Simulation-based Assessment to Identify Critical Gaps in Safe Anesthesia Resident Performance

Richard H. Blum, M.D., John R. Boulet, Ph.D., Jeffrey B. Cooper, Ph.D., Sharon L. Muret-Wagstaff, Ph.D.; for the Harvard Assessment of Anesthesia Resident Performance Research Group*

ABSTRACT

Background: Valid methods are needed to identify anesthesia resident performance gaps early in training. However, many assessment tools in medicine have not been properly validated. The authors designed and tested use of a behaviorally anchored scale, as part of a multiscenario simulation-based assessment system, to identify high- and low-performing residents with regard to domains of greatest concern to expert anesthesiology faculty.

Methods: An expert faculty panel derived five key behavioral domains of interest by using a Delphi process (1) Synthesizes information to formulate a clear anesthetic plan; (2) Implements a plan based on changing conditions; (3) Demonstrates effective interpersonal and communication skills with patients and staff; (4) Identifies ways to improve performance; and (5) Recognizes own limits. Seven simulation scenarios spanning pre-to-postoperative encounters were used to assess performances of 22 first-year residents and 8 fellows from two institutions. Two of 10 trained faculty raters blinded to trainee program and training level scored each performance independently by using a behaviorally anchored rating scale. Residents, fellows, facilitators, and raters completed surveys.

Results: Evidence supporting the reliability and validity of the assessment scores was procured, including a high generalizability coefficient ($\rho^2 = 0.81$) and expected performance differences between first-year resident and fellow participants. A majority of trainees, facilitators, and raters judged the assessment to be useful, realistic, and representative of critical skills required for safe practice.

Conclusion: The study provides initial evidence to support the validity of a simulation-based performance assessment system for identifying critical gaps in safe anesthesia resident performance early in training. **(ANESTHESIOLOGY 2014; 120:129-41)**

IDENTIFYING gaps in the competency of anesthesia residents in time for intervention is critical to patient safety and an effective learning system, yet comprehensive, reliable assessment approaches remain elusive. Currently available instruments are useful for testing specific procedural and teamwork skills.[1–7] However, the majority of direct observation assessment tools in medicine have not been subjected to proper validation studies,[8] and fewer than one third of 609 simulation studies in a recent meta-analysis offered any evidence to support validity of the performance measures.[9] In addition, few available instruments relate to complex behavioral performance or provide descriptors (rather than rank-ordered ratings) that could inform subsequent feedback, individualized teaching, remediation, and curriculum revision. These unmet needs may hinder Accreditation Council for Graduate Medical Education plans to "accelerate the Accreditation Council for Graduate Medical Education's movement toward accreditation on the basis of educational outcomes."[10–12]

What We Already Know about This Topic

- Although early identification of gaps in anesthesia resident competency is important to patient safety, currently available tools have not been subjected to proper validation

What This Article Tells Us That Is New

- A multiscenario, simulation-based performance assessment of 22 first-year residents and 8 fellows from two institutions showed high reliability, validity, and generalizability
- A majority of trainees and faculty judged the assessment to be useful, realistic, and representative of critical skills required for safe practice

This article is featured in "This Month in Anesthesiology," page 1A. Corresponding article on page 18. Presented in part at the International Meeting on Simulation in Healthcare, Phoenix, Arizona, January 25, 2010; International Meeting on Simulation in Healthcare, New Orleans, Louisianna, January 25, 2011; Post Graduate Assembly of the New York State Society of Anesthesiologists, New York, New York, December 11, 2011; International Meeting on Simulation in Healthcare, San Diego, California, January 31, 2012; International Assembly for Pediatric Anesthesia, Washington, D.C. October 11, 2012.

Submitted for publication April 18, 2013. Accepted for publication September 18, 2013. From the Department of Anesthesiology, Perioperative and Pain Medicine, Boston Children's Hospital, and Harvard Medical School, Boston, Massachusetts (R.H.B.); Foundation for Advancement of International Medical Education and Research, Philadelphia, Pennsylvania (J.R.B.); Center for Medical Simulation, Charlestown, Massachusetts, and Harvard Medical School and Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, Massachusetts (J.B.C.); Faculty Development and Innovation, Department of Anesthesia, Critical Care and Pain Medicine, Beth Israel Deaconess Medical Center, and Harvard Medical School, Boston, Massachusetts (S.L.M.-W.).

* Members of the Harvard Assessment of Anesthesia Resident Performance Research Group are listed in appendix 1.

 Anesthesiology 2014; 120:129-41

A behaviorally anchored simulation assessment tool designed to measure critical performance characteristics of first-year anesthesia residents in realistic patient care situations can, if properly validated, inform tailored learning and, ultimately, improve patient safety by ensuring that no resident graduates without demonstrating essential clinical skills.

In the current study, our aim was to design and test the use of a behaviorally anchored scale, as part of a multiscenario, simulation-based assessment system, to elucidate characteristics of high- and low-performing first-year Clinical Anesthesia year one [CA-1] anesthesia residents with regard to domains of greatest concern to expert anesthesiology faculty. We hypothesized that, based on various psychometric criteria, the assessment would yield scores that would be sufficiently precise and accurate for the identification of strengths and weaknesses in a resident's ability to provide safe and effective patient care.

Materials and Methods

Study Design

We designed tools and conducted a prospective observational study to systematically gather evidence for the validity of scores from a simulation-based assessment intended to identify critical gaps in first-year anesthesia resident performance. We addressed the first three of five validation arguments:[13–15]

1. *Construct representation:* Do the tasks (scenarios) elicit performances that reflect the intended constructs (provision of safe anesthesia care)?
2. *Scoring:* Are the scores (ratings) dependable (reliable and meaningful) measures of the intended constructs?
3. *Generalization:* Do the tasks adequately sample the constructs that are set out as important (*i.e.*, does the number and selection of tasks/scenarios provide for an adequate sampling of the measured constructs [safe care])?
4. *Extrapolation:* Are the constructs sampled representative of competence in the wider subject domain (*i.e.*, can simulation performance be linked to performance with real patients)?
5. *Decision-making:* Is guidance in place so that stakeholders know what scores mean and how the outcomes should be used?[15]

Assessments were confidential; no results were provided to training programs. The study was approved by the Institutional Review Boards of Boston Children's Hospital, Beth Israel Deaconess Medical Center, and Partners Healthcare, Boston, Massachusetts.

Participants

Twenty-two first-year (CA-1) residents (7 females and 15 males) from one Accreditation Council for Graduate Medical Education–accredited anesthesia residency program and eight pediatric fellows (F) (three females and five males) from an anesthesia fellowship program at a second institution (comparison group) were invited to participate based on clinical rotation schedules; all provided written informed consent. Given the exploratory nature of this pilot investigation, sample size was limited to the number of participants who could be accommodated feasibly in the study period (March to June 2009). Thirteen board-certified, practicing anesthesiologists were recruited and trained as either raters (five females and five males) or scenario facilitators (two females and one male). To protect confidentiality and minimize bias, these faculty members came from two academic medical centers different from the two institutions of the residents and fellows. Simulation sessions were conducted at the Center for Medical Simulation, then located in Cambridge, Massachusetts.

Scenario Design, Instrument Development, and Scoring Rubric

We addressed *construct representation* through the expert-based iterative design of the scenarios, assessment instruments, and scoring rubric. Data from participant surveys and analyses of relationships among dimension scores helped support the construct validity argument.

A panel of seven board-certified anesthesiologists with 5 yr or more of clinical experience and significant participation in resident education (*e.g.*, program director, clinical competency committee) were asked to describe critical skill deficits observed in anesthesia resident performance. Individual panel members wrote responses to the question, "What traits characterize residents who, upon graduation, have not achieved a minimum level of competency?" Through two rounds of a modified Delphi process, 27 original responses were reduced to five key behaviors that are lacking in underperforming senior residents: (1) Synthesizes information to formulate a clear anesthetic plan; (2) Implements a plan based on changing conditions; (3) Demonstrates effective interpersonal and communication skills with patients and staff; (4) Identifies ways to improve performance; and (5) Recognizes own limits.

Seven scenarios appropriate to the CA-1 training level were designed based on these five behavioral domains. We incorporated related material by using American Board of Anesthesiology examination content outlines[16] and Accreditation Council for Graduate Medical Education core competencies[17] in design of the scenarios. Scenarios featured anesthesia care in the following situations: (1) preoperative assessment of a patient scheduled for urgent exploratory laparotomy; (2) operative management of a patient with perforated ulcer and hemorrhage; (3) monitored anesthesia care for a patient with discomfort during basal cell carcinoma surgery; (4) postanesthesia care for a patient with aspiration after basal cell carcinoma surgery; (5) management of anaphylaxis in a patient with transurethral resection of the prostate and bladder biopsy; (6) care for a patient with

delayed awakening in the operating room after transurethral resection of the prostate; and (7) identification and management of mainstem intubation secondary to coughing in a patient undergoing total thyroidectomy. The seven scenarios spanned pre-to-postoperative care. For six of the seven scenarios, the same patient was managed in two related scenarios (*e.g.*, holding area, operating room) to engage the trainee for sustained periods and reduce unnecessary logistical complexity. Scenario design included specific scripts, cues, timing, and events to elicit complex behaviors. For example, in scenario 3 (appendix 2), the participant might rapidly gather and synthesize information from multiple sources and articulate a basic plan that takes into account the patient's request to be awake and her history of postoperative nausea and vomiting (domain 1). As the patient becomes anxious with conscious sedation, the participant is expected to demonstrate an alternative plan under changing conditions, considering such factors as maintaining the surgical field if an airway problem arises (domain 2). Communication with the patient and team members (confederates) is needed to manage the situation effectively (domain 3). The participant may identify ways to improve performance (domain 4) in response to the dynamic situation in the scenario (*e.g.*, by reflecting and acknowledging to team mates that a particular action is not working and stating an improved plan rather than denying or perseverating) or in postscenario questions or both. The participant's recognition of his or her own limits (domain 5) can be assessed, for example, by actions such as calling the pharmacy to get needed information about an unfamiliar drug or by incorrectly proceeding with general anesthesia without an attending physician present. Facilitators and staff rehearsed in pilot sessions with nonparticipating residents until scenarios were refined to maximize realism, performances clearly could be scored in all five domains, and scenarios could be delivered in a repeatable way as planned. By standardizing scenarios with respect to confederates, mannequins, equipment, and scenario flow, we aimed to maximize *scoring* accuracy and fairness of the assessment. We used a SimMan® 3G (Laerdal, Wappingers Falls, NY) mannequin. Scenarios were video recorded for viewing by raters using SimCapture® web-based software (B-Line Medical, LLC, Washington, DC) with three camera views, including one of the physiological monitors.

Each trainee's session lasted approximately 3 h and began with an introduction, obtaining informed consent, 15-min structured orientation, and a 3-min hands-on introduction to the simulation environment and equipment including the opportunity to ask questions, all led by the facilitator. A majority of residents and fellows had been exposed to one previous high-fidelity simulation session as a first-year requirement in both programs. Seven 15-min scenarios were then conducted, with a brief, scripted introduction to each case. After each scenario, the facilitator asked the trainee three questions in an adjacent room: (1) "I noticed that… [clinically significant occurrence, *e.g.*, "the patient had oxygen desaturation during the procedure"]; I'm wondering what your differential diagnosis and your management plan were;" (2) "There was a lot going on in this case. Could you tell me about any times that you felt challenged either in being able to think things through or to get things done?" (3) "If you were presented with this case again, is there anything that you would do differently?" Postscenario questions and responses were video recorded to enable raters to understand the behaviors from the trainee's perspective. At the end of the session, the facilitator held a 15-min educational debriefing with the trainee, and both completed a survey.

A behaviorally anchored rating scale (appendix 3) was developed to contribute to *scoring dependability* through its clarity in support of rater scoring decisions; to provide ratings that would be helpful for diagnostic feedback; and to enable faculty to tailor educational interventions. A 7-point scale was chosen to maximize reliability, validity, and discriminating power.[18] Consistent with validity-establishing practices for educational tests and scale development,[19,20] we reviewed literature and assessment tools in medicine, psychology, education, and business to gain conceptual coherence regarding the five constructs identified by our expert panel as worrisome gaps. We then interviewed anesthesiologists who were asked to describe an actual resident performance in each domain that was outstanding, inadequate, and typical. In addition, we observed residents from anesthesia and other subspecialties. Many interviewee verbatim phrases were retained. Descriptors were again refined based on feedback from the 10 raters before completion of their training.

Rater Training and Scoring

Raters participated in a 3-h group training and calibration session. Raters were blinded to participant training level (CA-1, F) and institutional affiliation. Each scenario, including postscenario questions, was scored independently by each of two raters who viewed recordings *via* a secure Internet server. No two raters were paired for more than one scenario. Each rater typically scored two scenarios on the five domains. Finally, raters completed a survey evaluating the simulation-based assessment.

For each of the seven scenarios, domain scores were averaged overall, and then over raters, to produce a total scenario score. This average total score, which weights each domain equally, was used as a measure of overall ability in managing each of the simulated patient's conditions. The total score for each candidate was calculated by averaging scenario scores. Survey response frequencies were calculated by group (trainee, facilitator, and rater). Scenarios that were not recorded due to logistical problems and "not applicable" ratings were treated as missing data and not included in the calculations.

Statistical Analysis

Descriptive statistics (means, SDs) were calculated for individual scenarios, performance domains, by candidate, and

by provider type (CA-1, F). To assess the magnitude of associations between individual domain and scenario scores, Pearson correlations were computed. The ability of individual scenarios to discriminate between low- and high-ability candidates was assessed by calculating discrimination indices (correlation between individual scenario scores and overall score—D statistic). Survey data were summarized by frequency counts and means.

To gather further evidence to support the *construct representation* of the assessment scores, performance of fellows was compared with that of first-year residents. A repeated measures ANOVA was conducted to test the hypothesis that there was no difference in scores based on experience. The independent variables were trainee level (CA-1, F) and scenario (repeated measure). The dependent variable was the summary scenario score.

To estimate *scoring reliability and generalizability*, a Generalizability (G) study of the assessment scores was conducted.[21,22] Variance components were estimated based on a Person (P) by Rater (R) nested in Task (T) design. That is, each of the trainees (Person, n = 30) was rated by two independent raters who were assigned to specific scenarios (Task, n = 7). All statistical tests were performed by using the software SAS version 9.1 (SAS Institute, Cary, NC).

Results

Descriptive Statistics and Correlations

Descriptive statistics (total score, average of domain scores), by scenario, are presented in table 1. Minimum and maximum scores show that performances varied across the range of the 7-point behavioral scales. On the basis of the average performance for all participants, scenario 5 (transurethral resection of the prostate, anaphylaxis) (mean = 4.2, SD = 1.2) was the most difficult and scenario 6 (transurethral resection of the prostate, delayed awakening in the operating room) was the easiest (mean = 5.4, SD = 1.4).

Intercorrelations among individual scenario scores (total) are provided in table 2. Although some associations were statistically significant (*e.g.*, scenario 2 to 5; $r = 0.66$; $P < 0.01$), the magnitude of most associations was only moderate (*i.e.*, <20% variance shared between scenario total scores). All scenario scores were highly correlated with the total score (D ≥ 0.55) indicating that the scenarios are able to discriminate between low- and high-ability practitioners.

Scoring Reliability

The estimated variance components for the G study are presented in table 3. The Person (resident/fellow) variance component is an estimate of variance across trainees in trainee-level mean scores. Ideally most of the variance should be here, indicating that individual abilities account for differences in observed scores. In this nested design (raters were only allowed to rate specific cases), the other "main effect" variance component is Task (scenario). The Task component is the estimated variance of scenario mean scores. Because the estimate is greater than zero, we know the seven tasks vary somewhat in average difficulty (table 1). The Rater nested in Task (Rater:Task) variance component is an estimate of variance of rater scores for each task. The relatively small magnitude of this component suggests that, for a given scenario, raters have similar mean scores. The large interaction variance component, Person × Task, suggests that there are considerably different rank orderings of trainee mean scores for each of the simulation scenarios. The final variance component, Error, is the residual variance that includes the P×R:T interaction and all other unexplained sources of variation.

On the basis of the G study (P×R:T), the variance components were used to estimate the reliability of trainee scores for various measurement designs. This process, known as a Decision (D) study, allows one to design the most efficient measurement procedures for future operations. For the simulation assessment, we want to generalize the trainees' scores to the universe that includes many other tasks (scenarios) and many other raters (trained to score specific scenarios). For a situation in which each trainee is rated in each of the seven scenarios (tasks; $n_t = 7$) and each of the tasks is rated by two independent raters ($n_r = 2$), sample sizes for the D study are the same as those for the G study. On the basis of the estimated variance components for $n_t = 7$ and $n_r = 2$, the Generalizability coefficient (ρ^2) is 0.81. The Dependability

Table 1. Mean Scores for All Participants on Seven Scenarios

Scenario	N	Mean	SD	Minimum	Maximum
1. Preoperative assessment/laparotomy	27	4.56	0.96	2.80	6.50
2. Laparotomy/hemorrhage	27	4.60	1.21	1.60	6.10
3. Basal cell carcinoma/discomfort	27	4.65	1.25	2.00	6.60
4. Basal cell carcinoma PACU/aspiration	27	4.58	1.26	2.10	6.50
5. TURP/anaphylaxis	27	4.19	1.22	2.20	6.80
6. TURP/delayed awakening in OR	26	5.38	1.41	2.30	6.90
7. Thyroidectomy/ETT displacement	30	4.58	1.18	2.10	6.70
Total	30	4.65	0.85	2.91	6.14

Mean performance by scenario for all participants. Individual scenario scores are based on the average of the five domain scores, averaged over two raters. Sample size not equal to 30 for each scenario because of missing data.

ETT = endotracheal tube; N = number; OR = operating room; PACU = postanesthesia care unit; TURP = transurethral resection of the prostate.

Table 2. Intercorrelations among Scenario Scores

Pearson Correlation Coefficients Probability > |r| under H0: Rho = 0 Number of Observations	Scen1	Scen2	Scen3	Scen4	Scen5	Scen6	Scen7	Total
1. Preoperative assessment/laparotomy	1.00 27							
2. Laparotomy/hemorrhage	0.32 0.10 27	1.00 27						
3. Basal cell carcinoma/ discomfort	0.28 0.18 24	0.57 0.003 24	1.00 27					
4. Basal cell carcinoma PACU/aspiration	0.45 0.028 24	0.20 0.34 24	0.39 0.04 27	1.00 27				
5. TURP/anaphylaxis	0.39 0.059 24	0.66 0.0005 24	0.46 0.03 24	0.51 0.01 24	1.00 27			
6. TURP/delayed awakening in OR	0.12 0.60 23	0.43 0.04 23	0.41 0.05 24	0.24 0.26 24	0.32 0.12 26	1.00 26		
7. Thyroidectomy/ETT displacement	0.28 0.16 27	0.49 0.01 27	0.43 0.03 27	0.17 0.41 27	0.50 0.008 27	0.34 0.09 26	1.00 30	
Total	0.58 0.001 27	0.78 <0.0001 27	0.74 <0.0001 27	0.63 0.0004 27	0.80 <0.001 27	0.65 0.003 26	0.67 <0.001 30	1.00 30

Pearson correlations between scenario scores. Total score calculated as the mean of scenario scores. Coefficients based on all available data (N ≤ 30).
ETT = endotracheal tube; OR = operating room; PACU = postanesthesia care unit; scen = scenario; TURP = transurethral resection of the prostate.

(D) coefficient, which takes into account rater stringency and scenario difficulty as potential sources of error in estimation of trainee ability, was $\varphi = 0.79$. The Generalizability and Dependability coefficients for a design that incorporates $n_t = 7$ tasks (scenarios) and only a single ($n_r = 1$) rater are estimated to be $\rho^2 = 0.75$ and $\varphi = 0.79$, respectively. The final column in table 3 provides the estimated variance components for a hypothetical design involving 14 tasks (scenarios) and a single rater. On the basis of these variance component estimates, $\rho^2 = 0.89$ and $\varphi = 0.88$.

Although reliability of the trainee scores is most dependent on the number of scenarios, it is still important to quantify the association between scores provided by the two independent raters (*i.e.*, interrater reliability). Correlations between scores of rater pairs, by scenario, were moderately high, ranging from $r = 0.48$ to $r = 0.79$.

Generalizability

Overall domain scores were calculated as the mean of the seven scenario scores, averaged over the two raters (table 4). On the basis of the magnitude of correlations among domain scores, the five performance domains are moderately related. Overall, domain 4 (identifies ways to improve performance) was the least related to the other domains (<35% of the variance in domain 4 ratings could be explained by any of the other domain scores). All summary domain scores were highly correlated with the total score. This indicates that the domains are related and that participants with problems in one area were likely to have problems in others.

On the basis of the repeated measures ANOVA, there is a significant group (CA-1, F) × scenario interaction ($F_{1,6} = 2.91$; $P < 0.05$). This indicates that average performance, by group level, was not the same for some scenarios. Across all scenarios, with the exception of scenarios 1 and 4, fellows, on average, outperformed first-year residents. There was also a significant main effect attributable to scenario ($F_{1,6} = 4.6$; $P < 0.01$), indicating that, averaged over trainee experience levels, the scenarios were not of equivalent difficulty. Based on a *post hoc* analysis of individuals' scenario scores, there was a statistically significant difference between CA-1 and fellow performance on scenario 6 ($F_{1,25} = 5.2$; $P < 0.05$; table 1).

Survey Results

Twenty-nine of 30 trainees (97%), 3 facilitators for 29 of 30 sessions (97%), and 10 of 10 raters (100%) completed surveys. Trainees reported that simulation scenarios demanded skills that practitioners at their level would be expected to have attained (100% agree/strongly agree); simulation experiences were sufficiently realistic to allow them to act as if they were in actual patient care situations (27 of 29, 93%); the experience was useful for resident

Table 3. Estimated Variance Components

Variance Component	Estimate	% of Total Variance	D Studies		
			$n_t = 7$, $n_r = 2$	$n_t = 7$, $n_r = 1$	$n_t = 14$, $n_r = 2$
Person—resident	0.54	27.9	0.54	0.54	0.54
Task—scenario	0.07	3.5	0.01	0.01	0.005
Rater:Task	0.06	3.0	0.004	0.008	0.002
Person × Task	0.54	27.9	0.08	0.08	0.039
Error (P×R:T)	0.73	37.7	0.52	0.11	0.026
Total	1.95				
Generalizability (G) Coefficient (ρ^2)			0.81	0.75	0.89
Dependability (D) Coefficient (φ)			0.79	0.73	0.88

Variance components for a Person (P) by Rater (R) nested in Task (T) design.
n_t = number of tasks (scenarios).

training and a valuable use of educational time (28 of 29, 96%); and they received sufficient and useful feedback (28 of 29, 96%). Facilitators reported that they administered scenarios as intended in 24 of 29 sessions (83%); logistical problems or equipment failures accounted for disruptions in an individual scenario within several sessions. They reported that scenarios were realistic and accurately represented resident responses seen in actual clinical cases (22 of 29, 76%). Facilitators found the system useful for assessing the participant's performance in most sessions (27 of 29, 93%). Similarly, all raters agreed or strongly agreed that resident/fellow performances were realistic and were representative of performance of residents the raters had observed in actual clinical situations. All reported that the scoring system domains and descriptors represent behaviors that are critical to patient safety as well as to advancement, successful completion of residency training, and safe independent clinical practice. All found the system to be a unique and useful in addition to currently available assessment tools, and one that makes it possible to tailor feedback and educational interventions.

Discussion

This pilot study provides initial evidence to support the validity of a simulation-based assessment system for identifying critical gaps in safe anesthesia resident performance early in training. In addition, based on the Generalizability study, reasonably precise measures of overall ability can be procured with seven simulation encounters ($\rho^2 = 0.81$). The study addresses the current "paucity of evidence to guide best practices of remediation in medical education at all levels."[11] In the high-risk field of anesthesiology, rigorous validation studies to support appropriate interpretation and use of training assessments are urgently needed to accelerate learning and patient safety efforts.

In this study, we addressed *construct representation* first through an iterative design process synthesizing expert opinions and literature review, assuring that the scenarios and scoring methods captured behaviors that reflect essential aspects of the five targeted constructs. This is consistent with previous work[23,24] and corroborated by participant, facilitator, and rater survey results indicating that the scenarios and rubrics are realistic, representative, and critical to safe independent practice. The scenarios elicited performances that were scored over the full range of behavioral descriptors. The assessment system was not intended to provide a single, all-inclusive assessment of attainment of overall training goals, teamwork, procedural skills, or areas covered by other assessment modalities such as the MiniCEX,[25] Anaesthetists' Non-Technical Skills,[26] and Mayo High Performance Teamwork Scale.[7] Rather, our focus was on critical gaps in safe resident performance determined by experienced anesthesiologists at local university-affiliated institutions. Others might choose to use similar methods to target other clinical tasks or domains.[27]

This approach with the use of a behaviorally anchored rating scale complements task-based rubrics (*e.g.*, checklists) and reflects recent competency-based medical education recommendations to improve measures by "rethinking the structure of the tools we are using, to ensure that the instruments authentically represent the way in which faculty functionally conceptualize their residents' clinical competence on a day-to-day basis"[28] in an integrative rather than reductionist way.[29]

Raters *scored* the performances after viewing videos of the scenarios and postscenario questions. This strategy was used to allow raters to make better judgments concerning performance in some of the domains (*e.g.*, identifies ways to improve performance) and to enable constructive formative feedback. Further studies will be needed comparing scores with and without postscenario questions to determine the unique, independent contributions and possible interaction effects in scoring of each of the interrelated performance domains. It should be noted that the total scenario score, used for many of the analyses, was based on an average of the domain scores. Although this is practical, it ignores the fact

Table 4. Domain-level Correlations (Person-level)

Pearson Correlation Coefficients, N = 30 Probability > |r| under H0: Rho = 0	m_dom1	m_dom2	m_dom3	m_dom4	m_dom5	Total
Synthesizes information to form a clear anesthetic plan	1.00					
Implements a plan based on changing conditions	0.94 <0.0001	1.00				
Demonstrates effective interpersonal and communication skills	0.92 <0.0001	0.92 <0.0001	1.00			
Identifies ways to improve performance	0.78 <0.0001	0.80 <0.0001	0.83 <0.0001	1.00		
Recognizes own limits	0.81 <0.0001	0.85 <0.0001	0.78 <0.0001	0.70 <0.0001	1.00	
Total	0.96 <0.0001	0.97 <0.0001	0.96 <0.0001	0.88 <0.0001	0.89 <0.0001	1.00

Pearson correlations between domain scores. Domain (dom) scores based on average score over scenarios.

that the domains may be related hierarchically and, as a measure of overall ability, it allows residents to compensate for poor performance in one domain with better performance in another. Future studies, incorporating larger samples, are needed to determine the specific structure of domain scores, whether they should be differentially weighted, and how this might vary by scenario.

Results of *scoring generalizability* and decision analyses demonstrate that critical anesthesia resident skills can be measured reliably using the assessment system. The generalizability coefficient of 0.81 meets or exceeds that reported in comparable simulation-based assessment studies and high-stakes certification examinations (range, 0.56 to 0.80).[5,30–35] Similar to previous investigations,[5,30,35] our results indicate that additional measurement precision could best be achieved by increasing the number of scenarios, and not the number of raters per scenario. The system design, based on multiple inputs regarding behaviors that are critical to safe performance, combined with the iterative development of detailed and anchored scoring rubrics, attention to scenario standardization, and methodical rater training, was effective in minimizing measurement errors.

Scenario-level differences between CA-1 and fellow performances provide some, although moderate, evidence to support validity of the scores. On five of seven scenarios, fellows (on average) outperformed first-year residents. Although only one of these scenario-level comparisons was statistically significant, our findings were based on a relatively small group of participants.

More importantly, we observed wide and overlapping variability of scores by scenario of all participants, regardless of training level. As such, the scores may reflect accurate measurement of skills that are not effectively taught and learned currently in residency programs. Previous simulation-based anesthesia studies also have documented widespread variability in performance both within and across practitioners with different levels of experience.[5,30,33,36,37] Others have demonstrated that learning curves differ among individuals, and that experience alone does not reliably lead to expertise without accompanying motivation, excellent teaching and feedback associated with practice, and focused environmental resources.[38–41] Although individual performance may be influenced by many factors, including motivation and realism of the simulated environment, large differences in performance within both resident and fellow groups, especially among individuals in the same or similar training programs, suggest the need for improved educational programs.[42]

An intensive use of resources was required to carry out resident assessments. Further study is needed to streamline the process without compromising validity. For example, our analyses show that the assessment system is sufficiently robust that the number of raters could reasonably be cut in half with little impact on scoring reliability. Given an estimated annual cost of $130,000 to train one resident,[43] an investment in ensuring safety and salvaging underperforming residents early through evidence-based remediation may be warranted and consonant with anesthesia's historic leadership in patient safety.

This study has two important limitations. First, performance data were obtained from a relatively small group of trainees at two institutions. Therefore, our findings may not be generalizable to other programs. Multiinstitutional studies, with larger participant cohorts, are certainly needed. Second, in addition to the validation evidence for *construct representation, scoring*, and *generalization* presented here, two further areas of validation remain. As in any simulation-based assessment, continuing studies are needed to address *extrapolation* of simulation-based scores to performance in "real-world" clinical settings,[44] and to establish scoring guidelines to support *decision-making* about interpretation and use of scores.[45]

In summary, our study provides evidence to support the validity of scores gathered *via* a simulation-based

anesthesia resident performance assessment system. Adequate assessment instruments, ones that yield valid and reliable scores, are necessary for identifying skill deficiencies, providing meaningful individual feedback, establishing remediation programs, and ultimately, ensuring fully competent independent practitioners.[9,46,47] Development of psychometrically defensible instruments is recognized as a high-priority need in the field.[48–50] The ability to identify and remediate poorly performing residents early in training is a necessary step in improving the quality and safety of patient care.

Acknowledgments

The authors thank the residents and fellows who participated; the faculty and department chairs who enabled the study; and Anthony Dancel, C.P.T., and Karen Nadelberg, R.N., Center for Medical Simulation, Charlestown, Massachusetts, for technical expertise and assistance.

Supported by the Anaesthesia Chairs' Education Fund, Beth Israel Deaconess Medical Center, Brigham and Women's Hospital, Boston Children's Hospital, Massachusetts General Hospital, Boston, Massachusetts; Anesthesia Patient Safety Foundation, American Society of Anesthesiologists Endowed Research Award; Ellison C. Pierce, Jr., M.D., Research Award; The Cathedral Fund, Newton Centre, Massachusetts; and the authors' respective departments.

Competing Interests

The authors declare no competing interests.

Correspondence

Address correspondence to Dr. Blum: Department of Anesthesiology, Perioperative and Pain Medicine, Boston Children's Hospital, 300 Longwood Avenue, Boston, Massachusetts 02115. richard.blum@childrens.harvard.edu. This article may be accessed for personal use at no charge through the Journal Web site, www.anesthesiology.org.

References

1. Mudumbai SC, Gaba DM, Boulet JR, Howard SK, Davies MF: External validation of simulation-based assessments with other performance measures of third-year anesthesiology residents. Simul Healthc 2012; 7:73–80
2. Berkenstadt H, Ben-Menachem E, Dach R, Ezri T, Ziv A, Rubin O, Keidan I: Deficits in the provision of cardiopulmonary resuscitation during simulated obstetric crises: Results from the Israeli Board of Anesthesiologists. Anesth Analg 2012; 115:1122–6
3. Ben-Menachem E, Ezri T, Ziv A, Sidi A, Brill S, Berkenstadt H: Objective Structured Clinical Examination-based assessment of regional anesthesia skills: The Israeli National Board Examination in Anesthesiology experience. Anesth Analg 2011; 112:242–5
4. Flin R, Patey R, Glavin R, Maran N: Anaesthetists' non-technical skills. Br J Anaesth 2010; 105:38–44
5. Fehr JJ, Boulet JR, Waldrop WB, Snider R, Brockel M, Murray DJ: Simulation-based assessment of pediatric anesthesia skills. Anesthesiology 2011; 115:1308–15
6. Blum RH, Raemer DB, Carroll JS, Dufresne RL, Cooper JB: A method for measuring the effectiveness of simulation-based team training for improving communication skills. Anesth Analg 2005; 100:1375–80
7. Malec JF, Torsher LC, Dunn WF, Wiegmann DA, Arnold JJ, Brown DA, Phatak V: The mayo high performance teamwork scale: Reliability and validity for evaluating key crew resource management skills. Simul Healthc 2007; 2:4–10
8. Kogan JR, Holmboe ES, Hauer KE: Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. JAMA 2009; 302:1316–26
9. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, Erwin PJ, Hamstra SJ: Technology-enhanced simulation for health professions education: A systematic review and meta-analysis. JAMA 2011; 306:978–88
10. Nasca TJ, Philibert I, Brigham T, Flynn TC: The next GME accreditation system—Rationale and benefits. N Engl J Med 2012; 366:1051–6
11. Hauer KE, Ciccone A, Henzel TR, Katsufrakis P, Miller SH, Norcross WA, Papadakis MA, Irby DM: Remediation of the deficiencies of physicians across the continuum from medical school to practice: A thematic review of the literature. Acad Med 2009; 84:1822–32
12. Zbieranowski I, Takahashi SG, Verma S, Spadafora SM: Remediation of residents in difficulty: A retrospective 10-year review of the experience of a postgraduate board of examiners. Acad Med 2013; 88:111–6
13. Kane MT: Validation, Educational Measurement, 4th edition. Edited by Brennan RL. Westport, Praeger, 2006, pp 17–64
14. Boulet JR, Jeffries PR, Hatala RA, Korndorffer JR Jr, Feinstein DM, Roche JP: Research regarding methods of assessing learning outcomes. Simul Healthc 2011; 6(suppl): S48–51
15. Shaw S, Crisp V, Johnson N: A framework for evidencing assessment validity in large-scale, high-stakes international examinations. Assess Educ Princ Pol Pract 2012; 19:159–76
16. Joint Council on Anesthesiology Examinations; American Board of Anesthesiology, American Society of Anesthesiologists: Content Outline Basic/Advanced. Raleigh, American Board of Anesthesiology, 2012
17. Accreditation Council for Graduate Medical Education: ACGME Program Requirements for Graduate Medical Education in Anesthesiology, effective July 1, 2008, rev. July 1, 2011. Chicago, ACGME, 2011
18. Preston CC, Colman AM: Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. Acta Psychol (Amst) 2000; 104:1–15
19. American Educational Research Association, American Psychological Association, National Council on Measurement in Education: Standards for Educational and Psychological Testing. Washington, American Educational Research Association, 1999, pp 7–68, 137–50
20. DeVellis RF: Scale Development. Theory and Applications, 3rd edition. Los Angeles, Sage, 2012, pp 31–114, 185–92
21. Brennan RL: Generalizability Theory. New York, Springer-Verlag, 2001, pp 4–20, 53–140
22. Boulet JR: Generalizability theory: Basics, Encyclopedia of Statistics in Behavioral Science. Edited by Evritt BS, Howell DC. Chichester, Wiley, 2005, pp 704–11
23. Mercer SJ, Moneypenny MJ, Fredy O, Guha A: What should be included in a simulation course for anaesthetists? The Merseyside trainee perspective. Eur J Anaesthesiol 2012; 29:137–42
24. Glavin RJ: Excellence in anesthesiology: The role of nontechnical skills. Anesthesiology 2009; 110:201–3
25. Weller JM, Jolly B, Misur MP, Merry AF, Jones A, Crossley JG, Pedersen K, Smith K: Mini-clinical evaluation exercise in anaesthesia training. Br J Anaesth 2009; 102:633–41

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. Goldszmidt M, Minda JP, Bordage G: Developing a unified list of physicians' reasoning tasks during clinical encounters. Acad Med 2013; 88:390–7
28. Regehr G, Ginsburg S, Herold J, Hatala R, Eva K, Oulanova O: Using "standardized narratives" to explore new ways to represent faculty opinions of resident performance. Acad Med 2012; 87:419–27
29. Carraccio CL, Englander R: From Flexner to competencies: Reflections on a decade and the journey ahead. Acad Med 2013; 88:1067–73
30. 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
31. Adler MD, Vozenilek JA, Trainor JL, Eppich WJ, Wang EE, Beaumont JL, Aitchison PR, Pribaz PJ, Erickson T, Edison M, McGaghie WC: Comparison of checklist and anchored global rating instruments for performance rating of simulated pediatric emergencies. Simul Healthc 2011; 6:18–24
32. Margolis MJ, Clauser BE, Swanson DB, Boulet JR: Analysis of the relationship between score components on a standardized patient clinical skills examination. Acad Med 2003; 78(10 suppl):S68–71
33. 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
34. van Zanten M, Boulet JR, McKinley D: Using standardized patients to assess the interpersonal skills of physicians: Six years' experience with a high-stakes certification examination. Health Commun 2007; 22:195–205
35. McBride ME, Waldrop WB, Fehr JJ, Boulet JR, Murray DJ: Simulation in pediatrics: The reliability and validity of a multiscenario assessment. Pediatrics 2011; 128:335–43
36. DeAnda A, Gaba DM: Role of experience in the response to simulated critical incidents. Anesth Analg 1991; 72:308–15
37. 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
38. Friedman Z, Siddiqui N, Katznelson R, Devito I, Davies S: Experience is not enough: Repeated breaches in epidural anesthesia aseptic technique by novice operators despite improved skill. Anesthesiology 2008; 108:914–20
39. Ericsson KA, Nandagopal K, Roring RW: Toward a science of exceptional achievement: Attaining superior performance through deliberate practice. Ann N Y Acad Sci 2009; 1172:199–217
40. Fraser SA, Feldman LS, Stanbridge D, Fried GM: Characterizing the learning curve for a basic laparoscopic drill. Surg Endosc 2005; 19:1572–8
41. Choudhry NK, Fletcher RH, Soumerai SB: Systematic review: The relationship between clinical experience and quality of health care. Ann Intern Med 2005; 142:260–73
42. Weinger MB: Experience not equal expertise: Can simulation be used to tell the difference? Anesthesiology 2007; 107:691–4
43. Steinmann AF: Threats to graduate medical education funding and the need for a rational approach: A statement from the alliance for academic internal medicine. Ann Intern Med 2011; 155:461–4
44. Holmboe E, Rizzolo MA, Sachdeva AK, Rosenberg M, Ziv A: Simulation-based assessment and the regulation of healthcare professionals. Simul Healthc 2011; 6(suppl):S58–62
45. 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
46. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ: A critical review of simulation-based medical education research: 2003–2009. Med Educ 2010; 44:50–63
47. Steadman RH, Huang YM: Simulation for quality assurance in training, credentialing and maintenance of certification. Best Pract Res Clin Anaesthesiol 2012; 26:3–15
48. Dieckmann P, Phero JC, Issenberg SB, Kardong-Edgren S, Ostergaard D, Ringsted C: The first Research Consensus Summit of the Society for Simulation in Healthcare: Conduction and a synthesis of the results. Simul Healthc 2011; 6(suppl):S1–9
49. Boulet JR, Murray DJ: Simulation-based assessment in anesthesiology: Requirements for practical implementation. Anesthesiology 2010; 112:1041–52
50. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T: Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach 2011; 33:206–14

Appendix 1. Harvard Assessment of Anesthesia Resident Performance Research Group

Core Investigators

Richard H. Blum, M.D., M.S.E. (Principal Investigator), Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts

John R. Boulet, Ph.D., Foundation for Advancement of International Medical Education and Research, Philadelphia, Pennsylvania

Jeffrey B. Cooper, Ph.D., Center for Medical Simulation, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts

Sharon Muret-Wagstaff, Ph.D., M.P.A., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts

Facilitators and Program/Site Directors

Keith H. Baker, M.D., Ph.D., Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts

Galina Davidyuk, M.D., Ph.D., Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts

Jennifer L. Dearden, M.D., Children's Hospital Boston, Harvard Medical School, Boston, Massachusetts

David M. Feinstein, M.D., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts

Stephanie B. Jones, M.D., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts

William R. Kimball, M.D., Ph.D., Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts

John D. Mitchell, M.D., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts

Robert L. Nadelberg, M.D., Center for Medical Simulation, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts

David B. Waisel, M.D., Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts
Sarah H. Wiser, M.D., Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts

Rating Team

Deborah J. Culley, M.D., Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts
Lauren J. Fisher, D.O., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts
Rikante O. Kveraga, M.D., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts
Shannon S. McKenna, M.D., Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts
John D. Mitchell, M.D., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts
John B. Pawlowski, M.D., Ph.D., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts
Robert N. Pilon, M.D., Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts
Douglas C. Shook, M.D., Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts
David A. Silver, M.D., Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts
Carol A. Warfield, M.D., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts

Appendix 2. Sample Scenario

Elderly patient for resection of a facial basal cell carcinoma under monitored anesthesia care.

Scenario #3					
Case Title *Do Not* read to participants	Basal Cell Carcinoma This scenario is Part 1 of 2. For Part 2 please see Scenario 4.				
Total time	18–20 min: Case presentation (1 min), in-room freeze time (1–2 min), scenario (10–12 min), postscenario questions (5 min).				
Patient information	Name: Maria Elias		Age: 75 yr		Sex: Female
	Weight: 150 pounds		Height: 5'6"		Race: N/A
Case presentation To be read to participants	• This is an elderly, but otherwise healthy, patient having an excision of a basal cell carcinoma on the left side of her face under MAC. • During the preoperative discussion, the patient strongly preferred MAC due to history of significant PONV. The surgeon routinely performs these cases in this manner and booked the case under MAC. • The surgeon has just finished prepping and draping. So far, midazolam 1 mg and fentanyl 50 μg have been administered. • We will enter the room together. Time will be "frozen" for approximately 1 to 2 min so you can get oriented to the room and case. You may ask questions during this time. • Once I leave the room, the case will begin.				
Past medical, surgical, and family history	• Allergies: None • Medications: None				
Diagnostic tools	None				
Narrative case description Describe how the case unfolds, including major patient trends and consequences of interventions	1. At 1 min, patient becomes agitated in spite of sedation and gradually gets out of control, complaining of generalized discomforts (back pain, neck pain, and claustrophobia). • If sedation is given → Patient first becomes more disoriented and moves a lot, then hypoventilates and/or becomes apneic and the Spo_2 decreases to low 90s depending on how much sedation is given; patient very sensitive to medications in general and due to age. • If no sedation is given → Patient complains of generalized discomforts and moves excessively. 2. Surgeon unable to operate in this environment and pushes resident to "do something." • If asked, surgeon can give more local but unsure of how much more effect he is going to get because he has infiltrated the entire field already.(Will give more once, but not after that. Will point out that she is not complaining of face pain.) 3. Nurse offers comfort to the patient but that just exacerbates the situation and makes the patient more upset. 4. Resident must deal with the situation.				
Clinical diagnosis, treatment options	1.Recognize the limits of MAC anesthesia. 2.Consult with attending physician before proceeding with general anesthesia or aborting the procedure.				
Teaching/debriefing points	1.Resident should assess the limits of the patient's reserve. 2.Resident should attempt to sufficiently allay patient's anxieties and control environment including interruptions from surgeon and nurse.				
Staffing Roles—participants needed:	1. *Standardized Patient* Mannequin	2. *Circulator*	3. *Surgeon*	4. *Attending* Direct scenario and serve as patient voice if needed; silent if patient is apneic	5. *Patient Voice* Attending w/ female voice changer; sim tech if patient still speaking, attending in OR

(*Continued*)

Appendix 2. (*Continued*)

Scenario #3	
Learners	One first-year resident or one anesthesia fellow
Props needed	• Normal OR setup, patient chart including info re: patient, surgical history and physical, laboratory data, and consent forms. • Info on white board in OR. • Syringe and needle for bupivacaine for surgeon to infiltrate into the field. • Stethoscope. • Facemask on the patient with oxygen administration. • Carbon dioxide monitoring connected to mask. • Sedative agents available (midazolam, fentanyl, remifentanil, propofol [60 cc syringe], morphine, ondansetron, dexamethasone, haloperidol, plus other standard drugs). • Two Baxter pumps and IV tubing for sedative drugs—do not have medications loaded, just available. • Half of face should be draped. • Check camera angles to ensure good viewing. • Mannequin sound check for voice.
Script essentials: Specific lines to be delivered. Triggers highlighted in blue.	
General description	The patient is under the drapes and the surgeon is working on the left cheek. The patient is moving and moaning. The surgeon is complaining. If the resident continues sedating the patient, she becomes briefly apneic, Spo_2 decreases and the setting is even more disruptive to the procedure. The resident has options to give the patient a general anesthetic (by securing the airway or doing IV general anesthesia with natural airway), continue inadequate sedation, or stop the case. This may require calling the attending and plan to alter the course of the anesthetic.
Cues for patient	1. If not sedated enough… → Say: "My back really hurts," "Can you get these drapes off my face? Can you scratch my face?" "Can I turn my head for just a little bit? My neck is bothering me," and move a lot. 2. If receiving excessive sedation... → Say nothing and do not move. 3. In response to the nurse's comments… → Say loudly, "I'm so angry when you keep telling me everything is going to be ok! You are not the one being cut open and hurting!" 4. When surgeon says that he wants general anesthesia→ If awake enough say, "I'll be better, I promise. I'll try harder to stay still. I really don't want to go to sleep! I don't want gas!"
Cues for circulator	1. You are trying hard to make the patient feel comfortable. 2. You are holding the patient's hand. Say phrases such as…→ "Everything is fine, don't worry," "Just hold still now and you can get through this," "Let the doctor finish this part, don't worry." 3. The patient responds with hostility to your words. The attention only serves to give the patient an ear to complain to more vociferously. If resident asks for attending…→ Go to the phone and pretend calling. Then say "Your attending is in the middle of another case and will be here in a few minutes."
Cues for surgeon	1. When patient is agitated and alert… → Be understanding but firm, turn toward the patient and say: "You've got to hold still, we talked about this already," "Just stay still for a few more minutes now, hang in there." 2. In response to the patient complaining… → Look to the resident and say: "Can you do anything to help me out here? Can we work together on this?" If resident tries to call for attending early, say, "Can you use something IV? By mask? How about some nitrous?"
Postscenario notes Facilitator	Facilitator: "This scenario is over." Then escort resident to debriefing room. "I will now ask you three questions. Please express all of your thoughts so we can identify your thought processes." At the discretion of the facilitator, the following question can be used at any time: "Would you please elaborate?" Ask these three questions: 1. I noticed that this patient was a challenge to keep comfortable. I am wondering what your options were and what your management plan was. 2. There was a lot going on in this case. Could you tell me about any times that you felt challenged either in being able to think things through or to get things done? 3. If you were presented with this case again, is there anything that you would do differently?

IV = intravenous; MAC = monitored anesthesia care; N/A = not applicable; OR = operating room; PONV = postoperative nausea and vomiting; Sim tech = simulation technician; Spo_2 = oxygen saturation.

Appendix 3. Scoring Rubric—Behavioral Domain Descriptors

	Low (1, 2)	Middle (3, 4, 5)	High (6, 7)
Domain 1: Synthesizes information to formulate a clear anesthetic plan. • Gathers data • Synthesizes • Formulates plan adapted to patient and situation	Misses key data elements. Incorrect or absent prioritization of what is most important. Seldom asks follow-up questions. Rote, algorithmic anesthetic management. Does not apply book knowledge to real world. Plan may be inappropriate or may not be articulated.	Gathers and distills major elements of relevant information. Asks some follow-up questions. Articulates an appropriate, basic plan that generally is adapted to the patient and situation.	Efficient in gathering relevant information. Synthesizes and prioritizes effectively. Articulates an appropriate anesthetic plan that is highly tailored to patient and situation. Develops plan with input from patient and colleagues. Initial plan includes anticipatory planning for contingencies.
Domain 2: Implements a plan based on changing conditions. • Situational awareness • Rapid and frequent reassessment • Adaptable • Prioritizes multiple tasks • Flow • Decisive • Manages time, personnel, resources	Pays poor attention to details of anesthesia procedures and safety: suction, drug labels, patient history. Absent or slow response to alarms, vital signs changes, surgical events. Exhibits poor judgment. Does not plan thoroughly for worst-case possibilities. Does not act as part of the system, *e.g.*, not ensuring antibiotics are given. Poor sequencing. Perseverates. Becomes flustered and cannot integrate information into comprehensive plan. Does not pursue diagnostic possibilities systematically. Fixated on one part of the problem; stuck; unable to alter plan.	Adapts plan to changing circumstances before patient is jeopardized. Occasional lapses, but generally able to follow clinical situation. Flexible, *e.g.*, uses incremental dosing and observes response. Plans ahead but may not consider contingencies. Performs individual role responsibilities, but may not engage the team.	Recognizes emerging problems quickly, *e.g.*, bleeding, hypertension. Articulates a complete and coherent plan under changing circumstances, *e.g.*, "Here's where we are, here's what I'm going to do." Mobilizes human resources effectively, *e.g.*, asks staff to call others, get equipment. Smooth flow, anticipates next step. Coherent, systematic pursuit of priorities. Alert to changing circumstances, continuously reassessing. Decisive. Nimble. Coordinates with the rest of the room. Exhibits leadership, controls the room.
Domain 3: Demonstrates effective interpersonal and communication skills with patient and staff. • Clear and assertive • Caring and respectful • Elicits others' views • Listens • Interactive	Vague, lacks assertiveness. Silent. Communication is not tailored to patient's level of understanding. Fails to address patient concerns. No eye contact, flipping through chart. Arrogant, condescending, argumentative, defensive, manipulative, dismissive, sarcastic, rolls eyes.	Greets and introduces (patient, staff); addresses others by name. Listens, clarifies. Explains. Usually makes eye contact. Stays on track. Adapts to patient's level of understanding. Picks up on patient concerns, *e.g.*, if patient complains of IV pain, trainee says, "I'll take a look." Initiates communication with surgeon, nurse.	Clear, articulate. Assertive, speaks up. Establishes rapport. Sensitive and responsive to others. Goal-directed. Respectful, warm, caring. Consistently makes eye contact. Acknowledges others' views. Transparent—says what she or he is thinking. Checks understanding of both parties, clarifies. Interactive chains of two-way communication, consistently closes the loop. Welcomes patient comments and encourages patient to ask questions.
Domain 4: Identifies ways to improve performance. • Acknowledges feedback • Uses data to self-assess Articulates plan for future challenge	Unwilling to accept criticism. Blames others for own deficiencies. Lack of insight. Stubborn. Does not learn from experience. Refuses to change mind in the face of contrary evidence. May be passive or defensive. Not open to patient management discussion.	Accepts positive and negative feedback about performance, although may not describe or elaborate on what went well or what did not. Identifies at least one way to strengthen performance.	Readily recognizes and acknowledges errors. Recognizes seriousness of mistakes, *e.g.*, drug error. Uses objective information to evaluate own performance. Accepts and synthesizes feedback. Articulates a plan or intention to translate feedback into action in a specific way. Recognizes positive performance.

(*Continued*)

Appendix 3. (*Continued*)

	Low (1, 2)	Middle (3, 4, 5)	High (6, 7)
Domain 5: Recognizes own limits. Knowledge limits Capability limits Physical limits Seeks assistance	Does not call for help in a timely way when needed, *e.g.*, not asking Pharmacy how fast to give an unfamiliar drug. Institutes inappropriate therapies unsupervised, *e.g.*, chest compressions. Task-overloaded without calling on others. Fails to use resources in the room, *e.g.*, ask nurse to set up IV.	Recognizes personal limits. Readily says, "I don't know but I will find out" when faced with a new situation. Seeks information during the case if needed, *e.g.*, readily asks or looks up drug dose or rate of administration of an unfamiliar antibiotic. Calls for help when needed.	Gathers needed information before the case starts. Immediately and calmly seeks information or calls for assistance in an appropriate and effective way. Able to gather information from various resources if doing something she or he is not familiar with or has never done before. Uses information from team members and patient if faced with a challenging situation with which she or he has little experience or is high risk. Consistently practices within the realm of his or her own specific knowledge, competence, experience, and circumstances.

For further information please contact:
Harvard Assessment of Anesthesia Resident Performance Research Group
Richard H. Blum, M.D., M.S.E., PI, Department of Anesthesiology, Perioperative and Pain Medicine, Boston Children's Hospital; Harvard Medical School; richard.blum@childrens.harvard.edu
Jeffrey B. Cooper, Ph.D., Co-I, Center for Medical Simulation; Harvard Medical School; jcooper@partners.org
Sharon Muret-Wagstaff, Ph.D., Co-I, Faculty Development and Innovation, Department of Anesthesia, Critical Care and Pain Medicine, Beth Israel Deaconess Medical Center; Harvard Medical School; smuret@bidmc.harvard.edu