ECONOMICS AND HEALTH SYSTEMS RESEARCH
SECTION EDITOR
RONALD D. MILLER

A Comparison of the Training Value of Two Types of Anesthesia Simulators: Computer Screen-Based and Mannequin-Based Simulators

Anne-Sophie Nyssen, PhD*, Robert Larbuisson, MD†, Marc Janssens, MD†, Philippe Pendeville, MD‡, and Alain Mayné, MD‡

*Department of Work Psychology, University of Liège; †Department of Anesthesia and Intensive Care Medicine, CHU Sart-Tilman, Liège; and ‡Department of Anesthesiology, Cliniques Universitaires St. Luc, Bruxelles, Belgium

In this study, we compared two different training simulators (the computer screen-based simulator versus the full-scale simulator) with respect to training effectiveness in anesthesia residents. Participants were evaluated in the management of a simulated preprogrammed scenario of anaphylactic shock using two variables: treatment score and diagnosis time. Our results showed that simulators can contribute significantly to the improvement of performance but that learning in treating simulated crisis situations such as anaphylactic shock did not significantly vary between full-scale and computer screen-based simulators. Consequently, the initial decision on whether to use a full-scale or computer screen-based training simulator should be made on the basis of cost and learning objectives rather than on the basis of technical or fidelity criteria. Our results support the contention that screen-based simulators are good devices to acquire technical skills of crisis management. Mannequin-based simulators would probably provide better training for behavioral aspects of crisis management, such as communication, leadership, and interpersonal conflicts, but this was not tested in the current study.

(Anesth Analg 2002;94:1560–5)

The use of full-scale (FS) simulators for training in anesthesia is becoming a topic of great interest and a source of controversy in the anesthesia profession. One reason is their purchase and maintenance costs. Another is their effectiveness compared with other less-expensive training methods. Yet, Chopra et al. (1) have shown that anesthesiologists trained on mannequin-based simulators perform better when handling simulated crisis situations a second time, than those not trained on the simulator. However, can the large cost of a mannequin-based simulator be justified or can the same training effect be obtained by using a less expensive computer screen-based simulator? Is the training effect the same for novices and more experienced anesthesiologists? The acid test to answer these questions would be to determine whether patients treated by simulator-educated physicians would have a better outcome or if simulator training could reduce the cost of training or health care. These direct measurements of simulator training effectiveness might be virtually impossible in anesthesia because of the large variability in work situations. As an alternative, the effectiveness of training simulators could be tested by measurement of the progress achieved through a simulator training program, not by comparison between simulated and real practice settings. In this study, we compared the impact on training of two types of anesthesia simulators, a computer screen-based simulator and a mannequin-based simulator, to better document their effectiveness as training tools. Because of the similar training that is involved in both simulators, both training simulators might be expected to improve performance. However, because of the increased complexity of the mannequin-based environment, we hypothesized that the progress of performance in this training environment would be less than that on the screen-based simulator.

This work was supported by the Belgian State, Prime Minister's Federal Offices for Scientific, Technical, and Cultural Affairs, Program of Worker Protection in the Area of Health-Phase I-Interuniversity Pole of Attraction, and by a research grant from the North Atlantic Treaty Organisation.

Accepted for publication February 5, 2002.

Address correspondence and reprint requests to A.-S. Nyssen, PhD, Department of Work Psychology, Bat. B-32 FAPSE, University of Liège, 4000 LIEGE Sart-Tilman, Belgium. Address e-mail to asnyssen@ulg.ac.be.

Materials and Methods

After approval of the Department of Anesthesia and Intensive Care Medicine Board, we studied 40 anesthesia trainees (20 subjects per training environment)

0003-2999/02

from the University Hospital of Liège, Belgium. The computer screen-based simulator or training device (TD) used in this study was the Anesthesia Simulator Consultant developed by Schwid and O'Donnell (2). It consists of a standard personal computer with one video monitor. In the program, a graphic interface displays the patient and monitors. The anesthesiologists can obtain physical examination data, select airway management options, and administer drugs and fluids by using mouse-controlled input and menu navigation. The preoperative assessment, physical examinations, and surgeon's messages are presented in the menu. The program includes an expert consultant function that presents learning objectives for each case, suggested management of the case, and emergency, physiologic, and pharmacologic information for the simulated case. This expert consultant function was not available during our testing.

The mannequin-based simulator, or FS simulator, used in this study was part of the CASE (Comprehensive Anesthesia Simulation Environment) series designed by Eagle Simulation, Inc. (Binghamton, NY). It is presented in the form of a real operating room (OR) equipped with all of the recommended equipment for an operation, but the patient has been replaced by a mannequin that best represents the clinical reality. The capabilities of the simulator have been described by Gaba (3). The anesthesiologist can listen to the cardiac and lung sounds and take the pulse in the neck; he or she can inject drugs to which the mannequin responds in real time in exactly the same way as the TD, which uses many of the same mathematical models. Before the simulation, patient history is given to the trainee following the same format as the historical data presented on the TD. During the simulation, one team member played the role of "primary anesthesiologist" whereas a second member acted as the OR nurse. The surgeon was played by the instructor of the simulator. No help was available from the instructor or from the team member during the testing.

Our aim was to evaluate and compare the effectiveness of two training simulators at improving the performance of both novices and more-experienced anesthesia trainees. The criterion of evaluation was performance at treating a simulated anaphylactic shock (AS) (test scenario). The study was designed to compare pre- and posttraining. A pretest in simulators was not feasible for practical reasons (high cost). Without a pretest measure, changes of performance between the before and after training may not be attributed to the effect of training. The two most serious reasons cited in the literature for this are the large individual variations usually found in problem solving and the equivalence of tests (4). We used different measures to minimize this bias. First, we used the results of a previous anesthesia examination to choose subjects with similar backgrounds in theoretical knowledge to assure similar competencies between members of the groups of comparison. After the testing in simulators, we asked all participants whether they had encountered any similar clinical situations in the OR and retained in the experimental design only those trainees who had no experience to reduce contaminating intervening events between the pre- and posttest. Participants were requested specifically not to discuss the simulation case with other anesthesiologists. Second, we used the same sequence of training material for the before and after training to guarantee, as far as possible, the test equivalence.

We studied 40 anesthesia trainees from the University Hospital of Liège, Belgium according to the design shown in Figure 1. They were divided into two subgroups of 20 subjects per training environment. The subjects with different levels of training (10 with 2 to 3 years of anesthesia training and 10 with 4 to 5 years of anesthesia training) were distributed equally between the two simulator environments, and each subject was exposed to only one of the two types of simulators.

Each simulation session began with activities designed to familiarize the participants with the simulator environment they would be tested on. They inspected the simulator, learned how it could be manipulated, and their questions concerning the use and the limitations of the simulator were answered. On each simulator, during Phase 1, the participants were randomly exposed to the test scenario (a simulated AS [Group A]) or to the control scenario (a simulated malignant hyperthermia [MH] [Group B]). Because both simulators use many of the same models, they respond to the same solution with the same time lag. The anaphylactic reaction of the test scenario was severe. Modeled histamine release occurred a few minutes after the first injection of drug, resulting in tachycardia, hypotension, arterial blood pressure changes, increased ventilatory pressures, wheezing, desaturation, and cardiac arrhythmias. During this time, the participants could obtain information about the skin color changes. The participants were instructed to diagnose and treat the problem presented as they would in real life. We used the same patient historical data on both simulators, but changed patient data between Phases 1 and 2 to avoid basic recognition of patterns. The sessions were videotaped with a superimposed stop watch for evaluations. At the end of these sessions, identical structured debriefing was organized on each simulator to present the appropriate therapeutic actions. All participants were requested specifically not to discuss the session outside the training environment. After 1 mo, the participants of both groups (A and B) were exposed to the test scenario (AS) during Phase 2. These sessions were also videotaped and evaluated as for Phase 1. The participants

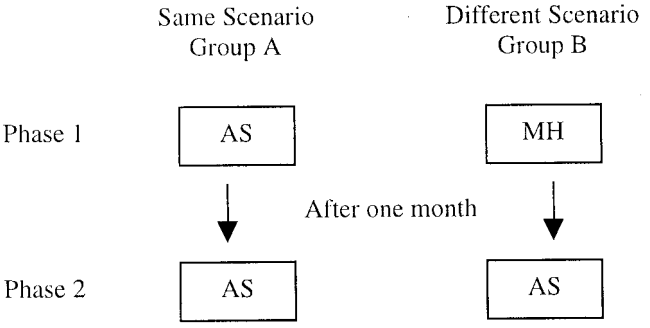

Figure 1. Representation of the experimental design used on both simulators. AS = anaphylactic shock, MH = malignant hyperthermia.

were unaware of the type of problem with which they would be presented during Phases 1 and 2.

Performance in the management of the test scenario (AS) was evaluated by using two variables: the treatment score and the diagnosis time. For the treatment score, a scoring system was developed by two experienced anesthesiologist instructors. It consisted of a list of points values assigned to appropriate medical and technical therapeutic actions (Table 1). Each treatment score was expressed as a fraction of a maximal possible score of 100. This scoring system is similar to that used by Chopra et al. (1) to assess technical performance in simulated crises. Two raters who did not know the trainees or whether it was Phase 1 or 2, initially recorded independently the presence or absence of each therapeutic action during the test scenario (AS), then compared the scores and reviewed the tape to resolve discrepancies. Thus, each subject received one single total point treatment score. Diagnosis times (in minutes), from the start of the target event to when the correct diagnosis was stated, were noted in the same manner.

Means and standard deviations for treatment scores and diagnosis times of all subjects were calculated on both simulators. On each simulator, we compared treatment scores and diagnosis times between Phases 1 and 2 in the "same-scenario" group (Group A) to measure the technical performance progress and between Group A and the "different-scenario" group (Group B) to distinguish the familiarization effect or the learned generic skills with the specific scenario training effect. We used a multivariate analysis with 4 independent variables: 1) level of experience (from novice to more experienced), 2) type of simulation (screen only and mannequin-based), 3) Phase (Phase 1 and Phase 2) and 4) Group (A and B). $P < 0.05$ was considered significant. The variability of performance scores between subjects was assessed before and after training using the test of variance homogeneity. We used the intraclass correlation coefficient to measure the interrater agreement.

Table 1. Scoring System for the Test Scenario: the anaphylactic shock

Treatment variables	Score
Ventilate with 100% oxygen	15
Vaporizer off	15
Patient in Trendelenburg	10
Expand fluid volume	15
Adrenaline administration	20
Treatment of arrhythmias	05
Check for bronchospasm	05
Administer H antagonist	05
Treatment of hypoxemia	02
Treatment of acidosis	02
Treatment of hyperkalemia	02
Blood sampling	02
Administer corticosteroids	02
Total treatment score	100

Results

Twenty subjects were trained on the Mannequin-based simulator (FS) and another 20 subjects were trained on the Computer screen-based simulator (TD). The two raters highly agreed on the presence and absence of a technical action and on the diagnosis time: the interclass correlation coefficient was 0.99. Treatment scores and diagnosis times (in minutes) for the test scenario (AS) on both simulators are shown in Table 2. We found no effect of the "type of simulator" and of the "level of experience" on the average treatment scores and on the diagnosis times comparing Phase 1 or Phase 2 between TD and FS.

On the Mannequin-based simulator, within the same-scenario group (Group A), there was a significant improvement in the diagnosis times between Phase 1 and Phase 2 ($F[1,19] = 14.46$, $P < 0.05$); the diagnosis time was shorter during Phase 2 (Phase 1, 2.98 ± 1.13; Phase 2, 1.96 ± 0.8). However, there was no significant difference in the diagnosis times and the treatment scores between the same-scenario group (Group A) and the different-scenario group (Group B) during Phase 2 on the mannequin-based simulator.

On the Computer screen-based simulator, there was a significant difference in the treatment scores between the same-scenario group (Group A) and the different-scenario group (Group B) during Phase 2 ($F[1,19] = 6.28$, $P < 0.05$); the treatment score was larger in the same-scenario group during Phase 2 (Group 1, 81.4 ± 7.0; Group 2, 68.8 ± 14.3). Within the same-scenario group, there was a significant improvement in the treatment scores between Phases 1 and 2 ($F[1,19] = 10.21$, $P < 0.05$); the treatment score was better during Phase 2 (Phase 1, 63.8 ± 16.44; Phase 2, 81.4 ± 7.0).

When comparing the same-scenario Group A Phase 1 and the different-scenario Group B Phase 2, we found no significant difference in the treatment scores, suggesting that the improvement of performance in

Table 2. Treatment Scores and Diagnosis Times

	Same-scenario Group A		Different-scenario Group B
	Phase 1	Phase 2	Phase 2
Treatment scores			
Mannequin-based			
Novices	67.4 (17.21)	74.2 (6.94)	71.6 (9.96)
Experienced	69.6 (21.03)	77.4 (9.40)	70.0 (16.34)
All	68.5 (18.16)	75.8 (7.97)	70.8 (12.79)
Computer-based			
Novices	65.8 (10.76)	77.4 (4.88)	65.0 (11.85)
Experienced	61.8 (21.96)	85.4 (6.80)	72.6 (16.79)
All	63.8 (16.44)	81.4 (7.00)	68.8 (14.27)
Diagnosis times (min)			
Mannequin-based			
Novices	2.34 (0.918)	2.12 (0.949)	2.41 (1.801)
Experienced	3.95 (0.594)	1.81 (0.690)	2.20 (1.697)
All	2.98 (1.135)	1.96 (0.800)	2.35 (1.629)
Computer-based			
Novices	3.96 (0.735)	2.67 (1.424)	3.23 (0.880)
Experienced	4.43 (1.768)	3.34 (1.649)	2.58 (0.918)
All	4.427 (1.429)	3.01 (1.496)	2.86 (0.894)

Data are presented as mean (SD) for the test scenario (anaphylactic shock).

the same-scenario group was attributable to training on the specific actions required for treating that scenario rather than to familiarity with the simulator or to learning generic skills of clinical problem solving.

On both simulators, the treatment scores between subjects in the same-scenario group (Group A) varied substantially during Phase 1 ($\chi^2 = 10.16111$, $P < 0.01$) (Fig. 2). In the second phase, the intersubject variability was reduced in the same-scenario group (Group A).

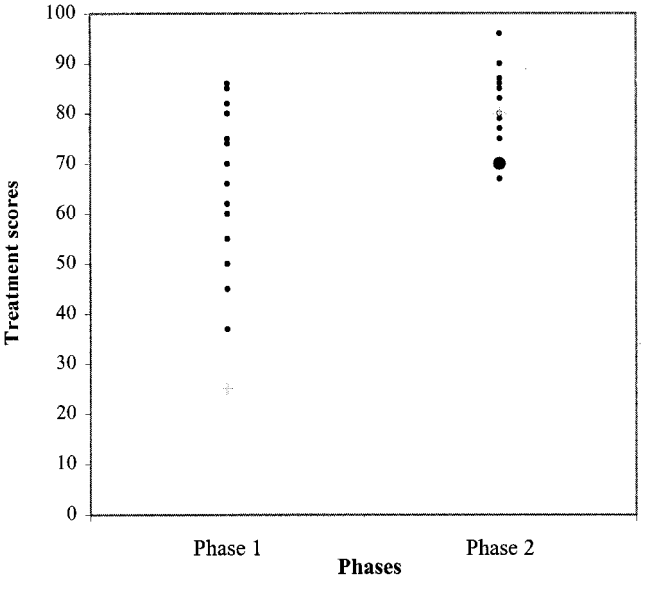

Figure 2. A comparison of treatment score variance between Phases 1 and 2 for the same-scenario Group A in the mannequin-based simulator and in the computer-based simulator. On both simulators, the treatment scores between subjects in the same-scenario group (Group A) varied substantially during Phase 1 ($\chi^2 = 10.16111$, $P <$ 0,01). In the second phase, the intersubject variability was reduced in the same-scenario group (Group A).

Discussion

Our aim was to compare two different training simulators (computer screen-based versus FS) relative to training effectiveness in anesthesia residents. This is an important issue, given the extensive use and the high cost of mannequin-based simulators. Previous studies conducted in aviation show that a higher fidelity of the simulation does not always lead to better performance (5–11). In anesthesiology, the effects of simulator training on job performance are more difficult to demonstrate because of the diversity of the situations, and the best use of different types of simulators for training has not been established. Our study is the first to compare two different types of training simulators. This was possible because of the similarity of the physiologic and pharmacologic models of the two simulators. The results showed that learning in managing simulated crisis situations such as AS did not vary significantly between FS and computer screen-based simulators. Furthermore, this study supports the use of computer screen-based training to achieve technical skills in treating simulated crisis situations. Consequently, the initial decision of whether to use FS or computer screen-based training simulators should be made on the basis of cost and learning objectives rather than on the basis of technical or fidelity criteria. The current

study provides potentially useful information to elucidate the training value of the two types of simulator.

Our results showed that simulators can contribute significantly to the improvement of performance. Schwid et al. (12) found that the use of a computer trainer improves retention of medical guidelines better than standard textbook review. Chopra et al. (1) found that anesthesiologists trained on an FS simulator in handling an MH deviate less from the accepted guidelines and perform better than those who are not trained on simulators. In comparison, the FS simulator in the current study produced an improvement in treatment and diagnosis times between phases, but the progress was not significant between groups. This discrepancy between the two studies raises the issue of the generalization of the findings beyond the specifics of the test scenario. In our case study, we observed that some subjects using mannequin-based simulators instituted the important recovering tasks for AS, such as the administration of epinephrine, before they stated the diagnosis, getting a "good" treatment score. In the case study by Chopra et al. (1), performing recovering tasks for MH, such as the administration of dantrolene, required a good prior diagnosis. This might explain the difference in findings in the two studies.

These results also raise the question of whether simulators can be used to assess performance because simulated situations, even highly realistic ones, are always a simplification of naturalistic situations, and the anesthesiologist's performance in handling a crisis in the OR involves a wider range of skills than diagnosis, retention of guidelines, and technical skills (13). In fact, it is not clear from our results whether diagnosis time is a useful concept if the proper management is performed. Gaba et al. (14) suggested the need to evaluate the anesthesiologist's performance during crisis management using both technical and behavioral aspects. Behavioral aspects refer to team behaviors such as communication, leadership, and interpersonal conflict. These have been reported as the most critical factors associated with human error accidents (15–17). We speculate that the FS simulator might be better for training in these aspects than the screen-based simulator, but this was not tested in the current study. In other professions, this distinction between behavioral and technical aspects of performance is at the origin of how simulators are used (10,18).

Our results showed that intersubject variability was reduced in the second phase in the same scenario. In many high-risk domains, there is a trend to reduce intersubject variability with procedures, regulations, and automation (19,20). Simulator training seems to be a good tool to achieve this goal, and thus can help in maintaining human and system reliability. However, although this finding indicates that simulator training produces greater consistency in performance between subjects, it does not indicate how the skill is developed by each subject. Is the improvement attributable to the fact that the trainees had encountered the problem situation? Is it attributable to the fact that all sessions were associated with a debriefing, or to other factors? Additional studies must address these questions.

The present study has some limitations, including the small size of the sample, the complexity of the design, and an increased expectation of problems during testing. Moreover, the novice group of residents using mannequin-based simulators had a markedly decreased diagnosis time during Phase 1 than the experienced group or all residents training on the screen-based simulator. This is surprising considering the increased complexity of the mannequin-based simulator. We believe that some novices did have knowledge of the scenarios that were likely to be presented despite our methodological precautions. This might have influenced their performance in diagnosis times during the evaluation sessions and invalidated the results.

We must develop more appropriate measures for evaluating the effectiveness of simulation-based training programs. A key element for future studies is to collect and compare data on human performance from the field and from simulators (21). The goal is not to improve the physical or functional fidelity of the simulators. Rather, the juxtaposition of data allows a better understanding of the demands of the task situation, the sources of errors, and the ability of simulators to meet the requirements for training to allocate the training objectives to the most effective and least-expensive setting.

References

1. Chopra V, Gesink BJ, De Jong J, et al. Does training on anesthesia simulator lead to improvements in performance? Br J Anaesth 1994;73:293–7.
2. Schwid HA, O'Donnell D. The Anesthesia Simulator-Recorder: a device to train and evaluate anesthesiologists' responses to critical incidents. Anesthesiology 1990;72:191–7.
3. Gaba DM. The human work environment and anesthesia simulators. In: Miller RD, ed. Anesthesia. 5th ed. New York: Churchill Livingstone, 1999:2613–68.
4. Patrick J. Training: research and practice. London: Academic Press, 1992.
5. Lintern G, Thomley-Yates KE, Nelson BE, et al. Content, variety and augmentation of simulated visual scenes for teaching air-to-ground attack. Hum Factors 1987;29:45–59.
6. Lintern G, Sheppard DJ, Parker DL, et al. Simulator design and instructional features for air-to-ground attack: a study of transfer. Hum Factors 1989;31:87–99.
7. Flexman RE, Stark EA. Training simulators. In: Salvendy G, ed. Handbook of human factors. New York: John Wiley & Sons, 1987:1012–38.
8. Jentsch F, Bowers CA. Evidence for the validity of PC-based simulations in studying aircrew coordination. Int J Aviat Psychol 1998;8:243–60.

9. Taylor HL, Lintern G, Koonce JM, et al. Scene content, field of view and amount of training in first officer training. In: Jensen RS, Neumester D, eds. Proceedings of the seventh international symposium on aviation psychology. Columbus: The Department of Aviation, The Ohio State University, 1993:753–7.
10. Salas E, Bowers CA, Rhodenizer L. It is not how much you have but how you use it: towards a rational use of simulation to support aviation training. Int J Aviat Psychol 1998;8:197–208.
11. Leplat J. Simulation and simulators in training: some comments, developing skills with information technology. In: Bainbridge L, Ruiz Quintanilla A, eds. Developing skills with information technology. London: John Wiley & Sons, 1989.
12. Schwid HA, Rooke GA, Ross BK, Sivarajan M. Use of 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.
13. Nyssen AS, De Keyser V. Improving training in problem solving skills: analysis of anesthetist's performance in simulated problem situations. Trav Humain 1998;61:387–401.
14. Gaba DM, Howard SK, Flanagan B, et al. Assessment of clinical performance during simulated crises using both technical and behavioral ratings. Anesthesiology 1998;89:8–18.
15. Cooper JB. Toward prevention of anesthetic mishaps. Int Anesthesiol Clin 1984;22:167–83.
16. Schaefer HG, Helmreich RL. The importance of human factors in the operating room. Anesthesiology 1994;80:479.
17. Helmreich RL, Schaefer HG. Team performance in the operating room. In: Bogner MS, ed. Human error in medicine. Hillsdale, NJ: Lawrence Erlbaum Associates, 1994:225–53.
18. Schwid HA. Computer simulations and management of critical incidents. Acad Med 1994;69:213.
19. Eichhorn JH, Cooper JB, Cullen DJ, et al. Standards for patient monitoring during anesthesia at Harvard Medical School. JAMA 1986;256:1017–20.
20. Billings CE. Aviation automation: the search for a human-centered approach. Mahwah, New Jersey: Lawrence Erlbaum Associates, 1997.
21. Nyssen AS. Analysis of human errors in anaesthesia: our methodological approach—from general observations to targeted studies in laboratory. In: Vincent C, De Mol BA, eds. Safety in medicine. London: Pergamon, 2000:49–63.