Simulation-based Assessment to Reliably Identify Key Resident Performance Attributes

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

ABSTRACT

Background: Obtaining reliable and valid information on resident performance is critical to patient safety and training program improvement. The goals were to characterize important anesthesia resident performance gaps that are not typically evaluated, and to further validate scores from a multiscenario simulation-based assessment.

Methods: Seven high-fidelity scenarios reflecting core anesthesiology skills were administered to 51 first-year residents (CA-1s) and 16 third-year residents (CA-3s) from three residency programs. Twenty trained attending anesthesiologists rated resident performances using a seven-point behaviorally anchored rating scale for five domains: (1) formulate a clear plan, (2) modify the plan under changing conditions, (3) communicate effectively, (4) identify performance improvement opportunities, and (5) recognize limits. A second rater assessed 10% of encounters. Scores and variances for each domain, each scenario, and the total were compared. Low domain ratings (1, 2) were examined in detail.

Results: Interrater agreement was 0.76; reliability of the seven-scenario assessment was r = 0.70. CA-3s had a significantly higher average total score (4.9 ± 1.1 *vs.* 4.6 ± 1.1, $P = 0.01$, effect size = 0.33). CA-3s significantly outscored CA-1s for five of seven scenarios and domains 1, 2, and 3. CA-1s had a significantly higher proportion of worrisome ratings than CA-3s (chi-square = 24.1, $P < 0.01$, effect size = 1.50). Ninety-eight percent of residents rated the simulations more educational than an average day in the operating room.

Conclusions: Sensitivity of the assessment to CA-1 *versus* CA-3 performance differences for most scenarios and domains supports validity. No differences, by experience level, were detected for two domains associated with reflective practice. Smaller score variances for CA-3s likely reflect a training effect; however, worrisome performance scores for both CA-1s and CA-3s suggest room for improvement. **(ANESTHESIOLOGY 2018; 128:821-31)**

EVALUATING whether graduates of anesthesiology residency programs are competent is an essential goal.[1,2] Timely identification of performance concerns enables early remediation and higher likelihood that interventions will be effective.[3,4] Early identification also enables curricular improvement and increases the likelihood of graduating competent residents who can increase quality of care and patient safety. Recent studies of patient outcomes show undesirable variations in cost and quality of care provided by resident physicians.[5–9] Case-focused, simulation-based assessments suggest that some residents are unable to perform at expected levels.[10–13] Typical resident evaluations include periodic ratings by anesthesia staff, in-training examinations and, for some residencies, oral exams. In attempts to improve resident training and proficiency, the Accreditation Council for Graduate Medical Education (ACGME) provides a structure of six core competencies

What We Already Know about This Topic

- Establishing anesthesiology resident competency is important. However, all critical performance competencies may not be captured in a standardized way during the course of residency training.
- A simulation-based assessment of anesthesiology residents was previously developed and used to assess core anesthesiology cases for first-year residents and fellows.

What This Article Tells Us That Is New

- This investigation administered the seven-scenario assessment to a larger sample of first-year and third-year residents from three training programs in three different simulation centers.
- The assessment had good reliability and interrater agreement, further validating the technique.
- Third-year residents significantly outscored first-year residents for five of seven scenarios, and the vast majority of residents found the simulations more educational than an average day in the operating room.

Corresponding article on page 707.

Submitted for publication November 17, 2016. Accepted for publication October 26, 2017. From the Department of Anesthesiology, Perioperative and Pain Medicine, Boston Children's Hospital and Harvard Medical School, Boston, Massachusetts (R.H.B.); the Department of Surgery, Emory University School of Medicine, Atlanta, Georgia (S.L.M.-W.); the Foundation for Advancement of International Medical Education and Research, Philadelphia, Pennsylvania (J.R.B.); the Center for Medical Simulation, Charlestown, Massachusetts (J.B.C., R.H.B.); and the Department of Anesthesia, Critical Care and Pain Medicine (J.B.C.), the Department of Surgery and Massachusetts General Hospital Learning Laboratory (E.R.P.), Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts.

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

 Anesthesiology 2018; 128:821–31

<zdoi;10.1097/ALN.0000000000002091>

organized developmentally (Milestones).[14] We seek to aid in the ACGME goal to improve resident proficiency by further evaluation of a methodology to assess a resident's critical performance behaviors that are not typically captured in a standardized way during the course of residency training.[3]

Simulation-based performance assessment has made significant advances[15–17] and now is being used to evaluate technical[18–21] and nontechnical[22–25] skills in anesthesiology. Most of these tools have been used for formative purposes, and their application in summative assessment is increasing.[26–28] Despite these advances, few tools identify and characterize complex behavioral skills deemed essential for safe practice. Additionally, most direct observation assessment tools have not been adequately validated.[2,29,30]

We previously described the development of the Harvard Assessment of Anesthesia Resident Performance (HARP) instrument. Evidence was gathered to support validity of scores from a simulation-based assessment of anesthesia residents and to identify performance gaps early in training.[11] This previous study demonstrated feasibility to administer a seven-scenario assessment of core anesthesiology cases for first-year (CA-1) residents and fellows. Using a Delphi process involving experienced anesthesiology educators and program directors, we defined a five-domain performance rating scale based on key characteristics of graduates who may not have reached a minimally acceptable level of performance before program completion. Reliability in the previous study, as measured by a generalizability coefficient,[31,32] was high, exceeding 0.80; anesthesia fellows outperformed CA-1s, and residents showed a wide variation of scores on each domain, providing some evidence to support construct validity.

We sought additional evidence to support the validity of the HARP instrument by administering the assessment in three different simulation centers to a larger sample of CA-1s and a comparison group of third-year (CA-3) residents from three training programs. We studied whether the performance of CA-3s, on average, was better and less variable (potential training effect) than that of CA-1s. We also sought to explore use of the HARP assessment for identifying and characterizing performance in key domains and scenarios for residents at different training levels.

Materials and Methods

Study Design

This is a prospective observational study designed to identify and characterize key resident performance domains as a function of training experience and to obtain additional data to further validate the HARP assessment scores. We followed the validity paradigm proposed by Shaw *et al.*[33] and adapted from Kane,[34,35] in which validity is backed by evidence supporting a contextual argument of the purpose, processes, analysis, interpretation, and use of results from an assessment. The adaptation by Shaw *et al.* is organized around five components: construct representation, scoring, generalization, extrapolation, and decision-making. We report validity evidence for the HARP assessment following this framework.

Residents

With institutional review board approval, a convenience sample based on clinical rotation of 67 of 68 residents from three ACGME-accredited anesthesia residency programs at Harvard Medical School (Boston, Massachusetts) provided signed informed consent and participated in the assessment program. Fifty-one CA-1s from three programs, including 22 from the original study,[11] and 16 CA-3s from two programs were excused from predefined rotations by their institutions to participate in the educational program. Residents participated in the latter half of their respective training year in this cross-sectional study; no resident participated more than once or in more than one of their training years. No additional information was used to recruit or exclude participating residents.

Procedure

The study was conducted at three simulation sites, including the one reported in the original study.[11] Each resident's assessment took approximately 3 h. An attending anesthesiologist, trained in the assessment protocol, oriented and debriefed each resident. Facilitators provided a brief scripted introduction to the task and environment and answered resident questions before administering each scenario. Each resident played the role of the primary anesthesiologist in all seven, approximately 15-min, scenarios. Residents were instructed to perform as they do in their actual clinical practice. A confederate nurse was available to assist the resident. A confederate attending anesthesiologist, trained to offer general assistance, met the resident during the orientation but did not provide active guidance during the scenario; they were available for support if requested. Each scenario was recorded using SimCapture web-based software (B-Line Medical, LLC, USA) with three camera views, including one of the physiologic monitors.

To elicit each resident's clinical reasoning and awareness of areas for improvement (Domain 4) and of personal limitations (Domain 5), the facilitator asked the resident three scripted questions after each scenario: (1) "I noticed that [clinically significant occurrence]; 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?" and (3) "If you were presented with the case again, is there anything that you would do differently?" These interviews were video-recorded.

After the resident completed all seven scenarios, the facilitator provided approximately 15 min of formative feedback concerning the resident's overall performance. To mitigate potential bias, the facilitator and raters for each resident did not work in that resident's training program.

Performance Assessment

Five domains of resident performance were rated: (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 patient and staff, (4) identifies ways to improve performance, and (5) recognizes own limits. These domains were created using a Delphi process with seven senior anesthesiology faculty educators, program directors, and/or clinical competency committee chairs, who began the process by answering, "What traits characterize residents who, upon graduation, have not achieved a minimum level of competency?"[11] Each domain had a seven-point scale based on criterion-referenced behavioral descriptors (see HARP Instrument in appendix 2) and "not assessable." These were iteratively developed by key informant interviews of senior anesthesia faculty members, by resident observations, and from literature in multiple high-risk fields. A detailed description of HARP scales development is available.[11]

Scenarios

The seven scenarios were designed by a subset of HARP Research Group members (appendix 1) to elicit a broad range of anesthesia skills appropriate for CA-1s across multiple perioperative settings (table 1) and enable assessment of performance relevant to the five domains.[11] Scenario 1 included a trained standardized patient; scenarios 2 to 7 used a high-fidelity SimMan 3G (Laerdal, USA) mannequin.

Facilitator Actions and Training

Seven anesthesia faculty from the three residency programs participated in a 2-h facilitator training session, receiving a detailed manual of the assessment process (available from authors). Before the first scenario, a facilitator oriented the resident to the assessment environment, including the high-fidelity mannequin and confederates, and described the resident's role, emphasizing that the resident should perform as in actual practice under supervision of his or her attending. Facilitators were trained to administer three scripted questions probing into a resident's reflection on their performance after each scenario, and to provide overall feedback at the end of all scenarios.

Raters and Rater Training

Twenty board-certified anesthesia faculty members from three training programs served as raters. Raters were trained in small group, 3-h sessions using a training manual that included scenarios, performance domains, associated behavioral anchors, and instructions. Raters viewed high- and low-scoring performances of nonparticipating residents rated by investigator team consensus, then scored additional videos independently until all reached agreement within 1 point of the benchmark ratings.[11]

Standardization of Three Simulation Locations

To determine the feasibility of administering a highly realistic and complex simulation-based assessment at multiple sites, three locations were chosen: the independent simulation center previously reported[11] and two hospital-based simulation centers associated with two of the three training programs. All sites had a complete mock operating room and recording facilities. Simulation teams were trained using a detailed manual for setup, scenario administration, standardization of materials and equipment, confederate and facilitator training, technical operations, and equipment turnover. The confederate attending anesthesiologist and a confederate nurse were the same individuals at all sites. In addition, at least one of two of the study investigators supervised sessions to ensure consistency of scenario administration.

Data Collection and Scoring

To focus the rating task, raters were assigned to rate performance for two or three of the seven scenarios with respective postencounter questioning *via* web-based asynchronous video. To mitigate potential bias, no rater assessed residents from his or her program, and raters were blinded to the education level of the resident being assessed. Each of the 20 raters assessed approximately 25 participant scenarios (range, 15 to 44). The raters completed a paper copy of the five-domain HARP tool. Dependability calculations from the pilot study, in which each scenario was independently rated by two raters, indicated minimal loss of score precision if a single rater was used for each scenario. In addition to the 22 CA-1 residents who participated in the original validation study,[11] 45 residents (n = 301 scenarios, 14 missing) were rated by one rater; approximately 10% of these scenarios (n = 32), chosen at random, were independently scored by a second rater. This double scoring was done to estimate interrater agreement. When two raters scored a resident's performance on a given scenario, comparative analyses were based on the average score.

Table 1. Simulation Scenarios

Scenario No.	Brief Scenario Description
1	Preoperative assessment of a patient scheduled for urgent exploratory laparotomy for a perforated gastric ulcer.
2	Operative management of a patient with significant hemorrhage during repair of a perforated ulcer.
3	Monitored anesthesia care for a patient with discomfort during resection of a facial basal cell carcinoma.
4	Postoperative care of a patient with a pulmonary aspiration after basal cell carcinoma resection.
5	Management of anaphylaxis in a patient having a transurethral resection of the prostate and bladder biopsy.
6	Delayed awakening after transurethral resection of the prostate and bladder biopsy.
7	Identification and management of a mainstem intubation secondary to coughing in a patient undergoing a total thyroidectomy.

Seven scenario scores, five domain scores, and a total score were computed for each resident. A scenario score is the average of five domain ratings. Each domain score is the average of ratings across all seven scenarios for that domain. The total score is the average of seven scenario scores for a given resident. To identify residents with potentially problematic performance (labeled "worrisome"), we tabulated the total number of ratings 2 or lower for each resident. We also tallied the number of ratings 2 or lower for each scenario. The choice of 2 or lower was made to align the categorization of residents' performance with the rating behavioral descriptors (appendix 2).

Survey Data

Each resident completed a survey at the end of the assessment session. Raters and facilitators were surveyed after completing their tasks. Of interest were their perceptions of the authenticity of the scenarios, the appropriateness of scenarios for the skill level of a CA-1 resident, and the educational value of the assessment. Complete surveys are available on request (R.H.B.). Tables 2 and 3 show sample survey items and responses for residents and raters.

Statistical Analysis

To estimate overall reliability, variance components were calculated based on a seven-scenario assessment and a single (randomly chosen) rater. Interrater reliability was based on the correlation between scenario scores from two raters for the 10% sample (n = 32 scenarios). Scenario discrimination indices were calculated by correlating scenario-level scores with the total score. These indices provide information on how performance on individual scenarios relates to overall performance.

Descriptive statistics were used to summarize performance (mean ± SD). Variances (SDs) for scenario and domain scores were examined to explore potential training effects. To gather evidence to support construct validity, domain scores were correlated. To test for differences in performance by training year, six repeated-measures ANOVAs (RM-ANOVAs) were conducted. The repeated measure was the scenario (7), and the dependent variable was the scenario score or the individual domain score. All analyses were done with SAS version 9.4 (SAS, USA). Effect sizes (ESs) were calculated to quantify the magnitude of any differences that were found. While the performance ratings, especially for some of the individual scenarios or specific domains, may not be normally distributed, ANOVA was chosen because it is considered a robust test against the normality assumption.

To explore potential differences in low performance between the two groups, we tallied percentages of ratings 2 or lower for each cohort (CA-1, CA-3) on each of the seven scenarios by domain and overall. The Pearson chi-square test was used to test for differences between groups.

Table 2. Residents' Perceptions of HARP Assessment

Survey Item	No. Marking "Yes"	Percent
Do you feel that what you were expected to do in the simulation scenarios was representative of skills that resident at your level has the necessary training and experience to perform?	65 of 65	100%
Were the simulated experiences sufficiently realistic to allow you to act in ways that you think you would in an actual patient care situation?	61 of 64	95%
Was this overall experience useful for your resident training and a valuable use of your educational time?	64 of 65	98%
Did you receive sufficient and useful feedback about your performance during the evaluation session?	64 of 65	98%

HARP = Harvard Assessment of Anesthesia Resident Performance.

Table 3. Raters' Perceptions of HARP Assessment

Survey Item	No. Marking "Agree" or "Strongly Agree"*	Percent
The simulation scenarios are representative of situations and skills that are realistic to expect of a CA-1 resident.	20 of 20	100%
The scoring system domains and descriptors represent trainee behaviors that are critical to patient safety.	20 of 20	100%
The scoring system domains and descriptors represent trainee behaviors that are critical to advancement, successful completion of residency training, and safe independent clinical practice.	20 of 20	100%
The scoring system is a unique and useful addition to currently available tools to evaluate resident clinical performance.	20 of 20	100%
The scoring system makes it possible to tailor feedback and educational interventions for trainees early in their careers.	20 of 20	100%

*Four-point scale: strongly disagree (1), disagree (2), agree (3), and strongly agree (4).
CA-1 = first-year; HARP = Harvard Assessment of Anesthesia Resident Performance.

Table 4. Performance by Scenario and Training Year

Scenario	All Residents Mean (SD)	Cohort	No.	Mean	SD	Min.	Max.	Discrimination
1	5.0 (1.0)	CA-1	49	4.8	0.9	2.8	6.5	.54
		CA-3	16	5.5	1.1	2.8	6.8	
2	4.9 (1.1)	CA-1	49	4.8	1.1	1.6	6.5	.72
		CA-3	16	5.2	1.0	3.2	6.8	
3	4.4 (1.2)	CA-1	48	4.4	1.2	2.0	6.6	.59
		CA-3	16	4.5	1.1	2.6	6.2	
4	4.7 (1.1)	CA-1	49	4.7	1.1	2.4	7.0	.55
		CA-3	16	4.6	1.1	1.8	6.6	
5	4.3 (0.9)	CA-1	48	4.3	0.8	2.2	6.0	.53
		CA-3	16	4.5	1.1	2.8	6.2	
6	4.7 (1.4)	CA-1	47	4.5	1.5	2.0	7.0	.63
		CA-3	16	5.2	0.7	4.0	6.8	
7	4.6 (1.1)	CA-1	50	4.5	1.2	2.1	6.7	.63
		CA-3	16	4.9	1.1	3.0	6.8	

CA-1 = first-year; CA-3 = third-year; max. = maximum; min. = minimum.

Table 5. Domain Score Correlations

	Domain 1	Domain 2	Domain 3	Domain 4	Domain 5
Domain 1	1.00	0.91 < 0.0001	0.78 < 0.0001	0.59 < 0.0001	0.75 < 0.0001
Domain 2		1.00	0.76 < 0.0001	0.62 < 0.0001	0.72 < 0.0001
Domain 3			1.00	0.57 < 0.0001	0.64 < 0.0001
Domain 4				1.00	0.58 < 0.0001
Domain 5					1.00

Results

Four hundred fifty-two (96%) scenario ratings were obtained with 17 missing scenarios: four were a result of technical recording problems, and 13 scenarios were skipped to complete the assessment within the maximum 3 h allotted so the resident could return to duty or end the normal workday.

Psychometric Results

Interrater reliability for the total score based on the 10% sample was 0.76. The overall test reliability of the seven-scenario assessment was 0.70. Domain score reliabilities (1 to 5) over scenarios were 0.63, 0.55, 0.68, 0.62, and 0.61, respectively. Discrimination indices (table 4) for all scenarios were high (greater than or equal to 0.5).

Table 5 shows domain intercorrelations. Domain 1 (synthesizes information) was highly correlated with Domain 2 (modifies a plan; r = 0.91). Domain 4 (identifies ways to improve performance) had the lowest correlations with the other domains (average r = 0.59). Domain 3 (communication) showed modest correlations to Domain 4 (identifies ways to improve performance; r = 0.57) and Domain 5 (recognizes own limits; r = 0.64).

Comparison of Performance for CA-1s and CA-3s on Scenario, Domain, and Total Scores

The RM-ANOVA, based on the total scenario scores, yielded no significant year-by-scenario interaction. There was a significant scenario effect ($F = 4.0$, $P < 0.01$), indicating that, averaged over residents, the scenarios were not of equal difficulty. There was also a main effect attributable to residency year ($F = 4.1$, $P < 0.05$) where CA-3 residents (mean = 4.9 ± 1.1) outperformed CA-1 residents (mean = 4.6 ± 1.1; ES = 0.33). Statistically, CA-3s outperformed CA-1s for scenarios 3, 4, 5, 6, and 7. Table 4 shows descriptive data for each of the seven scenarios. No scenario had an average total score greater than 5. Minimum and maximum scenario scores indicate that there were low and high performers in both residency cohorts across all scenarios. Further, some CA-3s had scores lower than high-scoring CA-1s.

Results of the RM-ANOVAs for Domains 1, 2, and 3 were similar to those for scenario scores. That is, the mean domain scores were not equivalent, and the CA-3s, on average, outperformed the CA-1s. For Domains 4 and 5 (identifies ways to improve performance, recognizes own limits), there were no differences in scores by training year. Table 6 shows scores for each of five domains and a total score.

Low "Worrisome" Performance

Each resident had approximately 35 ratings (7 scenarios × 5 domains). Counting any rating 2 or lower as worrisome, CA-1s had a significantly higher proportion of worrisome scores than CA-3s (chi-square = 24.1, $P < 0.01$, ES = 1.50). Figure 1 shows the percentages of 2 or lower ratings for each cohort on each scenario. Scenario 3 (patient with discomfort

Table 6. Domain and Total Scores for CA-1s and CA-3s

Domain	CA Level	Mean	SD	Low	High
1	CA-1	4.5	1.3	3.0	6.0
	CA-3	5.0	1.3	4.4	5.7
2	CA-1	4.4	1.5	2.6	6.3
	CA-3	5.1	1.3	4.4	6.2
3	CA-1	4.7	1.4	2.3	6.4
	CA-3	5.0	1.4	4.3	5.9
4	CA-1	4.0	1.3	3.2	6.0
	CA-3	4.7	1.4	3.8	6.3
5	CA-1	4.5	1.5	2.5	5.8
	CA-3	4.8	1.4	3.1	6.1
Total score	CA-1	4.6	1.6	3.1	5.8
	CA-3	4.9	1.4	4.1	6.0

CA-1 = first-year; CA-3 = third-year.

during resection of a facial basal cell carcinoma) had the most worrisome ratings. Of particular note is Scenario 6 (delayed awakening), for which CA-1s had the highest percentage of worrisome total scores 2 or lower (26.8%) and CA-3s had their lowest percentage (1.5%).

Survey Data

All 20 raters (100%), 65 of 67 (97%) residents, and 7 of 7 (100%) facilitators completed surveys. Tables 2 and 3 show survey items and responses from residents and raters indicating that most residents and raters believed the HARP experience was realistic, appropriate for a CA-1, required key skills for safe practice, and provided a better educational experience than the time spent in the operating room on an average day.

Discussion

Results of this multisite, multiprogram study add validity evidence to a multiscenario simulation-based assessment that can characterize important attributes of resident performance and identify residents with performance gaps in selected areas, including some that are not typically captured in resident evaluation. These five performance domains and rubrics were developed *a priori* based on our expert anesthesiologists' top concerns about those residents who may not have reached a minimally acceptable level of performance before program completion. Our findings also suggest that the HARP assessment, if properly structured, can provide high-quality, detailed information about each resident's clinical skills, including the ability to reflect on their performance. At the individual level, such assessments can inform specific remediation efforts. At the program level, performance data can help guide curricular enhancement.

This study addresses multiple elements concerning assessment validity.[33–36] *Construct representation* is an aspect of validity that was supported through a stringent iterative process of expert consensus to develop the scenarios and domains that capture critical behaviors. Construct representation is further supported by positive responses from residents, facilitators, and raters of the authenticity of scenarios, typicality of the observed behaviors and usefulness of the experience. In terms of *scoring*, development of the behaviorally anchored rating scale also followed an iterative process using expert consensus coupled with observation and literature review. Reliability of scores for a seven-scenario assessment was reasonably high, was good for structured performance assessments,[30] and builds upon previous findings.[11] Raters used the full range of scales and reported that the scoring rubric captures skills and behaviors required for safe anesthetic care. Using the HARP scales, expected differences in performance by experience level were documented. Regarding *generalization*, the iterative design process using multiple sources to define critical anesthesia skills and settings led to tasks and scenarios that residents, facilitators, and raters reported to be representative of a range of anesthesia cases,

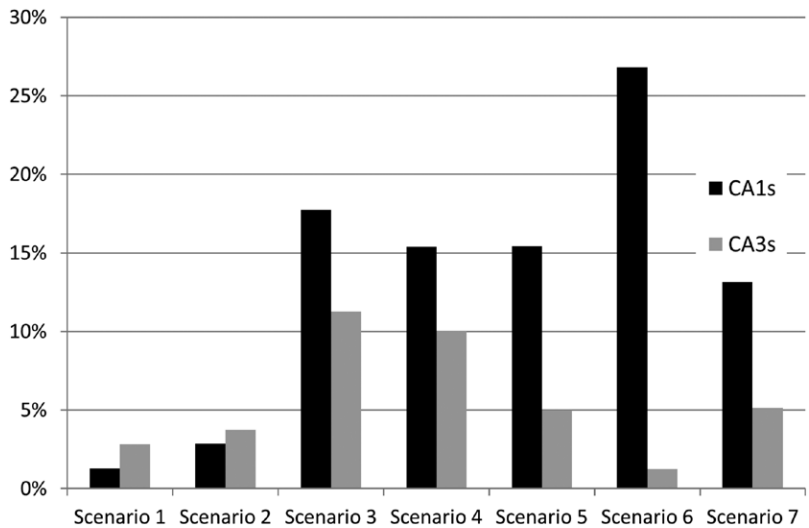

Fig. 1. Percentage of residents with scores 2 or lower for scenarios. CA-1 = first-year; CA-3 = third-year.

situations, and locations. Attention to scenario standardization and rigorous rater training were effective in controlling measurement error.

This study of the HARP assessment, extended to two other simulation facilities and teams, yields scores that are reasonably precise and can be used to identify specific content and skill-based performance issues. While a generalizability coefficient of 0.70 would not be adequate for a high-stakes assessment, it is acceptable for formative assessment purposes. The HARP assessment approach is feasible and provides a standardized assessment of important elements of anesthesiology resident performance. Further, it yields scores that, at least on average, can discriminate performance based on training level. Overall, psychometrically sound assessment data can be gathered with careful attention to scenario construction, scoring rubrics, and rater training.

Many anesthesiology assessment studies attempt to measure specific technical[18–21] or targeted behavioral skills like communication and teamwork.[22–25] Our simulation-based assessment and associated rating tools are designed to detect and characterize performance in specific cognitive/behavioral skills deemed critical to safe independent anesthesia practice. By recruiting a relatively large number of subjects and administering several scenarios to each participant, we obtained stable results characterizing the critical performance dimensions, providing additional evidence that supports the validity of the scores. This contrasts with many direct observation assessment tools that have not been subject to equally rigorous validation studies.[2,29,30] Validity evidence gathered in this study enabled further segmented depictions and comparisons of performances by domain, by scenario, and by high *versus* low ("worrisome") scores, outlined below.

Domain Scores

More experienced residents outperformed those with less experience on the more clinically oriented domains (1 to 3) that are typically addressed throughout training. This training effect is also evident by the smaller score variance for CA-3 residents compared to variance for CA-1s. Training should both increase performance scores and lead to greater homogeneity of resident performance. No training effect was detected for Domains 4 (identifying areas for improvement) and 5 (knowledge of one's limitations). Mastery of these skills, even during a 3-yr period, is likely to be more challenging.

Our assessment process was designed to explicitly elicit self-reflection on identifying areas for improvement and recognition of one's limitations. No significant differences in average scores between CA-1s and CA-3s in these two domains were observed. Interestingly, 7 to 9% of CA-3s' ratings for Domains 4 and 5 were 2 or lower, suggesting that, at least in a simulation environment with specific clinical interactions, some senior residents did not recognize their limitations or identify areas for improvement, despite the importance of these skills for reflective clinical practice.[37–40] We were encouraged that some CA-1s and CA-3s achieved high average scores for Domains 4 and 5. Several of our findings have been noted in other studies: performance variability among training level cohorts, overlap of simulation-based performance scores among novice and experienced clinicians, and associations between clinical skills or performance domains.[41]

Scenario Scores

Scores for CA-1s and CA-3s were not statistically different for Scenarios 1 (preoperative assessment) and 2 (intraoperative hemorrhage). These were the most straightforward scenarios with the highest mean scores. One reasonable explanation is that most CA-1s had already acquired skills necessary to manage these scenarios (ceiling effect). Scenario 6 (delayed awakening) was the most challenging case for CA-1s, requiring a higher level of clinical knowledge, experience, and communication skills to respond to production pressure from the surgeon to start the next case. CA-3s likely had substantially more opportunity to learn both the clinical management and communication skills to optimally manage this scenario. It is concerning, however, that some CA-3s did not do well on scenarios that CA-1 residents are expected to manage.

Low "Worrisome" Performance

Data indicated that CA-3s had a lower percentage of worrisome scores (rating 2 or lower) than CA-1s by group, by domain, and by scenario. However, it is concerning that any CA-3s had *any* worrisome scores and that several CA-3s had a total score less than the middle score (4) of the HARP rubric. While some may challenge our choice of ratings 1 and 2 as worrisome and argue for the generation of criterion-based performance standards, we based the choice on behavioral descriptors for ratings 1 and 2 for each domain. Additional research can help refine and validate these categorizations through formal standard setting procedures.[36] Nevertheless, given the alignment between our categorization of worrisome and the behavioral rating benchmarks, low performance of some CA-3 residents, albeit measured in the simulated environment, suggests opportunities for training improvement. Reasons for this low performance warrant further study.

Limitations

Residents and scoring metrics devised by the study faculty represent local programs and perhaps are not representative of all programs. Additionally, expected care for the conditions modeled in the scenarios may differ regionally. Although the simulated scenarios were developed to represent real patient care situations, the link between performance in the simulated environment and actual patient care situations has yet to be firmly established.[42–46] Further, our procedure to put CA-1s in the role of primary provider with somewhat passive attending supervision availability did not fully recreate the

actual clinical environment. Typically, anesthesia staff and senior residents working with CA-1s would take the lead in clinical decision-making.

To better understand reasons for the overlap of performance by some CA-3s with CA-1s, additional data are needed. Perhaps the scenarios were not challenging enough to allow for the detection of true ability differences. It is possible that other situational factors, such as motivation and fidelity of the simulations, may have confounded the results.

Further work is needed to address two important elements of the validity framework: (1) *extrapolation*: are the sampled constructs representative of competence in wider subject domains, *e.g.*, in real clinical settings? And (2) *decision-making*: is guidance in place so that stakeholders know what scores mean and how the outcomes should be used? Evidence to support these validity criteria are limited. For most simulation-based assessments, little work has been done to link performance in simulation to that in real clinical settings. Even for the few studies that document transfer of skills, most concern technical or procedural tasks.[18–20] Regarding decision-making, additional research is needed to map assessment scores to specific competency thresholds.

The HARP assessment approach yields psychometrically sound data, measures relevant patient care constructs, and can identify gaps in resident performance worthy of specific feedback that might not otherwise be identified early in training. The assessment can be replicated successfully at multiple sites. Additionally, this study reveals substantial performance variability within and among trainee groups at different levels, including locally defined worrisome scores in selected areas for some senior residents, and the provocative finding of no detectable training effects for reflective practice. Notably, these findings align with recent results for practicing physicians. About one fourth of 263 board-certified anesthesiologists participating in a simulation-based assessment received low overall ratings in managing simulated medical emergencies.[47] Our results build on previous work and provide a foundation for educators and researchers to replicate in larger samples that differ in various characteristics, pursue underlying causes, test correlations with actual perioperative performance, try new curricular revisions, and develop cut scores appropriate to diverse programs. As the field of simulation-based assessment matures, simulation-based scenarios with appropriate evaluation metrics should allow anesthesiology program directors to assess resident skills with more confidence and, where necessary, provide educational interventions that could improve the quality of their graduates' performance.

Acknowledgments

The authors thank the residents who participated; the faculty and department chairs who enabled the study; and staff members of the Center for Medical Simulation, Charlestown, Massachusetts, Shapiro Simulation and Skills Center, Boston, Massachusetts, and STRATUS (Simulation, Training, Research and Technology Utilization System) Center, Boston, Massachusetts, for their technical expertise and assistance.

Research Support

The study was supported by the Cathedral Fund, Newton Center, Massachusetts; the Branta Foundation, New York, New York; the Anesthesia Patient Safety Foundation, Rochester, Minnesota; American Society of Anesthesiologists, Schaumburg, Illinois; and the Departments of Anesthesia Chairs' Education Fund, Harvard Medical School, Boston, Massachusetts.

Competing Interests

Dr. Cooper was, during the time of this study, executive director of the Center for Medical Simulation, Charlestown, Massachusetts, a nonprofit organization that delivers simulation-based education to anesthesia professionals. Drs. Blum, Cooper, Petrusa, and Muret-Wagstaff were either directly or indirectly involved at the time of the study and currently in delivery of simulation-based education.

Correspondence

Address correspondence to Dr. Blum: Department of Anesthesiology, Pain and Perioperative Medicine, Boston Children's Hospital, 300 Longwood Avenue, Boston, Massachusetts 02115. richard.blum@childrens.harvard.edu. Information on purchasing reprints may be found at www.anesthesiology.org or on the masthead page at the beginning of this issue. Anesthesiology's articles are made freely accessible to all readers, for personal use only, 6 months from the cover date of the issue.

References

1. Hamstra SJ: Keynote address: The focus on competencies and individual learner assessment as emerging themes in medical education research. Acad Emerg Med 2012; 19:1336–43
2. Epstein RM: Assessment in medical education. N Engl J Med 2007; 356:387–96
3. 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
4. Rosenblatt MA, Abrams KJ; New York State Society of Anesthesiologists, Inc; Committee on Continuing Medical Education and Remediation; Remediation Sub-Committee: The use of a human patient simulator in the evaluation of and development of a remedial prescription for an anesthesiologist with lapsed medical skills. Anesth Analg 2002; 94:149–53
5. Asch DA, Nicholson S, Srinivas SK, Herrin J, Epstein AJ: How do you deliver a good obstetrician? Outcome-based evaluation of medical education. Acad Med 2014; 89:24–6
6. Asch DA, Nicholson S, Srinivas S, Herrin J, Epstein AJ: Evaluating obstetrical residency programs using patient outcomes. JAMA 2009; 302:1277–83
7. Newman D, Parente ST, Barrette E, Kennedy K: Prices for common medical services vary substantially among the commercially insured. Health Aff (Millwood) 2016; 35:923–7
8. Cooper Z, Craig S, Gaynor M and van Reenen J: The price ain't right? Hospital prices and health spending on the privately insured. National Bureau of Economic Research Working Paper 21815. 2015. Available at: http://www.nber.org/papers/w21815. Accessed June 10, 2016
9. Fisher ES, Bynum JP, Skinner JS: Slowing the growth of health care costs–lessons from regional variation. N Engl J Med 2009; 360:849–52

10. 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
11. 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
12. Hunt EA, Vera K, Diener-West M, Haggerty JA, Nelson KL, Shaffner DH, Pronovost PJ: Delays and errors in cardiopulmonary resuscitation and defibrillation by pediatric residents during simulated cardiopulmonary arrests. Resuscitation 2009; 80:819–25
13. 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
14. ACGME website. Available at: http://www.acgme.org/What-We-Do/Accreditation/Milestones/Overview. Accessed July 19, 2016
15. Boulet JR, Murray D: Review article: Assessment in anesthesiology education. Can J Anaesth 2012; 59:182–92
16. Boulet JR, Murray DJ: Simulation-based assessment in anesthesiology: Requirements for practical implementation. Anesthesiology 2010; 112:1041–52
17. 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
18. Giglioli S, Boet S, De Gaudio AR, Linden M, Schaeffer R, Bould MD, Diemunsch P: Self-directed deliberate practice with virtual fiberoptic intubation improves initial skills for anesthesia residents. Minerva Anestesiol 2012; 78:456–61
19. Wayne DB, Butter J, Siddall VJ, Fudala MJ, Wade LD, Feinglass J, McGaghie WC: Mastery learning of advanced cardiac life support skills by internal medicine residents using simulation technology and deliberate practice. J Gen Intern Med 2006; 21:251–6
20. Madenci AL, Solis CV, de Moya MA: Central venous access by trainees: A systematic review and meta-analysis of the use of simulation to improve success rate on patients. Simul Healthc 2014; 9:7–14
21. Sultan SF, Iohom G, Saunders J, Shorten G: A clinical assessment tool for ultrasound-guided axillary brachial plexus block. Acta Anaesthesiol Scand 2012; 56:616–23
22. 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
23. 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
24. Lorello GR, Cook DA, Johnson RL, Brydges R: Simulation-based training in anaesthesiology: A systematic review and meta-analysis. Br J Anaesth 2014; 112:231–45
25. Paige JT, Garbee DD, Kozmenko V, Yu Q, Kozmenko L, Yang T, Bonanno L, Swartz W: Getting a head start: High-fidelity, simulation-based operating room team training of interprofessional students. J Am Coll Surg 2014; 218:140–9
26. 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
27. Steadman RH, Huang YM: Simulation for quality assurance in training, credentialing and maintenance of certification. Best Pract Res Clin Anaesthesiol 2012; 26:3–15
28. Boulet JR: Summative assessment in medicine: The promise of simulation for high-stakes evaluation. Acad Emerg Med 2008; 15:1017–24
29. 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
30. Brannick MT, Erol-Korkmaz HT, Prewett M: A systematic review of the reliability of objective structured clinical examination scores. Med Educ 2011; 45:1181–9
31. Boulet JR: Generalizability Theory, Basics, Encyclopedia of Statistics in Behavioral Science. Edited by Evritt BS, Howell DC. Chichester, United Kingdom, Wiley, 2005, pp 704–11
32. Brennan RL: Generalizability Theory. New York, NY: Springer-Verlag, 2001, 4–20, 53–140
33. Shaw S, Crisp V and Johnson N: A framework for evidencing assessment validity in large-scale, high-stakes international examinations. Assess Edu Princ Policy Pract 2012; 19:159–76
34. Kane MT: An argument-based approach to validity. Psychol Bull 1992; 112:527–35
35. Kane MT: Current concerns in validity theory. J Educ Meas 2001; 38:319–41
36. 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
37. Schön, DA: Educating the Reflective Practitioner. San Francisco, Jossey-Bass, 1987
38. Dunne, M., Nisbet, G., Penman, M., McAllister, L: Influences and outcomes: A systematised review of reflective teaching strategies in student healthcare placements. Int J Pract Based Learn Health Soc Care 2016; 4: 55–77
39. Paget T: Reflective practice and clinical outcomes: Practitioners' views on how reflective practice has influenced their clinical practice. J Clin Nurs 2001; 10:204–14
40. Sargeant JM, Mann KV, van der Vleuten CP, Metsemakers JF: Reflection: A link between receiving and using assessment feedback. Adv Health Sci Educ Theory Pract 2009; 14:399–410
41. Weinger MB: Experience not equal expertise: Can simulation be used to tell the difference? Anesthesiology 2007; 107:691–4
42. Boet S, Bould MD, Sydor DT, Naik V, Friedman Z: Transfer of learning and patient outcome in simulated crisis resource management: A systematic review. Can J Anaesth 2014;61:571–82
43. Griswold-Theodorson S, Ponnuru S, Dong C, Szyld D, Reed T, McGaghie WC: Beyond the simulation laboratory: A realist synthesis review of clinical outcomes of simulation-based mastery learning. Acad Med 2015; 90:1553–60
44. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ: Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. Med Teach 2005; 27:10–28
45. Konge L, Bitsch M: Lack of correlation between performances in a simulator and in reality [article in Danish]. Ugeskr Laeger 2010; 172:3477–80
46. Domuracki KJ, Moule CJ, Owen H, Kostandoff G, Plummer JL: Learning on a simulator does transfer to clinical practice. Resuscitation 2009; 80:346–9
47. Weinger MB, Banerjee A, Burden AR, McIvor WR, Boulet J, Cooper JB, Steadman R, Shotwell MS, Slagle JM, DeMaria S Jr, Torsher L, Sinz E, Levine AI, Rask J, Davis F, Park C, Gaba DM: Simulation-based assessment of the management of critical events by board-certified anesthesiologists. Anesthesiology 2017; 127:475–89

Appendix 1. Harvard Assessment of Anesthesia Resident Performance Research Group

Core Investigators

Richard H. Blum, M.D. (Principal Investigator), Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts
Sharon Muret-Wagstaff, Ph.D., Emory School of Medicine, Atlanta, Georgia
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
Emil R. Petrusa, Ph.D., Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts

Facilitators and Program/Site Directors

Keith H. Baker, M.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
Sarah H. Wiser, M.D., Newton-Wellesley Hospital, Newton, Massachusetts

Rating Team

Meredith A. Albrecht, M.D., Ph.D, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts
Amanda K. Anastasi, M.D., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts
Ruma R. Bose, M.D., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts
Laura Y. Chang, 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, Boston, Massachusetts
Lauren J. Fisher, D.O., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts
Meera Grover, M.D., Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts
Suzanne B. Klainer, M.D., Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts
Rikante 0. Kveraga, M.D., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts
Jeffrey P. Martel, 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
Rebecca D. Minehart, M.D., Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts
John D. Mitchell, M.D., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts
Jeremi R. Mountjoy, M.D., Massachusetts General Hospital, 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
Katherine L. Zaleski, M.D., Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts

Appendix 2. Harvard Assessment of Anesthesia Resident Performance (HARP) Scoring Rubric—Behavioral Descriptors for Each Domain

	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 re-assessment • 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 re-assessing. 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 s/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.
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 s/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 s/he has little experience or is high risk. Consistently practices within the realm of his or her own specific knowledge, competence, experience, and circumstances.

Source: Blum R.H., Boulet J.R., Cooper J.B., Muret-Wagstaff S.L., for the Harvard Assessment of Anesthesia Resident Performance Research Group. Anesthesiology 2014; 120:129-41, appendix 3.