Anesthesiology 2006; 105:260–6

Development of an Objective Scoring System for Measurement of Resident Performance on the Human Patient Simulator

Barbara M. Scavone, M.D., Michele T. Sproviero, M.D.,† Robert J. McCarthy, Pharm.D.,‡ Cynthia A. Wong, M.D.,‡ John T. Sullivan, M.D.,* Viva J. Siddall, M.S.,* Leonard D. Wade, M.S.**

Background: **The decrease in the percentage of patients having cesarean delivery during general anesthesia has led some educators to advocate the increased use of simulation-based training for this anesthetic. The authors developed a scoring system to measure resident performance of this anesthetic on the human patient simulator and subjected the system to tests of validity and reliability.**

Methods: **A modified Delphi technique was used to achieve a consensus among several experts regarding a standardized scoring system for evaluating resident performance of general anesthesia for emergency cesarean delivery on the human patient simulator. Eight third-year and eight first-year anesthesiology residents performed the scenario and were videotaped and scored by four attending obstetric anesthesiologists.**

Results: **Third-year residents scored an average of 150.5 points, whereas first-year residents scored an average of 128 points ($P = 0.004$). The scoring instrument demonstrated high interrater reliability with an intraclass correlation coefficient of 0.97 (95% confidence interval, 0.94–0.99) compared with the average score.**

Conclusions: **The developed scoring tool to measure resident performance of general anesthesia for emergency cesarean delivery on the patient simulator seems both valid and reliable in the context in which it was tested. This scoring system may prove useful for future studies such as those investigating the effect of simulator training on objective assessment of resident performance.**

THE percentage of patients having cesarean delivery during general anesthesia has decreased, and this decline may result in an inadequate training experience for the current generation of anesthesiology residents.[1] At Prentice Women's Hospital in Chicago, Illinois, where approximately 10,000 deliveries take place each year, during 2004, only 81 of 2,384 cesarean deliveries (3.4%) were performed using general anesthesia. This translates to a median of 5 such cases per anesthesiology resident during the 3-yr training period in anesthesiology. In fact, the members of the anesthesiology resident graduating class of 2005 at our program encountered a range of 0–11 of this type of case during their residency. Inadequate proficiency regarding this anesthetic procedure may affect patient safety because the case fatality risk ratio of cesarean delivery during general anesthesia is 16.7 times that of cesarean delivery with regional anesthesia.[2] Therefore, educators have recommended the increased use of surrogate training modalities, such as simulation-based training.[3]

Although there has been a fair amount of material published on reliability and validity of assessing resident performance on the simulator, this has yet to be correlated with the ability to perform safe clinical care outside of the simulator environment. Optimally, these assessments should be made with an objective scoring system for measuring resident performance. To this end, we have developed an objective scoring system for measuring resident performance of general anesthesia for emergency cesarean delivery for use with a human patient simulator. We used a modified Delphi technique to create a consensus among several experts regarding a standardized scoring instrument to objectively evaluate resident performance of this simulated obstetric anesthetic emergency.[4] We then subjected the standardized scoring instrument to tests of validity and reliability. We hypothesized that the scoring tool would yield higher scores for more experienced, more able residents (validity) and that minimal variability in scores would exist among raters (reliability).

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

Materials and Methods

The Northwestern University Office for the Protection of Research Subjects Institutional Review Board, Chicago, Illinois, approved this study.

The Modified Delphi Process

The modified Delphi technique has been used as a means of obtaining a consensus on a subject among several experts.[4] It involves distributing a list of tasks, in this case, those tasks deemed necessary to properly perform general anesthesia for cesarean delivery, to several recognized experts (traditionally at least six). The list of tasks was generated by consensus among the obstetric anesthesiology faculty at the authors' institution. The experts were asked to rank each task according

* Assistant Professor of Anesthesiology, † Instructor in Anesthesiology, ‡ Associate Professor of Anesthesiology.

Received from the Department of Anesthesiology, Northwestern University Feinberg School of Medicine, Chicago, Illinois. Submitted for publication January 10, 2006. Accepted for publication April 3, 2006. Supported by the Eleanor Wood-Prince Grant Initiative: A Project of The Woman's Board of Northwestern Memorial Hospital, Chicago, Illinois. Awarded November 2003. Presented at the Annual Meeting of the Society for Obstetric Anesthesia and Perinatology, Palm Desert, California, May 4–7, 2005, and the Annual Meeting of the American Society of Anesthesiologists, Atlanta, Georgia, October 22–26, 2005.

Address correspondence to Dr. Scavone: Department of Anesthesiology, Northwestern University, Feinberg School of Medicine, 251 East Huron Street F5-704, Chicago, Illinois 60611. bscavone@nmff.org.

to importance on a five-point Likert scale (1 = not important and 5 = extremely important). Each expert has the opportunity to suggest the elimination or addition of tasks to the list and to make pertinent comments (round 1). The information is gathered by the investigator, and median scores and ranges for each task are calculated. These data and experts' comments are distributed, and each expert is afforded the opportunity to change any of his or her scores that deviate from the median. They are also given an opportunity to explain their reasoning if they do not wish to change their scores (round 2). Once again, the data are collected, medians and ranges are tabulated, and the information is forwarded to the experts, who may again change scores that deviate from the median (round 3). The process is repeated until an acceptably small level of variation is present. The end result is a list of tasks deemed necessary for proper performance of cesarean delivery during general anesthesia, each with a weight of importance, 1–5. When a resident performs the general anesthesia for cesarean delivery scenario on the simulator, each task is checked off as it is performed and multiplied by its weighting factor, and a sum of all the weighted scores provides a total score.

The experts in this case were six nationally recognized obstetric anesthesiologists who are members of the Society for Obstetric Anesthesia and Perinatology. Several geographic areas of the United States were represented; specifically, the East Coast, Midwest, Southwest, South, and West Coast. Two of the six are in private practice, and the others practice in university settings. The experts were blinded to each other's identities.

Simulator

The study was conducted in the Northwestern University Patient Safety Simulator Center, a simulated operating room environment which uses a high-fidelity life-sized human patient computerized mannequin manufactured by Medical Education Technologies, Inc. (Sarasota, FL) with HPS 6.1 software. The simulator software includes the program "Stanette," which mimics the physiology of normal pregnancy. A computer technician, who is seated behind one-way glass, inputs commands and manages the mechanical interfaces. The mannequin's voice is a speaker located inside the mannequin, with a microphone in the computer technician's control room behind the one-way glass. There is capability for communication between the simulated operating room and the control room *via* speakers. Three video cameras in the simulated operating room are controlled by the computer technician.

The Scenario

The scenario was developed for resident training purposes by the obstetric anesthesiology faculty for resident training purposes. Before the session, the resident is oriented to the simulator by the simulator technician. The resident anesthesiologist is paged to the simulated operating room for an emergency cesarean delivery ("start" time). In the room are a "surgeon" (actually a member of the department of anesthesiology) and a "circulating nurse" who acts as an assistant to the anesthesia resident (actually a member of the staff of the simulator center). The resident may ask questions of the obstetric surgeon, the nurse, or the patient, and those questions are answered (by the actors or, for the "patient," by operators in the computer control room). The resident is given information only if he or she asks for it. The obstetric history is that the patient has an umbilical cord prolapse with fetal heart tones present at 60 beats/min. The patient has no significant medical problems, is taking no medications, has no allergies, and has not had any food or drink by mouth for 6 h. Neither the patient nor her family members have ever had an adverse event associated with any type of anesthetic. If the resident performs an airway examination or inquires about the airway, he or she is informed that the patient has a normal Mallampati class I airway.

The resident chooses how to monitor the patient, how to induce general anesthesia and how to secure the airway. Immediately after induction of general anesthesia and muscle paralysis, the patient's oxygen saturation decreases mildly, as is expected of most obstetric patients. When the airway is secured and oxygenation and ventilation are begun, the patient returns to baseline. Immediately after the resident secures the airway, he or she should either tell the surgeon that he or she may proceed, or inquire as to whether he or she is proceeding. (Either is considered acceptable for credit on the checklist.) If the resident does not do this within 30 s of securing the airway, he or she will be prompted by the obstetric surgeon, who will inform the resident that surgery is beginning. (If the resident has to be prompted, he or she does not get credit for telling the surgeon to begin.) "Incision" time occurs when the resident tells the surgeons to begin or, in the case that the resident has to be prompted, when the resident is informed that surgery is beginning. The resident chooses how to provide maintenance anesthesia. Three minutes after the onset of surgery, the resident is told that the baby has been delivered, the cord has been clamped, and the placenta has been delivered. Two minutes later, if the resident has turned down the inhalational agent (commonly done to minimize obstetric hemorrhage due to relaxation effect of inhalational agents) but has not given additional anesthetic medications, the patient develops tachycardia, thus prompting the resident to give additional medications. An additional 3 min later, the resident is informed that the scenario is over.

Validity and Reliability

Eight third-year anesthesiology (CA-3) residents with extensive obstetric anesthesia experience and eight first-

year anesthesiology (CA-1) residents with little or no obstetric anesthesia experience gave informed consent to participate in the study. Note was made of how many times each resident had worked on the simulator during their residency from records that are kept by the Department of Anesthesiology, as well as the number of actual cesarean deliveries during general anesthesia that had been performed by each resident during their training to that point from a Department of Anesthesiology database of obstetric anesthesia cases. Before the simulation, each resident was asked to assess his or her own confidence level performing general anesthesia for emergency cesarean delivery, reported on a visual analog scale (100-mm line with endpoints labeled "not at all confident" and "extremely confident").

Each resident then performed the scenario and was videotaped. Each videotape was viewed and scored, by means of the weighted checklist, by four different attending obstetric anesthesiologists. The time interval from start until simulated surgical incision was recorded by the simulator computer technician, who made note of these events on the simulator computer, which has an internal clock. In an attempt to blind raters to resident identity, residents wore hoods, surgical masks, goggles, surgical gowns, and gloves. Raters viewed and scored the videotapes separately from each other and were blinded to each other's score. Study participants were asked not to discuss the study or the simulator scenario with each other or other residents.

Statistics

Agreement among the panel of experts participating in the modified Delphi process regarding weighting of tasks was assessed using the Kendall W statistic after each round. An overall concordance of 0.75 was required for acceptance of the scoring system. The content validity index of the final scoring system was determined by calculating the percentage of total items rated by the experts as either 4 or 5.

The number of previous simulation sessions, number of actual cesarean deliveries done during general anesthesia, and subject self-assessment of confidence were compared between CA-3 and CA-1 residents. The overall scores as well as the start to incision time intervals were also compared between groups to assess the validity of the scoring system. Nonparametric data (number of previous simulator sessions, number of actual cesarean deliveries done with general anesthesia, and confidence) were compared using the Mann–Whitney U test. Parametric data (scores and time intervals) were compared using a two-tailed t test. $P < 0.05$ was required to reject the null hypothesis.

Interrater reliability among the four scorers was assessed by the intraclass correlation coefficient. Factor analysis using principal component and varimax factor loading was used to determine the number of raters needed to explain more than 95% of the variation among the raters.

It was estimated that a sample size of 16 subjects with 4 observations per subject would achieve 80% power to detect an intraclass correlation of 0.85 under the alternative hypothesis assuming the intraclass correlation under the null hypothesis is 0.65 using an F test with a significance level of 0.05.

Results

The Modified Delphi Process

The list of tasks given to the panel of experts, medians, and interquartile ranges after the first and second rounds, and the final weighted scoring system are presented in table 1. One task was added to the system after the first round (use of an "appropriate tidal volume and rate" for mechanical ventilation). One panel member suggested eliminating both cycling the blood pressure cuff and placing the oral gastric tube; however, the other five panel members thought these tasks were important, and they were retained. Kendall W was 0.60 after one round and increased to the target concordance of 0.75 after two rounds. After two rounds, the scoring system included 51 observable tasks, each weighted by importance 1–5, for a total possible score of 207.5.

Several modifications were made to the scoring system when it was subjected to tests of validity and reliability. First, with use, it became evident that not all patients required nitrous oxide, opioids, hypnotics, and muscle relaxants. Therefore, the last four items on the checklist were combined into one category ("administer nitrous oxide, opioids, hypnotics, muscle relaxants as needed"), allowing the resident to choose from among these classes of medications. This task was assigned a weight of 3, which was equal to the weight assigned to each of these tasks individually. Second, a simulator malfunction prevented routine temperature monitoring, so this category was eliminated from the final scoring system. Therefore, the final weighted scoring system consisted of 47 tasks, each weighted by importance, for a total possible score of 196.5 points. The content validity index for the final scoring tool was 72%.

Validity and Reliability

Eight CA-3 and eight CA-1 residents performed the scenario, and four obstetric anesthesiology attending physicians independently viewed and scored each of the 16 videotapes. Because the simulator laboratory had been a recent addition to our department, there was no difference in the number of times that CA-3 *versus* CA-1 residents had worked on the simulator (table 2). CA-3 residents had more experience doing actual cesarean deliveries with general anesthesia, and they reported higher confidence levels doing the procedure. CA-3 res-

Table 1. Lists of Tasks Identified for Performance of General Anesthesia for Emergency Cesarean Delivery

	Median (Interquartile Range) after Round 1	Median (Interquartile Range) after Round 2	Final Weighted Scoring System*
Preoperative assessment			
Introduce self	3 (3–4)	3 (3)	3
Obtain pertinent obstetric history	5 (4–5)	4.5 (4–5)	4.5
Medical history	4 (3–4)	4 (4)	4
Medications	4 (4)	4 (4)	4
Allergies	5 (4–5)	5 (5)	5
Previous anesthetic/family anesthetic history	4 (4–5)	4 (4–5)	4
Airway examination	5 (5)	5 (5)	5
Preoperative patient care			
Administer sodium citrate	4 (3–4)	4 (3–4)	4
Administer 100% oxygen by mask	5 (4–5)	5 (5)	5
Left uterine displacement	5 (4–5)	5 (5)	5
Ensure working intravenous catheter	5 (5)	5 (5)	5
Apply blood pressure cuff	5 (3–5)	5 (3–5)	5
Apply pulse oximeter	5 (5)	5 (5)	5
Apply electrocardiogram leads	3 (3–5)	3 (3–5)	3
Equipment availability check			
Quick circuit check	5 (3–5)	5 (4–5)	5
Endotracheal tube	5 (4–5)	4.5 (4–5)	4.5
Syringe	4 (4–5)	4 (4–5)	4
Stylet	4 (4–5)	4 (4)	4
Laryngoscope with functional light	5 (4–5)	5 (5)	5
Functional suction	5 (4–5)	5 (4–5)	5
Induction/intubation			
Pulse oximeter audible	5 (3–5)	5 (3–5)	5
Blood pressure monitor cycling	3 (2–4)	3 (2–3)	3
Electrocardiogram functioning	3 (3–5)	3 (3–4)	3
Verify obstetric team readiness	5 (4–5)	5 (4–5)	5
Apply cricoid pressure	5 (4–5)	5 (4–5)	5
Administer induction agent	5 (5)	5 (5)	5
Administer succinylcholine	5 (5)	5 (5)	5
Wait for medication effect	5 (5)	5 (5)	5
Direct laryngoscopy	5 (3–5)	5 (5)	5
Pass endotracheal tube	5 (5)	5 (5)	5
Inflate cuff	5 (5)	5 (5)	5
Confirm presence of end-tidal carbon dioxide	5 (5)	5 (5)	5
Notify obstetrician to proceed	5 (4–5)	5 (4–5)	5
Release cricoid pressure	4 (2–4)	3.5 (3–4)	3.5
Confirm bilateral breath sounds	4 (3–4)	4 (3–4)	4
Secure endotracheal tube	4 (3–4)	3.5 (3–4)	3.5
Intraoperative management			
Before delivery			
Initiate mechanical ventilation	5 (4–5)	5 (5)	5
Appropriate tidal volume/respiratory rate†		3 (3)	3
Maintain $F_{IO_2} \geq 0.5$	4 (3–4)	4 (3–4)	4
Maintain inhaled agent ≥ 1 MAC	4 (3–4)	3.5 (3–4)	3.5
Protect eyes	4 (3–4)	3.5 (3–4)	3.5
Orogastric tube placed and suctioned	2 (1–3)	2 (1–2)	2
Esophageal stethoscope placed	2 (1–2)	2 (1–2)	2
Temperature monitored	2 (1–2)	2 (1–2)	Eliminated‡
Peripheral nerve stimulator placed	3 (2–3)	2.5 (2–3)	2.5
After delivery			
Oxytocin added to intravenous fluids	4 (3–5)	4 (3–5)	4
Decrease inhaled agent ≤ 0.5 MAC	3 (2–3)	3 (2–3)	3
Administer nitrous oxide	3 (2–4)	3 (2–3)	3§
Administer opioid as needed	3 (3–4)	3 (3)	
Administer hypnotic as needed	3 (3–4)	3 (3)	
Administer muscle relaxant as needed	3 (2–3)	3 (2–3)	
Total		207.5	196.5

Final score calculated from the sum of weights of each task properly performed by the resident.

* Final weighted scoring system used for validation. † New task added after round 1. ‡ Eliminated because of simulator malfunction. § Final four categories combined into one category.

F_{IO_2} = fraction inspired oxygen concentration; MAC = minimal alveolar concentration.

Table 2. Comparison of Performance during Cesarean Delivery Scenario by Resident Training Level

	CA-3	CA-1	*P* Value
Previous simulator sessions, n	3 (1–5)	2 (1–3)	NS
Previous experience providing general anesthesia for cesarean delivery, n	3.5 (0–5)	0 (0–0)	0.001
Prescenario self-assessment of confidence, 0–100 mm	67 (25–75)	2.5 (0–5)	0.001
Total score	151 ± 14	128 ± 13	0.004
Time interval from start of scenario to surgical incision, s	198 ± 24	380 ± 41	0.002

Data are expressed as median (range) or mean ± SD.

CA = clinical anesthesia year; NS = not significant.

idents scored an average of 150.5 of a total possible 196.5 points (77% of total possible; 95% confidence interval, 71–83%), whereas CA-1 residents scored an average of 128 of a total possible 196.5 points (65% of total possible; 95% confidence interval, 60–70%) (P = 0.004). Time interval from start to incision averaged 198 s for CA-3 *versus* 380 s for CA-1 residents (P = 0.002).

The scoring instrument demonstrated a high interrater reliability among the four raters with an intraclass correlation coefficient of 0.97 (95% confidence interval, 0.94–0.99) compared with the average score (fig. 1). Factor analysis showed that 96% of the variance from the mean score was explained by scores from two attending anesthesiologists, indicating the mean scores of two raters agreed with the mean scores of all four raters 96% of the time (fig. 2).

Discussion

This study demonstrates that it is possible to reach a consensus regarding the steps necessary for proper performance of a general anesthetic for emergency cesarean delivery. The weighted scoring system obtained through such a consensus seems to be both valid and reliable. To the authors' knowledge, this is the first description of a simulator scenario involving general anesthesia specifically for emergency cesarean delivery and the first attempt to develop a valid and reliable scoring instrument to measure resident performance of this specific anesthetic.

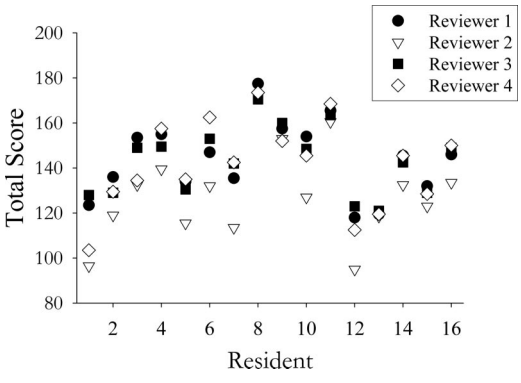

Fig. 1. Total scores achieved by resident trainees as evaluated by each of four reviewers.

Simulation-based training has been advocated for teaching and training of medical students and residents, especially for procedures that are less common, that residents may not gain adequate exposure to during the course of their residencies. Because providing general anesthesia for emergency cesarean delivery is a relatively high-risk procedure, to which contemporary residents may not gain adequate exposure, it was felt that this scenario provided necessary exposure to a relatively uncommon anesthetic technique, as anesthesiology educators have recommended.[3] Also, simulation has been suggested as a tool for evaluation of residents in competency-based curricula. Therefore, simulation has become more common in many teaching programs and is the subject of increasing numbers of research studies. A review of simulation research in anesthesia published in 2001 found that only 4 of 13 studies describing assessment of performance investigated the validity and reliability of the assessment systems, whereas a similar review published in 2004 described several additional studies investigating validity and/or reliability of assessment systems.[5,6]

Among the different types of validity, there exist face validity, content validity, and construct validity.[7] Face validity refers to whether the instrument seems as though it is measuring the appropriate construct and relies heavily on how realistic the simulator, scenario, and environment are to the participants. Modern era high-fidelity simulators allow for large degrees of face

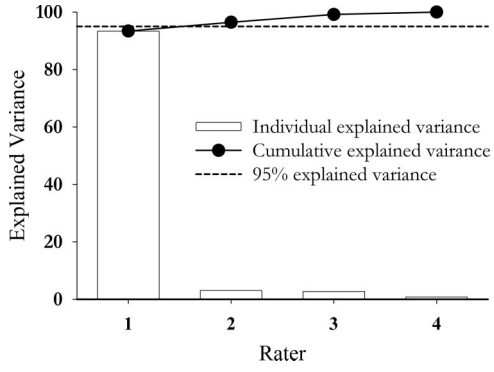

Fig. 2. Explained variance from the mean of the reviewers' evaluations as determined by factor analysis using principal component and varimax factor loading. Unexplained variance is decreased to less than 5% with two reviewers, which should be sufficient when using this scoring system.

validity. In two recent studies, participants rated the realism of simulator experience highly (7.8 on a 10-cm realism visual analog scale in one study, and 3.47 out of a possible 4 points for realism in the other study).[8,9]

Content validity refers to how representative the test is of learning objectives or, stated another way, whether the test measures skills important and valuable to learning the construct. Content validity is based on judgment, and it is common to use a panel of experts in a given content area when there is no established standard to evaluate new instruments. Because we used a modified Delphi technique to gain input from multiple experts, it is likely that our weighted scoring system represents an acceptable model for providing general anesthesia for emergency cesarean delivery and represents learning objectives valued by the anesthetic community as a whole. Indeed, after only two rounds, we were able to achieve the desired level of agreement in performance of tasks and their weighting: This may indicate that a substantial amount of agreement exists within the anesthetic community regarding performance of this particular anesthetic.

Assessment of our final scoring instrument yielded a content validity index of 72%, which is below the 80% considered to represent good content validity. Examination of the tasks revealed that those that did not have a consensus weight of at least 4 were primarily in the subcategory of intraoperative management. Although it is uncertain, lower importance assigned to some of these tasks may be a result of interindividual and interinstitutional practice differences. Furthermore, although some of the tasks in this subcategory are commonly performed (*e.g.*, orogastric tube placement, esophageal stethoscope placement), there are limited data regarding their importance. Elimination of some of these lesser importance checklist items may increase the content validity of the tool.

Construct validity represents how well the test measures what it claims to be measuring. It can be measured by how well the test results meet expectations of performance. As a means of testing construct validity of the weighted scoring system, we compared scores and start to incision time intervals between the two resident groups, an application of the known-groups technique for evaluating construct validity. For this to represent a test of construct validity, an assumption is made that CA-3 residents are indeed more knowledgeable and experienced, and more able to properly perform this particular anesthetic. There is evidence to support this assumption: Not only did the CA-3 residents have 2 yr more anesthesiology training than the CA-1 residents, they had more experience with this particular anesthetic and were appropriately more self-confident. Because there was no difference in the number of times the CA-3 *versus* the CA-1 residents had worked on the simulator, we were not likely measuring differences in simulator experience or ability to work on the simulator, but were seeing a reflection of differences in knowledge, clinical experience, and abilities *per se*. The CA-3 residents had both higher scores and shorter start to incision time intervals, suggesting that they worked both more thoroughly and more efficiently, as would be expected from their higher level of training. Therefore, our weighted scoring instrument seems to possess construct validity.

Previous authors have investigated construct validity in various ways. Schwid *et al.*[9] used a weighted scoring system to grade performance of a scenario that involved diagnosis and treatment of various intraoperative problems and found that they were able to discriminate among the performances of clinical base, CA-1, and more senior residents but were not able to distinguish between CA-2 and CA-3 residents. The weighted scoring system discriminated between resident levels more accurately than a shorter nonweighted checklist. The authors did not elaborate on how the weighted scoring system was obtained.[9]

In another series of studies, a weighted scoring system, obtained with input from faculty at the authors' institution, was used to grade performance of providing anesthesia in several scenarios, such as hemorrhagic hypotension, myocardial infarction, pneumothorax in closed-chest trauma, and others. The weighted scoring system was found to correlate with medical students' and residents' clinical backgrounds and experience.[10] These same authors later also found good correlation among various scoring systems, namely, a checklist, a system that measured when key diagnostic and therapeutic actions were performed, and a holistic global rating on a 10-cm visual analog scale.[11]

Finally, Forrest *et al.*[12] used a modified Delphi method to create a weighted scoring system for rapid sequence induction and intubation. They found that anesthesia attending physicians performed better than anesthesia novices and that the novices' performances improved over time.[12]

We also tested the interrater reliability of our scoring system, and it was found to be reliable, as evidenced by the high intraclass correlation coefficient. Other authors have found acceptable rates of interrater reliability for various scoring systems in use with scenarios on the simulator.[9–14] In addition, because factor analysis indicated that 96% of the variance from the mean score was explained by scores from two attending anesthesiologists, it seems that there is little gain in having more than two anesthesiologists view and score the videotapes. This may also reflect the objectivity of the scoring tool. Studies are conflicting regarding whether scores regarding technical actions may be more reliable than scores regarding behavioral performance, such as decision-making processes and team interaction.[15,16]

There are some potential limitations to this study. First, although participating residents were asked not to dis-

cuss the study, it is possible that some discussion took place, and this could have introduced bias, such as a resident "studying" and thereby improving his or her score. In addition, although we attempted to blind the raters to the residents' identities, it is possible that some residents were identified on the videotapes, including by voice recognition, and this could have introduced bias into the raters' evaluations of those residents. Last, it remains unclear whether performance in a simulated environment reflects performance during a live emergency. We would expect senior residents to perform better than junior residents in an actual operating room, and they did perform better in this simulated environment. This lends some support to the idea that one may extrapolate from the simulator to the operating room; however, we cannot draw definitive conclusions regarding performance in the operating room based on behavior in the simulator.

In summary, through a consensus of experts in the field, we have developed a valid and reliable weighted scoring instrument that can be used to measure resident performance of general anesthesia for emergency cesarean delivery on the human patient simulator. This scoring tool may prove useful for future studies such as those investigating the effect of simulator training on resident performance, both on the simulator and in the operating room environment.

The authors thank the Eleanor Wood-Prince Grant Initiative: A Project of The Woman's Board of Northwestern Memorial Hospital, Chicago, IL for their support of this project. The authors also thank Brenda Bucklin, M.D. (Associate Professor of Anesthesiology, University of Colorado Denver Health Sciences Center, Denver, Colorado), William Camann, M.D. (Associate Professor of Anesthesia, Harvard Medical School, Brigham and Women's Hospital, Boston, Massachusetts), Barbara Leighton, M.D. (Professor of Anesthesiology, Washington University School of Medicine, St. Louis, Missouri), Mark Norris, M.D. (Henry Medical Center, Atlanta, Georgia), and Alex Pue, M.D. (Mary Birch Hospital for Women, San Diego, California), for their assistance in obtaining the scoring system *via* the Delphi technique.

References

1. Bucklin BA, Hawkins JL, Anderson JR, Ullrich FA: Obstetric anesthesia workforce survey: Twenty-year update. ANESTHESIOLOGY 2005; 103:645–53
2. Hawkins JL, Koonin LM, Palmer SK, Gibbs CP: Anesthesia-related deaths during obstetric delivery in the United States, 1979–1990. ANESTHESIOLOGY 1997; 86:277–84
3. Lipman S, Carvalho B, Brock-Utne J: The demise of general anesthesia in obstetrics revisited: Prescription for a cure. Int J Obstet Anesth 2005; 14:2–4
4. Clayton MJ: Delphi: A technique to harness expert opinion for critical decision-making tasks in education. Educ Psychol 1997; 17:373–86
5. Byrne AJ, Greaves JD: Assessment instruments used during anaesthetic simulation: Review of published studies. Br J Anaesth 2001; 86:445–50
6. Wong AK: Full scale computer simulators in anesthesia training and evaluation. Can J Anaesth 2004; 51:455–64
7. Polit DF, Hungler BP: Nursing Research Principles and Methods, 6th edition. Philadelphia, Lippincott Williams & Wilkins, 1999, pp 407–36
8. Devitt JH, Kurrek MM, Cohen MM, Cleave-Hogg D: The validity of performance assessments using simulation. ANESTHESIOLOGY 2001; 95:36–42
9. Schwid HA, Rooke GA, Carline J, Steadman RH, Murray WB, Olympio M, Tarver S, Steckner K, Wetstone S: Evaluation of anesthesia residents using mannequin-based simulation: A multiinstitutional study. ANESTHESIOLOGY 2002; 97:1434–44
10. Boulet JR, Murray D, Kras J, Woodhouse J, McAllister J, Ziv A: Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. ANESTHESIOLOGY 2003; 99:1270–80
11. Murray DJ, Boulet JR, Kras JF, Woodhouse JA, Cox T, McAllister JD: Acute care skills in anesthesia practice: A simulation-based resident performance assessment. ANESTHESIOLOGY 2004; 101:1084–95
12. Forrest FC, Taylor MA, Postlethwaite K, Aspinall R: Use of a high-fidelity simulator to develop testing of the technical performance of novice anaesthetists. Br J Anaesth 2002; 88:338–44
13. Morgan PJ, Cleave-Hogg D: Evaluation of medical students' performance using the anaesthesia simulator. Med Educ 2000; 34:42–5
14. Morgan PJ, Cleave-Hogg DM, Guest CB, Herold J: Validity and reliability of undergraduate performance assessments in an anesthesia simulator. Can J Anaesth 2001; 48:225–33
15. Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R: Assessment of clinical performance during simulated crises using both technical and behavioral ratings. ANESTHESIOLOGY 1998; 89:8–18
16. Weller JM, Bloch M, Young S, Maze M, Oyesola S, Wyner J, Dob D, Haire K, Durbridge J, Walker T, Newble D: Evaluation of high fidelity patient simulator in assessment of performance of anaesthetists. Br J Anaesth 2003; 90:43–7