CLINICAL INVESTIGATIONS

Anesthesiology
1998; 89:8-18

Assessment of Clinical Performance during Simulated Crises Using Both Technical and Behavioral Ratings

David M. Gaba, M.D., Steven K. Howard, M.D.,† Brendan Flanagan, F.A.N.Z.C.A.,‡ Brian E. Smith, M.D.,§ Kevin J. Fish, M.D.,‖ Richard Botney, M.D.#*

Background: **Techniques are needed to assess anesthesiologists' performance when responding to critical events. Patient simulators allow presentation of similar crisis situations to different clinicians. This study evaluated ratings of performance, and the interrater variability of the ratings, made by multiple independent observers viewing videotapes of simulated crises.**

Methods: **Raters scored the videotapes of 14 different teams that were managing two scenarios: malignant hyperthermia (MH) and cardiac arrest. Technical performance and crisis management behaviors were rated. Technical ratings could range from 0.0 to 1.0 based on scenario-specific checklists of appropriate actions. Ratings of 12 crisis management behaviors were made using a five-point ordinal scale. Several statistical assessments of interrater variability were applied.**

Results: **Technical ratings were high for most teams in both scenarios (0.78 ± 0.08 for MH, 0.83 ± 0.06 for cardiac arrest). Ratings of crisis management behavior varied, with some teams rated as minimally acceptable or poor (28% for MH, 14% for cardiac arrest). The agreement between raters was fair to excellent, depending on the item rated and the statistical test used.**

Conclusions: **Both technical and behavioral performance can be assessed from videotapes of simulations. The behavioral rating system can be improved; one particular difficulty was aggregating a single rating for a behavior that fluctuated over time. These performance assessment tools might be useful for educational research or for tracking a resident's progress. The rating system needs more refinement before it can be used to assess clinical competence for residency graduation or board certification. (Key words: Evaluation; performance; simulation; teamwork; testing.)**

This article is accompanied by an Editorial View. Please see: Murray DJ: Clinical simulation: Technical novelty or innovation in education. ANESTHESIOLOGY 1998; 89:1-2.

Additional material can be found on the Anesthesiology Web Site. Go to the following address, and then scroll down to find the title link for this article.
http://www.anesthesiology.org/tocs/v89n1-TOC/cfm

* Staff Anesthesiologist, VA Palo Alto HCS; Associate Professor of Anesthesia, Stanford University School of Medicine.

† Staff Anesthesiologist, VA Palo Alto HCS; Assistant Professor of Anesthesia, Stanford University School of Medicine.

‡ Staff Anaesthetist, Monash Medical Centre, Clayton, Victoria, Australia.

¶ Staff Anesthesiologist, VA Palo Alto HCS; Clinical Instructor, Stanford University School of Medicine.

‖ Staff Anesthesiologist, VA Palo Alto HCS; Professor of Anesthesia, Stanford University School of Medicine.

Assistant Professor of Anesthesiology, Oregon Health Sciences University, Portland, Oregon.

Received from the Anesthesiology Service, VA Palo Alto Health Care System, and the Department of Anesthesia, Stanford University School of Medicine. Submitted for publication July 7, 1997. Accepted for publication March 6, 1998. Presented in part at the annual meeting of the American Society of Anesthesiologists, October 17, 1994, San Francisco, California. The study described herein was conducted using a noncommercial simulator manufactured in Dr. Gaba's laboratory. Dr. Gaba and his former research associate, John Williams, M.D., have received payments from Eagle Simulation, Inc. for licensing of their simulation technology, and they receive royalties on the sale of patient simulators by Eagle Simulation.

Address reprint requests to Dr. Gaba: Anesthesiology Service, 112A, VA Palo Alto HCS, 3801 Miranda Avenue, Palo Alto, California 94304. Address electronic mail to: gaba@leland.stanford.edu

TECHNIQUES to evaluate the clinical performance of anesthesiologists and certified nurse anesthetists during dynamic critical perioperative incidents (''crises'') are desirable for several purposes. These include assessing the educational progress or clinical competence of trainees or experienced practitioners and studying the efficacy of training methodologies. Performance assessment in anesthesiology has relied primarily on written and oral examinations and on evaluations of actual clinical performance. However, examinations cannot easily assess what the examinee would actually do in a critical event, nor can they determine how well the candidate would interact with the rest of the clinical team. Evaluations of clinical performance during actual events are

often made retrospectively from the evaluator's recollection of cases that have varied in complexity and in the types of events that occurred,[1] making it difficult to compare performance in managing a crisis from one clinician to another or against a specific standard. Realistic patient simulators offer a tool for the reproducible presentation of complex critical events. The advantages and disadvantages of simulation have been discussed previously.[2] Simulated patients (*i.e.*, actors) have been used to evaluate performance in history taking, physical examination, and other patient-interaction skills.[3,4] Mannequins for cardiopulmonary resuscitation have been used to evaluate performance of the technical actions of basic and advanced life support protocols.[5,6] Realistic simulators have also been used to assess the decision-making skills of anesthesiologists during simulated crises.[7-11]

Based on strong analogies to performance during management of critical events in other complex, dynamic domains such as aviation,[**12-17] we believe it is important to measure two separate aspects of skilled performance in managing crisis situations: implementing appropriate technical actions (technical performance) and manifesting appropriate crisis management behaviors (behavioral performance).[18] Technical performance concerns the adequacy of the actions taken from a medical and technical perspective. For example, chest compression and electric countershock are appropriate technical actions in cases of cardiac arrest and ventricular fibrillation. Behavioral performance concerns the decision-making and team interaction processes used during the team's management of a situation. Previously we described in detail these crisis management behaviors in a paradigm and curriculum we call Anesthesia Crisis Resource Management (ACRM).[10,18-20]

In this study, we used videotapes of different persons and teams while they managed realistic simulations of the same perioperative crises to assess their technical and behavioral performance. Then we evaluated the interrater variability of these assessments.

Methods

Rating Instruments††

Crisis Management Behaviors. Based on specific parallels between aviation and anesthesiology, we adapted an instrument used to rate flight deck crew behaviors, the Line/LOS‡‡ Checklist, which was developed by NASA and the University of Texas Aerospace Crew Performance Project.§§ Our instrument included ratings for 10 crisis management behaviors: orientation to case, inquiry/assertion, communication, feedback, leadership, group climate, anticipation/planning, workload distribution, vigilance, and reevaluation. We also included two summary ratings: overall performance of the primary anesthesiologist (primary overall), and overall performance of the anesthesia team (team overall).

For all behavioral ratings, a five-point ordinal scale was used as follows: 1 = poor performance; 2 = substandard (minimally acceptable) performance; 3 = standard performance; 4 = good performance; and 5 = outstanding performance. By convention, raters were instructed to rate the performance of the team as a whole (except for the marker of "primary overall") and also to round down to the next lower integer rating if they felt that performance was best described in between two scale points.

Technical Scoring. For each scenario we developed a list of the appropriate medical and technical actions for recognition, diagnosis, and therapy (table 1 shows an excerpt). Point values for successful implementation of each action were assigned subjectively by the investigators in advance. Raters recorded the presence or absence of each action during a scenario and then summed the point values for all actions recorded as present. Each technical score was expressed as the fraction of the maximum possible score (100 for cardiac arrest, 95 for malignant hyperthermia [MH]). This rating procedure is similar to that used by Chopra *et al.*[11] to analyze technical performance in simulated crises.

For each scenario a few actions were deemed so essential to success that even a perfect score for other activities could not compensate for a failure on one of them. For example, attempting electrical defibrillation

** Helmreich RL: Theory underlying CRM training: Psychological issues in flight crew performance and crew coordination, Cockpit Resource Management Training (NASA Conference Publication 2455). Edited by HW Orlady, HC Foushee. Washington, DC, National Aeronautics and Space Administration, 1986, pp 15-22.

†† The complete behavioral and technical rating forms can be found on the ANESTHESIOLOGY web page at http://www.anesthesiology.org

‡‡ Line checks involve a federally certified "check pilot" who examines the behaviors and technical flying skills of pilots during actual commercial flights; LOS stands for line-oriented simulation.

§§Helmreich RL, Wilhelm JA, Kello JE, Taggart WR, Butler RE: Reinforcing and evaluating crew resource management: Evaluator/LOS instructor reference manual (NASA/UT Technical Manual 90-2, revision 1). Austin, TX, NASA/University of Texas Aerospace Crew Performance Project, 1991.

Table 1. Excerpt from the List of Technical Actions and Point Values for the MH Scenario*

Action	Point Value
Initiation of MH protocol	
Notifies surgeon of MH emergency	5
Requests MH box	5
Calls for help, or already present	5
Terminates triggering agent within 1 min of notifying surgeon or requesting MH box	EI
Dantrolene administration	
Gives dantrolene ≥ 20 mg within 10 min of MH box arrival	EI
Gives dantrolene ≥ 40 mg by end of scenario	10
Uses correct diluent for all vials	8
Uses >30 ml diluent for all vials	6
Request additional dantrolene (only 80 mg in box)	6
Ventilation and oxygenation	
Uses $F_{I_{O_2}}$ of 1.0	5
Hyperventilates by ventilator or bag	5
Clears triggering agent with high flows or non-rebreathing circuit	5
Metabolic management	
Requests insertion of urinary catheter	1
Gives mannitol and/or furosemide	2
Checks blood for potassium level	3
Sends ABG for management, not for diagnosis of possible MH	2
Hyperthermia management	
Removes drapes	2
Calls for and places ice	2
Requests lower OR temperature	1
Requests supplies for cold lavage of stomach or peritoneum	2
Requests cold IV solutions	2
Considers planning for CPB	1
Miscellaneous management	
Treats PVCs with antiarrhythmic drug	2
Places arterial line	1
Reviews MH checklist (in box)	1
Suggests terminating surgery as soon as possible	1
Requests ICU bed	1
Contacts MH hotline	1

EI = essential item; failure to perform any essential item resulted in a technical score of zero for that scenario; MH = malignant hyperthermia; CPB = cardiopulmonary bypass.

* The entire list for the MH scenario contained 33 items totalling 95 points.

to treat cardiac arrest with ventricular fibrillation is an essential item. Thus, if at least two raters scored a team as failing to implement any essential items, the team was scored as "deficient" and the point score was not totalled. This procedure precluded a team from accruing a good point score after performing many appropriate actions while neglecting something essential.

Simulation Sessions

The simulation sessions were conducted in 1992 as part of a project to train faculty anesthesiologists from five training hospitals affiliated with a single medical school to teach the ACRM simulation curriculum. The project has been described previously in detail.[19] The simulation-based ACRM course was taught to 72 participants (37 CA2-CA4 residents, 31 faculty, and 4 certified nurse anesthetists) in 18 teams of 4 each. Physicians were never mixed with certified nurse anesthetists on a team, and with one exception residents were not mixed with faculty. Each team participated in a 2.5-h simulation session involving five different scenarios, two of which we chose for the current study. One scenario involved the development of myocardial ischemia, arrhythmias, and cardiac arrest during gastric surgery in a patient with a history of coronary artery disease. In the other scenario, an otherwise healthy patient with gastroesophageal reflux developed malignant hyperthermia (triggered by succinylcholine and isoflurane). During each simulation scenario one team member was the "primary anesthesiologist" while a second was sequestered in another room and could be called in to help unaware of previous events. If more help was requested, the other two team members (who had been observing the simulation) could be activated to join in. Team members rotated through these roles during different scenarios. Operating room nurses acted their roles and a nurse or surgical resident acted as the surgeon. To enhance the confidentiality offered to participants, the level of training of participants was not recorded with the videotape of their simulation sessions.

Human Subjects Approval, Training of Raters, and Rating Procedure

This study was conducted using existing videotapes made for the debriefings in the ACRM courses, and the identity of the clinicians was not recorded. Thus it qualified for and was granted exempt status by the Stanford University Institutional Review Board. Raters were board-certified anesthesiologists based at Stanford University who had each conducted and debriefed these scenarios during ACRM courses. To train the five investigators who rated crisis management behaviors, four (of 18) videotapes were chosen at random, rated independently, and then discussed by the raters. The remaining 14 videotapes were placed in random order. The five raters separately viewed and rated the videotapes in the same order to minimize any effect of view-

ing order. Raters scored each team for 12 crisis management behaviors separately for two time periods defined in advance for each scenario. Time period one began at the start of the scenario and ended when the team explicitly declared an emergency. Time period two covered the rest of the scenario, during which the declared crisis was actively managed.

After the behavioral ratings were completed, three raters (two who had also conducted behavioral ratings and one who had not) separately viewed each tape to conduct the technical scoring. Since this involved only identifying the presence or absence of defined clinical actions, no training procedure was employed in advance.

Thus the data set consisted of (1) a rating by each of five raters on 12 behaviors for each team during two time periods in two scenarios; and (2) a technical score by each of three raters for each team in two scenarios.

Statistical Analysis

Although raters independently evaluated the same teams, the variability between ratings cannot be considered a true measure of the interrater reliability of the rating scales, because the data set from this experiment did not include predefined levels of performance covering the spectrum from poor to outstanding. Nonetheless, with this caveat, standard statistics of interrater reliability provide an index of the level of agreement between raters in this experiment. Several statistical tests, each with its own limitations, have been used by investigators doing analogous studies. For this reason we present four statistics (as described in the appendix) to categorize the level of agreement more completely, and to facilitate comparison of our results with others:

- S_{av} and a variant we call S_{avr}—both are generalized forms of the family of tests of which the more familiar "kappa" statistic is a special case[21,22];
- the Intraclass Correlation Coefficient (ICC)[23]; and
- the Within-group Interrater Reliability Coefficient (r_{wg}).[24]

The ICC and the r_{wg} are correlation techniques that typically are applied to interval data. However, a study of interrater reliability of assessments of flight crew behaviors|||| applied r_{wg} to the same ordinal scale we used for rating behaviors; thus we report its value for both the behavioral ratings and the technical ratings. For all four statistics, a value of 0.0 indicates agreement between raters at the level of "chance" and 1.0 indicates perfect agreement (but see Appendix 1 on the meaning of "chance" for the different tests).

Hypothesis testing of whether a statistic of interrater reliability differed from zero used a t-statistic for S_{av} or S_{avr} and Fisher's Z transformation for ICC and r_{wg}.[23] Statistical significance at $P < 0.05$ and $P < 0.005$ was assessed ($P < 0.005$ is approximately equivalent to a Bonferroni correction for the rating of 12 related behaviors). One-tailed tests of significance were used because a negative value of any of the interrater reliability statistics indicates agreement worse than chance, which we included in the null hypothesis. Correlation between ratings of different behaviors were assessed using the Spearman Rank Coefficient.[23]

Results##

Technical Ratings

The technical scores were fairly high, as shown in figures 1 and 2, and there was only moderate variation between teams. No team qualified for an official rating of "deficient" (by at least two raters), although one rater did score a single team as deficient due to ambiguity in the exact time the MH box was requested.

On average, all three raters agreed about the presence or absence of a technical action 86% of the time for cardiac arrest and 83% of the time for MH. Figure 3 shows the variability in technical rating point scores from all three raters for each team for the MH scenario. The r_{wg} was 0.96 for both scenarios, but these values are artificially high because no team had a score <0.6. When the r_{wg} was computed based on the actual range of scores, the corrected values are 0.65 for cardiac arrest and 0.62 for MH (both significantly greater than zero, $P < 0.0002$). The ICCs were 0.57 for MH and 0.36 for cardiac arrest. The value for MH was significantly greater than zero ($Z = 2.64$, $P < 0.005$), but the value for cardiac arrest was not significant ($Z = 1.62$, $P = 0.054$).

Behavioral Ratings

As shown in figures 1 and 2, the mean ratings of overall team crisis management behavior showed sub-

|||| Law JR, Sherman PJ: Do raters agree? Assessing inter-rater agreement in the evaluation of air crew resource management skills. Proceedings of the 8th Ohio State Symposium on Aviation Psychology, April 24–27, 1995, Columbus, Ohio, pp 608–12.

The raw data tables and other results not presented here can be found on the ANESTHESIOLOGY web page at http://www.anesthesiology.org

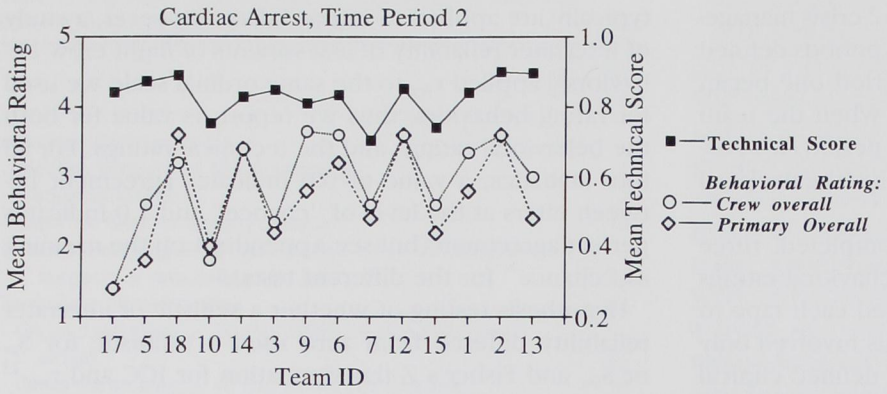

Fig. 1. Comparison of performance between teams for the cardiac arrest scenario. The figure shows both the technical score (0.0 to 1.0) and the behavioral ratings (only for time period 2—after declaration of cardiac arrest) for team overall and primary overall (1 = poor; 2 = substandard, minimally acceptable; 3 = standard; 4 = good; 5 = outstanding). "Team ID" is the numeric identifier of the participant team. The Team IDs are shown in the random order in which the tapes were viewed.

stantial differences between the teams. Several teams had mean overall team ratings at the level of "minimally acceptable" or below (14% for cardiac arrest, 28% for MH), and the performance of the primary anesthesiologist was rated at or below this level even more frequently (21% for cardiac arrest, 35% for MH). The ratings for specific crisis management behaviors showed similar patterns. For most teams, overall performance followed that of the primary anesthesiologist, with team ratings being slightly better than individual ratings in nearly all cases.

Figure 4 illustrates the interrater variability for behavioral scores, showing the ratings of all five raters for each team on the "team overall" behavior for time period two of the MH scenario (*i.e.*, after declaring an MH emergency). Although raters tended to agree, the deviations between them are apparent in the figure, and in one case (Team ID 13) one rater evaluated the team's performance as poor while another rated it as good.

As tables 2 and 3 show, the value of the conservative S_{av} statistic for the 12 behaviors rated ranged from 0.00 to 0.55 (0.32 – 0.54 for the two "overall" ratings). Twenty-five of 46 of the S_{av} values were significantly different than zero at the $P < 0.005$ level (corrected for multiple comparisons). The values of the more liberal S_{avr} statistic ranged from 0.6 to 0.9 and were significantly greater than zero ($P < 0.001$) for all behaviors in both time periods of both scenarios. Values of the r_{wg} ranged from 0.6 to 0.93, all of which were significantly different than zero at $P < 0.005$.

Based on all ratings of behaviors, there was a probability of 0.14 that any single rating would deviate by more than 1.0 rating scale point from the mean of all five raters for that item (one point is sufficient to move a rating from one scale classification to the next). There was a much lower probability of 0.015 that the mean rating of any two raters chosen at random would deviate by more than 1.0 from the mean score of all five raters. These probabilities are important in judging whether the rating of a single rater or of a pair of raters is a reliable predictor of the mean of the five raters.

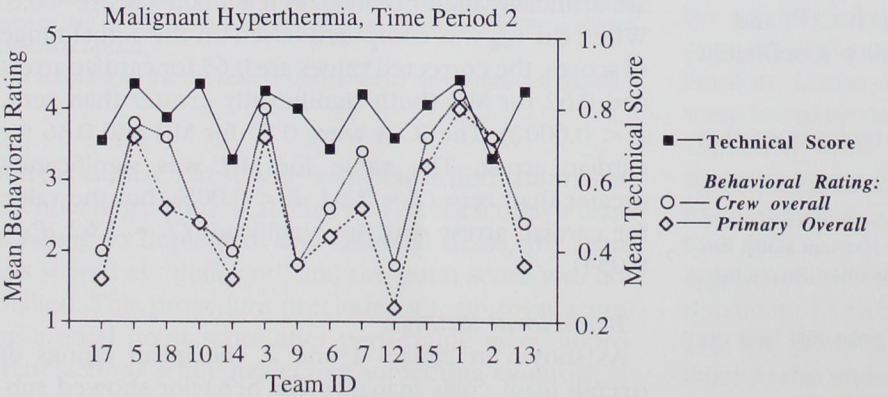

Fig. 2. Comparison of performance between teams for the malginant hyperthermia scenario (behavioral ratings are for time period 2 only—after declaration of a malignant hyperthermia emergency).

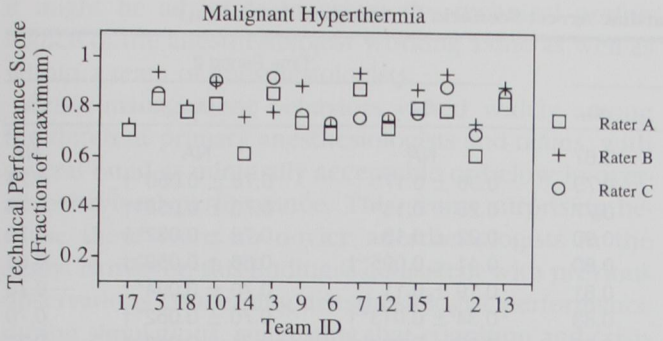

Fig. 3. The technical score assigned by each of three raters for all teams during the malignant hyperthermia scenario.

There were strong correlations between the mean rating of the overall team behavior and overall primary anesthesiologist behavior (R = 0.93 to 0.95) and between these overall ratings and specific ratings of important behaviors taught in ACRM, such as leadership, communications, and distribution of workload (R = 0.84 to 0.96).

Discussion

This study shows the feasibility of rating both technical and behavioral performance of anesthesiologists based on videotapes of simulated crisis events. The teams were successful at implementing appropriate technical actions, in general performing >80% of the checklist actions, and never missing an essential item. This should not be surprising because (1) participants had at least 2 yr of postgraduate medical training and all had previously received ACLS certification, (2) the scenarios portrayed in this study have well-known treatment protocols and the most critical items could be accomplished by only one or two persons who knew exactly what to do, (3) the ACRM training paradigm encourages distribution of workload and mobilization of help, which tended to "level out" the technical performance.

There was good interrater agreement on technical performance. The lower reliability for each scenario as measured by ICC compared with the r_{wg} was due primarily to the fact that ICC (but not r_{wg}) compares the variability between raters with the variability between teams. The low variability in technical performance between teams, especially in the cardiac arrest scenario, would tend to overstate the relative disagreement between raters, leading to a lower value of the ICC.

Other groups have also evaluated technical scoring based on videotapes of simulated crises. Chopra *et al.*[11] assessed technical performance in simulations of MH and anaphylaxis for anesthesiologists using a checklist of appropriate actions. However, they used oniy a single rater and could not measure the variability between raters. Devitt *et al.*[25] recently studied the interrater reliability of assessments of technical performance using videotapes of simulated cases managed by an anesthesiologist working alone. Their ratings used a simple three-point scale: 0 = no response; 1 = compensating intervention; and B2 = definitive management. They produced scripted test tapes in which an anesthesiologist-actor responded at each of the three predetermined levels of performance to 10 different clinical problems ranging in difficulty from very simple (bradycardia during peritoneal traction) to very difficult (anaphylaxis). The interrater reliability between two anesthesiologist raters was excellent (kappa = 0.96), whereas the level of agreement of technical scores in our study was not as high. However, this study differed substantially from ours. First, to allow a true assessment of interrater reliability, Devitt *et al.*'s scenarios were acted out to show clearly different levels of performance consistently for the duration of each clinical problem. In addition, in Devitt *et al.*'s study raters could have discerned that each clinical problem was portrayed exactly once at each performance level, making it easier for them to classify the performances. In contrast, we evaluated interrater variability from simulation runs with autonomous participants whose performance varied unpredictably between teams and fluctuated over time for single teams.

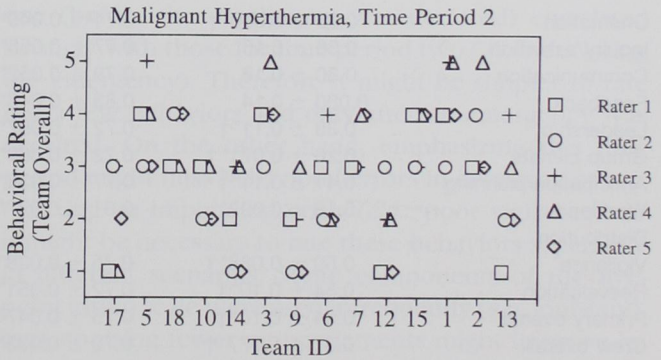

Fig. 4. The ratings of each of five raters for the team overall behavior of all teams during time period 2 of the malignant hyperthermia scenario. Raters are identified by numbers rather than by letters because three of these raters were different than those who performed the technical ratings.

Table 2. Interrater Reliability Statistics for Behavioral Markers: Cardiac Arrest Scenario (± Standard Deviation)

Behavior	Time Period 1			Time Period 2		
	S_{av}	S_{avr}	r_{wg}	S_{av}	S_{avr}	r_{wg}
Orientation	0.46 ± 0.13*†	0.67 ± 0.073*†	0.67	NA	NA	NA
Inquiry/assertion	0.28 ± 0.13*	0.79 ± 0.032*†	0.79	0.36 ± 0.17*	0.78 ± 0.050*†	0.78
Communication	0.31 ± 0.12*	0.77 ± 0.047*†	0.77	0.25 ± 0.13*	0.70 ± 0.058*†	0.70
Feedback	0.00 ± 0.00	0.9 ± 0.036*†	0.90	0.22 ± 0.13	0.78 ± 0.032*†	0.79
Leadership	0.40 ± 0.13*†	0.80 ± 0.029*†	0.80	0.41 ± 0.095*†	0.66 ± 0.059*†	0.66
Group climate	0.39 ± 0.18*	0.81 ± 0.041*†	0.81	0.19 ± 0.11	0.70 ± 0.054*†	0.70
Anticipation/planning	0.55 ± 0.11*†	0.83 ± 0.029*†	0.83	0.34 ± 0.013*†	0.70 ± 0.062*†	0.70
Workload	0.36 ± 0.12*	0.86 ± 0.044*†	0.86	0.30 ± 0.084*†	0.71 ± 0.053*†	0.71
Distribution						
Vigilance	0.29 ± 0.082*†	0.76 ± 0.049*†	0.76	0.21 ± 0.090*	0.65 ± 0.081*†	0.65
Reevaluation	0.30 ± 0.11*	0.75 ± 0.051*†	0.75	0.27 ± 0.085*†	0.68 ± 0.062*†	0.68
Primary overall	0.41 ± 0.18*	0.82 ± 0.030*†	0.82	0.46 ± 0.10*†	0.76 ± 0.039*†	0.76
Crew overall	0.32 ± 0.18*	0.84 ± 0.031*†	0.84	0.49 ± 0.14*†	0.80 ± 0.027*†	0.80

Time Period 1 = prior to the subject's declaration of a cardiac arrest; Time Period 2 = after the declaration; NA = not applicable in this time period; S_{av} and S_{avr} = tests of interrater reliability; r_{wg} = Within Group Interrater Reliability Coefficient.

* $P < 0.05$ (one-tailed).

† $P < 0.005$ (one-tailed).

Our technical scoring system may be a useful component of performance assessment. Persons or teams who do not execute one or more essential actions can be identified clearly as deficient, those who execute only the essential actions can be classified as minimally acceptable, and those who execute many appropriate actions can be distinguished as having performed well. In our study the primary anesthesiologist was able to call in other anesthesiologists for help; the helpers often performed actions that might have been forgotten otherwise (a benefit of teamwork). Yet anesthesiologists often work alone without the availability of skilled help; they may need to manage a crisis without assistance if necessary. In future applications of our rating system,

Table 3. Interrater Reliability Statistics for Behavioral Markers: Malignant Hyperthermia Scenario (± Standard Deviation)

Behavior	Time Period 1			Time Period 2		
	S_{av}	S_{avr}	r_{wg}	S_{av}	S_{avr}	r_{wg}
Orientation	0.35 ± 0.080*†	0.74 ± 0.040*†	0.74	NA	NA	NA
Inquiry/assertion	0.36 ± 0.16*	0.77 ± 0.055*†	0.77	0.19 ± 0.090*	0.64 ± 0.052*†	0.64
Communication	0.30 ± 0.18	0.79 ± 0.052*†	0.79	0.29 ± 0.12*	0.65 ± 0.044*†	0.65
Feedback	0.080 ± 0.14	0.83 ± 0.043*†	0.83	0.15 ± 0.11	0.60 ± 0.070*†	0.60
Leadership	0.39 ± 0.11*†	0.72 ± 0.067*†	0.72	0.42 ± 0.10*†	0.64 ± 0.051*†	0.64
Group climate	0.26 ± 0.071*†	0.72 ± 0.063*†	0.72	0.32 ± 0.088*†	0.73 ± 0.056*†	0.73
Anticipation/planning	0.41 ± 0.11*†	0.77 ± 0.036*†	0.77	0.51 ± 0.10*†	0.72 ± 0.037*†	0.72
Workload	0.18 ± 0.092*	0.81 ± 0.037*†	0.81	0.43 ± 0.13*†	0.66 ± 0.061*†	0.66
Distribution						
Vigilance	0.50 ± 0.088*†	0.75 ± 0.038*†	0.75	0.36 ± 0.12*	0.69 ± 0.045*†	0.69
Reevaluation	0.54 ± 0.10*†	0.72 ± 0.051*†	0.72	0.52 ± 0.086*†	0.75 ± 0.040*†	0.75
Primary overall	0.54 ± 0.10*†	0.75 ± 0.047*†	0.75	0.49 ± 0.098*†	0.66 ± 0.055*†	0.66
Crew overall	0.37 ± 0.077*†	0.79 ± 0.033*†	0.79	0.48 ± 0.063*†	0.66 ± 0.041*†	0.66

Time Period 1 = prior to the subject's declaration of an MH emergency; Time Period 2 = after the declaration; NA = not applicable in this time period; S_{av} and S_{avr} = tests of interrater reliability; r_{wg} = Within Group Interrater Reliability Coefficient.

* $P < 0.05$ (one-tailed).

† $P < 0.005$ (one-tailed).

it might be advisable to assess the technical performance of the anesthesiologist working alone as well as within a team of anesthesiologists.

Crisis management behaviors varied widely among the different primary anesthesiologists and teams, with several rated as minimally acceptable or below in overall behavioral performance. This seems surprising because there were no novice anesthesiologists in the study. However, this finding is consistent with previous observations concerning anesthesiologist performance during simulations, suggesting that cognition and crisis management behaviors vary considerably.[7-10,26] It was not possible in this study to determine whether performance was related to the level of experience, because this was not recorded to help preserve confidentiality. This is an important limitation of the study.

Although no formal scale of "agreement" is defined for the measures of interrater variability we used, Landis and Koch[27] give arbitrary interpretations of different levels of kappa that might be applicable to S_{av} and S_{avr}. For crisis management behaviors, the conservative S_{av} statistic showed "fair" (0.2 to 0.4) to "moderate" (0.4 to 0.6) agreement for the behaviors we consider to be most critical (leadership, workload distribution, primary overall, team overall), and all S_{avr} assessments of interrater variability were highly significant and demonstrated "good" agreement by analogy to the Landis and Koch nomenclature ($S_{avr} > 0.6$). The values of the r_{wg} showed good agreement (on the whole equivalent to 74% less variance than expected by chance).

The behavioral ratings showed greater interrater variability than did the technical ratings. However, given that this study used actual simulation runs, and considering the different statistical measures of interrater agreement, the level of agreement on behavioral ratings was satisfactory. The variability between raters was small enough to allow clear distinctions between particularly poor performances and particularly good performances. We showed that using at least two raters rather than just one can reduce the likelihood of a major error in performance assessment. Much as the American Board of Anesthesiologists oral examination and others evaluations[28] rely on pairs of examiners, performance assessments using simulations should avoid using only a single rater.

The assessment of anesthesiologist crisis management behaviors has not been attempted before. Therefore, how do our results compare with those from commercial aviation? Surprisingly, given that the Line/LOS checklist is used by Federal Aviation Administration check airmen to conduct formal certification evaluations of airline pilots, there has been little research on its interrater reliability. Law and Sherman## conducted a study in which crew resource management (CRM) evaluators-in-training completed the Line/LOS checklist (a revised 1994 version***) for two videotaped simulator runs, one of which was a preselected average/above-average performance (24 raters), the other a preselected below-average performance (10 raters). The authors noted that "many fewer raters rated the same CRM behavioral items for any given phases of flight," which means that reliability could be evaluated only for two overall behaviors rather than for any individual ones. They found good agreement, with a r_{wg} of 0.82 to 1.0 for "overall crew effectiveness" and 0.78 to 0.94 for "technical proficiency" (typically our values of the same statistics were between 0.6 and 0.8). However, this study differed markedly from ours. The raters watched and rated the tapes in a group setting, and the authors stated "it is not known whether or how much discussion occurred among the raters during viewing or rating of these videotapes." Discussion by the raters would almost certainly have increased their level of agreement.

The interrater agreement for behavioral ratings of anesthesiologists can be improved. Ratings of many behaviors were correlated with each other and with the ratings of overall performance. Ratings for time period one (before the emergency was declared) correlated strongly with those for time period two (after declaring the emergency). Therefore it might be simpler to rate just a few behaviors and only after an emergency was declared. On the other hand, emphasizing this time period might miss behavioral errors in detecting or recognizing an impending crisis (*e.g.*, poor vigilance), so it might be necessary to rate these behaviors separately in different scenarios. Some components of the five-point rating scale were used infrequently (*e.g.*, outstanding), so using fewer rating elements might improve interrater agreement (the latest version of the NASA Line/LOS Checklist has reduced the number of rating points to four***).

Debriefings of the behavioral raters revealed one com-

*** Helmreich RL, Butler RE, Taggart WR, Wilhelm JA: The NASA/University of Texas/FAA Line/LOS Checklist: A behavioral marker based checklist for CRM skills assessment. Aerospace Crew Research Project Technical Paper 94-02 (revised 12/8/95). (Also available on the World Wide Web at: www.psy.utexas.edu/psy/helmreich/llcinst1.htm)

mon difficulty. Even a 15-min period contains many behaviors with rapidly fluctuating levels of performance. For example, the team might be communicating well at one instant and then be observed mumbling queries "into thin air" that were unheeded a few minutes later. In assigning an aggregate rating for the 15-min period, one rater might happen to focus on the moments of good performance while another might focus on the moments of poor performance. This problem of "aggregation over time" is not seen when raters score "exemplar" test tapes showing uniform performance for an entire scenario (as in Devitt *et al.*'s study[25]). This problem can be addressed in various ways, including (1) scoring scenario-specific behaviors as well as scenario-specific technical actions; (2) developing specific conventions for aggregating scores over time; (3) counting the number of occurrences of especially poor or good behaviors; (4) performing ratings continuously using a joystick on a linear analog scale (oral communication, Yan Xiao, University of Maryland at Baltimore, May 15, 1997) with aggregation *via* root-mean-square or other mathematical transformations; and (5) developing complex scenarios in which only sustained good behavioral performance could allow the team to accomplish the critical technical tasks.

Allowing discussion by the raters could result in better agreement and could allow for multiple perspectives on performance. This approach could also ameliorate the problem with aggregation over time because raters could discuss which occurrences were important enough to affect the overall assessment. On the other hand, with this technique a forceful rater could dominate less assertive ones.

This experiment shows that a perfect scoring system for complex clinical behaviors and technical performance during the management of perioperative crises is difficult to achieve. In addition, because there is no standard measurement of clinical performance in actual practice, we do not know how well performance in the simulator predicts performance during real crises. The performance assessment tools we describe might be applied to research on crisis management and as a component of performance evaluation of trainees or of practitioners undergoing remedial training. Further refinements of this approach are also likely as the use of simulators for training and evaluation increases. Our study suggests that achieving equitable and meaningful evaluations is likely to require using multiple raters, assessing both behavioral and technical performance, and providing several opportunities to be evaluated. The use of a similar performance assessment system for formal certification or credentialing must await further refinement and testing.

Future studies of performance assessment tools might include the following: (1) a multicenter trial with a larger number of participants; (2) teams of known and widely differing levels of experience; (3) more tightly controlled simulation scenarios; (4) the use of both catastrophic (*e.g.*, MH and cardiac arrest) and noncatastrophic clinical scenarios (*e.g.*, atelectasis, bronchospasm); (5) evaluation of the effects of team size and composition and of the presence or absence of skilled help; and (6) evaluation of intrarater variability of repeat assessments of the same tape.

Statistical Appendix: Interrater Reliability Statistics

The S_{av} Measure of Interrater Reliability and its Variant, S_{avr}

S_{av} is a general observer-agreement measure that can be applied to individuals and groups. S_{av} is a general form that reduces to other well-known measures of interrater reliability, such as kappa and weighted kappa when appropriate constraints are applied.[21,22] S_{av} was originally developed by O'Connell and Dobson[21] and used by Posner *et al.*[22] in the American Society of Anesthesiologists Closed Claims Study analysis. We wrote software in the APL language (APL.68000 for the Apple Macintosh; MicroAPL, London, UK) to calculate S_{av} and related statistics and tested our software against the "Weighted S_{av} Program" software written by John Reed and provided to us by Posner et al.

To calculate S_{av}, one first calculates a set of S_i—the agreement between N raters rating a single subject, i, as follows:

$S_i = 1 - (F(d, rj, rj') + E(d, rj, rj', phi, phi'))$ (Equation A1), where:

F = a function that sums all actual disagreements between the ratings of pairs of raters j & j'

d = the metric of disagreement (see below)

rj & rj' = the ratings of raters j & j'

E = a function that sums the disagreement expected by chance between all pairs of raters j & j'

phi, phi' = the marginal distributions of rater j & j' respectively (the marginal distribution is the underlying probabilities with which each observer uses the rating scale categories).

The S_i for all subjects in a block of data can be averaged to calculate the average agreement for that block of data: S_{av}. The general form of S_{av} reduces to other tests depending in part on the metric of disagreement d used. For example, the standard "kappa statistic" which is appropriate for two raters and a nominal scale, uses a metric of:

d = {1 if rj is equal to rj'}

{0 if rj is not equal to rj'}

In our study a five-point ordinal scale was used. We used the following metric which is sensitive to the extent of disagreement between raters:

$d = (rj - rj')^2$

Given the chosen metric, we can compute the sum total of dis-

agreement to be expected by chance from the N raters for a single case: E(d, rj, rj′, phi, phi′). The definition of the term *by chance* is critical to understanding the statistic S_{av}, in that the function E uses the actual marginal distributions for raters j and j′.

In some cases S_{av} will be markedly conservative, underestimating the level of agreement relative to intuitive notions of "agreement" and "chance." This problem has also been discussed by Posner *et al.*[22] in terms of the assessment of agreement of raters in the ASA Closed Claims Study: ". . . the high level of *expected agreement* [*i.e.*, low disagreement expected by chance] in cases of . . ." poor care but severe injury" limits the possible agreement beyond chance that we may observe. This leads to small values of S_{av} . . . Thus, a conclusion of poor reliability in such cases may be unfounded." This is similar to the "base rate" problem for kappa as discussed by Spitznagel and Helzer[29] (who studied interrater reliability of psychiatric diagnoses) in which agreement is lessened as the rate of a condition in the population is lessened.

To deal with this phenomenon, we constructed a variant of S_{av}, which we term call S_{avr}, in which the function E is constructed assuming that "by chance" means that each rater would select each rating purely at random, as in drawing a number from a hat. Note that the same assumption is made for some other interrater reliability statistics (such as the r_{wg}). Both S_{av} and S_{avr} can range from a negative value (maximum disagreement, even greater than that expected by chance) to 0.0 for agreement at the level of chance to 1.0 for perfect agreement. Although S_{av} has been shown to be equivalent to intraclass correlation, this is not true for the variant statistic S_{avr}.

The Within-group Interrater Reliability Coefficient

For a single case i, the r_{wg} statistic = $1 - V_i^2/\rho_E^2$, where V_i^2 is the variance of the ratings for case i, and ρ_E^2 is the expected variance for the rating scale if the ratings are chosen purely at random (ρ_E^2 is calculated to equal 2.0 for five raters using all elements of a five-point ordinal scale). The notation V_i^2 for the variance is used instead of the traditional "S_x^2" to avoid confusion with the "S" values for the S_{av} and S_{avr} statistics. Just as S_{av} and S_{avr} are the averages of individual group S_i across the 14 groups rated, we calculated (using software written in APL) and report the average r_{wg} across all 14 groups.

The S_{avr} statistic and r_{wg} thus compare agreement relative to pure chance. Thus S_{av} might be considered a conservative lower bound for agreement, whereas S_{avr} and r_{wg} might be considered upper bounds. In deriving and presenting S_{av}, S_{avr}, and r_{wg}, we hope to categorize the level of agreement more completely and to be able to compare our results with other similar studies.

The Jackknife Standard Deviation of S_{av} and S_{avr}

There is no closed analytical form for the standard deviation of the statistics S_{av} and S_{avr}. Therefore, to calculate this measure of dispersion for purposes of testing the hypothesis that S_{av} or S_{avr} differs from zero, the jackknife procedure is used.[22] The jackknife procedure is one of the "resampling statistics." Briefly, ". . .jackknifed statistics are developed by systematically dropping out subsets of the data one at a time and assessing the variation in the [function] that results."[30] We used the formula for the jackknife estimate of the standard deviation as given in Efron[31] and checked our program (in APL) with test data against results of the Weighted S_{av} Program.

Interrater Reliability Statistics for Interval Data

The r_{wg} just described also can be used for interval data. When applied to three raters using a numerical scale ranging from 0 to 1.0, ρ_E^2 for r_{wg} is calculated (using a Monte Carlo technique) to equal 0.085. Because the technical scores in this experiment ranged from 0.6 to 0.94, we also computed ρ_E^2 for r_{wg} as 0.0096 when restricting the comparison random distribution to come from that range. We report r_{wg} values for both assumptions.

The ICC uses one-way analysis of variance to compare between-group and within-group variance: ICC = (groups MS − error MS) + (groups MS + error MS).[23,32] We programmed this in Microsoft Excel 4.0 (Redmond, WA) based on the formulas in a standard statistical reference[23] and tested our program against a worked example. The ICC is strongly affected (reduced) when there are between-group similarities, whereas r_{wg} is independent of between-group similarities.[32] Thus, because the teams studied here showed substantial similarities in technical performance, the ICCs are expected to be markedly lower than the r_{wg}. The ICC and r_{wg} have been used as complementary measures of interrater reliability for interval data in other studies.[32]

The authors thank Professor Robert Helmreich of the NASA/University of Texas Aerospace Crew Performance Project for providing the Line/LOS Checklist materials and that project's research publications; and the organizers, instructors, staff, and participants in the Simulator Instructor Training Project for conducting the simulation sessions.

References

1. Slogoff S, Hughes FP, Hug CC, Longnecker DE, Saidman LJ: A demonstration of validity for certification by the American Board of Anesthesiology. Acad Med 1994; 69:740–6

2. Gaba DM, DeAnda A: A comprehensive anesthesia simulation environment: Re-creating the operating room for research and training. ANESTHESIOLOGY 1988; 69:387–94

3. Grand'Maison P, Brailovsky CA, Lescop J, Rainsberry P: Using standardized patients in licensing/certification examinations: Comparison of two tests in Canada. Fam Med 1997; 29:27–32

4. Beullens J, Rethans JJ, Goedhuys J, Buntinx F: The use of standardized patients in research in general practice. Fam Pract 1997; 14:58–62

5. Brennan RT, Braslow A, Batcheller AM, Kaye W: A reliable and valid method for evaluating cardiopulmonary resuscitation training outcomes. Resuscitation 1996; 32:85–93

6. Quiney NF, Gardner J, Brampton W: Resuscitation skills among anaesthetists. Resuscitation 1995; 29:215–18

7. Gaba DM, DeAnda A: The response of anesthesia trainees to simulated critical incidents. Anesth Analg 1989; 68:444–51

8. DeAnda A, Gaba DM: Unplanned incidents during comprehensive anesthesia simulation. Anesth Analg 1990; 71:77–82

9. DeAnda A, Gaba DM: The role of experience in the response to simulated critical incidents. Anesth Analg 1991; 72:308–15

10. Howard SK, Gaba DM, Fish KJ, Yang GS, Sarnquist FH: Anesthesia crisis resource management training: Teaching anesthesiologists to handle critical incidents. Aviat Space Environ Med 1992; 63:763–70

11. Chopra V, Gesink BJ, De Jong J, Bovill JG, Spierdijk J, Brand R: Does training on an anaesthesia simulator lead to improvement in performance? Br J Anaesth 1994; 73:293–7

12. Billings CE, Reynard WD: Human factors in aircraft incidents: Results of a 7-year study. Aviat Space Environ Med 1984; 55:960–5

13. Gaba DM, Maxwell M, DeAnda A: Anesthetic mishaps: Breaking the chain of accident evolution. ANESTHESIOLOGY 1987; 66:670–6

14. Gaba DM: Dynamic decision-making in anesthesiology: Cognitive models and training approaches, Advanced Models of Cognition for Medical Training and Practice. Edited by DA Evans, VL Patel. Berlin, Springer-Verlag, 1992, pp 122–47

15. Jensen RS, Biegelski CS: Cockpit resource management, Aviation Psychology. Edited by RS Jensen. Aldershot, Gower Technical, 1989, pp 176–209

16. Helmreich RL, Foushee HC: Why crew resource management? Cockpit Resource Management. Edited by EL Weiner, BG Kanki, RL Helmreich. San Diego, Academic Press, 1993, pp 3–46

17. Helmreich RL, Schaefer HG: Team performance in the operating room, Human Error in Medicine. Edited by MS Bogner. Hillsdale, NJ, Lawrence Erlbaum Associates, 1994, pp 225–53

18. Gaba DM, Fish KJ, Howard SK: Crisis Management in Anesthesiology. New York, Churchill-Livingstone, 1994

19. Holzman RS, Cooper JB, Gaba DM, Philip JH, Small S, Feinstein D: Anesthesia crisis resource management: Real-life simulation training in operating room crises. J Clin Anesthesia 1995; 7:675–87

20. Kurrek MM, Fish KJ: Anaesthesia crisis resource management training: An intimidating concept, a rewarding experience. Can J Anaesth 1996; 43:430–4

21. O'Connell DL, Dobson AJ: General observer-agreement measures on individual subjects and groups of subjects. Biometrics 1984; 40:973–83

22. Posner KL, Sampson PD, Caplan RA, Ward RJ, Cheney FW: Measuring interrater reliability among multiple raters: An example of methods for nominal data. Stat Med 1990; 9:1103–15

23. Zar JH: Biostatistical Analysis. Englewood Cliffs, NJ, Prentice-Hall, 1984

24. James LR, Demaree RG, Wolfe G: r_{wg}: An assessment of within-group interrater agreement. J Appl Psych 1993; 78:306–9

25. Devitt JH, Kurrek MM, Cohen MM, Fish K, Fish P, Murphy PM, Szalai JP: Testing the raters: Interrater reliability of standardized anaesthesia simulator performance. Can J Anaesth 1997; 44:924–8

26. Schwid HA, O'Donnell D: Anesthesiologists' management of simulated critical incidents. ANESTHESIOLOGY 1992; 76:495–501

27. Landis RJ, Koch GG: The measurement of observer agreement for categorical data. Biometrics 1977; 33:159–74

28. Burchard KW, Rowland-Morin PA, Coe NPW, Garb JL: A surgery oral examination: Interrater agreement and the influence of rater characteristics. Acad Med 1995; 70:1044–6

29. Spitznagel EL, Helzer JE: A proposed solution to the base rate problem in the kappa statistic. Arch Gen Psychiatry 1985; 42:725–8

30. Mooney CZ, Duval RD: Bootstrapping: A nonparametric approach to statistical inference (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-095). Newbury Park, CA, Sage Publications, 1993

31. Efron B: The Jacknife, the Bootstrap and Other Resampling Plans. Philadelphia, Society for Industrial and Applied Mathematics, 1982

32. Edmondson AC: Learning from mistakes is easier said than done: Group and organizational influences on the detection and correction of human error. J Applied Behav Sci 1996; 32:5–28