Comparative Research on Training Simulators in Emergency Medicine

A Methodological Review

Matthew Lineberry, PhD;

Melissa Walwanis, MS;

Joseph Reni, BS;

Naval Air Warfare Center Training Systems Division

Summary Statement: Recent reviews and research agenda in simulation-based training for health care have repeatedly called for more studies comparing the effectiveness of different simulation modalities and approaches. Although a small body of such comparative research does exist, it is apparent that this line of research comes with particular methodological challenges. In this review, 20 studies comparing simulation modalities in emergency medicine were analyzed in terms of key methodological factors. Results of this literature review show that (1) past studies are largely underpowered to detect effects of the magnitude likely to be found, (2) researchers seeking to demonstrate equivalence of training approaches do not use appropriate statistical tests of equivalence, and (3) studies often use performance criterion test beds that may not support valid conclusions about trainees' future performance on the job. We discuss strengths and weaknesses of reviewed studies along these issues and recommend options for improving future comparative research in simulation-based training.
(*Sim Healthcare* 8:253–261, 2013)

Key Words: Comparative research, Statistical power, Reliability, Validity, Null hypothesis statistical testing, Performance criteria, Simulation, Training effectiveness evaluation

Although hundreds of studies have shown that simulators in general are preferable to training without simulators,[1,2] very few studies actually compare 2 or more different simulation modalities.[3–5] This is a critical gap in our understanding of simulation-based training (SBT) for health care. Comparisons across SBT modalities are needed to help establish priorities for simulator feature development and to guide education decision makers in allocating resources between simulators of varying fidelity and expense. As an evidence base accumulates, both in comparisons of modalities and of other relevant instructional features,[6] the value of that research will be largely contingent on the rigor with which studies are designed and analyzed.

As such, this article reviews the small body of existing research comparing simulation modalities for SBT in health care with the goal of exposing major methodological issues. Rather than exhaustively addressing all possible deficiencies, we focus on 3 broad issues, which commonly arise in the reviewed studies and which pose, in our estimation, particularly grave threats to the validity of inferences based on the results of those studies. Although this endeavor requires us to be critical, we also applaud all the authors of studies reviewed in this article, who are at the vanguard of an important line of research in SBT for health care.

REVIEWED STUDIES

A literature review strategy was used whereby eligible studies were systematically identified and then critiqued against the taxonomy of experimental validity threats outlined by Shadish, et al[7]: (1) internal validity threats, for example, irregularities in participant selection, history, and maturation; (2) external validity threats, for example, use of students for studies in which experienced practitioners are the population of interest; (3) construct validity threats, for example, mislabeling a construct as "fidelity" that actually reflects one or more other constructs such as "affordance of instructional feedback"; and (4) statistical conclusion validity threats, for example, misinterpreting *p* values, low power to detect effects, or violation of statistical assumptions. Our primary goal was to collect and critique a representative sample of research comparing 2 or more simulators or simulation approaches for training of basic emergency procedures. Given the paucity of studies in this area, we therefore sought to be as exhaustive as our resources allowed.

We identified eligible studies primarily through keyword searches on PubMed and the Cumulative Index to Nursing and Allied Health Literature. Studies published before January 2012 were sought, using several combinations of the search terms *mannequin, manikin, animal, cadaver, simulat*, virtual reality, VR, emergency, trauma, compar*, versus,* and *VS* (where the *asterisk* indicates a wildcard search operator). In addition, all titles and abstracts from the journal *Simulation in Healthcare* were reviewed exhaustively through January 2012. References to and from relevant studies and reviews identified in this search were also searched for additional candidate studies. To maintain a coherent and manageable scope for the review, we only included studies focused on basic

From the University of Illinois at Chicago, Chicago, IL.

Reprints: Matthew Lineberry, PhD, Department of Medical Education, CME 973 University of Illinois at Chicago, Chicago, IL
(e-mail: MattL@uic.edu).

The authors declare no conflict of interest.

Copyright © 2013 Society for Simulation in Healthcare
DOI: 10.1097/SIH.0b013e31828715b1

Copyright © 2013 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.

TABLE 1. Comparative Effectiveness Studies in SBT for Emergency Medicine

Authors and Year of Publication	Procedure	Simulators Compared	Participants	Design
Ali et al[9] (2009)	Trauma evaluation and management	Standardized patient vs. manikin	Medical students	Between-subjects, pretest and posttest
Bowyer et al[10] (2005)	Diagnostic peritoneal lavage	VR simulator vs. anesthetized pig	Medical students	Between-subjects, pretest and posttest for knowledge, posttest only for performance ratings
Bowyer et al[11] (2005)	IV cannulation	Human volunteers vs. simulated limb vs. 2 different VR simulators	Medical students	Between-subjects, pretest and posttest
Brydges et al[12] (2010)	IV cannulation	VR simulator vs. manikin vs. progression through VR simulator, simulated limb, and manikin	Medical students	Between-subjects, posttest only
Chang et al[13] (2002)	IV cannulation	VR simulator vs. simulated limb	Nurses	Between-subjects, posttest only
Crofts et al[14] (2006)	Management of shoulder dystocia	Doll and pelvis trainers with vs. without strain gauge	Midwives and obstetricians	Between-subjects, pretest and posttest
Engum et al[15] (2003)	IV cannulation	VR simulator vs. simulated limb	Baccalaureate nursing students and medical students	Between-subjects, pretest and posttest
Hall[16] (2011)	Tube thoracostomy and cricothyroidotomy	Manikin vs. anesthetized pig	Air Force medical trainees	Between-subjects, posttest only
Hall et al[17] (2005)	Endotracheal intubation	On-the-job training vs. manikin	Paramedic students	Between-subjects, posttest only
John et al[18] (2007)	Cricothyroidotomy	Manikin, with or without environmental simulation	Anaesthetists	Within-subjects, posttest only
Knudson and Sisley[19] (2000)	FAST ultrasound	VR simulator vs. human volunteers	Interns and residents	Between-subjects, pretest and posttest
Lee et al[20] (2003)	Trauma evaluation and management	Manikin vs. standardized patients	Interns	Between-subjects, noncorresponding pretest and posttests
Owen et al[21] (2006)	Trauma evaluation and management	CPR trainers vs. addition of basic manikins vs. addition of manikin in simulated environment	Interns and residents	Between-subjects, pretest and posttest
Roberts et al[22] (1997)	Use of laryngeal mask airway	Manikin with or without subsequent on-the-job training	Nurses	Between-subjects, posttest only
Rodgers et al[23] (2009)	Trauma evaluation and management	Manikin with or without advanced features	Baccalaureate nursing students	Between-subjects, pretest and posttest
Salen et al[24] (2001)	FAST ultrasound	VR simulator vs. human volunteers	Emergency medicine residents	Between-subjects, posttest only
Scerbo et al[25] (2006)	Phlebotomy	VR simulator vs. simulated limb	Medical students	Between-subjects, pretest and posttest
Scerbo et al[26] (2006), experiment 1	Phlebotomy	VR simulator vs. simulated limb	Medical students	Between-subjects, pretest and posttest
Wandell[27] (2010)	Phlebotomy	VR simulator vs. simulated limb	Technician and certificate students	Between-subjects, pretest and posttest
Youngblood et al[28] (2008)	Trauma evaluation and management	Manikin in OR vs. virtual patient/environment	Medical students (including recent graduates)	Between-subjects, pretest and posttest

CPR indicates cardiopulmonary resuscitation; FAST, focused assessment using sonography for trauma; IV, intravenous; NA, not applicable; VR, virtual reality.

Copyright © 2013 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.

TABLE 1. *(Continued)*

n per Group	Performance Measure(s)	Reliability of Performance Measurement	Performance Test Bed	Summary of Findings
24	Knowledge test	Inferred parallel forms reliability from previous research	N/A	No difference between simulator groups in knowledge posttest
20	Knowledge test and 2 blinded raters' ratings of procedural step performance, trainees' apparent understanding, and confidence in trainees	Interrater reliability of ratings = 0.85	Manikin	No test for differential knowledge gains; VR simulator group scored higher on posttest for 2 of 5 understanding dimensions and on rater confidence; procedural ratings not mentioned
13, 6, 7, and 8	Procedural rating scale (rating protocol unknown)	Unreported	Simulated limb	Virtual IV led to greater improvement than simulated limb; all other comparisons not significant
15	Rating scales and checklists, 2 raters	Intraclass correlation coefficient = 0.55–0.71	Simulated limb similar to that in progressive condition	Progressive group associated with better outcomes than VR simulator group and, for some outcomes, manikin group
14	Success/failure (criteria unstated), procedural checklist rated by 1 of 2 blinded raters	Spearman's ρ for a subsample = 0.80	Human patients	No differences in performance between groups
68 and 64	Force data from simulator, 2 raters' log of trainee actions, time to complete, and success rate	Unreported	Doll and pelvis trainer with strain gauge	Training with strain gauge associated with higher success rate, shorter time to complete, less total force applied; however, lower likelihood to call for pediatric support
Unclear; total n = 163	Knowledge test, weighted procedural rating scales rated by 1 of 3 raters, quality rating scale completed by 1 of 163 volunteers	Interrater reliability and KR test mentioned but not reported	Human volunteers	Greater improvement in knowledge among nurse subsample in simulated limb condition; better documentation among simulated limb trainees
12	Time to complete, incision size, location, success/failure with stated criteria (rating protocol unknown)	n/a	Human cadaver	No differences between groups on performance measures
18	Intubation success rate, first-attempt success rate, complication rate, time to complete (data recorders not blinded)	n/a	Human patients	No differences between groups on performance measures
70	Time to complete; success/failure (with unstated criteria and rating protocol)	Unreported	Manikin, both with and without environmental simulation	Time to complete is increased under environmental simulation
37	Knowledge test	Unreported	N/A	No differences in learning gains between groups
Four groups of 15 (2 training conditions by 2 testing conditions)	Knowledge test; procedural rating scale rated by 2 raters	Main effect of judges estimated in regression model; correlation between judges' scores = 0.52	Either manikin or standardized patient	Simulator training led to higher scores
20	Knowledge test, performance checklist by one of several blinded raters	Unreported; cited other research demonstrating reliability with checklist scoring generally	Manikin in simulated environment	Manikin in simulated environment training group scored higher on 2 of 3 posttest scenarios
20 and 32	Trials to success/failure, with stated criteria	n/a	Human patients	No difference in success rates between groups
16 and 18	Knowledge test; rating scale, 3 raters	Squared correlations between raters = 0.24–0.54; internal consistency α = 0.83	Manikin with advanced features	Greater performance and knowledge gains using manikin with advanced features
Unclear; total n = 20	Image diagnosis test	Unreported	N/A	No difference between groups
9 and 11	Weighted procedural rating scale (rating protocol unknown)	Unreported	Simulated limb	Improvement only for simulated limb condition
12 and 14 (but only 8 and 8 for patient posttest)	Procedural rating scale (rating protocol unknown), automated performance metrics (for VR group only)	Unreported	Simulated limb and actual patients	Only simulated limb training led to performance gains on simulated limb test; also leads to greater transfer to performance on actual patients
13 and 12	Procedural rating scale (rating protocol unknown)	Unreported	Volunteers with training blocks strapped on	No difference in performance improvements between groups
14 and 16	Rating scale, 3 raters	Interrater reliability = 0.71; internal consistency α = 0.96	Each group's assigned simulator	No difference in pretest or posttest scores between groups

Copyright © 2013 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.

procedural skills relevant to emergency medicine, specifically limiting eligibility to training on needle access procedures, airway management, tube thoracostomy, Focused Assessment using Sonography for Trauma, and trauma evaluation and management. As such, the review excludes advanced surgical procedure training and interpersonal skills training. Studies were included if they experimentally compared the effectiveness of multiple SBT approaches or if they compared a simulation-based approach to quantified on-the-job training. We do not report studies' results related to trainee reactions such as satisfaction or confidence because trainee reactions do not relate well to learning or behavior[8] (although they may be important for other purposes).

In identifying key methodological issues, the authors first examined eligible studies for methodological strengths and deficiencies against the aforementioned taxonomy of validity threats. From this, all authors extrapolated overarching methodological themes that were relatively prevalent and of grave threat to the validity of studies' inferences. Given the small number of databases searched and the breadth of venues in which relevant research is published, it is certainly possible that this review is not exhaustive, although we believe it to be reasonably representative of published work available to research consumers.

Table 1 lists the reviewed studies and provides selected details for each. All but one study compared simulator conditions in a between-subjects experimental design, and as per the eligibility requirements, all studies assessed trainee posttraining performance. Eight studies used a posttest only design; all other studies included a pretest for at least one of the performance criterion posttests. The most common approach to measuring trainee posttraining performance was instructor or expert ratings using checklists or scales (12 studies). Eight studies used knowledge tests, 6 studies assessed success or failure of procedural performance, and 5 studies collected relatively objective data on student performance (eg, time to complete procedure, incision size).

CRITICAL METHODOLOGICAL CHALLENGES

After reviewing the eligible studies, 3 major challenges with comparative SBT research are apparent.

Power to Detect Small, Practically Significant Effects

A recent meta-analysis by Cook et al[2] suggests that, compared with no intervention, SBT in health care leads to moderate-to-large learning gains, with average effect sizes of $g = 1.09$ to 1.20 on trainee learning criteria, $g = 0.79$ to 0.81 for on-the-job behavior with actual patients, and $g = 0.50$ for patient outcomes (where g is the difference in group means divided by the groups' pooled SD; it may range from $-\infty$ to ∞, with a g of 1.0 indicating that the mean of group 1 is 1 SD larger than the mean of group 2).

Comparative SBT research involves comparison of conditions, that are all likely to be at least somewhat beneficial for learning. We should therefore not expect such large effect sizes as when SBT is compared with no intervention (the latter presumably having zero educational benefit). Such smaller effects can still have great practical significance, particularly if (1) the affected outcomes are highly important, for example, reduction in morbidity or mortality; (2) the experimental manipulation can benefit many individuals; and/or (3) the experimental manipulation can be exploited inexpensively. For instance, a large trial examining aspirin's effect on incidence of heart attack was halted early because it was deemed unethical to continue withholding aspirin from the control group. The correlation between aspirin use and heart attack incidence was minute—a correlation of only $\phi = 0.03$—but that small effect corresponded to 85 fewer incidences of heart attack in the treatment group (of 22,071 total participants).[29,30] Differences in SBT in health care will likely have less pronounced effects than this; however, over the long term, small effectiveness differences between training approaches could translate into substantial differences in clinical outcomes.

Unfortunately, small effect sizes are profoundly difficult to detect experimentally, requiring extremely sensitive experimental designs. Because researchers' choices regarding sample size, reliability of measurement, and statistical decision rules figure heavily into whether such small effects will be detected, we reviewed past comparative SBT research along those 3 factors.

Sample Sizes

The reviewed studies feature quite small sample sizes, ranging from 6 to 82 participants per experimental group with a median of 15 participants per group. (Where total sample sizes were reported but not group sizes, group sizes were assumed as equal as possible, eg, 163 participants of the study by Engum et al[15] are assumed to be distributed as 82 and 81 in each group. For the John et al[18] (2007) within-groups study, the total number is counted twice, once for each condition.) On the whole, these sample sizes correspond to poor power to detect differences between conditions. For instance, a hypothetical between-groups study attempting to detect an effect of $d = 0.2$ using $n = 20$ per group and the customary $\alpha = 0.05$ decision rule would have only a 13% chance of detecting the difference.[31]

Large n studies are absolutely necessary for comparative SBT research where effect sizes may be small but important. Therefore, rather than having many underpowered and uncoordinated studies, it would be preferable for researchers to collaborate and conduct coordinated experiments across multiple sites. Besides allowing for greater sample sizes, multicenter research could also yield interesting results if findings differ across locations, perhaps suggesting less reliable effects than one might expect from a single-location study or offering insight into organization-level variables, that relate to different training outcomes.

Recommendation 1

Sample sufficient numbers of participants to detect small but practically significant effects, engaging in multicenter collaboration where possible.

Reliability of Performance Ratings

The predominant mode of posttraining performance measurement is experts' ratings using checklists and/or rating

Copyright © 2013 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.

scales. This is generally a very reasonable approach to criterion measurement; however, expert ratings come with particular challenges that affect experimental power in complicated and generally adverse ways. Possible inconsistencies in raters' judgments are numerous[32]; raters may differ in scoring severity, use irrelevant information, make "snap" judgments of trainees, or be inconsistent with themselves over time, to name a few. Even when raters agree on a trainee's score, they may do so for different reasons, any of which may not be valid. The magnitude of rating inconsistencies in SBT research is likely to be substantial; in performance rating generally, rating errors are estimated to contribute twice as much variance as the residual error variance typically seen in ratings.[33]

Rater inconsistencies can crowd out true trainee performance variance, decreasing both validity and power. As such, studies that use performance ratings with low or unknown reliability should be interpreted with skepticism, especially when small samples are used. Studies for which rater inconsistencies are a potential concern must use all tactics available to increase reliability (eg, using multiple raters, holding multiple rating occasions, training raters to use a common frame of reference[34]) and must estimate reliability rigorously. Unfortunately, across the 12 studies reviewed which used performance ratings, reliability of measurement was not consistently given optimal attention. Six studies reported the use of 2 or more raters per participant, with the remainder using a single rater or not stating how rating was conducted. The former studies each reported either an interrater reliability coefficient or correlations between raters' ratings. Often, these estimates were somewhat low (eg, squared correlations between raters as low as 0.24; intraclass correlation coefficients as low as 0.55), consistent with a recent review that found generally low reliability for performance rating in objective structured clinical examinations.[35] Furthermore, as we will explain later, the statistical estimates reported do not comprehensively estimate reliability. The remaining 6 studies offered no estimate of reliability whatsoever, and no studies featured rating on multiple occasions, such that any rater inconsistencies across time cannot be estimated. Hence, measurement reliability in the reviewed studies is often low or at least partially unknown.

As mentioned earlier, 6 studies estimated interrater consistency, either using a simple correlation of raters' scores or an interrater reliability statistic. Statistical estimates such as these imply unfounded assumptions about ratings.[36] For instance, use of such estimates assumes that all rater disagreement should be considered random error variance. Although it is true that rater disagreement probably does not constitute true score variance, it is not necessarily random, and not all systematic errors affect reliability in the same way. For instance, rater differences in severity or leniency do not affect the rank order of the rated trainees and thus do not contribute to error variance for relative comparisons between trainees (assuming that all trainees are rated by the same raters).

One reliability estimation approach properly models the complexity of reliability in performance rating, known as *generalizability theory* (GT).[37] Generalizability theory studies model the multiple sources of unreliability simultaneously, just as those sources act in reality. The principles behind the approach also draw attention to the fact that reliability is not a single property of a measurement instrument; we cannot say a particular measure simply "is" reliable. Rather, reliability depends on the decision that measurements are used to make and the particular way in which the measure is used (eg, how many raters are used, how often they make ratings, etc).

Information from GT studies can be used to design highly reliable measurement procedures. Generalizability theory should thus be adopted as standard practice for reliability assessment and measurement planning when performance ratings are used. Generalizability theory studies are becoming more common in medical assessment research reports,[38–40] and accessible introductions to the use of GT are available.[41,42] Nonetheless, the approach is statistically sophisticated and can require adjustments depending on the research questions at hand. For instance, in comparative SBT research, reliability of group averages is the primary interest, requiring adjustments in D study computations.[43] Research teams would likely benefit from including a methodologist who can apply the approach in a given research context.

Recommendation 2.1

For all performance measures, report reliability using statistically appropriate estimates, for example, GT coefficients for performance ratings.

Recommendation 2.2

Use robust measurement protocols to ensure highly-reliable measurement; for example, use multiple raters, conduct frame-of-reference training for raters, and so on.

Weighted Versus Unweighted Procedural Checklists and Rating Scales

Two reviewed studies—the studies of Engum et al[15] and Scerbo et al[25]—reported the use of a weighting scheme for their performance measures, such that higher-priority steps within a procedure were assigned a greater number of points possible. We strongly encourage this approach, both to increase reliability of measurement and also to increase validity of inferences. Many procedures contain steps that vary in criticality; some steps may be mundane, whereas other steps may be extremely important for successful outcomes. Assuming that trainees are likely to vary more widely on highly critical steps, assigning greater weights to those steps in scoring protocols amplifies trainee performance differences and thus facilitates more reliable measurement. Assuming that high-criticality steps are also more closely associated with patient outcomes, such an increase in reliability will also be associated with higher criterion validity.

Copyright © 2013 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.

Recommendation 2.3

For performance ratings, weight scores for procedural steps according to each step's criticality to procedural outcome quality, with criticality estimated according to expert judgment and/or empirical data.

Type I and Type II Error Rates

A lesser known strategy for improving power also happens to be particularly appropriate for the research question of interest: choosing α levels rationally, according to the relative severity of type I versus type II errors in one's study.

There is a strong, arbitrary norm in research to choose $\alpha = 0.05$ as the accepted type I error rate. All of the studies reviewed here either explicitly or implicitly adopt this α level, which rarely represents a sensible balance of the inferential risks in comparative SBT research. In this line of research, a type I error occurs if a study incorrectly declares one simulation modality more effective, when in reality both techniques are equally effective. In that scenario, since both techniques are equally effective in reality, any resulting changes in simulator use—such as a decision to stop using the seemingly inferior simulator—would have no effect on trainee learning. It is not clear why we should consider such an error especially serious. If the more expensive option of the 2 was incorrectly favored, this might result in fewer simulators being purchased and thus fewer training opportunities. However, although type I errors seem prudent to avoid, they are not overly grave.

Conversely, a type II error—failing to find an effect when one exists—may have more negative consequences. For instance, if one simulation approach is superior to another but a study fails to detect this, other considerations may drive decisions to use either simulator type. If this resulted in the more effective approach being used less often or even dropped completely, trainee readiness would suffer, and by extension, so too would patient outcomes. Thus, in the current context, type II errors could be quite serious.

Unfortunately, researchers rarely consider relative error severity and rarely acknowledge type II errors at all. Researchers often implicitly treat type I errors as being drastically more serious than type II errors. For example, a study with power = 0.60 (therefore $\beta = 0.40$) and $\alpha = 0.05$ is implicitly treating type I errors as 8 times more serious than type II errors. Such an unthinking and imbalanced risk mitigation strategy makes little sense in the current domain. Consequently, methodologists have suggested that we should be more thoughtful about balancing α and β,[44] a call which we echo here.

Recommendation 3

Choose α levels to sensibly balance relative risks of type I and type II errors. Be explicit about rationales for chosen error rates.

Proper Tests of Equivalence of Conditions

In comparative SBT for health care, one possible research goal is to demonstrate that 2 approaches are equally effective, especially when an objectionable or impractical training practice (eg, use of live animals or human volunteers) might be replaced with another practice (eg, training using manikins). Six reviewed studies explicitly sought to demonstrate equivalence between 2 conditions and/or used language indicating that their results demonstrated such equivalence. However, none of those studies used appropriate tests of equivalence.

The previous sections on sample size, reliability, and error rates illustrate how a failure to detect differences between conditions can easily stem from methodological deficiencies, rather than from true equivalence of conditions. We must therefore be careful how we interpret studies that do not detect differences between experimental groups. For instance, Hall[16] found a 34% difference in posttraining cricothyroidotomy completion times between training conditions—an average of nearly a minute during which a hypothetical patient would presumably continue to suffer hypoxia. However, with small *n*'s and a posttest-only design, even that large difference did not reach statistical significance in the study. The author did acknowledge that there may be a difference but elsewhere concluded that the conditions were "reaching parity," an assertion we argue is not justified by the data or analyses.

In traditional null hypothesis statistical testing, one may reject the null hypothesis and declare a difference between groups or fail to reject the null and reserve judgment, which is not the same as acceptance of the null. To accept the null would imply a belief that the difference between conditions is a very specific value—zero—or at least that it is so close to zero that it is practically zero. Such a strict conclusion requires far higher parameter estimation precision than is the norm in research; it cannot be concluded from low-power experiments.

There is a strong, statistically appropriate test to prove the null, which consist of setting a region of practical equivalence around a difference of zero and testing whether the confidence interval or credibility interval for the estimated difference between groups falls entirely within that region.[45] If, for instance, it can be argued that a time-to-complete difference of 20 seconds or less between groups would be clinically insignificant for a given procedure, then a study could be designed to evaluate whether the only plausible population differences between groups are all within 20 seconds, that is, they are effectively zero. This approach has the added benefit of forcing researchers to be explicit about what effect sizes they would deem practically significant for their chosen comparison and why. The study of Hall et al[17] comes quite close to this approach, having explicitly specified a 10% difference in intubation success rates as their criteria for clinical significance. However, they do not justify why such a difference would be clinically acceptable; furthermore, although the authors claim that their nonsignificant results demonstrated equivalence, the confidence intervals in their study included differences of 10% and greater, contradicting that assertion.

Copyright © 2013 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.

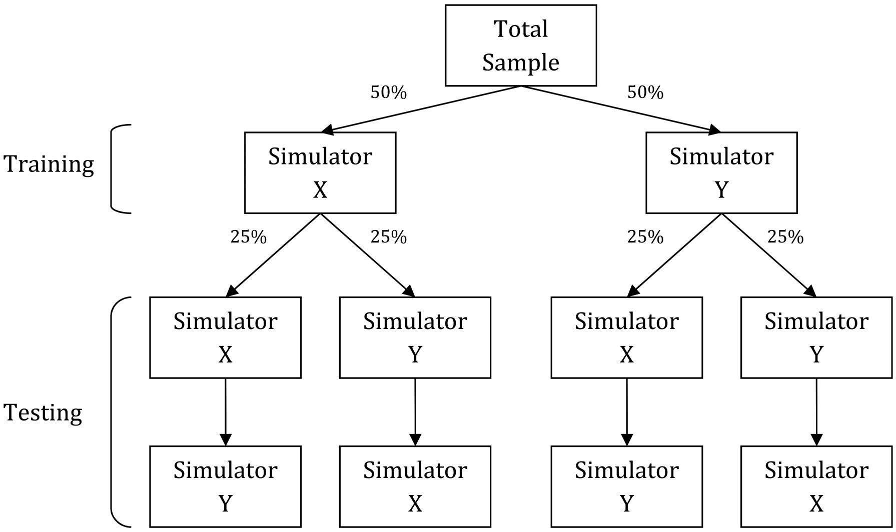

FIGURE 1. A crossed-criterion design for studies using simulators as criterion test beds.

Recommendation 4.1

For tests that fail to detect differences between experimental conditions, avoid erroneously accepting the null hypothesis.

Recommendation 4.2

When proving the null hypothesis is a research goal, use strong null hypothesis statistical testing by testing whether all plausible differences fall within a region practically equivalent to zero.

Performance Criteria—Relevance and Lack of Bias

Criterion relevance is the extent to which a criterion corresponds to the intended performance domain exhaustively and exclusively. That is, if performance of a cricothyroidotomy on a manikin features the same environmental cues, mental demands, physical actions, stressors, and so on as the performance domain of genuine interest (ie, performance of a cricothyroidotomy on an actual trauma patient), the criterion may be argued to have high relevance. Experimental data are more likely to reflect genuine effects on job performance when highly relevant posttraining performance criteria are used to compare groups. However, only 6 reviewed studies tested trainees after training on actual patients; in the majority of studies, trainees' ultimate criterion was performance on a simulator or on a multiple-choice test. Of these latter studies, 8 tested trainees on one of the simulators being compared during training, which is especially problematic for reasons we will discuss in the next section.

To their credit, a number of reviewed studies used criterion test beds that are likely to be highly relevant. Both the studies of Hall et al[17] and Roberts et al[22] are remarkable for comparing trainees on their patients' outcomes during actual medical treatment after training. Because they assessed direct samples of job performance, one can draw conclusions about the effect of their training manipulations with much greater confidence. That said, both studies rated trainees only on the outcome of their performance, not on the quality of their behavior; the former can be a crude indicator of trainees' long-term quality and safety of care simply because outcomes are not entirely under performers' control.[46]

The study of Engum et al[15] is also notable for its use of volunteer patients on whom trainees attempted to establish intravenous access. Although the use of volunteers is not feasible for more invasive procedures, it is an attractive option not only for the close approximation of actual patients but also for the possibility of having the volunteers rate the quality of trainees' performance.

Although excellent, even the 3 studies mentioned previously were also unrealistic in the sense that trainees were directed to perform a procedure, rather than recognizing that the procedure was indicated and deciding to act. Roberts et al[22] acknowledged this in their study, noting that their trainees were compelled to place the somewhat invasive but clinically preferred laryngeal mask airway. The authors thus feared that nurse trainees who did not feel confident using the more invasive airway might use less invasive, suboptimal alternatives after leaving the experimental environment. Thus, where possible, our research needs to demonstrate not only that trainees can perform a procedure after a training intervention but also that they will do so when appropriate.

Recommendation 5.1

To the extent possible, use criterion performance test beds that correspond highly to the conditions and demands, that trainees will face on the job.

Copyright © 2013 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.

The Crossed-Criterion Design for SBT Research

When a simulator is used as the criterion test bed in a comparative SBT study and that simulator is similar to or the same as the simulator used in one of the training conditions, any group differences after training could simply be due to differential familiarity with that test bed. Scerbo et al[25] acknowledged that this issue created a bias in their study, in which students learned phlebotomy using either a simulated limb or a virtual reality simulator but were all tested on the simulated limb. In a separate study, Scerbo et al[26] were able to test a small sample of students performing the procedure on actual patients and thereby argue more convincingly for the criterion validity of their earlier findings. However, it is not always practical to assess student performance on actual patients. One way to increase the inferential strength of a study without testing students on actual patients is to use a crossed-criterion design, where each group is tested on both the simulators that were used for training in counterbalanced order (Fig. 1). Although this design does not fully compensate for an inability to test trainees on actual patients, it does provide a more fair comparison of groups and should also yield more informative results. For instance, if learning under one simulation condition generalizes more readily to performance on the other condition, this may indicate that learning was more robust in that condition (assuming that performance assessment was equally reliable across the 2 conditions). A particular drawback of this design is that any pretesting would expose trainees to each of the simulators being compared. If pretest exposure were sufficient to derive significant benefits (eg, to give trainees their first exposure to live tissue), the experimental manipulations would be diffused across experimental groups. Some of the benefits of pretesting could be achieved by alternate experimental methods, such as statistical matching of groups. As with all research, researchers must balance competing priorities—in this case, the priority to use fair criteria versus to raise experimental power with a pretest.

Recommendation 5.2

When simulators must be used as criterion test beds, usecrossed-criterion designs to reduce bias and strengthen inferences.

CONCLUSIONS

Studies that compare different simulation approaches are the next logical step in the evolution of research in SBT for health care. In the case of emergency medical SBT, although a small body of research comparing modalities exists, it is difficult to draw meaningful overarching conclusions, owing in large part to the methodological issues addressed previously. In particular, studies have featured small sample sizes, and many studies used measures with low or unknown reliability. Combined with the adoption of strict, conventional type I error rates, this has very likely created a bias toward failing to detect effects. Occasionally, this bias has been compounded by researchers' misinterpretations of such failures to find effects; in relation to this, in cases where finding null effects is an explicit research goal, researchers have not used appropriate statistical tests. Furthermore, although some researchers have gone to great lengths to use demonstrably relevant criterion test beds, others have understandably used test beds that are more convenient but that are less clearly relevant to job performance.

The current review sacrifices comprehensiveness to focus briefly on particularly challenging methodological issues in this domain; the full breadth of potentially relevant methodological issues is beyond the scope of this article. For instance, Cook et al[6] point out other important methodological issues, including clear reporting of key aspects of experiments and the need for conceptual or theoretical rationales as to why a particular comparison might be expected to relate to outcomes. Useful and relatively comprehensive inventories of methodological factors and reporting guidelines are available,[47,48] although we suspect researchers will generally need to consult or include a trained behavioral science methodologist to ensure studies are well designed and analyzed.

All authors in the research reviewed are to be commended for anticipating the need for comparative effectiveness research, which has only recently become more explicitly stated in the health care SBT literature. We believe that the recommendations in our review constitute practical actions that future researchers can take—and which research sponsors and journal editors should insist upon—to improve the value of subsequent research in this area.

REFERENCES

1. McGaghie WC, Issenberg B, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Acad Med* 2011;86:706–711.
2. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. *JAMA* 2011;306:978–988.
3. Issenberg SB, Ringsted C, Ostergaard D, Dieckmann P. Setting a research agenda for simulation-based healthcare education: a synthesis of the outcome from an Utstein style meeting. *Simul Healthc* 2011;6:155–167.
4. Patronek GJ, Rauch A. Systematic review of comparative studies examining alternatives to the harmful use of animals in biomedical education. *J Am Vet Med Assoc* 2007;230:37–42.
5. Byrne AJ, Pugsley L, Hashem MA. Review of comparative studies of clinical skills training. *Med Teach* 2008;30:764–767.
6. Cook DA, Hamstra SJ, Brydges R, et al. Comparative effectiveness of instructional design features in simulation-based education: systematic review and meta-analysis. *Med Teach* 2012;35:e1–e32.
7. Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-experimental Designs for Generalized Causal Inference.* Boston, MA: Houghton Mifflin; 2002.
8. Alliger GM, Tannenbaum SI, Bennett W, Traver H, Shotland A. A meta-analysis of the relations among training criteria. *Personnel Psychol* 1997;50:341–358.
9. Ali J, Al Ahmadi K, Williams JI, Cherry RA. The standardized live patient and mechanical patient models—their roles in trauma teaching. *J Trauma* 2009;66:98–102.
10. Bowyer MW, Liu AV, Bonar JP. Validation of SimPL—a simulator for diagnostic peritoneal lavage training. *Stud Health Technol Inform* 2005;111:64–67.

Copyright © 2013 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.

11. Bowyer MW, Pimentel EA, Fellows JB, et al. Teaching intravenous cannulation to medical students: comparative analysis of two simulators and two traditional educational approaches. *Stud Health Technol Inform* 2005;111:57–63.

12. Brydges R, Carnahan H, Rose D, Rose L, Dubrowski A. Coordinating progressive levels of simulation fidelity to maximize educational benefit. *Acad Med* 2010;85:806–812.

13. Chang KKP, Chung JWY, Wong TWS. Learning intravenous cannulation: a comparison of the conventional method and the CathSim Intravenous Training System. *J Clin Nurs* 2002;11:73–78.

14. Crofts JF, Bartlett C, Ellis D, Hunt LP, Fox R, Draycott TJ. Training for shoulder dystocia: a trial of simulation using low-fidelity and high-fidelity mannequins. *Obstet Gynecol* 2006;108:1477–1485.

15. Engum SA, Jeffries P, Fisher L. Intravenous catheter training system: computer-based education versus traditional learning methods. *Am J Surg* 2003;186:67–74.

16. Hall AB. Randomized objective comparison of live tissue training versus simulators for emergency procedures. *Am Surg* 2011;77:561–565.

17. Hall RE, Plant JR, Bands CJ, Wall AR, Kang J, Hall CA. Human patient simulation is effective for teaching paramedic students endotracheal intubation. *Acad Emerg Med* 2005;12:850–855.

18. John B, Suri I, Hillerman C, Mendonca C. Comparison of cricothyroidotomy on manikin vs. simulator: a randomised cross-over study. *Anaesthesia* 2007;62:1029–1032.

19. Knudson MM, Sisley AC. Training residents using simulation technology: experience with ultrasound for trauma. *J Trauma* 2000;48: 659–665.

20. Lee SK, Pardo M, Gaba D, et al. Trauma assessment training with a patient simulator: a prospective, randomized study. *J Trauma* 2003;55:651–657.

21. Owen H, Mugford B, Follows V, Plummer JL. Comparison of three simulation-based training methods for management of medical emergencies. *Resuscitation* 2006;71:204–211.

22. Roberts I, Allsop P, Dickinson M, Curry P, Eastwick-Field P, Eyre G. Airway management training using the laryngeal mask airway: a comparison of two different training programmes. *Resuscitation* 1997;33:211–214.

23. Rodgers DL, Securro S, Pauley RD. The effect of high-fidelity simulation on educational outcomes in an Advanced Cardiovascular Life Support course. *Simul Healthc* 2009;4:200–206.

24. Salen P, O'Connor R, Passarello B, et al. FAST education: a comparison of teaching models for trauma sonography. *J Emerg Med* 2001;20:421–425.

25. Scerbo MW, Bliss JP, Schmidt EA, Thompson SN. The efficacy of a medical virtual reality simulator for training phlebotomy. *Hum Factors* 2006;48:72–84.

26. Scerbo MW, Schmidt EA, Bliss JP. Comparison of a virtual reality simulator and simulated limbs for phlebotomy training. *J Infus Nurs* 2006;29:214–224.

27. Wandell HF. Using a virtual reality simulator in phlebotomy training. *Lab Med* 2010;41:463–466.

28. Youngblood P, Harter PM, Srivastava S, Moffett S, Heinrichs WL, Dev P. Design, development, and evaluation of an online virtual emergency department for training trauma teams. *Simul Healthc* 2008;3:146–153.

29. Steering Committee of the Physicians Health Study Research Group. Preliminary report: findings from the aspirin component of the ongoing physicians' health study. *N Engl J Med* 1988;318:262–264.

30. Rosenthal R. How are we doing in soft psychology? *Am Psychol* 1990;45:775–777.

31. Cohen J. *Statistical Power for the Social Sciences*. 2nd ed. New York, NY: Taylor Francis; 1988.

32. Landy FJ, Farr JL. Performance rating. *Psychol Bull* 2008;87:72–107.

33. Scullen SE, Mount MK, Goff M. Understanding the latent structure of job performance ratings. *J Appl Psychol* 2000;85:956–970.

34. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: a quantitative review. *J Occup Organ Psychol* 1994;67:189–205.

35. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Med Educ* 2011;45:1181–1189.

36. Murphy KR, DeShon R. Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychol* 2000;53:873–900.

37. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. *The Dependability of Behavioral Measurements: Theory of Generalizability of Scores and Profiles*. New York, NY: Wiley; 1972.

38. Reinders ME, Blankenstein AH, van Marwijk HWJ, et al. Reliability of consultation skills assessments using standardised versus real patients. *Med Educ* 2011;45:578–584.

39. Wass V, Wakeford R, Neighbour R, van der Vleuten CPM. Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component. *Med Educ* 2003;37:126–131.

40. Donoghue A, Ventre K, Boulet J, et al. Design, implementation, and psychometric analysis of a scoring instrument for simulated pediatric resuscitation: a report from the EXPRESS Pediatric Investigators. *Simul Healthc* 2011;6:71–77.

41. Crossly J, Davies H, Humphris G, Jolly B. Generalizability: a key to unlock professional assessment. *Med Educ* 2002;36:972–978.

42. Shavelson RJ, Webb NM, Rowley GL. Generalizability theory. *Am Psychol* 1989;44:922–932.

43. Kane MT, Brennan RL. The generalizability of class means. *Rev Educ Res* 1977;47:267–292.

44. Murphy K. Using power analysis to evaluate and improve research. In: Rogelberg SG, ed. *Handbook of Research Methods in Industrial and Organizational Psychology*. Malden, MA: Blackwell; 2004:119–137.

45. Kruschke JK. Bayesian assessment of null values via parameter estimation and model comparison. *Perspect Psychol Sci* 2011;6:299–312.

46. Binning JF, Barrett GV. Validity of personnel decisions: a conceptual basis of the inferential and evidential bases. *J Appl Psychol* 1989;74: 478–494.

47. Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. Association between funding and quality of published medical education research. *JAMA* 2007;298:1002–1009.

48. APA Publications and Communication Board Working Group on Journal Article Reporting Standards. Reporting standards for research in psychology: why do we need them? What might they be? *Am Psychol* 2008;63:839–851.

Copyright © 2013 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.