

Patient Simulator Competency Testing: Ready for Takeoff?

Patricia A. Kapur, MD, and Randolph H. Steadman, MD

Department of Anesthesiology, University of California, Los Angeles, California

The number of whole-body computerized patient simulators in use worldwide continues to grow. More than 70 such simulators are now active for training purposes, compared with less than 20 simulators 2 yr ago. Approximately one third of the currently available simulators are in United States academic anesthesiology training programs. The remainder are in allied health programs, such as respiratory therapy, paramedic, or nursing programs, in the United States, or are located at international sites, primarily anesthesiology programs, such as the site of the study conducted by Devitt et al. (1), published in this issue of *Anesthesia & Analgesia*. Two commercially distributed simulator models make up the majority in use, but several venues, primarily in Europe, have developed their own units on site.

The rapidly growing interest in potential applications of patient simulator technology in the field of anesthesiology is also evidenced by the recent appointment of an *ad hoc* committee on simulators to investigate the current state of the art for the American Society of Anesthesiologists. Obvious potential applications include training, research, and competency testing. (2-4) It is the latter that undoubtedly raises the most anxiety among practitioners, as questions of consistency, reproducibility, validity, and relevance of responses in the simulator environment to true performance in the clinical arena all have yet to be answered. Which testing methodology will best predict clinical performance versus simply representing understanding of the limitations and familiarity with the simulator milieu? Will patient simulator testing (real-time coordination of knowledge with eye-hand responses) evolve into a third component, along with written (fundamental body of factual knowledge) and oral (integrative application of knowledge) competency testing methods? These and other questions are far from being answered.

Certainly, simulators have a wide range of teaching applications. In our institution, under the direction of

one of the authors (RHS), the whole-body patient simulator is used for critical incident training of advanced anesthesiology residents, for introduction to anesthesiology clinical care for first-month anesthesiology residents, for introduction to clinical hemodynamic and respiratory management of the perioperative patient for medical students rotating on anesthesiology electives, for hemodynamic and pulmonary physiology modules for physiology courses for all medical students, for clinical management scenarios for pre-matriculation medical students before enrollment to whet their appetites for clinical medicine, for critical care modules for the nursing staff, and as a component of advanced cardiac life support (ACLS) training. It is also being proposed for special clinical exercises for internal medicine, surgery, or emergency medicine trainees. Of all of these training activities, the only aspect that encompasses an evaluation with outcome consequences is the observation of the correct responses for the megacode portion of ACLS training, on which basis the granting of the ACLS certificate in part depends.

What are some of the reasons that evaluation of performance in response to simulated perioperative patient scenarios is problematic at this time? First and foremost is the difficulty of developing appropriate testing scenarios, which is the focus of the article by Devitt et al.(1) This difficulty is part of the nature of critical care itself, of which perioperative medicine is one aspect. The timecourses of the interactions between patient and clinician are often very rapid. Therapeutic approaches to the "best" treatment may be adjusted midstream if original premises regarding problem etiology are not borne out by subsequent patient responses.

It is quite a challenge to script this aspect of patient presentation, intervention, modified presentation, second intervention, additionally modified presentation, modified intervention, etc., for the instructor running the simulator. For simplicity, most simulator scenarios are designed to highlight only one problem at a time. For training purposes, this approach has merit, because it allows the instructor to focus on the problem, its etiology, the typical patient presentation, and then a hierarchy of potential responses with the subtleties

Accepted for publication February 20, 1998.

Address correspondence to Patricia Kapur, MD, Department of Anesthesiology, UCLA School of Medicine, Box 951778, Los Angeles, CA 90095-1778.

of an algorithm of when to choose each type of therapeutic response. However, in an evaluative situation, this scenario design paradigm could conceivably lead to the test candidate being perfectly able to solve isolated problems in the simulator environment, yet not being competent in the real world of multifaceted, overlapping, simultaneous clinical derangements. This would yield a so-called false-positive result, in which such candidates would be deemed competent by simulator testing, when in fact they may not be able to perform satisfactorily in the less straightforward real world.

The simulator environment may prove to be intimidating to candidates at first. The presence of video cameras, evaluators, scripted roles for co-actors in the scenes, and limited flexibility of the programmed scenarios to accept alternative therapeutic pathways or alternative thought processes that avoid harm and achieve an acceptable patient outcome, could potentially lead to a false-negative test, in which a candidate could be deemed incompetent in the simulator situation, yet be an entirely acceptable clinical anesthesiologist under less artificial conditions.

Some of these issues were faced by Devitt et al. The exact scripts of the five-problem scenarios were not provided, just the topics (A1-A5 and B1-B5 in their Table 1), along with the information that the order of the five topics for each of the two scenarios was scrambled among the participants. The exact presentation of each problem certainly can affect the ability of the candidate to understand the point of the exercise. Because the problems were presented one at a time and in different order among the examinations, could there have been some clinical flow inconsistencies from one problem to another in each examination, so that the candidates may have had difficulty in recognizing the logic of the next development? Those who participate in simulated oral examinations realize all too clearly that smooth transitions from topic to topic are critical for avoiding an "intellectual disconnect" in the examinee if a seemingly illogical subsequent event is presented. If an examiner is not attentive to transition elements, a resultant loss of time in the examination can result, as the examiner must retreat to smooth out the disjointed presentation. In the study by Devitt et al., it is not clear whether the testing methodology took into account the swiftness of the response, although it seems not to have. Nevertheless, randomization of the order of presentation of problem events without a clinical logic to their progression could confuse qualified clinicians who would expect a closer correlation to a true clinical evolution of events.

The difficulty of constructing reproducible scenarios is further highlighted in this work by the fact that 40% of the problems performed poorly and had to be dropped from analysis just to obtain a Cronbach's coefficient α of only 0.66. The fact that the faculty

scored worse than the residents on the identification and treatment of such a critical anesthetic complication as coronary ischemia could potentially lower the confidence of the reader in the authors' definition of the validity of the testing methodology, which was defined as the superior scoring by experienced anesthesiologists compared with trainees. Experienced simulator instructors, however, note that performance can be uneven among individuals and that poor choices can be made during simulator scenarios, even by very experienced practitioners. Thus, faculty performing better or worse than residents neither proves nor disproves the validity of the test. The reasons for the differences (e.g., inadequate recognition of the problem versus inadequate response among varying years since training) may, however, be important findings of a simulator examination. Nevertheless, the ultimate test of validity, i.e., whether simulator testing has anything to do with true patient care outcomes, requires a large-scale analysis comparing the clinical outcomes of candidates who have met the success criteria of a simulator evaluation versus the clinical outcomes of candidates who did not meet the success criteria of a simulator evaluation.

It is also interesting that test candidates performed poorly on Problem B5, which was a presentation of anuria for which obtaining the maximal score required the simple unkinking of a urinary catheter. It is entirely possible that the simulator-naive test subjects were predisposed to believe that the expected therapeutic choices involved only medication or anesthesia machine manipulation, which are indeed the prominent features of the simulator environment. It is interesting, however, that the test candidates also scored poorly on the machine fault problems. Perhaps the realm of possible problems to be encountered was not fully appreciated by the test candidates in this study.

It is not clear whether this scoring methodology is truly appropriate for the authors' definition of internal consistency, i.e., "all items...measure the same attribute...internal test items act as a series of repeated measures of the same factor..." Because the score choices were "no action," "compensating intervention," or "definitive measurement," it does not necessarily follow that a resident who cannot recognize a missing inspiratory valve would be consistent by not being able to recognize and appropriately treat anuria. Missing valves and anuria are not related pieces of knowledge; thus, a trainee candidate could very well receive a 0 score on B1 and receive a score of 2 on B5 without any inconsistency whatsoever. The premise of this analysis would not yield useful information about progress in training if consistency of answers across widely varying topics is required. On the other hand, if all topics must be equally mastered by those in practice, consistency that a correct answer to one topic would lead to an expectation of a correct answer to

other topic may be a reasonable construct with regard to those who have already completed training. On this issue, the authors have not addressed the matter of consistency within the faculty versus within the trainees, but the numbers of candidates may be too small for a breakout analysis of this nature.

An entire additional set of questions for those designing future studies on the topic of simulator competency evaluation is further related to the appropriateness of the evaluation scoring methodology. How many scorers are needed? Is it appropriate to compute an arithmetic average of multiple evaluators' scores, or is being above a threshold score on each scorer's tally sufficient? Should some items be more heavily weighted than others (e.g., omissions on cardiovascular/respiratory issues carry more significance to the pronouncement of a successful examination result)? Must the scorers be blinded to the identity and level of experience of the candidates?

A very important question is whether it is ever inappropriate for the scorers to be participants in any way in the test administration. There are many ways that participating evaluators could already be making subtle judgements about the adequacy of the test candidate's responses during the test that could affect the outcome. This could occur if: 1) scorers were involved in the test by playing the roles of the surgeons or nurses, with the potential to provide more or less helpful ancillary information or positive/negative body language during the test progress; 2) the evaluator were operating the scenario computer program and selecting different responses for the mannequin; or 3) an evaluator was also the person who prompts the actors in different ways by means of the radio frequency communication system. It would seem that, similar to cardiovascular studies using blind cardiologists to read Holter monitor tapes after the fact, blind simulator evaluators should read the videotapes *post hoc* if such simulator competency testing may be used to limit clinical privileges.

In summary, the article by Devitt et al. (1) serves to highlight many of the challenges facing investigators studying human performance issues in the specialty of anesthesiology and related multidisciplinary fields of critical life support, focusing on the potential applications of the newly available patient simulators to benefit patient care. The improved realism made possible by the new generation of simulators has expanded the scope of possible individual and team training opportunities that have already been enthusiastically embraced around the globe.

However, as this article also brings to light, the role of such patient simulators in clinical competency evaluation, insofar as performance in the simulator environment predicts future patient outcomes, is not at all clear. Much future research is needed for the development of predictive patient problem solving scenarios, objective test administration strategies, distinction in testing methodology for the evaluation of progress in training versus the evaluation of adequacy of performance in the years posttraining, and unbiased scoring methodology that recognizes that performance in the simulator setting encompasses a multitude of cognitive, deductive, inferential, intuitive, and psychomotor skills, some of which are difficult to quantify.

References

1. Devitt JH, Kurrek MM, Cohen MM, et al. Testing internal consistency and construct validity during evaluation of performance in an anesthesia simulator. *Anesth Analg* 1998;86:1160-4.
2. Gaba DM, DeAnda A. The response of anesthesia trainees to simulated critical incidents. *Anesth Analg* 1989;68:444-51.
3. DeAnda A, Gaba DM. Role of experience in the responses to simulated critical incidents. *Anesth Analg* 1991;72:308-15.
4. Chopra V, Gesink BJ, deJong J, et al. Does training on an anaesthesia simulator lead to improvement in performance? *Br J Anaesth* 1994;73:293-7.