MILITARY MEDICINE, 178, 10:64, 2013

Evaluation of Medical Simulations

William L. Bewley, PhD; Harold F. O'Neil, PhD†*

ABSTRACT Simulations hold great promise for medical education, but not all simulations are effective, and reviews of simulation-based medical education research indicate that most evaluations of the effectiveness of medical simulations have not been of sufficient technical quality to produce trustworthy results. This article discusses issues associated with the technical quality of evaluations and methods for achieving it in evaluations of the effectiveness of medical simulations. It begins with a discussion of the criteria for technical quality, and then discusses measures available for evaluating medical simulation, approaches to scoring simulation performance, and methodological approaches. It concludes with a summary and discussion of future directions in methods and technology for evaluating medical simulations.

INTRODUCTION

Since the first written clinical simulations were used for assessment nearly 50 years ago, simulations have become common in medical education.[1] Defined broadly as a "person, device, or set of conditions which attempts to present evaluation problems authentically,"[2] medical simulations emulate patients, anatomical areas, or clinical tasks. They include standardized patients,[3–8] part-task trainers (e.g., pelvic replicas),[9–14] virtual reality systems,[15] computer simulations[16,17] and games,[18] mannequins,[19–22] and even multiple-choice questions presenting information on a case to be evaluated.[23] Simulations can be used for instruction or assessment, and are currently used by many medical schools for end-of-course comprehensive examinations,[24] by the Medical Council of Canada as part of the licensure process[25] and as part of the United States Medical Licensing Examination, among many others.[26,27]

Simulation-based training has become popular because it is usually less costly, and it provides experiences without risk to patients.[28] In addition to the benefits of cost and risk avoidance, there are also benefits to learning.[29] Training can be directed at specific knowledge and skills, especially procedures and higher level cognitive processes, and some simulations can unobtrusively collect detailed data providing assessment information that can be used to automatically score performance and diagnose learning problems.[30] Simulations can also be used to provide experiences not possible in the real environment, such as repeated practice on parts of a task that cannot be isolated in the real world (e.g., intubation, venipuncture, tying surgical knots, or incision and drainage of abscesses). This is not to say that simulation-based training can replace training with real patients supervised by a knowledgeable instructor—nobody would want a surgeon trained only on simulations—but a useful level of knowledge and skill can be developed cost-effectively and safely with simulation-based training in preparation for training in the real environment. Medical simulations have great promise, but not all simulations are effective, and, unfortunately, reviews of simulation-based medical education research indicate that most evaluations of the effectiveness of medical simulations have not been of sufficient technical quality to produce trustworthy results.[31–34] This article discusses issues associated with technical quality and methods for achieving it in evaluations of the effectiveness of medical simulations. Note that the focus is on effectiveness, not cost. The article in this supplement by Fletcher and Wind[35] describes approaches to economic analyses that, with data on effectiveness using methods discussed in this article, can be used to determine cost-effectiveness or cost-benefit.

The article begins with a discussion of the criteria for technical quality, the measures available for evaluating medical simulations, approaches to scoring simulation performance, methodological approaches, and then describes an evaluation model. It concludes with a summary and discussion of future directions in methods and technology for evaluating medical simulations.

TECHNICAL QUALITY OF EVALUATIONS

Evaluations must satisfy two major criteria for technical quality: reliability and validity. This section discusses each. There are also two lesser but, nevertheless, important criteria that warrant mentioning in brief: fairness and usability. Fairness is an aspect of validity, and its absence is discussed later as a "threat to validity." Fairness means that inferences based on the results of the evaluation are appropriate for most people, of most backgrounds. In the measurement literature,[36] fairness is defined in terms of four properties:

- The test is free of bias.
- There is equal opportunity to show proficiency.
- In tests of knowledge and skill, there is equal opportunity to learn.
- Score distributions are as equal as possible across different groups.

*National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles, 10945 Le Conte Avenue, Suite 1400, Mailbox 957150, Los Angeles, CA 90095-7150.

†Rossier School of Education/National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of Southern California, 15366 Longbow Drive, Sherman Oaks, CA 91403.

The findings and opinions expressed here do not necessarily reflect the positions or policies of the Office of Naval Research.

doi: 10.7205/MILMED-D-13-00255

Of the four properties, bias has received the most attention in the measurement literature. Bias is defined as any construct-irrelevant source of variance that systematically affects the performance of different groups of examinees, e.g., groups defined by gender, ethnic or cultural background, socioeconomic status, or age.[37] Usability refers to practical considerations in conducting the evaluation, such as the cost of implementation as well as time requirements, ease of administration, and the comprehensibility of results to the intended audience. Usability is important, but not as important as reliability and validity.

Reliability

Reliability concerns the consistency of measurement, e.g., internal consistency or test/retest. It requires that results are consistent from one measurement to another, e.g., at different times, with different raters, or even with different (but considered equivalent) tasks. It requires that the evaluation methodology give the same result each time it is used. This is achieved through the use of well-defined and standardized procedures and measurement instruments.

Perfect consistency is not possible because people are not perfectly consistent. Simulation users may have learned or forgotten things, or may be under more or less stress on different days. Raters may not agree on interpretations of all judgment criteria, and a rater's criteria may change over time. Tasks may be more or less difficult for different users, depending on prior experience. All these factors introduce measurement error into evaluation results. Methods for determining reliability are based on determining the measurement error. The greater the consistency of results, the smaller the measurement error, and thus the greater the reliability.[36]

These methods are based on traditional psychometrics or classical test theory,[38] which is based on assumptions about how a test is constructed: linear, static, and homogeneous, providing many samples of behavior, and focused on between-individual differences[39]—think standardized tests, such as the Scholastic Aptitude Test.[40]

Most simulations, however, have fewer of these characteristics. Simulations are nonlinear, i.e., with more than one pathway to success or failure. They are frequently short, dynamic, adaptive, and heterogeneous, and provide relatively few samples of behavior. Finally, these assessment simulations are often focused on within-individual differences, including changes in performance during use of the simulation, as well as interindividual differences. In addition, classical test theory is not well suited for handling the complex correlations often found in data produced by simulations, for providing the real-time scoring and feedback often required for simulation-based assessments, or for providing measures of changes in proficiency over time.

In this supplement, Li Cai[41] describes alternatives to classical test theory appropriate for the psychometrics of medical simulation. These alternatives are based on a new generation of latent variable models applying Bayesian inferential methods to make inferences about latent variables from observed variables.

Simulations provide one long or a few short samples of behavior, rather than answers to many short questions (i.e., multiple choice), making the usual approaches to reliability inappropriate. As a result, approaches to reliability for simulations (and all performance assessments) have focused on the reliability of judges or raters scoring the performance rather than the "score" reliability of individuals.[42]

As noted earlier, the use of judges or raters introduces a source of error, along with characteristics of simulation users, the tasks, factors associated with testing occasion, e.g., time of day, and interactions of these sources. Generalizability theory is designed to allow identification of the sources of error and estimation of the contribution of each to a behavioral measurement.[43–45] Sources of error are called facets of the measurement. To evaluate the reliability of a measurement, a generalizability study is conducted to estimate the contribution of each facet and the interaction of facets. A decision study is then conducted to determine elements of a measurement procedure that minimizes error. For example, we can use generalizability theory to determine how many judges we need to make reliable assessments of performance. If judges differ in their interpretation of criteria or the evaluation is complex, more judges are needed to obtain an accurate measurement. But if judges agree on criteria or the evaluation is simple, fewer judges will be required.

In addition, because computer simulations are complex and take longer to complete, it may be the case that a small number of simulation trials can be administered in the time available for collection of data. This limits the generalizability of the results because, unlike selected response tests that provide equivalent forms, the problem of designing equivalent simulation scenarios (tasks) has not been solved. If time is available for only one assessment task, there is uncertainty as to whether performance on a different task thought to require the same knowledge and skills would provide the same results. Performance in one scenario will not necessarily be a good predictor of performance in another.

Validity

Validity is the degree to which evidence supports the interpretations and uses of results. Of the two major criteria for technical quality, reliability and validity, validity is the most important. The consistency measured by reliability makes it possible to have validity, but it is possible to have consistent results that are not valid.[36]

Validity is not a property of the evaluation; it is a property of the inferences made based on the results.[36] Validation should be thought of as an argument presenting evidence to make a case, and not, as with reliability, the calculation of a statistic. A validity argument must be developed that marshals a wide range of evidence to make the case.[36,37] This argument is very different from early conceptions of validity[46] in which specific validity types are considered, e.g., face validity (Does the test performance look like what is supposed to be measured?), content validity (Is the performance measured related to content goals or domains?), predictive validity (Do people with higher scores do better on a future criterion measure?), and criterion validity (Does performance on the new measure relate in predictable ways to an existing measure of known quality?). Although all these questions may be considered in making a validity argument, one no longer looks at a list of validity types and chooses 1 or 2 as most appropriate or, more likely, easiest to implement.

According to *Standards for Educational and Psychological Testing*,[36] there are five major sources of evidence that might be used to support a validity argument: evidence based on content, response processes, internal structure, relations to other variables, and consequences of testing. These are described below, along with two additional sources of evidence: threats to validity and sensitivity to instruction and experience.

–Evidence based on content. This is the weakest form of evidence for a validity argument. It is concerned with the representativeness of the content on which the simulation is based, not with examinee performance or the interpretation of the meaning of the performance.

–Evidence based on response processes. This has to do with the validity of interpreting examinee performance as evidence for the cognitive processes the examinees use when responding, e.g., some aspect of simulation performance is taken as evidence for situation assessment or problem solving skills. Evidence about response processes might be obtained by questioning the examinee about strategies used, or by using think-aloud protocols.[36]

–Evidence based on internal structure. Simulations are often designed to provide instruction and/or assessment on several knowledge or skill dimensions, such as situation awareness, planning, decision making, and communication. Evidence that these dimensions could be reliably distinguished based on examinee performance, by using the results of a confirmatory factor analysis,[47] would support the validity argument.

–Evidence based on relations to other variables. Correlations of examinee performance with other measures thought to be related also provide support for the validity argument.[36] Such evidence includes predictive accuracy, in which scores are correlated with a criterion measure that simulation performance is intended to predict, e.g., diagnosis performance with a standardized patient[3–8] and subsequent diagnosis with a real patient. Other examples are correlations with other measures designed to measure the same knowledge or skill, e.g., diagnosis performance with a standardized patient correlated with performance on a multiple-choice test presenting cases for diagnosis. Lack of correlation with measures designed to measure different knowledge or skill is another source of evidence. An example would be the relation of diagnosis performance with a standardized patient to intubation performance with a mannequin.

–Evidence based on consequences of testing. Use of a simulation has consequences for the examinee, especially when it is used for assessment. If results are due to knowledge or skills the simulation was designed to assess, this obviously supports the validity argument. If, however, results are due, at least in part, to knowledge or skills unrelated to what is to be assessed, such as a lack of computer skills interfering with performance on a computer simulation, validity should be questioned. This is an example of a "threat" to validity—an alternative explanation for good and poor performance. It is also an example of a lack of validity due to consequences of testing if it can be linked to an examinee characteristic that has nothing to do with the goal of the assessment, including membership in a particular socioeconomic group.

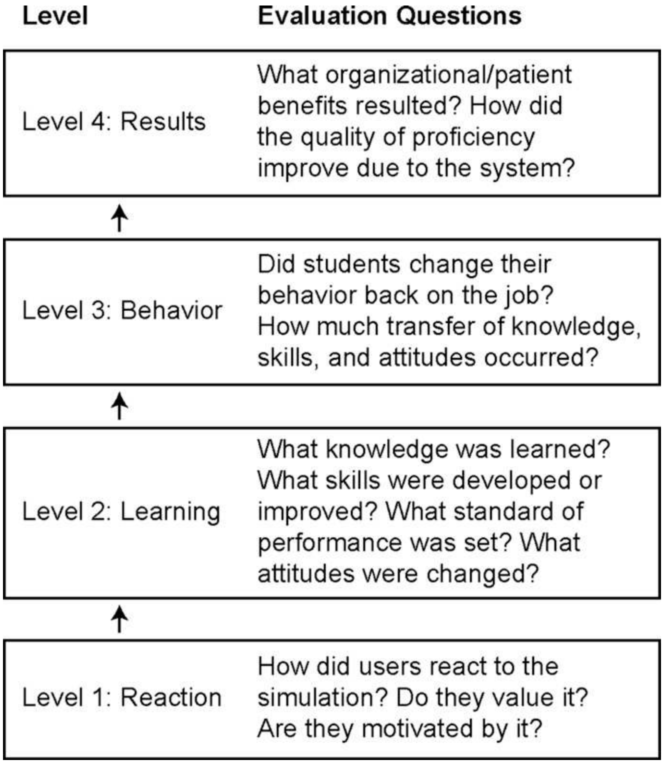
–Threats to validity. A validity argument is weakened by "threats" to validity, alternative explanations for good and poor performance unrelated to the knowledge or skill that is to be assessed. There are many potential threats: poor reliability; misalignment of the simulation experience and the knowledge/skill objectives; misalignment of the measures and objectives of the simulation; inadequate instructions, user interface defects, or lack of computer skills for computer simulations; unfair use of administration, such as inadequate instructions or time; inappropriate scoring models, e.g., scoring that does not accommodate all acceptable strategies; poor examinee sampling; and poor scenario selection (content sampling). To support the validity argument, all threats to validity should be identified and eliminated.

–Sensitivity to instruction and experience. A valid simulation should be sensitive to instruction and experience, eliciting higher scores for people who have received instruction or who have more experience or acknowledged expertise in the targeted knowledge or skill.

KIRKPATRICK MODEL

The Kirkpatrick model[48,49] is an evaluation framework that supports the idea of marshaling evidence to make a validity argument. It is also an approach for evaluation that has been successful in many different training and educational settings, and has become an industry standard in the training world. It has been adapted and modified over time, but the basic structure has not changed. As shown in Figure 1, the model describes four levels of evaluation. The levels are intended to represent a sequence of evaluation questions, each level providing information that affects the next level.

An evaluation is conducted at each level, beginning at Level 1 and moving up. Each level provides evidence for a validity argument and information supporting interpretation of results at the next level. For example, if there is no evidence for student learning at Level 2, reactions at Level 1 may tell us why—students may not be motivated to learn from the



Level | Evaluation Questions

Level 4: Results | What organizational/patient benefits resulted? How did the quality of proficiency improve due to the system?

Level 3: Behavior | Did students change their behavior back on the job? How much transfer of knowledge, skills, and attitudes occurred?

Level 2: Learning | What knowledge was learned? What skills were developed or improved? What standard of performance was set? What attitudes were changed?

Level 1: Reaction | How did users react to the simulation? Do they value it? Are they motivated by it?

FIGURE 1. The Kirkpatrick evaluation model.

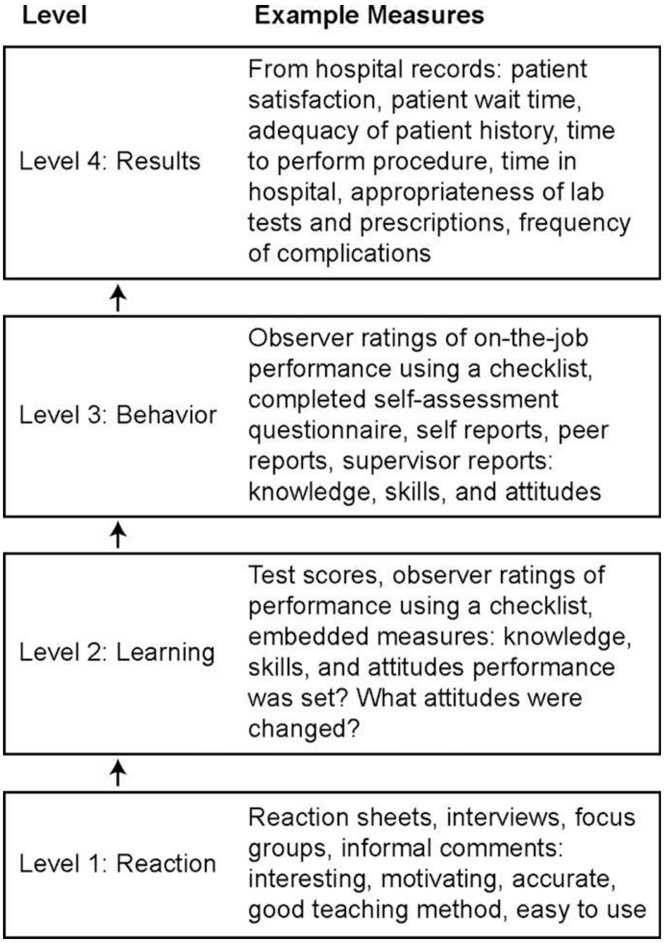
simulation. Similarly, a failure at Level 3 (no behavior change back on the job) may be explained by an absence of learning at Level 2. Difficulty increases as you move up, but the value of information also increases at each level. Kirkpatrick recommends evaluating at all levels, but in practice, because the difficulty and cost increase at each level and because Level 3 and especially Level 4 may be difficult in the work environment, it may be tempting to stop at Level 2, or even Level 1, but Kirkpatrick emphasizes the impact of misalignment of measures to goals on validity. For example, if the objective is transfer of knowledge, skills, or attitudes to performance on the job, you need to go to Level 3 for a valid evaluation. And if the objective is organizational/patient benefit, a Level 4 evaluation is required.

SIMULATION PERFORMANCE MEASURES (PROCESS VS. OUTCOME)

A measure is a number indicating the presence and amount of something, such as the number of errors, time, or ratings of some aspect of simulation performance on a five-point scale. McNulty et al[50] provide an excellent overview of computer-based testing in the medical curriculum. We will focus on computer simulations. One of the great advantages of a simulation is the ability to measure knowledge and skills in performing procedures and higher level cognitive processes. This measurement is based on the examinee's actions as the task is performed, in addition to measures focused on the outcome of the process such as a rating of overall success, for example, measurement of the value of a physiological indicator like blood glucose level, albumin level, or blood pressure. As noted earlier, a key requirement for achieving validity is the use of appropriate measures aligned with the intended objectives of the simulation, usually related to knowledge and skill required to perform the simulated task. This seems obvious, but there are many examples of misalignment of measures with objectives. An extreme example is the evaluation that measures learning using reaction forms or opinion surveys asking students how much they learned.[51,52] This provides information on how much students think they learned, not how much they actually learned.

Figure 2 shows examples of measures for each Kirkpatrick level.

Measures must tap the entire range of knowledge and skills at the same level of complexity addressed by the simulation, and they must be validated for the purposes and situations to which they are applied. Swick et al[53] provide an excellent treatment of assessing the Accreditation Council



Level | Example Measures

Level 4: Results | From hospital records: patient satisfaction, patient wait time, adequacy of patient history, time to perform procedure, time in hospital, appropriateness of lab tests and prescriptions, frequency of complications

Level 3: Behavior | Observer ratings of on-the-job performance using a checklist, completed self-assessment questionnaire, self reports, peer reports, supervisor reports: knowledge, skills, and attitudes

Level 2: Learning | Test scores, observer ratings of performance using a checklist, embedded measures: knowledge, skills, and attitudes performance was set? What attitudes were changed?

Level 1: Reaction | Reaction sheets, interviews, focus groups, informal comments: interesting, motivating, accurate, good teaching method, easy to use

FIGURE 2. Typical measures for Kirkpatrick evaluation model levels.

for Graduate Medical Education competencies in psychiatric programs. Brünken et al[54] provide indicators for measuring cognitive load, and Hays[55] provides various rating scales for evaluating computer-based instruction. To evaluate simulations targeting procedural or higher level knowledge and skills, measures derived from simulation performance are desirable. There are two sources of measures: (1) human raters score performance using checklists based on scoring rubrics, and (2) automated scoring based on measures embedded in the simulation itself. For example, in tasks performed by manipulating objects on a computer screen, a mannequin, or an anatomic model, it may be possible to record the actions of the examinee in performing the task, including mouse clicks on a computer screen or actions on a physical device, with the associated location, time, and task context as appropriate.[56]

CHECKLISTS

The easiest and most widely used approach to scoring (and the only feasible approach when automated scoring is not possible) is to use checklists consisting of explicit outcome and/or process criteria. Scoring rubrics are used to assign scores to each item, and the scores can be weighted to account for the importance of the item. Checklists are used with standardized patient-based tests (e.g., Swanson,[57] van der Vleuten and Swanson[58]) with written and computer-based clinical simulations or computer-based case simulations, also called patient management problems,[1] and with mannequins.[59–61] The standardized patients may do the rating in standardized patient-based tests. People with clinical expertise serve as raters for the other simulation types and for some standardized patient-based tests. Ratings can be done live or by reviewing videotapes.

In addition to being the only feasible approach when automated scoring using embedded measures is not possible, checklists have the benefit of being objective for recording clearly observable examinee actions such as questions and physical examination maneuvers. Rater training is required, and with training, raters can be very accurate.[62] Inter-rater reliability, the degree of agreement among raters, should always be measured.

Potential problems with checklists include the difficulty in developing rubrics that appropriately reward different strategies that are similar in quality and similar strategies that differ in quality.[1] Also, it can be difficult to develop weights to accommodate more and less important actions, and if weights are large or negative, scoring can be complex, which can lead to inconsistencies that compromise reliability, and the examinee could get a high or low score based on a single action. Holistic scoring, focusing on the outcome or process as a whole rather than breaking it into separate parts (i.e., analytic scoring) has also been used. It has been criticized as subjective, but, with good rater training, has been shown to work.[62,63]

AUTOMATED SCORING

There have been multiple frameworks for evaluation and use of automated scoring (see Williamson et al[64] and Shermis and Burstein[65]). We organize the literature into three major approaches: expert-based methods, data-driven methods, and domain-modeling methods.

Expert-Based Methods

There are two expert-based methods: using expert performance and modeling expert judgment. In the first approach, actual expert performance is considered the gold standard against which student performance is compared,[66,67] not what experts say should be competent performance or how experts rate student performance. This approach has been used to develop tasks for content understanding using essays[67] and knowledge maps.[68]

A related approach is to model experts' rating of examinees' performance on various task variables. Expert judgment is considered the gold standard against which student performance is compared, not actual expert performance. This scoring approach has been used successfully to model expert and rater judgments in a variety of applications including essays[69] and patient management skills.[30]

One of the major issues with expert-based scoring is the selection of the expert.[70,71] Problems include experts' biases, the influences of the experts' content and world knowledge, linguistic competency, expectations of student competency, and instructional beliefs.[72]

Data-Driven Techniques

In data-driven techniques, performance data are subjected to statistical or machine-learning analyses (e.g., artificial neural networks with hidden Markov models). Using artificial neural network and hidden Markov model technologies, Ron Stevens et al[73] have developed a method for identifying learner problem-solving strategies and modeling learning trajectories, or sequences of performance states. Applying the method to chemistry, they were able to identify trajectories revealing learning problems that include not thoroughly exploring the problem space early, reaching a performance state that makes it unlikely to reach a more desirable end state, and reaching a state from which the learner could transition to a better or worse state with equal likelihood. With this information, it may be possible to perform a fine-grained diagnosis of what learners do not know and to use learning trajectories to guide the sequence of instruction and the type and form of remediation, and to do it impromptu.

Validation of data-driven methods is complicated because there is no a priori expectation of what scores mean and no inherent meaning of the classification scheme. Interpretation is post hoc, which creates the potential for the introduction of bias in assignments to groups after the groups have been defined.[74] A second problem is that machine learning techniques can be highly sample-dependent and the scoring process is driven by statistical rather than theoretical issues.[71] Because of these issues, validity evidence is particularly important when using data-driven techniques to score student responses.

Domain Modeling

This approach attempts to model the cognitive demands of the domain itself. The model specifies how knowledge and skills influence each other and the task variables on which observations are being made. The approach relies on a priori linking of student performance variables to hypothesized knowledge and skill states. Student knowledge and skills are then interpreted in light of the observed student performance. This approach has been used successfully in a variety of domains and modeling types, from canonical items (e.g., Hively et al[75]); to Tatsuoka's rule-space methodology;[76] to the use of Bayes nets to model student understanding in domains such as Web searching,[77] rifle marksmanship,[78] hydraulic troubleshooting,[79] dental hygiene skills,[80] network troubleshooting,[81] and circuit analyses.[82]

The most important issue in domain modeling is identifying the essential concepts and their interrelationships. This can be mitigated through cognitive task analyses and direct observation of performance, but it is critical to gather validity evidence to validate the structure of and inferences drawn by the Bayes net. For examples of empirical validation techniques, see Chung et al[78] and Williamson et al.[83]

METHOD SELECTION

For evaluations conducted at each Kirkpatrick level, the methods used are important because they affect the quality of the evaluation. Method selection and design are not easy tasks because medical simulation evaluation is very difficult, for all the reasons any educational research is difficult, and there are additional obstacles that come with the use of technology. The effectiveness of a simulation is due to a combination of factors, not one, and these factors may interact in complex ways. The instructional experience depends on many variables, including instructor background, teaching philosophy, training, and experience; the support of school management; and characteristics of the students.[84] And when technology is part of the experience, there are additional variables, including availability of hardware, software, and technical support; curriculum integration strategies; students' prior experience with and expertise in using technology; and instructor expertise in technology and skill in implementing the simulation.[84]

This section presents an overview of three major methodological approaches, the random-assignment experiment, quasi-experiments, and alternatives based on qualitative methods, and then we discuss combined methods. We end with a discussion of heuristics for matching methods to situations (or research questions). For an excellent and detailed treatment of these issues see Shadish et al.[85]

Random-Assignment Experiments

A random-assignment experiment requires random assignment of the unit of treatment application, e.g., students, instructor, or the school, to experimental and control groups. The unit of treatment application is the unit of analysis, and it defines the sample size. Random assignment is required to achieve equivalent groups in terms of variables not explicitly controlled by the evaluator. Variables explicitly controlled by the evaluator are the treatment—the introduction of the simulation—and all measures and procedures that may affect the results.

For examples of the use of random-assignment experiments see Adler et al,[86] Boulet and Swanson,[23] and Robinson et al.[87] The argument for the use of random-assignment experiments is that they provide better evidence for causal inferences than any other method. This is true, assuming that the conditions required for experiments are met. The difficulty of meeting these conditions has led to strong objections to experiments in education research, including simulation evaluations. The key problem is the requirement for random assignment to experimental groups. Medical schools do not typically assign students to classrooms and instructors randomly, and students and instructors are not randomly assigned to schools. It is also difficult to meet the requirement for a control group not receiving the treatment. Students (and instructors) do not readily accept withholding the use of technology for the sake of an experiment. It may also be the case that simulation use in other classes is so widespread that it is difficult or impossible to have a control group with no experience that might be relevant. And many argue that the goal of simulation is to provide experiences not possible without the simulation, which means that it is impossible to have a control group receiving the same experience but without the simulation.

A related problem is the need for an adequate sample size. The point of conducting an experiment, either a random-assignment experiment or a quasi-experiment as described below, is to detect a difference between groups in the study sample when a difference actually exists in the populations from which the samples are drawn. The probability of detecting such a difference is called the power of a statistical test. Obviously, the power should be high, so that if there is no difference between groups in the experiment, it is reasonable to conclude that there is no difference in reality. The power of a study depends on several factors, including the statistical test, significance criterion, measurement error, and the size of the experimental effect, but the general approach to increasing power is to increase the sample size. Despite this, as reported by Moher et al,[88] researchers often use sample sizes too small to achieve power adequate to detect real effects, and most do not even report a sample size calculation. For information on calculating sample size, see Cohen[89,90] and Lenth.[91] Lenth[92] provides an online tool for power and sample size calculations.

Another criticism of the experimental approach is that although it provides better evidence for causal inferences, it does not provide information on why the simulation had its effects. The argument is that the experiment is a black box that provides evidence of connections between causes and effects, but does not provide information on the processes inside the box that explain why the simulation caused the effects, many of which are based on the context of the simulation.

Finally, there are the practical problems of cost and time. Experiments are expensive and time-consuming. They may require all the funds available for evaluation and take so long to complete that decisions are made before results are available. Whether this is unique to random-assignment experiments is arguable, but it is a common criticism nonetheless.

Quasi-Experiments

Quasi-experiments have many of the features of experiments except random assignment to experimental and control groups and appropriate control of selected variables, such as the timing of exposure to the simulation.[85] One example is the time-series experiment, in which periodic measurements are taken over time and an experimental change is inserted at some point in the time series of measurements. Changes after insertion may indicate an effect caused by the experimental change, but may also be caused by other events occurring during the time series because there is no control over events other than the introduction of the experimental change.

Another example is the nonequivalent control group design, one of the more common designs in educational research. There is an experimental group and a control group. Both are given a pretest and a posttest, but only the experimental group receives the experimental treatment between the two tests. This is similar to an experimental design, but students are not randomly assigned to each group. Causation can be inferred if there is an experimental versus control group difference in the posttest score. Because the two groups are naturally assembled, e.g., students in two different classes, not randomly assigned, they cannot be considered equivalent; and it is possible that some difference affecting the groups other than the experimental treatment could be the cause. Although this may seem unlikely, it is possible. The point is that the evidence from quasi-experiments is not as strong as the evidence from random-assignment experiments, but it is also true that quasi-experiments are usually more feasible and practical in an education setting. For an example of a quasi-experiment, see the article by Giuliano et al.[93]

Qualitative Methods

Qualitative methods do not attempt to compare experimental and control groups at all, or to control variables. They investigate the simulation through observation, review of artifacts, and interviews, studying cases in their natural setting to consider variables as they appear in all the complexity of the context.[94] These methods are very popular in education research, including evaluation of simulations, due in part to the difficulties in doing experimental research in educational settings, and in part to the desire to obtain information on why the simulation had its effects—the processes and mechanisms that lead from specifics of the simulation to effects—and the contextual conditions under which the simulation is more or less effective. The focus is on the context of the simulation, such as local engagement, collaboration, and feedback, and investigating why those results occurred. Understanding the cause of the result involves developing a theory of change, a description of the processes through which the effects are produced. Qualitative methods are weak on causal inference, but the contextualization makes them very useful to decision makers by providing models (theories of change) describing how and why the simulation works or does not work in the existing system and information needed to decide whether, how, and when to use the simulation.

Qualitative methods are especially useful for studying a broad range of naturally occurring practices found in many different parts of the school, not from a particular simulation, which would usually be evaluated with an experiment. Such studies are often descriptive, interested in the frequency of various instructional technology uses and practices, not their effects. Some correlate descriptive data with student outcomes to attempt to identify relationships, if not causes. Concluding anything about causation from correlations is, of course, problematic. For an example of a technology evaluation using qualitative methods, see the article by Overly et al.[95]

Combined Methods

As is usually the case when there is a debate over the merits of radically different points of view, the practical truth lies somewhere in between. There is no one right way to do technology evaluation. The approach depends on the purpose of the evaluation, the nature of the simulation, and the context in which it is situated. Some will require quantitative methods, some will require qualitative methods, and usually the evaluation will benefit from a combination providing both quantitative and qualitative data on student learning and attitude outcomes, context, the military environment, and the implementation of the simulation.

Selecting Methods

This section describes a heuristic process for deciding when to use what research methods and combinations of methods. The decision depends on the purpose of the evaluation, the nature of the simulation, the context in which it is situated, and practical constraints including site cooperation and available time, funding, equipment, and support resources. The choice need not be limited to a single design. Depending on the purpose, simulation, context, and practical constraints, the evaluation may and usually should consist of a combination of methods.

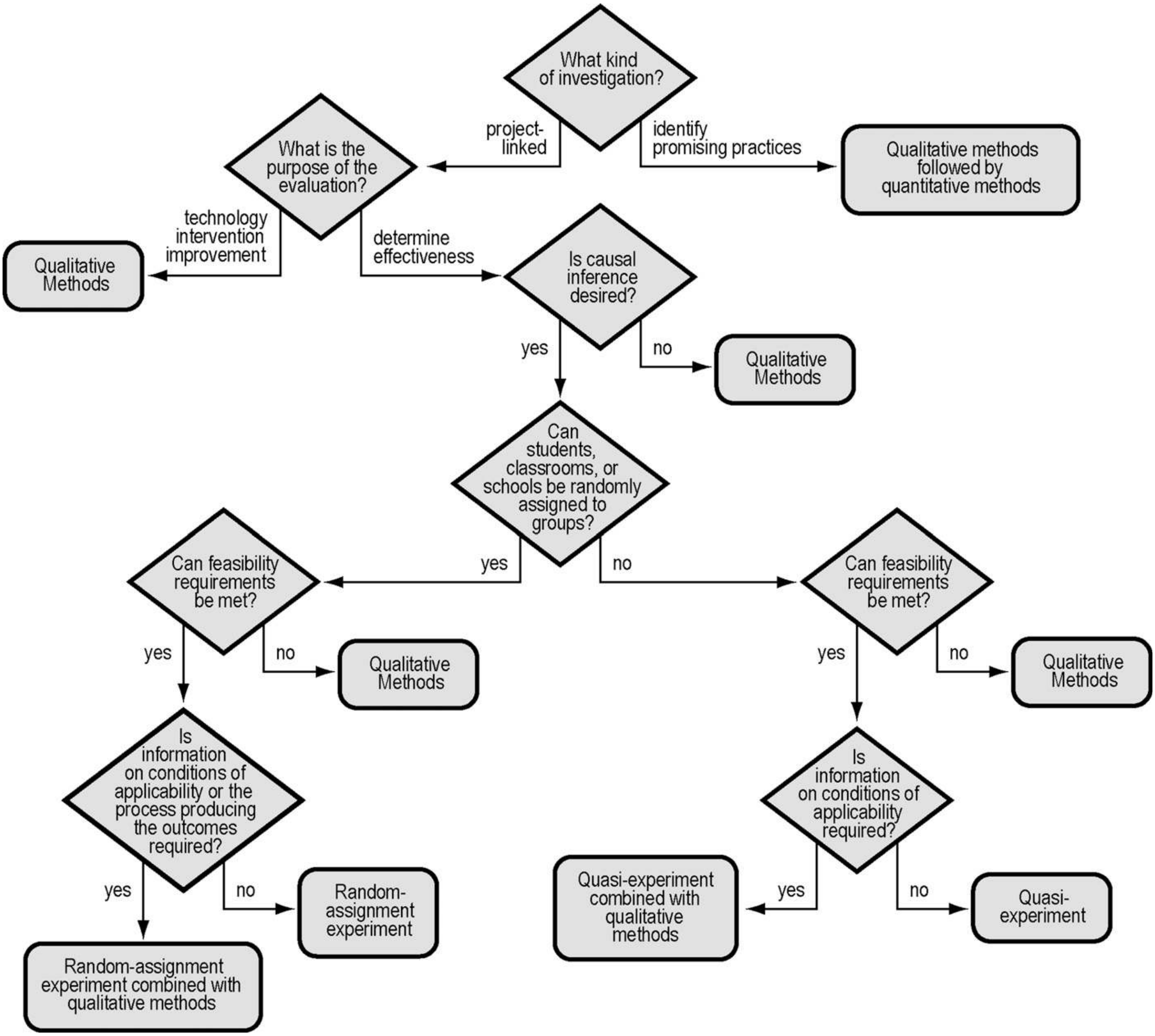

What kind of investigation?
project-linked
identify promising practices
Qualitative methods followed by quantitative methods
What is the purpose of the evaluation?
technology intervention improvement
Qualitative Methods
determine effectiveness
Is causal inference desired?
yes
no
Qualitative Methods
Can students, classrooms, or schools be randomly assigned to groups?
yes
no
Can feasibility requirements be met?
yes
no
Qualitative Methods
Can feasibility requirements be met?
yes
no
Qualitative Methods
Is information on conditions of applicability or the process producing the outcomes required?
yes
no
Random-assignment experiment
Random-assignment experiment combined with qualitative methods
Is information on conditions of applicability required?
yes
no
Quasi-experiment combined with qualitative methods
Quasi-experiment

FIGURE 3. A heuristic process for selecting evaluation methods.

Figure 3 summarizes a heuristic process for matching evaluation methods to situations and requirements.

The process is organized into the following set of guidelines, presented as questions followed by recommendations.

1. Is the evaluation concerned with the impact of a specific simulation or with identifying promising practices?

   –If it is identifying promising practices, the evaluation should start with a quantitative study to identify successful sites based on some measure, and then qualitative methods should be used to understand the differences between successful and unsuccessful sites and the practices related to success.

   –If the investigation is concerned with a specific simulation, there is a question on the purpose of the evaluation—question 2.

2. Is the purpose of the evaluation to improve the simulation or determine its effectiveness?

   –If the purpose is to improve the simulation, the evaluation is a "formative" evaluation. Formative evaluations are used to improve early stage projects by collecting information that can be used to guide the development and implementation of the intervention. This requires the use of qualitative methods to provide information on how the simulation works. The evaluator will be

interested in how features of the environment interact with features of the simulation, and how features of the simulation will influence behavior.
–If the purpose is to determine the effectiveness of the simulation, the evaluation is a "summative" evaluation. In this case there is a question on the need for causal information—question 3.

3. Is causal information needed?

   –If causal information is not needed, qualitative methods are appropriate.
   –If causal information is needed, quantitative methods are indicated. Random-assignment experiments are best for determining causation and should be considered first, but before selecting an experiment, there is a question on the feasibility of random assignment—question 4.

4. Is it possible for students, classes, or schools to be randomly assigned to conditions?

   –If the answer is yes, a random-assignment experiment may be possible, depending on the answer to question 5.
   –If the answer is no, a quasi-experiment may be possible, depending on the answer to question 5.

5. Is an experiment feasible? Before selecting a random-assignment experiment or quasi-experiment, the feasibility of conducting either must be determined. For either experiment type to be feasible, it must satisfy the following requirements:

   —Use of the simulation must be different from standard practice in order to achieve a meaningful comparison.
   —Use of the simulation must be maintainable, that is, it must continue unchanged for the course of the experiment.
   —Participation must not deny students access to an entitlement, e.g., access to an instructional experience.
   —Human subjects protection requirements must be met.
   —Participants and the site must be willing to cooperate.
   —An adequate sample size must be available.
   —Time, funding, equipment, and support resources must be available.
   –If feasibility requirements cannot be met, qualitative methods should be used.
   –If feasibility requirements can be met for either experiment type, there is a question on the need for information on context—question 6.

6. Is there a requirement for information on conditions of applicability or the process producing the outcomes?

   –If the answer is yes, and this should usually be the case, an experiment (random-assignment or quasi-experiment, whichever is indicated in question 4) combined with qualitative methods for the contextual information is appropriate.
   –If the answer is no, the experiment is sufficient.

If random assignment is not possible, but feasibility requirements can be met, and there is a requirement for information on conditions of applicability or the process producing the outcomes, a quasi-experiment combined with qualitative methods would be appropriate. If there is no requirement for conditions of applicability or process, which should not be the usual case, a quasi-experiment is appropriate. And as with the random-assignment experiment branch of the method selection process, if the quasi-experiment or quasi-experiment/qualitative method combination are not appropriate, qualitative methods are the choice.

SUMMARY AND DISCUSSION

This article has presented an overview of issues and approaches relevant to evaluating medical simulations. It discusses criteria for the technical quality of evaluations, and methods for achieving it. It introduces the Kirkpatrick model, a proven evaluation model supporting the idea of marshaling evidence to make a validity argument. It discusses measures, approaches to scoring, and research methods used to provide evidence, with guidelines for selecting appropriate methods.

Takeaway Message

Medical simulations have great promise for training complex high-value tasks at less cost and without risk to patients. However, great promise and impressive technical capability are not sufficient to conclude effectiveness. To realize the promise, practitioners must assess the systems and the learning they help produce, and the evaluations must have technical quality. The article's central takeaway message is the importance of technical quality—reliability and, especially, validity—as the fundamental requirement for any evaluation. The message is linked to two supporting ideas:

1. Validity is not a general quality of an evaluation. An evaluation's validity depends on the context of its use and the inferences to be drawn based on the results. A validity argument must be made using a wide range of evidence for the appropriateness of the inferences for the particular context.[36]
2. Begin with a definition of the objectives. The first step in evaluation design is to define the objectives of the simulation—the knowledge and skill required for success. This leads to defining measures, operationalizing the scoring, and then validating the approach with empirical evidence.
3. Align measures, scoring, and research methods with the objectives. Validity requires alignment with the objectives. Evaluate at all levels of the Kirkpatrick model if possible, but always at the level matching the objectives.

Future Directions

Although not widely used in current medical simulations, we expect greater use of automated scoring based on measures embedded in the simulation itself. Because of the growing sophistication of computationally supported data collection, and the importance of formative information about the trainee's process during learning, in the future outcome measures will merge with process measures to create learner profiles rather than scores or classifications. We anticipate that these will have domain-independent components that may predict learners' likely success in a range of other tasks. We see the study of expertise continuing to add to our knowledge of performance measurement and its validity, and we also predict an increased use of artificial intelligence and advanced decision analysis techniques to support assessment and evaluation. These include ontologies, Bayes nets, artificial neural networks, hidden Markov models, lag sequential analysis, and constraint networks.

Test development guidelines have been developed from lessons learned in the assessment of clinical competence literature.[96] The same is needed for medical simulation design and evaluation based on lessons learned in the evaluation of medical simulations. The Federal Medical Simulation Training Consortium, a partnership of the Department of Defense and other federal institutions involved in medical training and education, is taking a major step in this direction, working with the University of California, Los Angeles Center for Research on Evaluation, Standards, and Student Testing to develop a framework to guide evaluation and refinement of existing curricula (including but not limited to simulations) and development of new curricula, and a set of training effectiveness metrics to allow comparison of curricula.

ACKNOWLEDGMENTS

The work reported herein was supported by a grant from the Office of Naval Research, Award Number N00014-10-1-0978.

REFERENCES

1. Swanson DB, Norcini JJ, Grosso J: Assessment of clinical competence: written and computer-based simulations. Assess Eval Higher Educ 1987; 12: 220–46.
2. McGaghie WC, Issenberg SB: Simulations in professional competence assessment: basic considerations. In: Innovative Simulations for Assessing Professional Competence, pp 7–22. Edited by Tekian A, McGuire CH, McGaghie WC. Chicago, Department of Medical Education, University of Illinois at Chicago, 1999.
3. Barrows HS, Abrahamson S: The programmed patient: a technique for appraising student performance in clinical neurology. J Med Educ 1964; 39: 802–5.
4. Collins JP, Harden RM: The Use of Real Patients, Simulated Patients and Simulators in Clinical Examinations (AMEE Medical Education Guide, No. 13). Dundee, UK, Association for Medical Education in Europe, 2004.
5. Wilson L, Rockstraw L (editors.): Human Simulation for Nursing and Health Professions. New York, Springer, 2012.
6. Gerner B, Sanci L, Cahill H, et al: Using simulated patients to develop doctors' skills in facilitating behaviour change: addressing childhood obesity. Med Educ 2010; 44: 706–15.
7. Betcher DK: Elephant in the room project: improving caring efficacy through effective and compassionate communication with palliative care patients. Medsurg Nurs 2010; 19: 101–5.
8. Safdieh JE, Lin AL, Aizer J, et al: Standardized patient outcomes trial (SPOT) in neurology. Med Educ Online 2011; 16(1): 1–6.
9. Marecik SJ, Prasad LM, Park JJ, et al: A lifelike patient simulator for teaching robotic colorectal surgery: how to acquire skills for robotic rectal dissection. Surg Endosc 2008; 22: 1876–81.
10. Crochet P, Aggarwal R, Dubb SS, et al: Deliberate practice on a virtual reality laparoscopic simulator enhances the quality of surgical technical skills. Ann Surg 2011; 253(6): 1216–22.
11. Lee JT, Son JH, Chandra V, Lilo E, Dalman RL: Long-term impact of a preclinical endovascular skills course on medical student career choices. J Vasc Surg 2011; 54: 1193–200.
12. Privett B, Greenlee E, Rogers G, Oetting TA: Construct validity of a surgical simulator as a valid model for capsulorhexis training. J Cataract Refract Surg 2010; 36: 1835–8.
13. Coles TR, John NW: The Effectiveness of Commercial Haptic Devices for Use in Virtual Needle Insertion Training Simulations. In: 2010 Third International Conference on Advances in Computer-Human Interactions, pp 148–53. Piscataway, NJ, The Institute of Electronic and Electrical Engineers, 2010. Available at http://www.computer.org/csdl/proceedings/achi/2010/3957/00/3957a148-abs.html; accessed May 7, 2013.
14. Barsuk JH, McGaghie WC, Cohen ER, O'Leary KJ, Wayne DB: Simulation-based mastery learning reduces complications during central venous catheter insertion in a medical intensive care unit. Crit Care Med 2009; 37: 2697–701.
15. Ahlberg G, Enochsson L, Gallagher AG, et al: Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. Am J Surg 2007; 193: 797–804.
16. Cook DA, Triola MM: Virtual patients: a critical literature review and proposed next steps. Med Educ 2009; 43(4): 303–11.
17. Cendan JC, Lok B: The use of virtual patients in medical school curricula. Adv Physiol Educ 2012; 36(1): 48–53.
18. Cannon-Bowers JA, Bowers C, Procci K: Using video games as educational tools in healthcare. In: Computer Games and Instruction, pp 47–72. Edited by Tobias S, Fletcher JD. Charlotte, NC, Information Age Publishing, 2011.
19. 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: 1477–85.
20. Alinier G, Hunt WB, Gordon R: Determining the value of simulation in nurse education: study design and initial results. Nurse Educ Pract 2004; 4(3): 200–7.
21. Radhakrishnan K, Roche JP, Cunningham H: Measuring clinical practice parameters with human patient simulation: a pilot study. Int J Nurs Educ Scholarsh 2007; 4: Article 8.
22. Cendan JC, Johnson TR: Enhancing learning through optimal sequencing of web-based and manikin simulators to teach shock physiology in the medical curriculum. Adv Physiol Educ 2011; 35(4): 402–7.
23. Boulet JR, Swanson DB: Psychometric challenges of using simulations for high-stakes assessment. In: Simulators in Critical Care Education and Beyond, pp 119–30. Edited by Dunn WF. Des Plaines, IL, Society of Critical Care Medicine, 2004.
24. Scalese RJ, Obeso VT, Issenberg SB: Simulation technology for skills training and competency assessment in medical education. J Gen Intern Med 2008; 23(Suppl 1): 46–9.
25. Dauphinee WD, Reznick R: A framework for designing, implementing, and sustaining a national simulation network: building incentive-based network structures and iterative processes for long-term success: the case of the Medical Council of Canada's Qualifying Examination, Part II. Simul Healthc 2011; 6(2): 94–100.
26. Dillon GF, Boulet JR, Hawkins RE, Swanson DB: Simulations in the United States Medical Licensing Examination (USMLE). Qual Saf Health Care 2004; 13(Suppl 1): i41–5.
27. Dillon GF, Clauser BE: Computer-delivered patient simulations in the United States Medical Licensing Examination (USMLE). Simul Healthc 2009; 4: 30–4.

28. Bradley P: The history of simulation in medical education and possible future directions. Med Educ 2006; 40: 254–62.
29. Larsen CR, Soerensen JL, Grantcharov TP, et al: Effect of virtual reality training on laparoscopic surgery: randomised controlled trial. BMJ 2009; 338: b1802.
30. Margolis MJ, Clauser BE: A regression-based procedure for automated scoring of a complex medical performance assessment. In: Automated Scoring of Complex Tasks in Computer-Based Testing, pp 123–67. Edited by Williamson DM, Behar II, Mislevy RJ. Mahwah, NJ, Erlbaum, 2006.
31. Issenberg SB, McGaghie WC, Petrusa ER, Gordon DL, Scalese RJ: Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach 2005; 27(1): 10–28.
32. Bordage G, Caelleigh AS, Steinecke A, et al: Review criteria for research manuscripts. Acad Med 2001; 76: 897–978.
33. Lurie SJ: Raising the passing grade for studies of medical education. JAMA 2003; 290: 1210–2.
34. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ: A critical review of simulation-based medical education research: 2003-2009. Med Educ 2010; 44: 50–63.
35. Fletcher JD, Wind AP: Cost considerations in using simulations for medical training. Mil Med 2013; 178(10)(Suppl): 37–46.
36. American Educational Research Association, American Psychological Association, and National Council for Measurement in Education: Standards for Educational and Psychological Testing. Washington, DC, American Educational Research Association, 1999.
37. Miller MD, Linn R, Gronlund N: Measurement and Assessment in Teaching, Ed 11. Upper Saddle River, NJ, Prentice Hall, 2012.
38. Gulliksen HO: Theory of Mental Tests. New York, John Wiley, 1950.
39. Nunnally JC, Bernstein IH: Psychometric Theory, Ed 3. New York, McGraw-Hill, 1994.
40. Liu J, Harris DJ, Schmidt A: Statistical procedures used in college admissions testing. In: Handbook of Statistics, Volume 26: Psychometrics, pp 1057–94. Edited by Rao CR, Sinharay S. New York, Elsevier, 2007.
41. Cai L: Potential applications of latent variable modeling for the psychometrics of medical simulation. Mil Med 2013; 178(10)(Suppl): 115–20.
42. Patz RJ, Junker BW, Johnson MS, Mariano LT: The hierarchical rater model for rated test items and its application to large-scale educational assessment data. J Educ Behav Stat 2002; 27(4): 341–84.
43. Shavelson RJ, Webb NM: Generalizability Theory: A Primer. Thousand Oaks, CA, Sage, 1991.
44. Brennan RL: Generalizability Theory. New York, Springer-Verlag, 2001.
45. Chiu CWC: Scoring Performance Assessments Based on Judgements: Generalizability Theory. New York, Kluwer, 2001.
46. Messick S: Validity. In: Educational Measurement, Ed 3, pp 13–103. Edited by Linn R. Phoenix, AZ, The Oryx Press, 1993.
47. Thompson B: Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications. Washington, DC, American Psychological Association, 2004.
48. Kirkpatrick DL, Kirkpatrick JD: Evaluating Training Programs: The Four Levels, Ed 3. San Francisco, Berrett-Koehler, 2006.
49. Kirkpatrick DI: Evaluating Training Programs: The Four Levels, ED 2. San Francisco, Berrett-Koehler, 1998.
50. McNulty JA, Halama J, Espiritu B: Evaluation of computer-aided instruction in the medical gross anatomy curriculum. Clin Anat 2004; 17: 73–8. doi: 10.1002/ca.10188
51. Via DK, Kyle RR, Trask JD, Shields CH, Mongan PD: Using high-fidelity patient simulation and an advanced distance education network to teach pharmacology to second-year medical students. J Clin Anesth 2004; 16(2): 144–51.
52. Fitch MT: Using high-fidelity emergency simulation with large groups of preclinical medical students in a basic science course. Med Teach 2007; 29: 261–3.
53. Swick S, Hall S, Beresin E: Assessing the ACGME competencies in psychiatry training programs. Acad Psychiatry 2006; 30: 330–51.
54. Brünken R, Seufert T, Paas F: Measuring cognitive load. In: Cognitive Load Theory, pp 181–202. Edited by Plass J, Moreno R, Brünken R. New York, Cambridge University Press, 2010.
55. Hays RT: The Effectiveness of Instructional Games: A Literature Review and Discussion. Technical report 2005–004. Orlando, FL, Naval Air Warfare Center Training Systems Division, 2005. Available at http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA441935; accessed May 7, 2013.
56. Bewley WL, Chung GKWK, Delacruz GC, Baker EL: Assessment models and tools for virtual environment training. In: The PSI Handbook of Virtual Environments for Training and Education: Developments for the Military and Beyond, Vol. 1, pp 300–13. Edited by Schmorrow D, Cohn J, Nicholson D. Westport, CT, Greenwood Publishing, 2009.
57. Swanson DB: A measurement framework for performance-based tests. In: Further Developments in Assessing Clinical Competence, pp 13–45. Edited by Hart I, Harden R. Montreal, Can-Heal Publications, 1987.
58. van der Vleuten C, Swanson DB: Assessment of clinical skills with standardized patients: state of the art. Teach Learn Med 1990; 2: 58–76.
59. Morgan PJ, Cleave-Hogg D, DeSousa S, Tarshis J: High-fidelity patient simulation: validation of performance checklists. Br J Anaesth 2004; 92(3): 388–92.
60. 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.
61. 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.
62. Boulet JR, McKinley DW, Whelan GP, Hambleton RK: Quality assurance methods for performance-based assessments. Adv Health Sci Educ Theory Pract 2003; 8: 27–47.
63. Regehr G, MacRae H, Reznick R, Szalay D: Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med 1998; 73: 993–7.
64. Williamson DM, Xi X, Breyer FJ: A framework for evaluation and use of automated scoring. Educ Meas 2012; 31(1): 2–13.
65. Shermis MD, Burstein JC (editors): Automated Essay Scoring: A Cross-Disciplinary Perspective. Mahwah, NJ, Erlbaum, 2003.
66. Baker EL: Model-based performance assessment. Theory Pract 1997; 36(4): 247–54.
67. Baker EL, Freeman M, Clayton S: Cognitive assessment of history for large-scale testing. In: Testing and Cognition, pp 131–53. Edited by Wittrock MC, Baker EL. Englewood Cliffs, NJ, Prentice-Hall, 1991.
68. Herl HE, O'Neil HF Jr., Chung GKWK, Schacter J: Reliability and validity of a computer-based knowledge mapping system to measure content understanding. Comput Human Behav 1999; 15: 315–33.
69. Burstein J: The e-rater scoring engine: automated essay scoring with natural language processing. In: Automated Essay Scoring: A Cross-Disciplinary Perspective, pp 113–22. Edited by Shermis MD, Burstein JC. Mahwah, NJ, Erlbaum, 2003.
70. Bennett RE, Bejar II: Validity and automated scoring: it's not only the scoring. Educ Meas 1998; 17(4): 9–17.
71. Bennett RE: Moving the field forward: Some thoughts on validity and automated scoring. In: Automated Scoring of Complex Tasks in Computer-Based Testing, pp 403–12. Edited by Williamson DM, Behar II, Mislevy RJ. Mahwah, NJ, Erlbaum, 2006.
72. Baker EL, O'Neil HF Jr.: Performance assessment and equity. In: Implementing Performance Assessment: Promises, Problems, and Challenges, pp 183. Edited by Kane MB, Mitchell R. Mahwah, NJ, Erlbaum, 1996. p. 183–99.
73. Stevens R, Soller A, Cooper M, Sprang M: Modeling the Development of Problem Solving Skills in Chemistry with a Web-Based Tutor, pp 580–91. Proceedings of the 7th International Conference on Intelligent Tutoring Systems. Berlin, Springer-Verlag, 2004.
74. Baker EL, Chung GKWK, Delacruz GC: Design and validation of technology-based performance assessments. In: Handbook of Research on Educational Communications and Technology, Ed 3, pp 595–604.

Edited by Spector JM, Merrill MD, van Merriënboer JJG, Driscoll MP. Mahwah, NJ, Erlbaum, 2008.

75. Hively W, Patterson HL, Page SH: A “universe defined” system of arithmetic achievement tests. J Educ Meas 1968; 5: 275–90.
76. Birenbaum M, Kelly AE, Tatsuoka KK: Diagnosing knowledge states in algebra using the rule-space model. J Res Math Educ 1993; 24: 442–59.
77. Bennett RE, Jenkins F, Persky H, Weiss A: Assessing complex problem solving performances. Assess Educ 2003; 10: 347–59.
78. Chung GKWK, Delacruz GC, Dionne GB, Bewley WL: Linking assessment and instruction using ontologies. Proceedings of the I/ITSEC 2003; 25: 1811–22. Available at http://ntsa.metapress.com/link.asp?id=td7v9u19wddex1dd; accessed May 7, 2013.
79. Mislevy R, Gitomer DH: The role of probability-based inference in an intelligent tutoring system. User Model User-adapt Interact 1996; 5: 253–82.
80. Mislevy RJ, Steinberg LS, Breyer FJ, Almond RG, Johnson L: Making sense of data from complex assessments. Appl Meas Educ 2002; 15: 363–89.
81. Williamson DM, Almond RG, Mislevy RJ, Levy R: An application of Bayesian networks in automated scoring of computerized simulation tasks. In: Automated Scoring of Complex Tasks in Computer-Based Testing, pp 201–57. Edited by Williamson DM, Behar II, Mislevy RJ. Mahwah, NJ, Erlbaum, 2006.
82. Darwiche A: A differential approach to inference in Bayesian networks. J ACM 2003; 50: 280–305.
83. Williamson DM, Almond RG, Mislevy RJ: Model criticism of Bayesian networks with latent variables. In: Uncertainty in artificial intelligence: Proceedings of the 16th conference, pp 634–43. Edited by Boutilier C, Goldzmidt M. San Francisco, CA, Morgan Kaufmann, 2000.
84. Haertel GD, Means B (editors): Evaluating Educational Technology: Effective Research Designs for Improving Learning. New York, Teachers College Press, 2003.
85. Shadish WR, Cook TD, Campbell DT: Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston, Houghton-Mifflin, 2002.
86. Adler MD, Vozenilek JA, Trainor JL, et al: Development and evaluation of a simulation-based pediatric emergency medicine curriculum. Acad Med 2009; 84(7): 935–41.
87. Robinson JD, Bray BS, Willson MN, Weeks DL: Using human patient simulation to prepare student pharmacists to manage medical emergencies in an ambulatory setting. Am J Pharm Educ 2011; 75(1): Article 3.
88. Moher D, Dulberg CS, Wells GA: Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994; 272: 122–4.
89. Cohen J: Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ, Erlbaum, 1988.
90. Cohen J: A power primer. Psychol Bull 1992; 112(1): 155–9.
91. Lenth RV: Some practical guidelines for effective sample size determination. Am Stat 2001; 55: 187–93.
92. Lenth RV: Java applets for power and sample size [Computer software]. 2006. Available at http://www.divms.uiowa.edu/~rlenth/Power/; accessed May 7, 2013.
93. Giuliano KK, Johannessen A, Hernandez C: Simulation evaluation of an enhanced bedside monitor display for patients with sepsis. AACN Adv Crit Care 2010; 21(1): 24–33.
94. Herman JL, Morris LL, Fitz-Gabbon CT: Evaluator’s handbook. In: Program Evaluation Kit. Edited by Herman JL. Newbury Park, CA, Sage, 1, 1987.
95. Overly FL, Sudikoff SN, Shapiro MJ: High-fidelity medical simulation as an assessment tool for pediatric residents’ airway management skills. Pediatr Emerg Care 2007; 23(1): 11–5.
96. Newbie D, Dawson B, Dauphinee D, et al: Guidelines for assessing clinical competence. Teach Learn Med 1994; 6: 213–20.