 2013; 22:672–80
55. Ranum D, Ma H, Shapiro FE, Chang B, Urman RD: Analysis of patient injury based on anesthesiology closed claims data from a major malpractice insurer. J Healthc Risk Manag 2014; 34:31–42
56. Johnston MJ, Arora S, King D, Bouras G, Almoudaris AM, Davis R, Darzi A: A systematic review to identify the factors that affect failure to rescue and escalation of care in surgery. Surgery 2015; 157:752–63
57. Duffy FD, Holmboe ES: Self-assessment in lifelong learning and improving performance in practice: Physician know thyself. JAMA 2006; 296:1137–9
58. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L: Accuracy of physician self-assessment compared with observed measures of competence: A systematic review. JAMA 2006; 296:1094–102
59. Gaba DM, Fish KJ, Howard SK, Burden A: Crisis Management in Anesthesiology, 2nd edition. Philadelphia, Elsevier Saunders, 2014
60. Weinger MB, Slagle JM, Kuntz AH, Schildcrout JS, Banerjee A, Mercaldo ND, Bills JL, Wallston KA, Speroff T, Patterson ES, France DJ: A multimodal intervention improves postanesthesia care unit handovers. Anesth Analg 2015; 121:957–71
61. Bruppacher HR, Alam SK, LeBlanc VR, Latter D, Naik VN, Savoldelli GL, Mazer CD, Kurrek MM, Joo HS: Simulation-based training improves physicians' performance in patient care in high-stakes clinical setting of cardiac surgery. Anesthesiology 2010; 112:985–92
62. Bion JF, Abrusci T, Hibbert P: Human factors in the management of the critically ill patient. Br J Anaesth 2010; 105:26–33
63. Greenfield S, Kaplan SH, Kahn R, Ninomiya J, Griffith JL: Profiling care provided by different groups of physicians: Effects of patient case-mix (bias) and physician-level clustering on quality assessment results. Ann Intern Med 2002; 136:111–21
64. Cook TM, Woodall N, Harper J, Benger J; Fourth National Audit Project: Major complications of airway management in the UK: Results of the Fourth National Audit Project of the Royal College of Anaesthetists and the Difficult Airway Society–Part 2: Intensive care and emergency departments. Br J Anaesth 2011; 106:632–42
65. Mitchell I, Schuster A, Smith K, Pronovost P, Wu A: Patient safety incident reporting: A qualitative study of thoughts and perceptions of experts 15 years after 'to err is human'. BMJ Qual Saf 2016; 25:92–9
66. de Feijter JM, de Grave WS, Muijtjens AM, Scherpbier AJ, Koopmans RP: A comprehensive overview of medical error in hospitals using incident-reporting systems, patient complaints and chart review of inpatient deaths. PLoS One 2012; 7:e31125
67. Neily J, Mills PD, Young-Xu Y, Carney BT, West P, Berger DH, Mazzia LM, Paull DE, Bagian JP: Association between implementation of a medical team training program and surgical mortality. JAMA 2010; 304:1693–700
68. Salas E, Rosen MA: Building high reliability teams: Progress and some reflections on teamwork training. BMJ Qual Saf 2013; 22:369–73
69. Rosen MA, Salas E, Wilson KA, King HB, Salisbury M, Augenstein JS, Robinson DW, Birnbach DJ: Measuring team performance in simulation-based training: Adopting best practices for healthcare. Simul Healthc 2008; 3:33–41