The Validation Process of Incorporating Simulation-Based Accreditation into the Anesthesiology Israeli National Board Exams

Haim Berkenstadt MD[1,2,3], Amitai Ziv MD MD[2], Naomi Gafni PhD MD[4] and Avner Sidi MD[1,3]

[1]Israeli Board Examination Committee in Anesthesiology, Scientific Council, Israel Medical Association, Ramat Gan, Israel
[2]Israel Center for Medical Simulation, Sheba Medical Center, Tel Hashomer, Israel
[3]Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv, Israel
[4]National Institute for Testing and Evaluation, Jerusalem, Israel

Key words: simulation, examination, assessment, accreditation, anesthesiology, validity, reliability, realism

Abstract

Background: The Israeli Board of Anesthesiology Examination Committee added a simulation-based Objective Structured Clinical Evaluation component to the board examination process. This addition was made in order to evaluate medical competence and considers certain domains that contribute to professionalism. This unique and new process needed to be validated.

Objectives: To validate and evaluate the reliability and realism of incorporating simulation-based OSCE into the Israeli Board Examination in Anesthesia.

Methods: Validation was performed before the exam regarding Content Validity using the modified Delphi technique by members of the Task Force of the Israeli Board Examination Committee in Anesthesiology.

Results: The examination has been administered six times in the past 3 years to a total of 145 examinees. The pass rate ranged from 62% (trauma) to 91% (regional anesthesia). The mean inter-rater correlations for the total score (all items), for the Critical checklist items score, and for the Global (General) rating were 0.89, 0.86 and 0.76, respectively. The inter-correlations between the five OSCE stations scores were significant ($P < 0.01$) only between Trauma & Ventilation for the Total score ($r = 0.32$, n=63), and between Resuscitation & Regional and OR-crisis for the Global score ($r = 0.42$ and 0.27, n=64 and 104, respectively). The correlation between the OSCE examination score and the success rate at each of the eight different clinical domains of the oral board examination did not reach statistical significance. Most participants (70–90%) found the difficulty level of the examination stations reasonable to very easy. All major errors, which were identified in the initial two exam periods, disappeared later in the next two exam periods.

Conclusions: The exam has gradually progressed from being an optional part of the oral board examination to a prerequisite component of this test. Other anesthesiology programs or medical professions can adopt the model described here.

IMAJ 2006;8:728–733

Professional competence assessment should include both formative (training) and summative (testing) functions [1]. The modern model of medical competence [2] considers certain domains that contribute to professionalism – knowledge, technical skills, clinical reasoning, communication, and emotions [3]. In the model of medical competence, each of the domains contributing to professionalism should be evaluated according to the four stages of competence, which are defined as "knows," "knows how," "shows how," and "does." The upper levels of competence in each domain evaluated or tested are "shows how" (i.e., oral exams) and "does" (i.e., practical exams).

The mandatory practical step of recognizing operating room equipment in the oral exams of the Israeli Board of Anesthesiology (introduced in the 1980s and 1990s) falls into the category of "know how" rather than "show how" or "does" stages of competence. In an attempt to assess to what extent residents in anesthesiology are competent in the "shows how" and "does" stages, we introduced the use of high fidelity medical simulation. In other words, the mandatory practical step of recognizing OR equipment in the oral exams was abandoned because it gave no information with regard to "shows how" and "does" stages.

The *inter-rater reliability* [4] and *construct validity* (progression of simulator scores with the level of training) [5,6] of simulation-based scenarios have been demonstrated, and a multi-institutional study has validated simulation-based scenarios as an effective tool for the evaluation of residents [7]. Simulation has also been described as a practical tool for accreditation, but in only a few and in limited fields or organizations, for example, the New York State Society of Anesthesiologists for anesthesiologists with lapsed medical skills [8]. Another example is the European Board Vascular Examination, which has its Part II exam divided into four parts, one of which is a technical skill evaluation session comprising three bench stations (knots dexterity, junction dissection, anastomosis) [9]. Other examples are the Department of Family Practice at the University of Kentucky, which conducts computer-based testing in family practice certification and recertification [10]; the University Hospital in Heidelberg, Germany, which uses full-scale simulation for the accreditation of anesthesia nurses [11]; the Department of Anesthesiology in Rochester, for first year

OR = operating room
OSCE = objective structured clinical evaluation

anesthesiology residents to take overnight call duty [12]; and the University of Pittsburgh Department of Anesthesiology for difficult airway management training [13].

The model presented here is not necessarily feasible for implementation in other countries. The large number of candidates that the American Board of Anesthesiology examines annually renders objective structured clinical evaluation logistically problematic. However, the National Board of Medical Examiners in the USA has recently introduced the simulation-based clinical skills examination [14], reflecting a major shift in the field of medical accreditation and licensure toward acknowledging the crucial role of performance assessment as an important component of professional accreditation.

Nonetheless, only 7–14% of simulation centers use advanced simulation for competence evaluation [15] and there have been no other additional reports (except from this group) [16] of the implementation of simulation-based performance assessment in high-stakes board examinations in anesthesiology. Worldwide, board examinations in anesthesiology are predominantly based on the traditional paradigm of oral examinations and multiple-choice questionnaires. Generally, these examinations evaluate the cognitive aspects of the profession but they lack the ability to appraise performance and practical skills. These domains are left to non-structured "on the job" evaluation in the apprentice clinical environment.

Acknowledging the fact that the Israeli board examination in anesthesiology lacked a performance evaluation element, and that this element had not been a substantial part of training programs in the country, the Israeli Board of Anesthesiology Test Committee decided to explore the potential of adding a simulation-based objective structured clinical examination component to the board examination process [16]. We previously described in detail the development of the unique process whereby simulation-based OSCE was incorporated into the Israeli board examination in anesthesiology. In the present article we provide additional and updated information of the process, which has not been published previously, including descriptive statistics for the scenarios, summary of descriptive statistics for one examination term (five examination stations), and summary of all pass rates for all five examination stations in six test periods.

In the process of assessing validation of a teaching/testing tool like simulation-based medical education, or high-stake exams, the validation assessment should take into consideration the case selection, case-subject interaction, rater variability, and the rating system. The following parameters can be evaluated:

- *Subjective parameters*
  a) Satisfaction (subjective interpretation of the trainees)
  b) Sense of realism (subjective rating of examinees) such as familiarity/comfort, which has no effect on objective performance [6]
- *Objective parameters*
  a) Outcome (cost-effectiveness, patient outcome)
  b) Reliability (measure of the reproducibility or consistency of a test, including inter-rater reliability comparing rater variance [5]; internal consistency (inter-case reliability evaluation using Cronbach-alpha statistics or other statistical methods of measuring intra-class correlation coefficient, e.g., reliability coefficient: ratio of a variance to total variance) [17]; inter-simulator reliability (effect of the simulator type).
  c) Validity (measured or "Face" validity) including context/content validity of pretest experts' consensus using the modified Delphi technique [18]; construct validity (construct-related: evaluating the progression of simulator scores with the level of training); criterion-related (convergent) validity comparing simulator scores with other forms of evaluation (i.e., written exams, mock orals, faculty evaluations).
  d) Efficacy (effect of simulation on performance), including error reduction and predictive validity – i.e., transferability (to clinical practice) or sustainability (for longer time) [19].

We conducted and evaluated the following: satisfaction and realism, inter-rater and inter-case reliability, content validity and convergent validity, and error reduction.

Methods

Content validation

The development of the OSCE component of the examination was based on a collaborative effort led by the Israeli Board of Anesthesiology Test Committee with the assistance of simulation experts from the Israel Center for Medical Simulation and experts in psychometrics and performance assessment from Israel's National Institute for Testing and Evaluation.

The content of the examination was defined and developed according to the steps and criteria recently described by Newble [20]. First, clinical problems or conditions that residents nearing the end of their training are required to handle competently were defined on the basis of the expert opinion of members of the national "Task Force Examination Committee." Given the relative benefits of medical simulation and the capabilities of the available simulation platforms in Anesthesia, five crucial clinical conditions were selected as general tasks: trauma management, resuscitation, operating room crisis management, regional anesthesia, and mechanical ventilation.

The second step of the process involved the definition of tasks for each of the five crucial clinical conditions. The tasks were selected on the basis of minimum requirements, which were decided upon on the basis of over 80% consensus among the members of the task force. The expert committee used a variation of the Delphi technique [18] and critical items were decided only on the basis of expert consensus. We built our exam on testing minimal requirement task performance; most of the tasks were within a framework of basic knowledge and technical aspects. Examples of basic knowledge/technical tasks include the following: preoperative patient assessment, equipment checks, laboratory data acquisition and interpreta-

tion, airway assessment and intubation, treatment according to algorithms of Advanced Cardiac Life Support (ACLS) and Advanced Trauma Life Support (ATLS), algorithms of handling intraoperative changes in blood pressure or airway pressure, ventilator setup and parameter changes, and regional block description. Only a few of the tasks were in the framework of advanced knowledge and behavior, which included: acquisition of all available information, anticipation and plans, reevaluation of the situation with a new complication, prioritizing concise directed instructions, communication, and teamwork within the trauma/resuscitation team.

In the third step of the process, the tasks were incorporated into five simulation-based stations in the OSCE format: trauma management, resuscitation, OR crisis management, mechanical ventilation, and regional anesthesia.

Assessment

The assessment of examinee performance in each scenario was based on a checklist comprising 12–20 items. Each checklist was approved by the task force committee of experts, and critical items were determined on the basis of expert consensus. During the examination, each checklist was completed by two independent examiners, and the data were collected for further analysis by the examination committee. Examinees received a "pass" score on the scenario if they successfully performed 70% of the station's checklist items, including all critical actions/items. Examiners also graded the examinees' performance independently and holistically (for a "general" impression) on a scale of 1 to 4. The global (general impression) score was used to override a borderline (70%) checklist result.

The examination was administered six times during 2003–2005. The first two test periods of the examination, before the exam became obligatory, were defined as a transition period. In the next four periods, passing the "practical" exams was mandatory for being admitted to the "oral" exams.

Statistics

The checklists completed by the examiners were analyzed using SAS software. For each item in each of the scenarios we calculated the following: the error rate (the degree to which the two examiners both agreed that the examinee did not perform the item satisfactorily), the incongruence rate (the degree to which the two examiners did not agree), and mean difficulty (the proportion of examinees who performed an item well or satisfactorily). In addition, errors that occurred frequently during each test period were analyzed and their frequency was compared between periods.

The following scores were computed for each examinee and for each scenario: a) "proportion correct" (Total) across all items included in the checklist, across the two examiners (1 for correct performance, 0 otherwise); b) proportion correct (Critical): same rules for assigning scores as above for the critical items included in the checklist; and c) mean general (Global grading) evaluation across the two examiners.

Using Pearson's correlation, the correlation between the proportions of correct items across all items, the proportion of correct critical items and attainments on the general evaluation were calculated for the 104 examinees who participated in the four test periods. The inter-correlations between the five simulation stations for the proportion of correct scores and general evaluation scores were also calculated. The correlation between the practical OSCE component and the results of the oral board examination were assessed for the third and fourth test periods of the examination.

The examinees completed feedback questionnaires on the difficulty of each of the scenarios and their subjective ability to express their knowledge as compared to conventional oral examinations. These were analyzed using Microsoft Excel software.

Table 1. Descriptive statistics for one of the resuscitation scenarios (n=11 examinees)

Item	Critical	No. of errors	Error rate	No. of incongruencies	Incongruence rate	Mean difficulty
Address the patient		1	1/11	1	1/11	0.91
Assessment		4	4/11	0	0	0.64
Give oxygen	Yes	0	0	0	0	1.00
Medications		1	1/11	3	3/11	0.91
Consider non-invasive ventilation (NIV)		1	1/11	0	0	0.91
Correct application of NIV		3	3/11	0	0	0.73
ECG		2	2/11	0	0	0.82
Chest X-ray 1		0	0	1	1/11	1.00
Chest X-ray 2		1	1/11	1	1/11	0.91
Recognition of heart rhythm		0	0	1	1/11	1.00
Assessment		0	0	3	3/11	1.00
Defibrillation	Yes	6	6/11	0	0	0.45
Assessment		1	1/11	0	0	0.91
Equipment for tracheal intubation	Yes	2	2/11	0	0	0.82
Medications for tracheal intubation		0	0	2	2/11	1.00
Recognition of heart rhythm		0	0	0	0	1.00
Assessment		1	1/11	3	3/11	0.91
Amiodarone		2	2/11	1	1/11	0.82
Sequence of action – 1st cardiac arrhythmia	Yes	6	6/11	1	1/11	0.45
Sequence of action – 2nd cardiac arrhythmia	Yes	1	1/11	1	1/11	0.91
All		32		18		0.85

No. of errors = the degree to which the two examiners both agreed that the examinee did not perform the item satisfactorily

No. of incongruencies = the number of errors that only one examiner indicated, but the two examiners did not agree.

Error rate = the rate at which the two examiners both agreed that the examinee did not perform the item satisfactorily.

Incongruence rate = the rate at which the two examiners did not agree.

Mean difficulty = the proportion of examinees who performed an item well or satisfactorily.

Results

The examination has been administered six times (=test periods) over 2 years, to a total of 145 examinees. Table 1 presents an example of descriptive statistics – including the number and rate of errors and incongruence, and the difficulty grade – for one of the resuscitation scenarios. In this example, a high rate of error was manifest in the initial assessment of the patient, in the critical action of defibrillation, and in the critical item of sequence of actions during the treatment of the first arrhythmia. The numbers of incongruencies in these items were, however, low.

Table 2 presents an example of descriptive statistics for the scenarios in examination period 4. The table presents data pertaining to the different scenarios used in the examination, including the number of times a scenario was used and the number of items and critical items it contained. Scenarios used for the same clinical conditions had similar levels of difficulty. The rate of incongruencies between examiners ranged from 4 to 15% and the error rate ranged from 0 to 25% of items assessed. The mean inter-rater correlations for the total, critical, and general scores for the test period 4 were 0.89, 0.86 and 0.76 respectively.

Table 2. Summary of descriptive statistics for the 4th examination term (n=23)

Scenario	Resus 1	Resus 2	Trauma 1	Trauma 2	OR 1	OR 2	Vent 1	Vent 2	Regional 1	Regional 2
Examinees	14	9	14	9	14	9	14	9	14	9
No. of items	20	20	14	13	14	14	17	18	15	15
Critical items	5	5	4	3	4	4	1	1	1	1
Incongruencies	26/280 (9%)	12/180 (7%)	27/196 (14%)	17/117 (15%)	11/196 (6%)	9/126 (7%)	10/238 (4%)	10/162 (6%)	11/210 (5%)	14/135 (10%)
Errors (%)	23/280 (8%)	0	24/196 (12%)	19/117 (16%)	12/196 (6%)	31/126 (25%)	31/238 (13%)	12/162 (7%)	9/210 (4%)	15/135 (11%)
Mean difficulty (SD)	0.92 (0.07)	1.00 (0)	0.88 (0.14)	0.84 (0.17)	0.94 (0.09)	0.75 (0.15)	0.87 (0.16)	0.93 (0.10)	0.96 (0.11)	0.89 (0.18)
Mean general evaluation (SD)	2.11 (0.63)	3.17 (0.56)	2.14 (0.78)	1.78 (0.79)	2.50 (1.00)	2.00 (0.66)	1.89 (1.04)	2.78 (1.18)	2.36 (0.77)	2.83 (1.15)

No. of errors = the degree to which the two examiners both agreed the examinee did not perform item satisfactorily.
No. of Incongruencies = the number of errors which only one examiner indicated, but the two examiners did not agree.
% Error rate = the rate at which the two examiners both agreed the examinee did not perform the item satisfactorily.
% Incongruence rate = the rate at which the two examiners did not agree.
Mean difficulty = the proportion of examinees who performed an item well or satisfactory.
Mean general evaluation = mean holistic score across the two examiners, on a scale of 1 to 4, with 1 indicating insufficient and 4 indicating excellent performance.
Trauma = trauma management. Resus = resuscitation, OR = operating room crisis management, Vent = mechanical ventilation, Regional = regional anesthesia.

The pass rates in each OSCE station for the different test periods are presented in Table 3. The limited information available precludes the drawing of conclusions concerning success rate trends in the different stations. The pass rate ranged from 65% (trauma) to 91% (regional anesthesia). The overall pass rate (which is needed for admittance to the oral exams) for all six OSCE test periods is 70–80%. The passing rate for both OSCE and oral exams is 40–56%.

Table 3. Pass rate for the different examination stations by different test periods

Test period	Resuscitation	Trauma management	Crisis management in the OR	Mechanical ventilation	Regional anesthesia
1	23/34 (68%)	26/34 (76%)	–	–	–
2	19/26 (73%)	21/26 (81%)	–	–	–
3	11/21 (52%)	18/21 (86%)	16/21 (76%)	9/10 (90%)	10/11 (91%)
4	20/23 (87%)	17/23 (74%)	18/25 (78%)	16/23 (70%)	21/23 (91%)
5	15/20 (75%)	13/20 (65%)	15/20 (75%)	17/20 (85%)	14/20 (70%)
6	18/21 (86%)	13/21 (62%)	18/21 (86%)	18/21 (86%)	17/21 (81%)

Scores

The correlation between total (all items) and critical (critical items) was 0.58 ($P < 0.001$), between the total score (all items) and the global (general impression) score 0.72, and between critical and general 0.48.

Inter-rater reliability

The mean inter-rater correlations for the total score based on all checklist items, for the score based on the critical checklist items only, and for the global (general) rating were 0.89, 0.86 and 0.76, respectively. The inter-rater correlations for total, critical and global scores ranged from 0.75 to 0.81.

Inter-case reliability (internal consistency)

The inter-correlations were calculated between the five OSCE stations' scores for total and general scores. These correlations were significant ($P < 0.01$) only between trauma and ventilation for the total score ($r = 0.32$; n=63), and between resuscitation and regional and resuscitation and OR crisis for the global score ($r = 0.42$ and 0.27; n=64 and 104, respectively).

Criterion-related or convergent validity (comparing simulator scores with other evaluator forms)

The correlation between the OSCE examination scores and the success rate at eight different clinical domains of the oral board examination did not reach statistical significance ($P > 0.05$). The correlation r value between the OSCE examination scores and oral scores for the same examinees ranged from 0.14 to 0.54 (with total $r = 0.37$ between sum stations' scores).

Realism and satisfaction

According to a subjective feedback questionnaire, most (70–90%) participants found the difficulty level of the exam stations reasonable to very easy, and a minority (< 10%) did not prefer this method of examination to a conventional oral exam. The

realism (examinee's familiarity/comfort) was defined as high by the examinees (80–90%).

Efficacy

The simulator effect on error reduction was evaluated and found effective. All major errors that were identified in the initial two exam periods disappeared in the next two periods.

Discussion

This manuscript describes the process by which we validated and evaluated the reliability and realism of incorporating simulation-based objective structured clinical evaluation into the Israeli Board Examination in Anesthesiology.

In evaluating performance we need first to define what kind of performance is being evaluated. The way performance in anesthesiology is defined varies dramatically: from the vague ("Vigilance, data interpretation, plan formulation and implementation") [6] – to the much more organized, detailed and technical (preoperative evaluation, induction technique, intraoperative checks, postoperative management, airway assessment) [5,21]; from the combined performance of anesthesia non-technical skills, which include a variety of non-technical criteria [4] (knowledge of pre-op consideration, preparation, intra-op problem-solving), integrated with the technical skills demonstrated in the induction of anesthesia [22] – to completely separating the technical from the behavioral (decision-making and team interaction) [23]. Since we have built our exam on testing minimal requirement task performance, most of the tasks were within the framework of the basic knowledge and technical aspect. Only a few of the tasks were in the framework of ANTS (advanced knowledge and behavioral) (see Methods).

The process of developing the examination involved the definition of assessment conditions, tasks and scenarios on the basis of wide national consensus ensuring high content validity of the exam. The leadership of the national task force, supported and endorsed by the Scientific Council of the Israel Medical Association, and the standardized preparation of both examiners and examinees contributed not only to the fairness and objectivity of the exam but also to its positive reception in the Israeli anesthesiology community. This process was further supported by positive feedback from examiners and examinees who participated in the examination. The sense of realism, which was high (see Results), is a subjective rating of examinees regarding their familiarity or comfort. It is an important variable in validation of any assessment although it has no correlation with the actual performance score [6].

Psychometric evaluation conducted by experts from the National Institute for Testing and Evaluation was another major contributor to the fairness and objectivity of the examination. Comparison of the difficulty of the various scenarios was performed, incongruencies between examiners resulting from inadequate definition of accepted performance were highlighted and corrected, the incidence of common mistakes and errors during the examination was calculated, and information was shared with training program chairmen and the examinees.

In the validation process of our summative (testing) assessment we were not able to look for predictive validity, which is related to transferability (to clinical setup) and sustainability [19] (for a longer period), or for construct-related validity [5,6] (progression simulator scores with the level of training). However, we did look for the content validity (as described above, by the modified Delphi technique, introduced by Clayton [18]), and for the criterion-related or convergent validity comparing simulator scores with other forms of evaluation (traditional oral exams).

The reliability is a measure of the reproducibility or consistency of a test. One measure of reliability was the inter-rater reliability [5], comparing rater variance. In order to evaluate the internal consistency or the inter-case reliability we used inter-correlations between the five examination stations for the total (all items) and general (holistic) score. Thus, inter-case reliability was measured using inter-case correlation, similar to the way intra-class correlation coefficient reliability coefficient (a ratio of a variance to the total variance) is calculated, or instead of Cronbach-alpha statistics [17].

Other valuable psychometric information included the correlation between the different scores or assessment parameters (total, critical, and global scores), which is different from the inter-correlation between the stations (inter-case reliability), and from the correlation between the OSCE-based examination and the conventional oral board exam (criterion-related or convergent validity). The correlation between "proportions correct" score items across all items included in the checklist and the general evaluation (0.72) supported the conclusion that the evaluation techniques are similar but not identical, leading to the decision that the global (general) scores would be part of the pass/fail decision, with a low general score overriding the checklist score.

The low inter-correlations between the five stations support the conclusion that this examination has a limited degree of inter-case reliability. This limitation might be overcome by increasing the number of stations, or examinees. In order to assess this parameter properly in the future, more data would have to be collected from future examinations and more tasks from various additional clinical conditions incorporated.

The comparison of success rates on the newly developed OSCE and the conventional oral board examination (criterion-related or convergent validity) demonstrated a low correlation between the two modalities. Previous publications described the correlation between OSCE and written knowledge tests to be as high as 0.72 [24]. Others documented the correlation between simulation-based assessment and written tests at 0.19 [25], and between simulation-based assessment and faculty assessment (0.37), written examinations (0.44), and mock oral examinations (0.47), in the case of residents in anesthesiology [7]. These findings are similar to ours regarding low correlations between the simulator-based exam (see Results) and are not surprising in view of the fact that in this practical part of the exam we are evaluating certain functions of the examinee using the simulator-based assessment. These functions were: basic clinical

ANTS = anesthesia non-technical skills

performance, operating anesthesia machinery, interpretation of laboratory data, team work, and working under stress conditions. Those specific and unique functions are different from the functions tested in a conventional oral exam, namely priority and judgment in decision-making, application of scientific principles to clinical problems, adaptability to changing clinical situations, logical organization and effective presentation of information.
We contend that this novel and different modality constitutes a new examination tool. Thus, these medium-size correlations are commensurate with the assertion that different examination modalities assess different aspects of performance and, yet, are related to each other.

The process of incorporating OSCE-driven modalities in the certification of anesthesiologists in Israel is still incomplete and continuous evaluation and assessment is being undertaken. We hope that this new format of examination will play a formative (training) and a summative (testing) role, involving the Anesthesiology Board, anesthesia departments, the participating examiners and the examinees. The examination development process prompted a critical appraisal of the current training and assessment paradigm, and led to exploration, definition and prioritization of the critical clinical skills expected from a residency graduate. The examination also provided a rare glance at the authentic products of Israeli residencies – highlighting areas of strengths and weaknesses that could serve as guidelines to future modifications in the residency curriculum. The present process may evolve in the future not only as a constructive form of feedback for residency programs and means of establishing simulation-based training as part of the national residency curriculum, but also toward the adoption of full-scale simulation-based accreditation.

Acknowledgments. We thank all members of the Israeli Board Examination Committee in Anesthesiology and members of the Task Force of the Israeli Board Examination Committee in Anesthesiology.

References

1. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet* 2001;357:945–9.
2. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;65:S63–7.
3. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002;287:226–35.
4. Weller JM, Bloch M, Young S, et al. Evaluation of high fidelity patient simulator in assessment of performance of anaesthetists. *Br J Anaesth* 2003;90:43–7.
5. 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.
6. Devitt JH, Kurrek MM, Cohen MM, et al. Testing internal consistency and construct validity during evaluation of performance in a patient simulator. *Anesth Analg* 1998;86:1160–4.
7. Schwid HA, Rooke GA, Carline J, et al., for the Anesthesia Simulator Research Consortium. Evaluation of anesthesia residents using mannequin-based simulation: a multiinstitutional study. *Anesthesiology* 2002;97:1434–44.
8. Rosenblatt MA, Abrams KJ, for the New York State Society of Anesthesiologists, Inc; Committee on Continuing Medical Education and Remediation; Remediation Sub-Committee. The use of a human patient simulator in the evaluation of and development of a remedial prescription for an anesthesiologist with lapsed medical skills. *Anesth Analg* 2002;94:149–53.
9. Bergqvist D, Liapis C, Wolfe JNH. The developing European Board Vascular Examination. *Eur J Vasc Surg* 2004;27:339–40.
10. Hagen MD, Summer W, Roussel, Rovinelli R, Xu J. Computer-based testing in family practice certification and recertification. *J Am Board Fam Pract* 2003;16:227–32.
11. Grube C, Sinner B, Boeker T, Graf BM. The patient simulator for taking examinations – a cost effective tool? *Anesthesiology* 2001;95: A1202.
12. Henson LC, Richardson MG, Stern DH, Shekhter I. Using human patient simulator to credential first year anesthesiology residents for taking overnight call [Abstract]. 2nd Annual IMMS, 2002.
13. Schaefer JJ. Mandatory competency-based difficult airway management training at the University of Pittsburgh Department of Anesthesiology [Abstract]. 4th Annual IMMS, 2004.
14. Papadakis MA. The step 2 clinical skills examination. N *Engl J Med* 2004;350:1703–5.
15. Morgan PJ, Cleave-Hogg D. A worldwide survey of the use of simulation in anesthesia. *Can J Anaesth* 2002;49:659–62.
16. Berkenstadt H, Ziv A, Gafni N, Sidi A. Incorporating simulation-based OSCE into the Israeli National Board Examination in Anesthesiology. *Anest Analg* 2006;102(3):853–8.
17. Morgan PJ, Cleave-Hogg D, DeSousa S, Tarshis J. High-fidelity patient simulation: validation of performance checklists. *Br J Anaesth* 2004;92:388–92.
18. McLaughlin CR, Sheldon A, Hansen RC, McIver BA. Management uses of the Delphi. Health Care Manage Rev. 1976 Spring;1(2):51-62.
19. 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;38:45–55.
20. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. *Med Educ* 2004;38:199–204.
21. Murray DJ, Boulet JR, Kras JF, Woodhouse JA, Cox TC, McAllister JD. Acute care skills in anesthesia practice. A simulation-based resident performance assessment. *Anesthesiology* 2004;101:1084–95.
22. 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.
23. 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.
24. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998;73:993–7.
25. Morgan PJ, Cleave-Hogg D. Evaluation of medical students' performance using the anesthesia simulator. *Med Educ* 2000;34:42–5.

Correspondence: Dr. A. Sidi, Dept. of Anesthesiology and Intensive Care, Sheba Medical Center, Tel Hashomer 52621, Israel.
Fax: (972-3) 535-1565
email: asidi@anest.ufl.edu