Construct Validity and Generalizability of Simulation-Based Objective Structured Clinical Examination Scenarios

Avner Sidi, MD
Nikolaus Gravenstein, MD
Samsun Lampotang, PhD

Abstract

Background It is not known if construct-related validity (progression of scores with different levels of training) and generalizability of Objective Structured Clinical Examination (OSCE) scenarios previously used with non-US graduating anesthesiology residents translate to a US training program.

Objective We assessed for progression of scores with training for a validated high-stakes simulation-based anesthesiology examination.

Methods Fifty US anesthesiology residents in postgraduate years (PGYs) 2 to 4 were evaluated in operating room, trauma, and resuscitation scenarios developed for and used in a high-stakes Israeli Anesthesiology Board examination, requiring a score of 70% on the checklist for passing (including all critical items).

Results The OSCE error rate was lower for PGY-4 than PGY-2 residents in each field, and for most scenarios within each field. The critical item error rate was significantly lower for PGY-4 than PGY-3 residents in operating room scenarios, and for PGY-4 than PGY-2 residents in resuscitation scenarios. The final pass rate was significantly higher for PGY-3 and PGY-4 than PGY-2 residents in operating room scenarios, and also was significantly higher for PGY-4 than PGY-2 residents overall. PGY-4 residents had a better error rate, total scenarios score, general evaluation score, critical items error rate, and final pass rate than PGY-2 residents.

Conclusions The comparable error rates, performance grades, and pass rates for US PGY-4 and non-US (Israeli) graduating (PGY-4 equivalent) residents, and the progression of scores among US residents with training level, demonstrate the construct-related validity and generalizability of these high-stakes OSCE scenarios.

Editor's Note: The online version of this article contains 6 appendixes: Israeli examination process development and adaptation; process validation of a simulation-based OSCE used in Israeli Anesthesiology Board Examination; examples of operating room, trauma, and resuscitation scenario checklists; examinee and examiners feedback; and calculated checklist formulas.

All authors are in the Department of Anesthesiology, University of Florida College of Medicine, and the Center for Safety, Simulation & Advanced Learning Technologies. **Avner Sidi, MD,** is Associate Professor of Anesthesiology; **Nikolaus Gravenstein, MD,** is Professor of Anesthesiology; and **Samsun Lampotang, PhD,** is Professor of Anesthesiology.

Funding: The authors report no external funding source for this study.

Conflict of Interest: The authors declare they have no competing interests.

Previous presentations (abstract, poster, meeting): Sidi A, Berkenstadt H, Ziv A, Euliano T, Lampotang S, White C. Evaluating Construct Validity of Simulation-Based OSCE for Summative Assessment in an Anesthesiology Teaching Program. Abstract A-151, ASA Annual Meeting, October 15, 2011.

The authors would like to thank the Israeli Board of Anesthesiology Examination Committee for its cooperation, and especially acknowledge the help of its present chairman, Dr Haim Berkenstadt; Isaac Luria, the Center for Safety, Simulation & Advanced Learning Technologies, University of Florida, College of Medicine; and Jonathan Sidi, Statistician Consultant, Department of Statistics, Hebrew University, Jerusalem, Israel.

Corresponding author: Avner Sidi, MD, Department of Anesthesiology, University of Florida College of Medicine, PO Box 100254 JHMHSC, 1600 SW Archer Road, Gainesville, FL 32610-0254, 352.413.8073, fax 352.392.7029, asidi@anest.ufl.edu

Received October 3, 2013; revisions received February 5, 2014, and March 15, 2014; accepted March 31, 2014.

DOI: http://dx.doi.org/10.4300/JGME-D-13-00356.1

Introduction

The Israeli Board Examination in Anesthesiology simulation-based objective structured clinical examination (OSCE) is well described and validated[1–3] using objective and subjective parameters.[2–4] The examination is administered only to graduating residents, and its construct-related validity (progression of OSCE scores across years of training)[4] has not been evaluated, as all Israeli participants were at the same level of training.

Since 2004, the National Board of Medical Examiners has required a simulation-based clinical skills examination for medical students.[5] The Accreditation Council for Graduate Medical Education Outcome and Milestone Projects[6–9] led the American Board of Anesthesiology (ABA) to give diplomates enrolled in Maintenance of Certification in Anesthesiology from 2000 through 2007 the option to complete an endorsed simulation course to support

their knowledge and skills.[10] More recently, the ABA implemented a simulation-based OSCE as part of its certification examinations.[11] As development of simulation-based OSCEs carries time and financial costs, sharing of validated scenarios across nations would facilitate their broader dissemination and implementation.

We assessed the construct validity of simulation-based OSCE summative assessment tools developed for the Israeli Board Examination, and their potential generalizability to a US training program for formative or summative assessment. We examined anesthesiology residents across all postgraduate years (PGYs) 2 to 4 at 1 institution. This validation could not be performed in the Israeli Board setup, which tested only graduating residents (equivalent to US PGY-4 residents). The other aim was to demonstrate the generalizability, using scenarios developed for an international examination in a US academic environment for formative (teaching) and summative (testing) assessment.

What was known

Objective Structured Clinical Examinations (OSCEs) are increasingly used to assess the competence of residents, including both formative and summative high-stakes assessments.

What is new

A set of simulation-based OSCE scenarios has construct-related validity and generalizability in the assessment of residents at different levels of training.

Limitations

Single-site study, lacking a formal power calculation; use of advancing clinical year as the sole discriminant variable for competence.

Bottom line

Reduction in error rates, and progressive score and performance for residents with advancing training level demonstrate construct-related validity and utility of OSCE scenarios.

Methods

The study was conducted at the University of Florida Anesthesiology Residency Program. In a simulated environment, 2 similar but not identical scenarios (to counter scenario content leakage and enhance content security) were used in each of 3 clinical domains: resuscitation, trauma, and operating room crisis management. The scenarios were originally developed by the Israeli Board of Anesthesiology Examination Committee.[2,3,12,13] Faculty members from the Department of Anesthesiology at the University of Florida, assisted by education and simulation experts, translated the scenarios and assessment tools, with maximal adherence to the original script,[2,3] scenario protocol, language, and assessment tools. No changes were made in scoring, assessment, pass/fail determinations, orientation of residents, or the examination process itself (see APPENDIXES 1 AND 2 for development and validation).

Individual performance in each scenario was assessed by using a 12- to 20-item scenario-specific checklist. Checklist items covered simple and more complex functions and were scored in a binary done/not done format. A set of criteria for a well-performed task was provided to the evaluator to standardize the assessment. All checklist items were weighted equally. Critical items carried the same score as noncritical items but were considered mandatory for "passing" the scenario.[2,3]

All resident participants were recruited by the chief residents and had participated in an orientation and prior sessions with the Human Patient Simulator (CAE Healthcare). We evaluated all PGY groups within a 3-month window.

Before participation each consenting resident received oral and printed materials explaining the study objectives and assurance that performance was confidential and had no impact on their residency program evaluations. The examiner followed a predetermined script sequence and checklist. The target time to work through a scenario was 20 to 30 minutes (APPENDIX 3).

All residents were evaluated by an experienced evaluator (A.S.), an original member of the Israeli Examination Committee,[2,3] who had no clinical interaction with the residents during the study period and was blinded to the residents' PGY level until after assessment and scoring were completed. The assessor followed the original process of the Israeli examination[2,3]: repeating questions only twice, playing a passive role, avoiding prompting, and responding to examinee-specific requests for action or information only when the script/checklist allowed it.

Scores were computed for each resident and each scenario: (1) proportion correct (total) across all items in the checklist (the final score was the items correctly performed out of the total possible in a scenario); (2) proportion correct (critical); and (3) critical errors (the critical items error rate was the rate at which examinees did not perform critical items). The residents also completed questionnaires on realism of each scenario, including the perceived relevance of the scenario(s) and the residents' satisfaction from their performance in the simulation (APPENDIX 4).

We elicited feedback with reviewers familiar with the training and performance of US-based examinees (APPENDIX 5). For every item in each of the scenarios, the following parameters were calculated as previously published[2,3] and compared between PGY groups (APPENDIX 6):

Item Performance Grade: Fraction of residents who performed the item satisfactorily.

TABLE **Distribution of Residents in PGY 2 to 4, in Each Domain, and in Each Scenario (No. 1 and No. 2) Within a Domain**

Scenario	OR1	OR2	Total	N[a]	Tr1	Tr2	Total	N[a]	Res1	Res2	Total	N[a]
PGY-2 residents	3	5	8		3	5	8		4	3	7	
PGY-3 residents	5	4	9		3	4	7		3	4	7	
PGY-4 residents	4	3	7	24	4	2	6	21	4	3	7	21
Items	20	20	40		14	12	26		17	17	34	
Critical items	2	3	5		5	4	9		11	11	22	

Abbreviations: OR, operating room; Tr, trauma; Res, resuscitation.

[a] Number of PGY-2 to PGY-4 residents tested in a domain.

Group (PGY) Error Rate: Number of errors as a fraction of the number of items answered by all tested for a given scenario in a PGY group. It offers information about the difficulty level of each scenario. Item performance grade gives information about the performance of examinees for each item and is inversely related to the difficulty.

Individual (Resident) Success Grade: Items answered satisfactorily by a resident as a fraction of the number of items answered in a scenario.

All checklists and scripts included identification of the resident PGY level. Results were analyzed by using statistical software (SAS Version 9.2, SAS Institute Inc, Cary, NC).

Critical Items

Critical items error occurrence was compared between and among PGY groups.

Scenarios

Scenario performance grade was calculated for all items in the scenario tested by a group of n residents as mean ± SD. Scenario pass grade was defined as a score of 70% as per the Israeli Board of Anaesthesiology of the checklist items successfully performed, including all critical items.[2,3] The pass rate was compared between PGY groups.

Statistics

A noninferiority test (for proportion correct scores) was conducted between the pairs of scenarios in each field to test equivalence, assuming an allowable difference of ≤ 30% in performance or difficulty grades, while checking power for range of difference.[14] This was done to ensure the 2 scenarios in each of the clinical areas (within the same type or field) were not inferior to each other. A subsequent equivalence test (for proportion correct scores) was conducted between each pair of scenarios to evaluate similarity.[15] Equivalence was accepted with 80% certainty if the ratio (log difference) of the grades was within 20% for the pair. The log difference was used because the grades distribute log normally. The 80% and 20% thresholds were used because these are accepted rates of equivalency tests.[15] Variables are presented as mean ± SD. Differences were considered significant when $P \leq .05$.

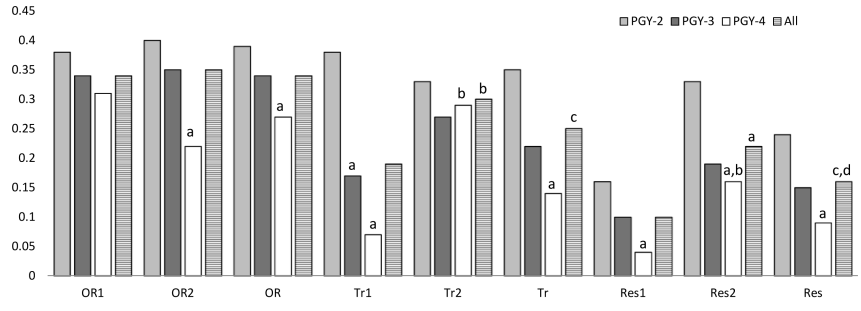
The questionnaires that analyzed the realism of the scenario, perceived relevance, and satisfaction from participants' performance in the simulation scored on a scale from 1 to 5, with 5 being the highest (APPENDIX 4, TABLE B). Correlations between resident satisfaction with their own performance in the simulation and both the total proportion-correct scores and the general scores were calculated (APPENDIX 4, TABLE C).

Variables were compared between groups by a random mixed-effect analysis of variance (ANOVA) model. We calculated means for PGY and field as random variables and the scenario as the fixed variable. The error and pass rates for scenarios were compared by using a 2-proportion z test.

Results

Fifty PGY-2 to PGY-4 residents participated in the study during 4 months. Five residents (out of 75 in the program) declined to participate, and since residents were selected in a random fashion on the basis of availability for the day of the test, only 50 residents were available for the study. The simulation-based OSCE was administered 66 times to 50 different residents (with approximately equal distribution from each PGY group). Each resident performed 1 or 2 different scenarios from 1 to 2 clinical domains. Thirty-four residents were tested on 1 scenario, and 16 were tested on 2 scenarios from 2 domains. The reason we did not limit the study to 1 scenario per resident was to overcome poor resident availability.

The TABLE shows the distribution of residents in each year in each domain and scenario, and the number of items and critical items (5–22) in each domain and scenario. The



Scenario	OR1	OR2	OR	Tr1	Tr2	Tr	Res1	Res2	Res
PGY-2	0.38	0.4	0.39	0.38	0.33	0.35	0.16	0.33	0.24
PGY-3	0.34	0.35	0.34	0.17[a]	0.27[b]	0.22	0.1	0.19	0.15
PGY-4	0.31	0.22[a]	0.27[a]	0.07[a]	0.29[b]	0.14[a]	0.04[a]	0.16[a,b]	0.09[a]
All	0.34	0.35	0.34	0.19	0.3	0.25	0.1	0.22[b]	0.16[c,d]

Note that the error rate for all residents in trauma and resuscitation domains was lower than the OR domain, and lower in the resuscitation domain than in the trauma domain.
[a]$P < 0.05$ compared to PGY-2
[b]$P < 0.05$ compared to scenario 1
[c]$P < 0.05$ compared to OR in the same PGY
[d]$P < 0.05$ compared to trauma in the same PGY
Abbreviations: OR, operating room; Tr, Trauma; Res, Resuscitation.

FIGURE 1 **Error Rate for Residents in Each Postgraduate Year (PGY) 2 to 4 in Each Domain, and in Each Scenario (No. 1 and No. 2) Within Each Domain**

ANOVA analysis of the different PGY levels was significant, and analysis revealed what drove those differences.

All scenarios were compared in the difficulty level (performance grade), and were not different among different PGYs and clinical domains tested. A noninferiority test[10] and a subsequent equivalence test[15] (for proportion correct scores) demonstrated the similarity between the 2 scenarios for the operating room (OR) and resuscitation. The corresponding *P* values to the equivalence tests are resuscitation, $P = .10$ (equivalent at 10% level); trauma, $P = .27$ (not equivalent at 10% level); OR, $P = .06$ (equivalent at 10% level); and overall, $P = .005$ (equivalent at 10% level). Thus, for trauma, equivalence could not be established, and scenario 1 has higher grades than scenario 2. Other than for trauma, our equivalence and noninferiority tests showed the 2 scenarios in each domain were not different in difficulty or performance level, and were treated as 1 unit for the OR and resuscitation domains. The evaluator had only a single level of variability because 1 experienced evaluator evaluated all residents. PGY and domain between groups were compared by using ANOVA. There were no significant differences in the performance grades (calculation of scenario difficulty) within any scenario pair in a domain.

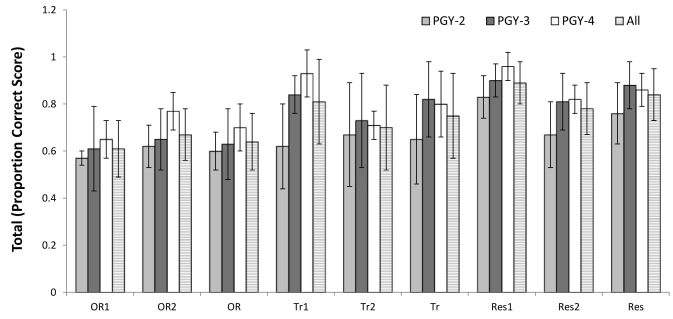
The error rate was lower for PGY-4 residents than PGY-2 residents in each domain and scenario, except for scenario OR No. 1 and trauma No. 2, where the error rate was relatively high for all participants regardless of PGY (FIGURE 1). When scenario No. 1 and No. 2 in each clinical domain was considered as 1 unit, the error rate was significantly lower in each domain for PGY-4 residents.

The error rate was significantly different between scenarios No. 1 and No. 2 in the trauma and resuscitation domains for PGY-4 residents (29% versus 7% and 16% versus 4%, respectively; $P < .05$ for both) and for all residents (30% versus 19% and 22% versus 10%, respectively; $P < .05$ for both). The error rate for all residents in the trauma and resuscitation scenarios was lower than that in the OR scenario (25% and 16%, respectively, versus 34%; $P < .01$ for both), and lower for resuscitation than for trauma (16% versus 25%; $P = .01$).

The total proportion correct scores (FIGURE 2) were not significantly different for PGY-3 and PGY-4 compared to PGY-2 residents. However, the critical items error rate (FIGURE 3A) was significantly lower for PGY-4 residents than PGY-3 residents in the OR domain, and was significantly lower for PGY-4 residents than PGY-2 residents in the resuscitation domain. The final pass rate was significantly higher for PGY-3 and PGY-4 residents than PGY-2 residents in the OR scenario (FIGURE 3B), and final pass rate was significantly higher for PGY-4 residents than PGY-2 residents when all 3 clinical domains were combined (11 of 22 = 0.50 versus 2 of 23 = 0.09).

Discussion

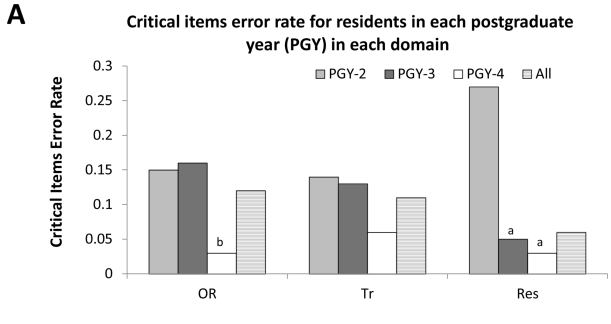
We demonstrated the utility and "generalizability" of scenarios developed for an Israeli examination for assessment in a US academic environment. Our study was



Scenario	OR1	OR2	OR	Tr1	Tr2	Tr	Res1	Res2	Res
PGY-2	0.57	0.62	0.6	0.62	0.67	0.65	0.83	0.67	0.76
PGY-3	0.61	0.65	0.63	0.84	0.73	0.82	0.9	0.81	0.88
PGY-4	0.65	0.77	0.7	0.93	0.71	0.8	0.96	0.82	0.86
All	0.61	0.67	0.64	0.81	0.7	0.75	0.89	0.78	0.84

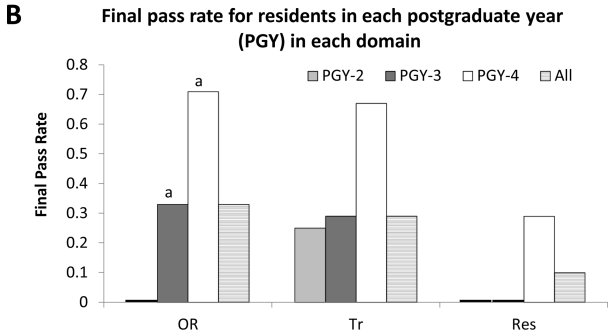
Abbreviations: OR, operating room; Tr, Trauma; Res, Resuscitation.

FIGURE 2 **Total (Proportion Correct Score) for Residents in Each Postgraduate Year (PGY) in Each Domain, and in Each Scenario**



Domain		OR	Tr	Res
Critical Items Error Rate	PGY-2	0.15	0.14	0.27
	PGY-3	0.16	0.13	0.05[a]
	PGY-4	0.03[b]	0.06	0.03[a]
	All	0.12	0.11	0.06

[a]$P < .05$ compared to PGY-2
[b]$P < .05$ compared to PGY-3



Domain		OR	Tr	Res
Final Pass Rate	PGY-2	0	0.25	0
	PGY-3	0.33[a]	0.29	0
	PGY-4	0.71[a]	0.67	0.29
	All	0.33	0.29	0.10

[a]$P < .05$ compared to PGY-2
[b]$P < .05$ compared to PGY-3

FIGURE 3 **Residents' Critical Items Error and Final Pass Rates**

performed in an academic department, which allowed assessment of the construct-related validity of the scenarios (ie, the ability of each scenario to differentiate between participants on the hypothesis that more senior residents will generally have better scores). Our findings are similar to the results of the original tests performed on graduating residents in the Israeli Examination,[2,3] with error rates of 0.08, 0.16, and 0.25; performance (difficulty) grade of 0.93, 0.83, 0.85; and pass rates of 0.68, 0.62, 0.75 for resuscitation, trauma, and OR crisis management, respectively. These results further support the feasibility of sharing scenarios between different residency programs.[16,17] Our results showed that PGY-4 residents had superior results compared to PGY-2 residents in error rate, total scenario score, general evaluation score, critical items error rate, and final pass rate.

There are inconsistencies in the statistical significance of different parameters, and various parameters do not necessarily correlate with each other. The practical inference is that the parameters complement each other, and that more than 1 should be used when making decisions using a simulation-based OSCE. For example, the proportion correct score does not represent the final score or the entire scope of the evaluation, because the critical errors score is not included. The final pass score is a result of the proportion correct *and* critical errors scores, and thus the error rate results may point in the opposite direction of the pass rate results (ie, the OR scenarios had higher error rates, but also a higher pass rate from fewer critical errors than the resuscitation scenario).

The final resuscitation scenarios' pass rates were surprisingly low for all residents tested. This suggests the obvious—there is value to the PGY-4 year—and that there

may be gaps in the curriculum or learning pertaining to resuscitation.

One possible added reason is the higher number of critical items in the resuscitation scenarios, which caused the pass rates to be: OR > trauma > resuscitation. Our findings should be further evaluated to pinpoint and define the learning deficiencies for trauma and for resuscitation.[18,19]

Although simulation in anesthesia has become part of the teaching curricula,[19–21] only 14% of simulation centers use it for evaluation of competence.[22] Reasons for this include lack of standardized, valid, and reliable tests.[22] Communication and collaboration among simulation centers, including sharing of validated scenarios, is important to the future of this technology and approach.[23]

Limitations of our study include that it was conducted at a single institution, we did not perform a comprehensive calculation of sample size, and we used a limited sample of residents. The study also is limited in its inability to differentiate learning from teaching, and its inability to consider indicators of competence other than advancing clinical year.

Conclusion

Our work confirms both the construct-related validity of incorporating a simulation-based OSCE in the assessment of anesthesiology residents, and the generalizability of these specific scenarios for formative and summative assessment.

Prospective evaluation of OSCE scenarios should be part of a joint effort of the health care simulation community, with future research including evaluation of OSCE scenarios to enable differentiating between technical and nontechnical skills, and identification and correction of performance deficiencies identified during these assessments.

References

1 Ziv A, Rubin O, Sidi A, Berkenstadt H. Credentialing and certifying with simulation. *Anesthesiol Clin*. 2007;25(2):261–269.

2 Berkenstadt H, Ziv A, Gafni N, Sidi A. Incorporating simulation-based objective structured clinical examination into the Israeli National Board Examination in Anesthesiology. *Anesth Analg*. 2006;102(3):853–858.

3 Berkenstadt H, Ziv A, Gafni N, Sidi A. The validation process of incorporating simulation-based accreditation into the anaesthesiology Israeli national board exams. *Isr Med Assoc J*. 2006;8(10):728–733.

4 Devitt JH, Kurrek MM, Cohen MM, Fish K, Fish P, Noel AG, et al. Testing internal consistency and construct validity during evaluation of performance in a patient simulator. *Anesth Analg*. 1998;86(6):1160–1164.

5 Papadakis MA. The Step 2 clinical-skills examination. *N Engl J Med*. 2004;350(17):1703–1705.

6 Accreditation Council for Graduate Medical Education. ACGME Outcomes Project, 2013. http://www.acgme.org/acgmeweb/Portals/0/PDFs/SlideDecks/SLIDEDECK-FDMilestones2013.pptx. Accessed May 23, 2014.

7 Accreditation Council for Graduate Medical Education. Implementing Milestones and Clinical Competency Committees. http://www.acgme.org/acgmeweb/Portals/0/PDFs/ACGMEMilestones-CCC-AssesmentWebinar.pdf. Released April 24, 2013. Accessed May 23, 2014.

8 Accreditation Council for Graduate Medical Education. *Journal of Graduate Medical Education* announces March 2014 release of Milestones for all Phase II Specialties. http://www.acgme.org/acgmeweb/Portals/0/PDFs/newsRel_06_10_13.pdf. Accessed May 23, 2014.

9 Objective Structured Assessment of Technical Skill (OSATS) in ACGME Outcomes Project Toolbox of Assessment Methods. http://www.acgme.org/acgmeweb/Portals/0/PFAssets/Nov4NASImpPhaseII.pdf and Reference No. 6: http://www.acgme.org/acgmeweb/Portals/0/PDFs/ACGMEMilestones-CCC-AssesmentWebinar.pdf. Accessed July 14, 2014.

10 MOCA Part IV requirements update. American Board of Anesthesiology, Inc. September 21, 2010. https://www.asahq.org/For-Members/Education-and-Events/Calendar-of-Events.aspx and https://www.asahq.org/education. Accessed July 14, 2014.

11 Applied Examination (new part 2 Oral Examination) announcement update. American Board of Anesthesiology (ABA), Inc. June 29, 2012. http://www.theaba.org/pdf/OSCE-Panel.pdf. Accessed November 19, 2012.

12 Schwid HA, Rooke GA, Carline J, Steadman RH, Murray WB, Olympio M, et al. Evaluation of anesthesia residents using mannequin-based simulation: a multiinstitutional study. *Anesthesiology*. 2002;97(6):1434–1444.

13 Scavone BM, Sproviero MT, McCarthy RJ, Wong CA, Sullivan JT, Siddall VJ, et al. Development of an objective scoring system for measurement of resident performance on the human patient simulator. *Anesthesiology*. 2006;105(2):260–266.

14 James Hung HM, Wang S-J, Tsong Y, Lawrence J, O'Neill RT. Some fundamental issues with non-inferiority testing in active controlled trials. In: James Hung HM, ed. *Proceedings of the Annual Meeting of the American Statistical Association*. http://www.amstat.org/sections/SRMS/Proceedings/y2001/Proceed/00534.pdf. Accessed May 23, 2014.

15 Lužar-Stiffler V, Stiffler C. Equivalence testing the easy way. *J Comput Inf Technol*. 2002;3:233–239.

16 Berkenstadt H, Kantor GS, Yusim Y, Gafni N, Perel A, Ezri T, et al. The feasibility of sharing simulation-based evaluation scenarios in anesthesiology. *Anesth Analg*. 2005;101(4):1068–1074.

17 Boulet JR, Murray DJ. Simulation-based assessment in anesthesiology: requirements for practical implementation. *Anesthesiology*. 2010;112(4):1041–1052.

18 Seropian MA. General concepts in full scale simulation: getting started. *Anesth Analg*. 2003;97(6):1695–1705.

19 Gaba DM. What makes a "good" anesthesiologist? *Anesthesiology*. 2004;101(5):1061–1062.

20 Blum RH, Raemer DB, Carroll JS, Sunder N, Felstein DM, Cooper JB. Crisis resource management training for an anaesthesia faculty: a new approach to continuing education. *Med Educ*. 2004;38(1):45–55.

21 Weller J, Wilson L, Robinson B. Survey of change in practice following simulation-based training in crisis management. *Anaesthesia*. 2003;58(5):471–473.

22 Morgan PJ, Cleave-Hogg D. A worldwide survey of the use of simulation in anesthesia. *Can J Anaesth*. 2002;49(7):659–662.

23 Girard M, Drolet P. Anesthesiology simulators: networking is the key. *Can J Anaesth*. 2002;49(7):647–649.