Journal of Educational Evaluation for Health Professions

J Educ Eval Health Prof 2009, 6: 3 DOI: 10.3352/jeehp.2009.6.3

RESEARCH ARTICLE

Open Access

Patient Simulation: A Literary Synthesis of Assessment Tools in Anesthesiology

Alice A. Edler[1,*], Ruth G. Fanning[2], Michael. I. Chen[2], Rebecca Claure[2], Dondee Almazan[2], Brain Struyk[3], Samuel C. Seiden[2]

[1]Department of Graduate Medical Education, Stanford Hospitals and Clinics, Stanford, CA; [2]Department of Anesthesia, Stanford University School of Medicine, Stanford, CA; [3]Department of Anesthesia, Children's Hospital of Philadelphia, Philadelphia, PA, USA

Abstract

High-fidelity patient simulation (HFPS) has been hypothesized as a modality for assessing competency of knowledge and skill in patient simulation, but uniform methods for HFPS performance assessment (PA) have not yet been completely achieved. Anesthesiology as a field founded the HFPS discipline and also leads in its PA. This project reviews the types, quality, and designated purpose of HFPS PA tools in anesthesiology. We used the systematic review method and systematically reviewed anesthesiology literature referenced in PubMed to assess the quality and reliability of available PA tools in HFPS. Of 412 articles identified, 50 met our inclusion criteria. Seventy seven percent of studies have been published since 2000; more recent studies demonstrated higher quality. Investigators reported a variety of test construction and validation methods. The most commonly reported test construction methods included "modified Delphi Techniques" for item selection, reliability measurement using inter-rater agreement, and intra-class correlations between test items or subtests. Modern test theory, in particular generalizability theory, was used in nine (18%) of studies. Test score validity has been addressed in multiple investigations and shown a significant improvement in reporting accuracy. However the assessment of predicative has been low across the majority of studies. Usability and practicality of testing occasions and tools was only anecdotally reported. To more completely comply with the gold standards for PA design, both shared experience of experts and recognition of test construction standards, including reliability and validity measurements, instrument piloting, rater training, and explicit identification of the purpose and proposed use of the assessment tool, are required.

Key Words : *High-Fidelity Patient Simulation; Anesthesiology; Patient Simulation; Performance Assessment; Systemic Review; Test Theory*

INTRODUCTION

"Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality" Thorndike EL [1].

Professional education, medicine included, recognizes that for an expert, patient/client-centered practice, competencies beyond purely fact-based knowledge and technical skills are required [2, 3]. These competencies span multiple professional fields and include communication, clinical reasoning and decision making, and reflection in daily practice.

The medical education profession has formalized the com-

*Corresponding email: edlera@aol.com

Received: Oct 17, 2009, Accepted: Dec 12, 2009, Published: Dec 20, 2009

This article is available from: http://jeehp.org/

binations of knowledge, skills and attitudes (KSA) into six constructs, medical knowledge, patient care, communication and interpersonal skills, professionalism, practice based learning and improvement, and systems based medicine [4]. However, medical educators are now struggling with methods of authentic assessment for not only those factual knowledge and technical skills but also for the more psychologically based constructs of communication, life long learning and interdisciplinary reasoning.

As our concept of occupational competence develops, so must our assessment tools; beyond pencil and paper tests of knowledge recall, to include higher level cognitive and behavioral assessments [5]. Such assessments require challenging educational techniques for teaching and sophisticated psychometric methods for assessment. Medical education, in particular in anesthesiology has taken a lead from the aviation industry and now is an avid consumer of multiple forms of patient simulation-based teaching and performance assessment (PA), such as high-fidelity patient simulation (HFPS) and mixed-modality simulation (HFPS paired with other simulation techniques such as standardized patients), to assess in complex, simulated, life-like healthcare situations. As use of these PA tools increases, the methodologies used for test construction need to be complete and robust.

In this manuscript we have systematically reviewed the current methodological approaches to PA tool construction using HFPS. We have chosen to limit our review to the field of anesthesiology in order to unify our discussion of progress and because, anesthesiology the longest and most productive use of HFPS for PA.

The purpose of our review is as follows: to identify available HFPS PA tools used in anesthesiology, to comment on the quality of each in terms of classic and modern test theory, and to identify areas of needed research and possible means of standardization of test construction methods.

MATERIALS AND METHODS

Search strategies

We reviewed the literature to identify studies in which HFPS was used to test performance in anesthesiology practice and education. We used methods of literary synthesis research or systematic non-statistical meta-analysis of research literature described by Bland et al. [6] and Slavin [7]. Literary synthesis is useful in reviews not amenable to statistical meta-analysis, where dependent and independent variables vary from study to study and data are collected using non-statistically compatible instruments.

A professional medical librarian helped design the sensitive search strategy. Subject headings included: "((anaesth*[ti] OR anesth*[it]) AND simulate*[it]) OR (("Anesthesiology" [Mesh]) AND ("Computer Simulation" [Mesh])) OR (("Anesthesiology" [Mesh]) AND ("Patient Simulation" [Mesh] OR "Models, Educational" [Mesh])) OR (("Patient Simulation" [MeSH] OR "Computer Simulation" [MeSH]) AND (anesth* OR anaesth*) AND medical education)." The search was not limited to date of publication, publication type, or language. While preparing this manuscript, we used the "MY MCBI" updating protocol monthly, continuing to collect data through September thirtieth, 2008. Inclusion and exclusion criteria are listed in Table 1.

Table 1. Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
1. Identified search strategies of : "((anaesth*[ti] OR anesth*[it]) AND simulate*[it]) OR (("Anesthesiology" [Mesh]) AND ("Computer Simulation" [Mesh])) OR (("Anesthesiology" [Mesh]) AND ("Patient Simulation" [Mesh] OR "Models, Educational" [Mesh])) OR (("Patient Simulation" [MeSH] OR "Computer Simulation" [MeSH]) AND (anesth* OR anaesth*) AND medical education)."	1. Manuscripts not containing a performance assessment tool or outcome
2. All abstracts containing "performance, performance assessment, reliability studies, educational assessment, assessment, or competence (ies)" the complete citation was reviewed.	2. Manuscripts dealing with simulation other than high-fidelity patient simulation
3. All manuscripts identified through bibliographic review and cross check with the original search of manuscripts identified in #1 and #2.	3. Manuscripts dealing with disciplines other than general clinical anesthesiology
4. Abstracts and manuscripts of unpublished research, identified in #2-3, for which the primary author provided a copy of the manuscript.	4. All editorials
5. Letters to the editor meeting the above qualifications.	

Assessment tool analysis

To assess the quality of test construction consistent with recognized test construction methods [8, 9], we analyzed the following:

1) Methods of item selection, including degree of theory-grounded selection of test items, identification of the knowledge domain and skills or behaviors to be assessed.

2) Elements of test construction, including test piloting, rater training, multiple parallel scenarios/testing occasions, the use of varied scoring systems (analytic, holistic or other).

3) Measurement of the score reliability, using reliability indexes from both classic test theory and modern test theory (MTT) methods [10].

4) Degree of appropriateness of the inferences regarding examinees' ability made from these scores. Based on reported conclusions and stated validity claims in terms of standard psychometric definitions of content, criteria, and construct validity.

5) Practicability and usability of tools (Table 2).

Data extraction and analysis

AE and RF (the initials refer to the authors) reviewed the abstracts for all citations and identified manuscripts for full review if they matched the inclusion criteria listed in Table 1. Then a research assistant reviewed the manuscripts' bibliographies and crosschecked with the original search to identify additional citations not found in the initial search. For abstracts of unpublished research, the primary author was contacted and asked for a copy of the manuscript.

AE and RF reviewed and coded all selected manuscripts. Four secondary reviewers (MC, DA, BS, and RC) again reviewed and confirmed the findings of methods of item selection, elements of test construction, and type of statistical reliability measures. Secondary raters received instruction on statistical and qualitative methods of analysis. Any further disagreements were resolved by consensus among the primary and secondary reviewers.

Statistical analysis

We used SPSS (SPSS Inc., Chicago, IL, USA) to obtain descriptive statistics and Coefficients of Correlation for agreement between primary and secondary reviewers.

RESULTS

Investigators retrieved 412 articles, whose abstracts they and other members of the Department of Anesthesia at Stanford School of Medicine translated from Danish, German, Italian, and Japanese into English. Fifty studies met the inclusion criteria [11-60]. The Coefficient of Correlation for rater agreement between primary and secondary rater in the initial comparison was 0.76.

Methods of item selection

In our review, item selection methods varied considerably, including: round-table discussion, reported as modified Delphi techniques; task analysis; formal Delphi consensus of expert opinion, and internal consistency with items shown to produce reliably scores on previous tests. The most commonly reported item-selection method was round-table discussion among test designers, frequently termed "modified Delphi technique" coupled with items from previous published PA tools. Thirty eight percent of the studied reported this method for item selections. In sixteen percent of the studies, items were chosen only from previously published HFPS assessment tools with out modification of items and twenty eight percent came from exclusively from roundtable discussion among text designers. Ten percent used anonymous Delphi methods for item identification. Eight percent of the studies used formal Task Analysis. However, a valuable tool for item selection and refinement, item response theory (IRT), was not reported in any of these investigations. IRT allows test designers to determine the relative difficulty of test items, completeness of KSA domain sampling. Both of which are

Table 2. Reliability estimations: methods of agreement/reliability estimations and their uses

Statistic reported	Analytic method	Used for evaluation of
Coefficient of Stability (r)	Correlation	Test retest score correlation
Half split estimations (r)	Correlation	Subtest correlations
Half split estimations with Spearman Brown Prophecy (r)	Correlation	Subtest correlations corrected for length of test
Coefficient (Cronbach) alpha (α)	Analysis of variance	Individual item variance
Intra class coefficient (ICC)	Analysis of variance	Individual item variance
Kuder Richardson (KR 20)	Analysis of variance	Individual item variance for dichotomously scored items
Kappa (κ)	Correlation	Inter rater agreement
Generalizability (G) theory (g)	Analysis of variance	Individual and relative item variance

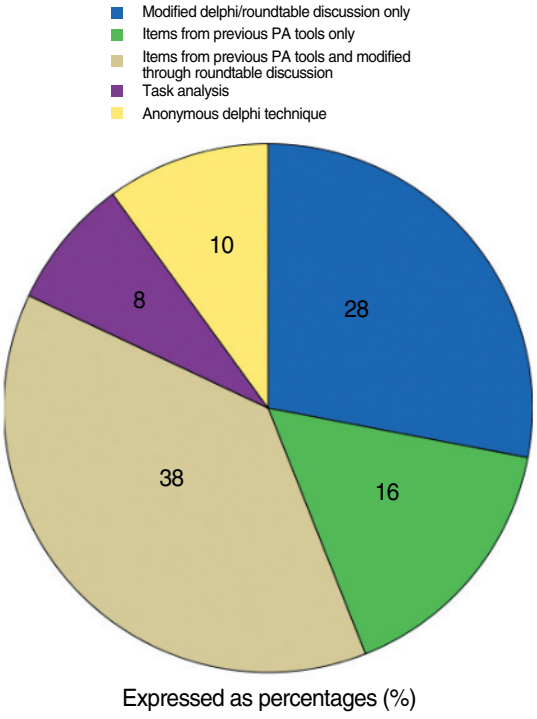

Fig. 1. Methods of item selection.

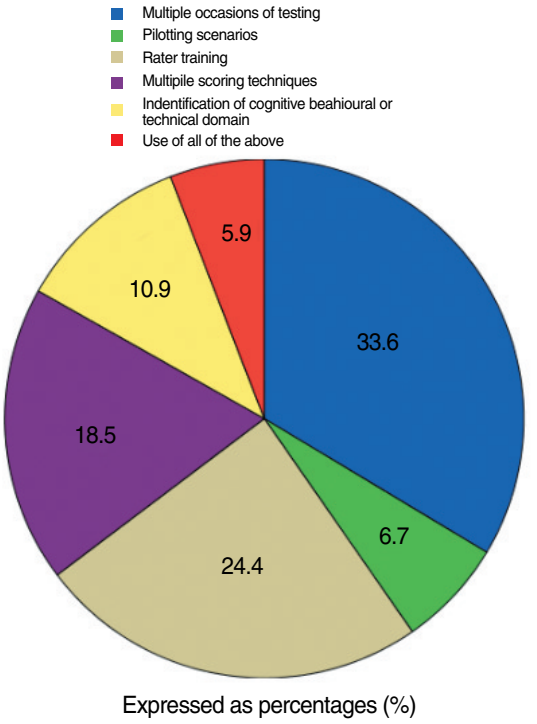

Fig. 2. Test refinement methods.

critical to the use of any PA for determination of minimal competency standards for credentialing (Fig. 1).

Elements of test construction

Performance assessments, which include multiple subtests, provide more information about the examinees true ability than those testing situations in which only one assessment is obtained. In forty percent of the studies multiple scenario/occasions of testing were the only method of item refinement used (Fig. 2).

Twenty eight percent of the studies reported rater training prior to actual scoring; Twenty-one studies used multimodal scoring techniques, of which the most common of which were analytic checklists or holistic rubrics for performance. Only fourteen percent of the investigators used scenario piloting to identify problems within the scenario itself.

Many of the investigators, twenty eight percent, used 2 or more methods for item refinement. The most common combination of techniques was multiple scenarios with rater training. However only seven percent used all of the above methods listed.

Measurement of the score reliability

Inter-rater agreement as the sole measure of reliability was noted in twenty four percent of the studies; most of these studies were preformed prior to the year 2000. A variety of rater agreement statistics were used. The most common was

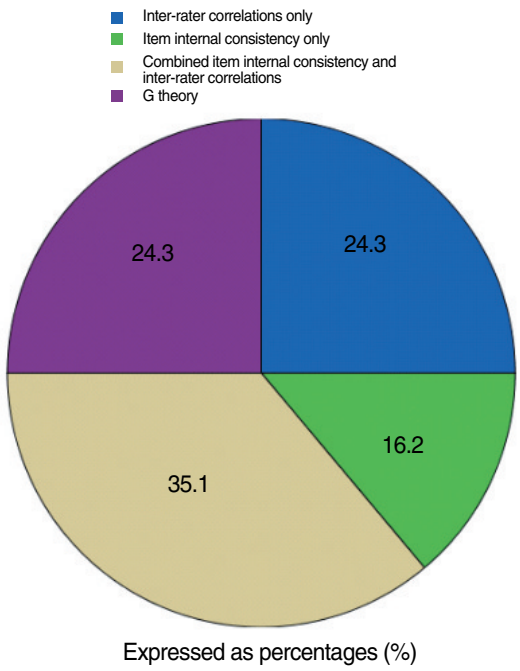

Fig. 3. Score reliability measures.

intra-class correlations for variance estimations. Other methods used included Kappa for rater agreement, Pearson's correlation, and simple percentage agreement. The degree of rater agreement varied moderately between studies (0.56-0.99). However, not all studies documented rater training, and it was not possible to examine the relationship between rater training and subsequent rater agreement statistics (Fig. 3).

Thirty five percent of the studies reported using measures of internal consistency of items or subtests along with inter-rater agreement. When we examined the methods used to analyze score reliability from individual items, again, the most common estimation was intra-class correlations. For the most part, these reliability results were moderate.

Only sixteen percent of the studies used MTT, in particular generalizability theory (G theory), to examine the relative internal consistency of items, including the interactions between raters, occasions of testing, and/or other covariates. G theory, unlike correlationally-based intra class coefficient (ICC) or Kappa, derives from analysis of variance, and it can statistically describe the individual components of score error that arise separately from the examinee, the raters, the test items or any number of other confounding conditions that may contribute to score error. Within the last reviewed year, 2007-2008, all published studies describing score reliability estimations have included G theory, vastly improving our ability to discriminate sources of error and assure that differences in scores are truly from differences in examinees' abilities.

An even greater value of G theory is factor analysis, which allows a second type of analysis, decision-making studies (D study). D studies estimate the reliability of the score if any of the sources of that score error are changed. For example, an increase or decrease in the number of raters, occasions of testing and so forth. D studies allow test designers to assure acceptable reliability measures prior to testing rather than post hoc as with ICC. However, only two studies used G theory decision-making analysis (D studies) to maximize the reproducibility of their scoring systems pre-testing.

For the most part over the years in which the studies were performed, we noted a progressive improvement of score-reliability measurements, including the use of combined rater and internal item consistency statistics or the use of analysis of score variance through Generalizability statistics.

Degree of appropriateness of the inferences regarding examinees' ability made from these scores

The second essential consideration for evaluating the quality of an assessment tool is validity; an attribute of the inferences about the examinees derived from the test scores and not the test, itself. Tests produce valid scores if the inferences about the examinee's ability made from those scores are correct. Though definitions of validity are currently evolving, for the purposes of this review, the authors will use classic validity definitions of criteria, content, or construct [61].

Early studies were limited to comments on *face-validity* conclusions, more recent investigators have reported *content, criteria,* and construct validity conclusions [10]. The most common method used to assess content validity, the adequate and complete representation of subject-matter content in test items, was expert opinion through round-table discussion or "modified Delphi technique". As described "modified Delphi methods" only roundtable discussion of items. Several investigators attempted to demonstrate content validity by comparing their tests with previously identified subject-matter-based tests used. On the whole, PA tools whose resultant scores were compared with only paper and pencil test scores faired poorly [35, 40, 48]. Only when a broader view of competency was taken to include both higher-level cognitive skills and technical skills [43-45], the agreement was improved.

Criteria validity, either concurrent or predictive, is the ability of the resultant scores to correspond with scores from other recognized assessments of similar KSA's. Criteria validity is used to assure that inferences about the abilities that the examinee currently demonstrates or will demonstrate are correct. Investigators in this review, most frequently reported concurrent criteria validity, matching simulation-based assessment scores with level of clinical anesthesia training. The results of these correlations were moderate to strong for criteria-based inferences [15, 20, 33, 42, 43, 45, 49, 50, 52]. This ability of simulation-based assessment to discriminate between levels of training was demonstrated but limited. Most simulation-based assessment scores could distinguish between early trainees and academic faculty, and some could distinguish between levels of anesthesia training or other professional anesthesia providers but not consistently. None of the reviewed assessment tools reported item difficulty indexes.

Construct validity of score inferences is the most difficult to conceptualize and assess; currently the very concept of construct validity is under question. Cronbach and Meehl [61] defines construct validity as "the ability to infer correct qualities [sic, of the examinee] which are not operationally defined". The difficulty lies in the fact that some constructs, such as teamwork, communication, or professionalism, are strongly influenced by culture, gender, or professional identity and cannot be easily and universally operationalized. Fletcher et al. [24] and Weller [58] have provided excellent models of construct validation in behaviorally-based assessment tools. In these studies, elements of teamwork were identified by task analysis and then examined statistically through factor analysis to see if each item correlated with others and with the test as a whole. All estimations of the final scoring systems displayed good to excellent psychometric qualities.

Practicability and usability of tools

The final two quality characteristics of quality test construction-practicality and usability-are external to the tool itself but juxtaposed [62, 63]. Though many studies reported likeability of instruments, the use of scenario piloting, rater training and multiple question formats to improve usability was not consistent. The information on the practicality of HFPS in was limited and contained in editorial comments about the cost/demands of an HPFS center and found principally in the non-anesthesiology literature [64-68]. We found no explicit cost/benefit literature in anesthesiology.

Finally tests are designed for a purpose, to identify areas of needed improvement (formative tests) or to assess minimally acceptable competence (summative tests). We noted that manuscripts published between 2007 and 2008, substantially improved in reporting the purpose of the test. Two excellent examples are use of HFPS PA as a method to improve the curriculum [59] and the use of HFPS testing for the determination of minimal competence or "cut scores" for summative assessment of examinees' ability [16].

DISCUSSION

We have found progressive and noteworthy improvement in quality of performance testing using anesthesiology-based HFPS over the past two decades with dramatic increases in the quality of item selection and test construction in the published tools since 2007. Since 2007, there appears to be a more universal acceptance of standard PA tools construction methods. Techniques for careful item selection and minimization of bias through piloting, rater training and multiple subtests/scenarios are also improving but inconsistently. For example, Task Analysis remains the gold standard for identifying skills and attitudes. However few of the studies employed Task Analysis for item selection [25, 30, 48, 54].

Likewise score reliability measures are improving, but the relatively weak measures of internal consistency when comparing scores across varied subtest/scenarios raises the question if examiners are choosing scenarios, which assess the same KSA across these varied subtest/scenarios. As an example, KSA needed for the correct management of ventilator settings in the patient with lung disease are not necessarily the same KSA needed for management of team efforts in the acute treatment of trauma from motor accident. This maybe the cause of lower correlations between scores on differing subtest/scenarios as seen in several of the studies. A more careful look is needed to identify testing scenarios that contain equivalent KSA's assessments. Here the introduction of MTT and IRT maybe of great help.

The recent literature shows a greater use of G theory for reliability estimates but still a severe underuse of D study estimations. D study estimations, though not statically a sample size estimation but conceptually similar, essentially improve the "bang for the buck" when choosing the number of raters, items, testing scenarios etc. A very important feature in these labor and cost intensive HFPS performance assessments.

In another encouraging work, investigators in Israel found reliable inter rater agreement between US and Israeli raters when using the same scenario set. This finding suggests that with proper PA construction and the use of decision-making studies to minimize score error, scenarios and PA tools may be shared in similar practice venues; a particularly pertinent point, as the construction of highly reliable and valid assessment tools that are practical to develop and cost efficient to implement. Developing such tests/tools that are universally applicable and shareable throughout the medical community will be invaluable.

Validity issues still plague the HFPS performance assessment literature. Several studies do demonstrate concurrent validity but only at a gross level, novice verses experienced practitioner. The reasons for this are unclear, however may include incomplete domain sampling. Other possible explanations include inappropriate scaling, low discrimination indexes of the items, or non-linear average item difficulty.

One fascinating point, which bridges both criteria and construct validity issues, was the finding by Devitt et al. [22] that simulation based performance assessment can differentiate between academic anesthesia trainees and their faculty but not between faculty and their private practitioners counterparts with equal levels of practice experience. This raises the question whether different sets of KSA are needed for academic anesthesia practice where trainees are present as compared with practices in which trainees are not. And the more profound question: if performance assessment tools used for credentialing are created by academic anesthesiologists, are the scores obtained from these tools equally valid for non-academic anesthesiologists? This perhaps is one of the most important issues to address prior to the use of simulation based PA in professional credentialing.

The exponential growth of HFPS for PA in anesthesiology during the past two decades has resulted in greater expertise in both test construction and execution. However, although the quality of PA tools has dramatically improved, they need further refinement. For example, test-construction methods, rater training, and piloting and scripting of scenarios should be standardized and uniformly applied. Combined scoring systems should address not only a checklist of technical skills but also global latent trait measurement. In addition, MTT could diminish sources of error variance and gaps in item discrimination ability. The relevance of educational and assessment methods has moved beyond the realm of me-

dical educationalists into the realm of mainstream practice. As we all face assessment of competence in dynamic environments over the next few decades, it is pertinent that we ensure the validity and reliability of test scores to adequately reflect examinees' true competence in the practice of medicine.

CONFLICT OF INTEREST AND DISCLAIMERS

There is no conflict of interest, no disclaimers and no financial support recordable.

REFERENCES

1. Thorndike EL. An introduction to the theory of mental and social measurement. New York: Science Press; 1904.
2. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002;287:226-35.
3. McClelland DC. Testing for competence rather than for "intelligence". Am Psychol 1973;28:1-14.
4. Accreditation Council for Graduate Medical Education. Outcomes Project [Internet]. Chicago (IL): Accreditation Council for Graduate Medical Education; c2009 [cited 2009 Jan 20]. Available from: http://www.acgme.org/outcome/.
5. Tetzlaff JE. Assessment of competency in anesthesiology. Anesthesiology 2007;106:812-25.
6. Bland CJ, Meurer LN, Maldonado G. A systematic approach to conducting a non-statistical meta-analysis of research literature. Acad Med 1995;70:642-53.
7. Slavin RE. Perspectives on evidence-based research in education: What works? Issues in synthesizing educational program evaluations. Educ Res 2008;37:5-14.
8. Fritz PZ, Gray T, Flanagan B. Review of mannequin-based high-fidelity simulation in emergency medicine. Emerg Med Australas 2008; 20:1-9.
9. Howard SK, Gaba DM, Fish KJ, Yang G, Sarnquist FH. Anesthesia crisis resource management training: teaching anesthesiologists to handle critical incidents. Aviat Space Environ Med 1992;63:763-70.
10. Crocker LM, Algina J. Introduction to classical and modern test theory. 2nd ed. Victoria: Thomson Wadsworth; 2006.
11. Berkenstadt H, Kantor GS, Yusim Y, et al. The feasibility of sharing simulation-based evaluation scenarios in anesthesiology. Anesth Analg 2005;101:1068-74.
12. Berkenstadt H, Ziv A, Gafni N, Sidi A. The validation process of incorporating simulation-based accreditation into the anesthesiology Israeli national board exams. Isr Med Assoc J 2006;8:728-33.
13. Blike GT, Christoffersen K, Cravero JP, Andeweg SK, Jensen J. A method for measuring system safety and latent errors associated with pediatric procedural sedation. Anesth Analg 2005;101:48-58.
14. Blum RH, Raemer DB, Carroll JS, Dufresne RL, Cooper JB. A method for measuring the effectiveness of simulation-based team training for improving communication skills. Anesth Analg 2005;100: 1375-80.
15. 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.
16. Boulet JR, Marray D, Kras J, Woodhouse J. Setting performance standards for mannequin-based acute-care scenarios: an examinee-centered approach. Simul Healthc 2008;3:72-81.
17. Byrne AJ, Jones JG. Responses to simulated anaesthetic emergencies by anaesthetists with different durations of clinical experience. Br J Anaesth 1997;78:553-6.
18. Byrne AJ, Sellen AJ, Jones JG. Errors on anaesthetic record charts as a measure of anaesthetic performance during simulated critical incidents. Br J Anaesth 1998;80:58-62.
19. Chopra V, Gesink BJ, de Jong J, Bovill JG, Spierdijk J, Brand R. Does training on an anaesthesia simulator lead to improvement in performance? Br J Anaesth 1994;73:293-7.
20. Devitt JH, Kurrek MM, Cohen MM, Cleave-Hogg D. The validity of performance assessments using simulation. Anesthesiology 2001; 95:36-42.
21. Devitt JH, Kurrek MM, Cohen MM, et al. Testing the raters: inter-rater reliability of standardized anaesthesia simulator performance. Can J Anaesth 1997;44:924-8.
22. 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.
23. Farnsworth ST, Egan TD, Johnson SE, Westenskow D. Teaching sedation and analgesia with simulation. J Clin Monit Comput 2000;16: 273-85.
24. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. Br J Anaesth 2003;90:580-8.
25. 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.
26. Gaba DM, DeAnda A. The response of anesthesia trainees to simulated critical incidents. Anesth Analg 1989;68:444-51.
27. 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.
28. Harrison TK, Manser T, Howard SK, Gaba DM. Use of cognitive aids in a simulated anesthetic crisis. Anesth Analg 2006;103:551-6.
29. Howard SK, Gaba DM, Smith BE, et al. Simulation study of rested versus sleep-deprived anesthesiologists. Anesthesiology 2003;98: 1345-55.
30. Hunt EA, Walker AR, Shaffner DH, Miller MR, Pronovost PJ. Simulation of in-hospital pediatric medical emergencies and cardiopulmonary arrests: highlighting the importance of the first 5 minutes. Pediatrics 2008;121:e34-43.
31. Jacobsen J, Lindekaer AL, Ostergaard HT, et al. Management of anaphylactic shock evaluated using a full-scale anaesthesia simulator. Acta Anaesthesiol Scand 2001;45:315-9.

32. Johnson KB, Syroid ND, Drews FA, et al. Part Task and variable priority training in first-year anesthesia resident education: a combined didactic and simulation-based approach to improve management of adverse airway and respiratory events. Anesthesiology 2008; 108:831-40.
33. Lorraway PG, Savoldelli GL, Joo HS, Chandra DB, Chow R, Naik VN. Management of simulated oxygen supply failure: is there a gap in the curriculum? Anesth Analg 2006;102:865-7.
34. Morgan PJ, Cleave-Hogg D. Evaluation of medical students' performances using the anesthesia simulator. Acad Med 1999;74:202.
35. Morgan PJ, Cleave-Hogg D. Evaluation of medical students' performance using the anaesthesia simulator. Med Educ 2000;34:42-5.
36. Morgan PJ, Cleave-Hogg D, Desousa S, Lam-McCulloch J. Applying theory to practice in undergraduate education using high fidelity simulation. Med Teach 2006;28:e10-5.
37. Morgan PJ, Cleave-Hogg D, DeSousa S, Tarshis J. Identification of gaps in the achievement of undergraduate anesthesia educational objectives using high-fidelity patient simulation. Anesth Analg 2003; 97:1690-4.
38. Morgan PJ, Cleave-Hogg D, DeSousa S, Tarshis J. High-fidelity patient simulation: validation of performance checklists. Br J Anaesth 2004;92:388-92.
39. Morgan PJ, Cleave-Hogg D, Guest CB. A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. Acad Med 2001;76:1053-5.
40. 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.
41. Morgan PJ, Pittini R, Regehr G, Marrs C, Haley MF. Evaluating teamwork in a simulated obstetric environment. Anesthesiology 2007; 106:907-15.
42. Murray D, Boulet J, Ziv A, Woodhouse J, Kras J, McAllister J. An acute care skills evaluation for graduating medical students: a pilot study using clinical simulation. Med Educ 2002;36:833-41.
43. Murray DJ, Boulet JR, Avidan M, et al. Performance of residents and anesthesiologists in a simulation-based skill assessment. Anesthesiology 2007;107:705-13.
44. Murray DJ, Boulet JR, Kras JF, McAllister JD, Cox TE. A simulation-based acute skills performance assessment for anesthesia training. Anesth Analg 2005;101:1127-34.
45. 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.
46. Olympio MA, Whelan R, Ford RP, Saunders IC. Failure of simulation training to change residents' management of oesophageal intubation. Br J Anaesth 2003;91:312-8.
47. Ringsted C, Ostergaard D, Ravn L, Pedersen JA, Berlac PA, van der Vleuten CP. A feasibility study comparing checklists and global rating forms to assess resident performance in clinical skills. Med Teach 2003;25:654-8.
48. Rosenblatt MA, Abrams KJ. 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.
49. Savoldelli GL, Naik VN, Joo HS, et al. Evaluation of patient simulator performance as an adjunct to the oral examination for senior anesthesia residents. Anesthesiology 2006;104:475-81.
50. Scavone BM, Sproviero MT, McCarthy RJ, et al. Development of an objective scoring system for measurement of resident performance on the human patient simulator. Anesthesiology 2006;105:260-6.
51. Schwid HA, O'Donnell D. Anesthesiologists' management of simulated critical incidents. Anesthesiology 1992;76:495-501.
52. Schwid HA, Rooke GA, Carline J, et al. Evaluation of anesthesia residents using mannequin-based simulation: a multiinstitutional study. Anesthesiology 2002;97:1434-44.
53. Schwid HA, Rooke GA, Ross BK, Sivarajan M. Use of a computerized advanced cardiac life support simulator improves retention of advanced cardiac life support guidelines better than a textbook review. Crit Care Med 1999;27:821-4.
54. Steadman RH, Coates WC, Huang YM, et al. Simulation-based training is superior to problem-based learning for the acquisition of critical assessment and management skills. Crit Care Med 2006;34:151-7.
55. Weller J, Merry A, Warman G, Robinson B. Anaesthetists' management of oxygen pipeline failure: room for improvement. Anaesthesia 2007;62:122-6.
56. 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.
57. Weller JM, Jolly B, Robinson B. Generalisability of behavioural skills in simulated anaesthetic emergencies. Anaesth Intensive Care 2008; 36:185-9.
58. Weller JM, Robinson BJ, Jolly B, et al. Psychometric characteristics of simulation-based assessment in anaesthesia and accuracy of self-assessed scores. Anaesthesia 2005;60:245-50.
59. Wheeler DW, Degnan BA, Murray LJ, et al. Retention of drug administration skills after intensive teaching. Anaesthesia 2008;63:379-84.
60. Yee B, Naik VN, Joo HS, et al. Nontechnical skills in anesthesia crisis management with repeated exposure to simulation-based education. Anesthesiology 2005;103:241-8.
61. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull 1955;52:281-302.
62. Shavelson RJ, Webb NM. Generalizability theory: a primer. Newbury Park: Sage Publications Inc.; 1991.
63. Wigdor AK, Green BF. Performance assessment for the workplace. Washington (DC): National Academy Press; 1991.
64. Hravnak M, Tuite P, Baldisseri M. Expanding acute care nurse practitioner and clinical nurse specialist education: invasive procedure training and human simulation in critical care. AACN Clin Issues 2005;16:89-104.
65. Kurrek MM, Devitt JH. The cost for construction and operation of a simulation centre. Can J Anaesth 1997;44:1191-5.
66. Lampotang S, Good ML, Westhorpe R, Hardcastle J, Carovano RG. Logistics of conducting a large number of individual sessions with a

full-scale patient simulator at a scientific meeting. J Clin Monit 1997; 13:399-407.
67. Seropian MA. General concepts in full scale simulation: getting started. Anesth Analg 2003;97:1695-705.
68. Tuoriniemi P, Schott-Baer D. Implementing a high-fidelity simulation program in a community college setting. Nurs Educ Perspect 2008; 29:105-9.