The National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at the University of California, Los Angeles (UCLA) has extensive experience in research, development, program management, and evaluation. A core capability is an integrated view of learning, instruction, human performance assessment, and psychometrics. CRESST is a pioneer of ontology-based designs for domain knowledge, assessment, and 21st century skills. CRESST is creating such ontologies for game-based STEM (Science, Technology, Engineering, and Mathematics) learning interventions funded by DARPA. For the Telemedicine and Advanced Technology Research Center (TATRC), CRESST has completed a state-of-the-art assessment of the use of medical simulation in medical education and training, which is the topic of this special issue. With Office of Naval Research (ONR) funding, CRESST examined learning and assessment issues in Navy training (e.g., Surface Warfare Officers School simulation-based training).

The Department of the Navy’s Office of Naval Research (ONR) provides the science and technology necessary to maintain the Navy’s and Marine Corps’ technological advantage. Through its affiliates, ONR is a leader in science and technology with engagement in 50 states, 70 countries, 1,035 institutions of higher learning, and 914 industry partners. ONR employs approximately 1,400 people, comprising uniformed, civilian, and contract personnel, with additional employees at the Naval Research Lab in Washington, D.C.

The Armed Forces Simulation Institute for Medicine (AFSIM) was established in 2010 as the laboratory activity for medical simulation at the Telemedicine and Advanced Technology Research Center (TATRC). AFSIM runs the TATRC Innovation Lab for prototype evaluation and conducts science activities in the field of medical education and medical simulation technology. The AFSIM also has advisory groups to bring industry, academia, and tri-service input into medical simulation technology development.
DESIGNING AND USING COMPUTER SIMULATIONS IN MEDICAL EDUCATION AND TRAINING

Guest Editors

Harold F. O’Neil
University of Southern California/CRESST

Kevin Kunkler
Telemedicine and Advanced Technology Research Center

Karl E. Friedl
Telemedicine and Advanced Technology Research Center

Ray S. Perez
Office of Naval Research

Foreword iv
Ray S. Perez

INTRODUCTION 1

Designing and Using Computer Simulations in Medical Education and Training: An Introduction
Karl E. Friedl and Harold F. O’Neil

DESIGN ISSUES

Cognitive Task Analysis-Based Design and Authoring Software for Simulation Training 7
Allen Munro and Richard E. Clark

Using Cognitive Task Analysis to Develop Simulation-Based Training for Medical Tasks 15
Jan Cannon-Bowers, Clint Bowers, Renee Stout, Katrina Ricci, and Annette Hildabrand

Use of Cognitive Task Analysis to Guide the Development of Performance-Based Assessments for IntraOperative Decision Making 22
Carla M. Pugh and Debra A. DaRosa

Balancing Physiology, Anatomy and Immersion: How Much Biological Fidelity Is Necessary in a Medical Simulation? 28
Thomas B. Talbot

Cost Considerations in Using Simulations for Medical Training 37
J. D. Fletcher and Alexander P. Wind

ASSESSMENT AND EVALUATION ISSUES

Assessment Methodology for Computer-Based Instructional Simulations 47
Alan Koenig, Markus Iseli, Richard Wainess, and John J. Lee

Application of National Testing Standards to Simulation-Based Assessments of Clinical Palpation Skills 55
Carla M. Pugh

Evaluation of Medical Simulations 64
William L. Bewley and Harold F. O’Neil
INSTRUCTIONAL STRATEGIES ISSUES

Prevention of Surgical Skill Decay 76
Ray S. Perez, Anna Skinner, Peter Weyhrauch, James Niehaus, Corinna Lathan, Steven D. Schwitzberg, and Caroline G. L. Cao

Effects of Simulation-Based Practice on Focused Assessment With Sonography for Trauma (FAST) Window Identification, Acquisition, and Diagnosis 87

Adaptive and Perceptual Learning Technologies in Medical Education and Training 98
Philip J. Kellman

PSYCHOMETRIC ISSUES

Evidence-Centered Design for Simulation-Based Assessment 107
Robert J. Mislevy

Potential Applications of Latent Variable Modeling for the Psychometrics of Medical Simulation 115
Li Cai

Use of the Assessment–Diagnosis–Treatment–Outcomes Model to Improve Patient Care 121
Kevin F. Spratt

Special Issue Editors and Acknowledgements 132
Foreword

The use of advanced education and training technologies has been common and widespread in the military for decades. Over two decades ago, the Department of Defense initiated large continuing investments in developing modeling and simulation (M&S) programs to improve training effectiveness while reducing their reliance on actual, costly hardware and software systems. The medical and health sciences community relatively recently began the systematic use of modeling and simulation for education and training of medical personnel. The Defense Advanced Research Projects Agency (DARPA) and the Office of Naval Research (ONR) were some of the very first agencies to provide substantial government funding for this purpose. Industry, recognizing its value quickly, picked up on this initiative and began to offer medical procedure simulators for specific training applications. The use of M&S in medical and health sciences training and education has progressed rapidly and this community is now one of the more enthusiastic adopters of M&S technologies for initial training as well as skill maintenance. Skill retention has recently, in the scientific literature and among medical and health sciences trainers, become a critical issue of concern regarding patient safety. M&S provides the means to educate military personnel, families, and colleagues about the impact of physical and psychological injury. We are seeing evidence that these technologies can also be deployed on stand-alone PCs, the Internet, game consoles, and a variety of mobile devices.

This use of simulation is also necessary across both civilian and military domains to refresh medical and first responder skills, to test their competencies, and to provide training anywhere and anytime. For the military and health sciences care professionals, medical simulators provide the unique opportunity to train skills that are not readily available within the civilian health care systems. Today’s medical simulators can serve as on-the-job training in combat zones and during humanitarian relief efforts. As M&S technologies continue to become more robust, acceptable, and affordable, the development of valid representations of the human patient will also expand to meet this demand. The medical and health sciences community recognizes these advances in M&S and the potential for use of computer graphics, natural language processing, and artificial intelligence technologies to deliver high-fidelity training.

While these advances in M&S for medical education and training have supported development of medical simulations, there has not been a similar focus on measuring and assessing performance while using these tools. There has been less research in the evaluation of simulation-based interventions, and validation of metrics to assess training outcomes and transfer of training from virtual to real world tasks. This research will require extensive attention to educational design principles, human factors issues, and rigorous attention to validation to ensure that M&S platforms are both safe and efficacious. Such evaluation of the performance outcomes is a critical focus to ensure positive medical and health sciences procedure outcomes and patient safety.

ONR and the Telemedicine and Advanced Technology Research Center (TATRC) asked the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) to conduct a series of workshops focused on the state of the art in teaching and assessing medical skills using M&S technology. The results of these workshops are presented in this supplement of Military Medicine. It is our belief that these findings are just the beginning of a robust military and civilian enterprise approach to examining and implementing meaningful, effective, and affordable metrics for assessing medical simulations and simulation-based training outcomes.

Ray S. Perez, PhD
Program Officer, Cognitive Science of Learning Program
Office of Naval Research
Designing and Using Computer Simulations in Medical Education and Training: An Introduction

COL Karl E. Friedl, PhD, MS USA*; Harold F. O’Neil, PhD†

ABSTRACT  Computer-based technologies informed by the science of learning are becoming increasingly prevalent in education and training. For the Department of Defense (DoD), this presents a great potential advantage to the effective preparation of a new generation of technologically enabled service members. Military medicine has broad education and training challenges that range from combat medic training to personal protective measures required for every service member. Many of these challenges can be met with gaming and simulation technologies that this new generation has embraced.1,2 The purpose of this supplement is to review the state of the science in computer games and simulations that could be applied to military education and training in medicine. Based on this supplement and DoD research policy, we plan to develop a road map for research and development in this area as well as develop a set of What Works guidelines for the design and evaluation of these technologies in military medicine.

INTRODUCTION
Computer-based technologies informed by the science of learning are becoming increasingly prevalent in education and training. For the Department of Defense (DoD), this presents a great potential advantage to the effective preparation of a new generation of technologically enabled service members. Military medicine, in particular, has broad education and training challenges that range from combat medic training to personal protective measures required for every service member. Many of these challenges can be met with gaming and simulation technologies that this new generation has embraced.1,2 The purpose of this supplement is to review the state of the science in computer games and simulations that could be applied to military education and training in medicine. Based on this supplement and DoD research policy, we plan to develop a road map for research and development in this area as well as develop a set of What Works guidelines for the design and evaluation of these technologies in military medicine.

THE NEED FOR MILITARY MEDICAL SIMULATION TRAINING
Key reasons why the DoD is invested in computer games and simulation technology go beyond the obvious advantages of modern training effectiveness and patient safety.3,4 There is also a need to dramatically reduce training costs, especially through better preparation of personnel before costly field training exercises and reduction in total training time. This technology should deliver training in a form that the current generation of recruits expects. There is a need to have just-in-time training and refresher training, which potentially would reduce the decay of critical skills. There is a need to reach individuals wherever they are (e.g., away from their duty station, as they mobilize for disaster relief, in remote areas; etc.), including team training of individuals even before the team is geographically assembled.

The use of virtual patients is particularly important to replace standardized patients. Such virtual patients could also be used to teach interpersonal skills and diagnostics and treatment skills based on symptom presentation and interviews.5 The use of virtual patients could broadly be applied, from pharmacy technicians to psychiatry residents learning to interact with standardized patients. It can improve the quality of medical training by providing training adapted to the needs and ability of the individual, update relevant scenarios to provide adaptive training in response to changing health threats, and provide a large number of variations of conditions and scenarios to expand the military capability of the individual or team.

These militarily unique computer training scenarios are critical to the integration of classroom and other medical training that will have to be applied in high-stress complicated environments (e.g., noisy, lowlight, airframe or vehicle vibration, threat of attack). An example is the training provided through the Army’s Medical Simulation Training Centers currently established at 18 military installations (http://www.peostri.army.mil/PRODUCTS/MSTC/). Most importantly, if appropriate competency outcome measures have been determined, computer-based games and simulation training can test and objectively score individuals for their competence.

There are other challenges in the application of computer-based training technologies that are important to the DoD.
The service members cannot become dependent on these technologies in an operational environment rather the technology should truly improve the independent capabilities of the trained individual. This becomes important for agility and survivability, especially with modern day threats aimed at the technology itself (e.g., electromagnetic pulse weapons that would shut down all electronics). Thus, this military medical training must balance the use of technology in training context, with its probable absence in operational environments. This is especially important for the DoD and others who may operate in remote and austere environments where technology including electronic systems of all kinds, including telementoring and decision support tools, may not be available.

For the DoD, there is also a need to provide elements of science, technology, engineering, and math (STEM) training for the pool of motivated recruits who may have inadequate background education in many areas critical to their skill set (e.g., math and physics for radiology technicians). Thus, STEM proficiency is also a military medical need (aside from the national security implications of a generation of underperforming U.S. high school students). One example of engaging game-based STEM training would be a game that incorporates positive role models and military scenarios. For example, use of physics and mathematical principles to secretly capture an enemy submarine and its code machines, revolving around a display such as the U-505 at the Chicago Museum of Science and Industry (http://www.msichicago.org/whats-here/exhibits/u-505/learning-tools/learning-games/). These games would help prepare future generations of K-12 students for critical STEM technical jobs including possible interest in military jobs.

The DoD must also produce qualified individuals within a constrained period of time, to reduce manpower costs and to improve availability of the needed specialists. Over-the-horizon concepts of more efficient and effective learning through electrical stimulation and pharmacological activation of specific brain centers may ultimately serve to explain how current technologies used in the entertainment industry are so effective in promoting behaviors that facilitate learning. For example, the same learning activation in classrooms may be purposefully accomplished through the use of engaging game technology if we understand the scientific underpinnings. The requirement to reduce training time through more effective training strategies also requires an appreciation of the training “dosage” as well as evaluation criteria that will determine the probability of properly trained medical personnel.

Such medical training for enlisted service members is conducted at San Antonio, Texas (http://www.metc.mil/). This command supports the largest single medical training campus in the world, i.e., Medical Education & Training Center. There are 21,000 students trained annually in more than 60 major medical specialty programs to serve the needs of the Army, Navy, and Air Force. In addition to this training campus for enlisted service members, the Uniformed Services University of the Health Sciences (USUHS) in Bethesda, Maryland trains physicians, nurses, and allied medical professionals (http://www.usuhs.mil/). The Army alone has approximately 5,000 physicians and 11,000 nurses, along with many other medical professionals. Since 1980, USUHS has trained 4,700 physicians, providing approximately 20% of the military medical corps physician accessions each year. These service members, and many more medical and allied science officers recruited from other sources, require specialty training in military medicine and refresher training within each of their specialties throughout their careers.

Beyond the specific specialty training needs of DoD medical personnel, all service members require common skills training in health topics to maximize their readiness status (i.e., their preparedness to perform their trained mission today) as well as to instill individual responsibility for their own health. This training includes a wide range of topic areas such as understanding signs and symptoms associated with chemical, biological, radiological, nuclear, and explosives attacks, recognizing behavioral health issues in themselves and their buddies, proper use of effective personal protective measures against disease-carrying vectors (e.g., malarial mosquitoes), and avoidance of personal health damaging behaviors (e.g., smoking and excessive alcohol consumption). One example of a computerized approach to providing this fundamental training are the small business grants awarded under the topic “Micro Games for Proactive Preventive Medicine” (http://www.sbir.gov/sbirsearch/detail/242049).

This latter topic of promoting individual responsibility for one’s own health is a key theme of the Army Surgeon General (LTG Horoho), who has described the need to develop training tools to help develop the “LifeSpace” of the soldier—that majority of the soldier’s time when they are not in contact with the medical health care system. This LifeSpace can be filled in, in part, through the use of medical education and training delivered via mobile health technologies especially for health behaviors such as weight regulation, activity and fitness habits, and smoking cessation.

Everyday off-the-shelf technologies such as personally owned smart phones and laptop computers can be an effective means to reach and educate today’s young service member. Technologies developed first in the entertainment industry, such as individual and multiplayer electronic games, interactive virtual worlds, virtual human technologies, and 3-dimensional interactive capabilities such as new motion sensing technologies that create a Star Trek-like “holodeck” (e.g., Kinect), can be harnessed and tested for military medical training. The challenge is not just the technology but the science of learning that should underlie technology use in medical education and training.

Military use of medical simulation technologies has to date relied heavily on a family of commercially available off-the-shelf human manikins and part-task training simulators (e.g., chest tube simulators, intubation and colonoscopy simulators, anesthesia simulators, etc.). Less than a decade ago, everyone had a “simulator in a closet,” representing
primarily an expensive, difficult to operate and maintain simulator with unknown training effectiveness, and thus, kept in the storage closet rather than actually used in any curriculum. Many military medical training simulation centers still have a myriad of different systems intended for different medical task training, most of which do not operate on other hardware or software systems and which are in various stages of development and validation. Unlike the Food and Drug Administration approval for new medical devices, simulation systems used in medical education and training do not fall under one governing body that would ensure that such training systems are efficient and effective. If it is not intended for medical diagnosis or treatment, but rather for training, these systems are not reviewed by the Food and Drug Administration. Thus, medical education and training simulators are currently a wild west of unregulated simulation systems and claims. There should be guidelines for the use of computer simulation in education and training.

Medical training goals, curricula, and performance standards as well as test standards should dictate the appropriate insertion points where more affordable and usable training simulations could greatly improve medical training. This overall goal is, in part, the specific goal of a new DoD intramural consortium led by the USUHS National Capital Area Medical Simulation Center (http://simcen.usuhs.edu/Pages/default.aspx).

For example, standards for physician training are most advanced in the surgical community where testing and training of the fundamentals of laparoscopic surgery have been pioneered. Laparoscopic surgery is well suited to the early adoption of virtual training and testing technologies because of the similarity between virtual and real procedures. The training emphasizes the development of psychomotor skills and dexterity necessary to master laparoscopic surgery.

Another DoD goal in the use of simulation technologies in medicine is to dramatically reduce the use of live animals in military medical training, an objective of the DoD medical simulation training research program. Simulations now provide a high level of training engagement and realism, which may allow some tasks with simulation to replace training with live animals. Development of such simulations needs to be guided by a science of learning approach that has been lacking in the earlier development of medical training including that involving live animals and cadavers. Other significant benefits of simulation vs. live animals are the ability to train to multiple presentations of a medical problem and with many repeated practice trials that would not be practical with animals. We believe that this simulation technology is most useful in the initial phase of skills development that leads up to apprentice-level competence after which a trainee can then participate and be mentored in real medical procedures using real patients. The current state of the art in either simulation or life tissue training is not an adequate substitute for the patients in the training of fully competent medical providers.

For all of the reasons above, it is important for the DoD to determine what criteria must be applied in the design and evaluation of computer-based medical training systems. The goal is to use new simulation opportunities to change the old discomfiting concept of training medical procedures from “see one, do one, teach one” to “see one, simulate many, do one competently, teach many.”

This supplement is based on the first of three workshops co-organized with the Office of Naval Research and the Telemedicine and Advanced Technology Research Center (TATRC) by UCLA/CRESST. The first workshop reviewed the state-of-the-science in medical simulation training, design, and evaluation from researchers in these areas. The second workshop focused on specific technology applications that could be designed and delivered to support the specific needs of the Medical Education & Training Center in combat medic training and other enlisted medical specialty training as well as training for Aeromedical Evacuation personnel. This second workshop focused on the development of a road map and topics for what works in computer simulations in gaming and simulations. It is expected that this latter book of guidelines would be published by a commercial publisher. The third workshop will finalize the road map and recommended guidelines for simulations developed for military medical education and training. This work directly supports DoD training requirements as dictated by DoD Instruction 1322.24 (Medical Readiness Training), which states that “Medical readiness training programs shall include realistic individual and collective medical skills training and shall maximize the use of emerging technology, including distance learning, simulation, and virtual reality.” It also supports another DoD Instruction, DoDI 3216.01 (Use of Animals in DoD Programs), which states that methods other than animal use shall be considered and used whenever possible to attain the objectives of training “if such alternative methods produce scientifically or educationally valid or equivalent results.”

CURRENT DO D RESEARCH INVESTMENTS IN MEDICAL SIMULATION

The current investment in military medical simulation education and training research is summarized in detail everywhere and some military studies on medical simulation training have been conducted and reported. In brief, new core funding has been made available in the DoD to support an organized research effort that includes at least four principal thrust areas. The most mature of these four areas is a fully funded multiyear academically based effort under the Combat Casualty Training Consortium initiative. This carefully crafted consortium specifically addresses issues that are currently highest priority in military medical training—i.e., combat medic skills training.

To the disappointment of some, however, this funding was not directed to high-powered laboratories specializing in surgical specialty simulation training, but instead focused...
on uniquely military problems to improve medics training and also to reduce reliance on live animals. In the civilian sector, there has been a great deal of good analytic work on the tradeoffs of using live animals. For example, the impact on the neurosciences by the International Animal Research Regulations was documented by the Institute of Medicine and the National Research Council.35

Three other DoD initiatives are partially funded and in development, including the Medical Practice Initiative for medical specialty training (e.g., mobile platforms for just-in-time training and refresher training for humanitarian mission and disaster response in the Mobile Learning Environment project) (http://www.mole-project.net/the-project/global-medaid-app); the Patient Focused Initiative that addresses the Army Surgeon General’s high-priority LifeSpace initiative (e.g., virtual human interactive behavioral health coaching [SimCoach] project)59; and the Developer Tools Initiative that will fill a critical need for open source physiology engines (models and artificial intelligence tools) to drive realistic and accurate medical simulation trainers.31

In addition to new DoD funding support, high interest in medical simulation and training is spurred in part by the recognition that current technologies can be brought to bear to address critical needs in medical education and training. Service members today tend to be equipped with personal smart phones and expect to find and receive information through these everyday technologies. The Defense Advanced Research Projects Agency has entered the field to promote development of game-based medical training software that will include development of deeper understanding of physiological principles that promote a broader ability to deal with real-life medical problems.36 In addition, the DoD has partnered with allies in a NATO workshop on Advanced Training Technologies for Medical Healthcare Professionals (NATO HFM-215); the results of this workshop will be reported elsewhere.

There is also much to learn from the rest of the DoD and the Human Systems Integration community that has embraced and continues to develop new training technologies as well as define how electronic systems integrate into military internet security systems.57 In particular, the Office of Naval Research has been a leader in advancing research in more effective education and training technologies, and the results of this effort provide a springboard to the development of guidelines for standardization and evaluation of new military medical simulations for education and training. Such military medical applications will also have many dual-use benefits in civilian medical training.

The science of learning specifically targeted toward the use of computer simulations in medical education and training is also a focus of this supplement. Mayer27 conceptualized the science of learning as the scientific study of how people learn. It is supported by the science of instruction (how to teach) where teaching is accomplished by humans or technologies. Mayer conceptualizes the science of assessment as the determination of what people know. Baker38 provides an excellent overview of assessment with a focus on reliable and valid measures.

The classification of learning outcomes as a family of cognitive demands has been provided by Baker and Mayer.39 These cognitive outcomes are the following: content or domain understanding, problem solving, self-regulation, communication, and collaboration/teamwork. The utility of this classification for computer games is provided by O’Neil et al.40 A complementary taxonomy of learning outcomes is provided by Anderson and Krathwohl.41 This taxonomy replaces Bloom’s taxonomy.42 An application of the science of learning to science education is provided by the National Research Council.43

An interesting metalook of how to conceptualize learning is provided by Bransford et al.44 They suggest an integration of three types of learning, i.e., informal learning, implicit learning, and formal learning. They define informal learning as learning that happens in non-school public settings such as museums, zoos, and after-school clubs or the learning that occurs in homes, on playgrounds, among peers, and in other situations (p. 216). Their definition of implicit learning is information that is acquired effortlessly and sometimes without conscious recollection of the learned information or having acquired it (p. 210). Formal learning is the learning that goes on in formal environments such as schools where the science of instruction should be explicit.

**ORGANIZATION OF THE SUPPLEMENT**

The organization of this supplement was partially based on the organization of a workshop organized by TATRC/ORNG/CREST on “Designing and Using Computer Simulations in Medical Education and Training.” This supplement is the documentation of this workshop that involved presentations, followed by the development of manuscripts.

The rationale for selection of authors was based on three underlying challenges, namely1 there are different research communities in simulation research in medical education and training, e.g., researchers from medical vs. nonmedical environments and defense vs. civilian research communities;2 the lack of infusing research in simulation for medical education and training from a science of learning perspective; and3 the lack of lessons learned from the use of simulation in other nonmedical fields (e.g., aviation). Unfortunately, there is minimal scientific contact between these different communities. For example, researchers tend to have specific professional identities contextualized in either military or civilian research organizations that result in different journals and conferences to report their research.

To partially address these challenges, we organized this journal supplement. Thus, this supplement presents work by authors from multiple disciplines whose expertise is using computer simulation to teach or assess in both medical and nonmedical environments. Experts from both military and civilian organizations are also represented. Two disciplines,
i.e., science of learning and psychometrics, were specifically overrepresented as there exists minimum impact of science of learning in the medical uses of simulation and there is little psychometric knowledge for estimating the reliability and validity of simulations used for teaching or assessment. The articles in the supplement also include original studies, reviews, and conceptual analyses. The supplement itself is organized into 4 major sections, representing 4 different issues, respectively: Design Issues, Assessment and Evaluation Issues, Instructional Strategies Issues, and Psychometric Issues.

**Design Issues**

We viewed critical design issues in computer simulation for medical education and training as the use of cognitive task analysis, how much fidelity is desirable, and costs. This cognitive task analysis focus is represented with three different but complementary cognitive task methodologies (Munro et al, Cannon Bowers et al and Pugh et al). Fidelity and cost considerations are represented by articles by Talbot and by Fletcher et al, respectively.

**Assessment and Evaluation Issues**

We viewed these issues as another area where nonmedical research has useful lessons learned for military medicine education and training. This section is represented by three articles: an assessment methodology derived from a research project focused on Navy but nonmedical training (Koenig et al), the application of national test standards to simulation of clinical palpation skill (Pugh), and lessons learned in nonmedical simulation evaluations applied to the evaluation of medical simulations (Bewley et al).

**Instructional Strategies Issues**

How to teach and assess are represented by three articles: what instructional and assessment strategies are effective in reducing skill decay (Perez et al); an instructional strategy using virtual trainers on FAST Window identification, acquisition, and diagnosis (Chung et al); and the strategies effective in adaptive and perceptual learning (Kellman).

**Psychometric Issues**

Psychometrics (the study of psychological measurement) deals with the technical quality of measures, which is the overall basis of the trustworthy interpretation of measurement situation(s). It includes validity, reliability, comparability, equating, scaling, and standardization as well as theoretical frameworks such as classical test theory, item response theory, and generalizability theory. This section is represented by three different approaches: an evidence-centered design one (Mislevy), a latent modeling one (Cai), and the Assessment-Diagnosis-Treatment-Outcome model that includes psychometric issues (Spratt).

In summary, this supplement presents an overview of the key considerations in simulation design and assessment for medical education and training. Multiple experts from multiple disciplines have contributed their specific ideas on how thoughtful simulation design should be planned and executed. These articles suggest how medical education and training simulations can be developed that are scientifically based, outcomes driven, and cost conscious.

**REFERENCES**


Cognitive Task Analysis-Based Design and Authoring Software for Simulation Training

Allen Munro, PhD; Richard E. Clark, EdD

ABSTRACT The development of more effective medical simulators requires a collaborative team effort where three kinds of expertise are carefully coordinated: (1) exceptional medical expertise focused on providing complete and accurate information about the medical challenges (i.e., critical skills and knowledge) to be simulated; (2) instructional expertise focused on the design of simulation-based training and assessment methods that produce maximum learning and transfer to patient care; and (3) software development expertise that permits the efficient design and development of the software required to capture expertise, present it in an engaging way, and assess student interactions with the simulator. In this discussion, we describe a method of capturing more complete and accurate medical information for simulators and combine it with new instructional design strategies that emphasize the learning of complex knowledge. Finally, we describe three different types of software support (Development/Authoring, Run Time, and Post Run Time) required at different stages in the development of medical simulations and the instructional design elements of the software required at each stage. We describe the contributions expected of each kind of software and the different instructional control authoring support required.

INTRODUCTION

The traditional medical teaching method used in the past was based on an outdated apprenticeship model. Hand-in-hand with this limitation was the necessity for experts to be the teachers and for the teaching moments to be moved along by the progress of a case, often without the option of redoing an error or stopping at a critical moment to ask questions. Recent advances in medical pedagogy, such as problem-based learning, have apparently not addressed these concerns. Increasing concerns about medical mistakes, patient safety, and demands on teaching faculty have led the medical community to supplement mentor-based training practices with medical simulators. The introduction of simulators provided some relief for medical faculty, by providing a convenient place for students to gain medical experience without putting patients at risk or delaying treatment for a teaching moment. As a result, medical educators have tended to simply insert simulators into medical training wherever possible. As simulators have been put into place and tested, we have realized that we must improve the way we design the software that supports or delivers the simulations and put more careful emphasis on the development and coordination of instructional content and teaching methods presented by simulators to produce maximum learning efficiency and effectiveness. Simulators are susceptible to the GIGO (garbage in, garbage out) problem when they present wrong or incomplete information and, in some cases, present it in ways that risk cognitively overloading students and so make learning ineffective and inefficient.

The development of more effective medical simulators requires a collaborative team effort where three kinds of expertise are carefully coordinated: (1) exceptional medical expertise focused on providing complete and accurate information about the medical challenges to be simulated; (2) instructional expertise focused on the design of simulation-based training and assessment methods that produce maximum learning and transfer to patient care; and (3) software development expertise that permits the efficient design and development of the software required to capture expertise, present it in an engaging way, and assess student interactions with the simulator. This article focuses on each of these three essential components of the design of effective and efficient medical simulations.

MEDICAL EXPERTISE FOCUSED ON THE MEDICAL CHALLENGES TO BE SIMULATED

Medicine relies heavily on instructional content drawn from the clinical expertise of medical experts. Physicians with successful clinical experience both teach and provide the medical expertise necessary for simulations. If their descriptions of medical procedures are complete and accurate, the designers of simulations would only have to incorporate them faithfully. Yet Clark and Elen have provided compelling evidence that the experiential knowledge acquired by all experts is largely “implicit,” automated, and nonconscious. When experts attempt to describe medical protocols to students or to the designers of simulations, evidence suggests that they unintentionally omit approximately 70% of the critical decisions and analysis required to succeed. A number of experiments in various medical areas have validated this “70% rule.” These omissions are the natural result of the way that all expertise is learned. To circumvent the brain’s limits on the amount of information that can be consciously processed at any one time, when decisions and analysis strategies are
practiced successfully, they gradually automate to make way for novel information that must be processed consciously. As experts describe what they do during a procedure to students or simulation specialists, their descriptions tend to be accurate about actions that they can observe themselves or others perform because they are stored in declarative memory—but decisions are not directly observable and so can only be inferred because they are stored in nonconscious procedural memory. As the decisions experts make tend to be largely implicit and nonconscious, they are only about 30% accurate when describing the critical decisions they must make to succeed at a medical procedure. Yet it appears that different experts are consciously aware of somewhat different decisions and so interviewing multiple experts with cognitive task analysis (CTA) gradually increases the capturing of critical decisions and reduces, but does not completely eliminate, omitted decisions.

Omissions may be the source of medical mistakes as students “fill in the blanks” through trial and error when treating patients and implicit expert knowledge may also cause incomplete information to be programmed into medical simulations. The only exception to the 70% rule identified to date was found in an experiment on a trauma procedure that was controversial because of documented medical mistakes and so was being discussed openly. Previous studies of this procedure had found the 70% omissions but in a later follow-up study conducted after the textbook approach to the procedure had been questioned and was being discussed widely, the omissions fell from 70% to 40%. Even in this unusual case, many critical decisions were not reported.

Medicine obviously requires a strategy for identifying and capturing critical but nonconscious expert knowledge and one viable option that is compatible with the design of simulators is a technique called CTA.

CTA
CTA refers to over 100 interview and observation techniques used to elicit and represent the knowledge, goals, strategies, and decisions that underlie observable task performance. Although there are many types of CTA methods, all share a common goal of capturing the knowledge of subject-matter experts (SMEs) who have shown reliable proficiency in performing a task over a long period of time. Yates reviewed all published CTA approaches and identified only five that seemed viable for capturing both the conceptual and implicit procedural knowledge required by medical simulators. An example of one of the five viable methods was developed and has been validated in multiple studies by one of the authors. The validation of other viable methods has been described by Catrambone and Lipshitz et al.

The CTA approach described here is called the Concepts, Processes, and Principles CTA and has proven to be effective in both capturing expert knowledge and contributing to the development of training simulators that reduce the time trainees need to meet their learning objectives and decrease the errors they make during transfer after training (e.g., see the study by Velmahos et al).

Concepts, Processes, and Principles CTA is most commonly performed in the following six stages—Interview SMEs with recent and consistently successful experience:

1. Ask about sequence of tasks that must be performed to complete the medical procedure to “outline” the procedure.
2. Capture “When and How” decision steps and all action steps for all tasks captured in step 1 and develop a document describing each step in the sequence they are performed.
3. Ask SMEs to edit their own and other expert procedure documents to correct and add missing steps so that analysts can combine different approaches into one “gold standard” approach that can be learned and performed by students.
4. For each task and procedure identified, describe critical concepts, processes, or principles that need to be learned to explain and perform the procedure.
5. Collect additional information about the various types of problems the procedure will solve; the material and equipment required as well as the indications, and contraindications needed to support their use with patients.
6. Pull the CTA protocol into an instructional design system that will reside inside the simulation.

The effort required to produce a viable final gold standard CTA is determined in large part by two factors: (1) the time required to conduct and analyze full interviews with each SME to capture all of the knowledge required for training novices; and (2) the number of SMEs who must experience full interviews to capture the most complete and accurate description of the skills and knowledge required to execute a given task. Research conducted by Chao and Salvendy and more recently by Crispin and Bartholio has provided evidence that three to four experts are necessary to capture the optimal amount of knowledge before reaching “the point of diminishing utility.” The payoff from this level of investment is an explicit record of nearly all critical decisions and actions, described at a level appropriate for students with adequate background knowledge. The CTA can be changed easily as the procedure is improved over time without the need for medical faculty to invest wasted effort repeatedly describing (an incomplete version of) the procedure.

A CTA-based protocol is the first stage in the design of an effective simulation. In the second stage, we need to format the expert-based CTA information into an instructional design that will transmit the content to students in a simulation.

INSTRUCTIONAL SIMULATION METHODS THAT PRODUCE MAXIMUM LEARNING AND TRANSFER TO PATIENT CARE
Whether simulators are used for instruction and/or for practice of skills that have been taught before the use of the
simulator must influence the design of the simulations. If they are used to teach new skills and knowledge, the evidence clearly supports the use of fully guided, explicit instructional processes. Merrill\textsuperscript{23} most evidence-based guided instructional design processes implement the list of features described by Merrill\textsuperscript{23} include the following:

1. Provide realistic field-based problems for students to solve (note: available from a CTA).
2. Give students analogies and examples that relate their relevant prior knowledge to new learning as the procedure is being introduced.
3. Offer clear and complete demonstrations of how to perform key tasks and solve authentic problems (note: available from a CTA).
4. Insist on frequent practice opportunities during training to apply what is being learned (by performing tasks and solving problems) while receiving corrective feedback.
5. Require application practice that includes “part task” (practicing small chunks of larger tasks) but also “whole tasks” (applying as much of what is learned as possible to solve the complex problems that represent challenges encountered in operational environments) both during and after instruction.

Items 1 to 3 are required for simulators that offer instruction, whereas items 4 and 5 are necessary for simulators that only provide opportunities for practice, feedback, testing, and transfer. The problems and tasks provided during practice exercises must be representative of the population of problems and tasks they will be expected to tackle after instruction. As most transfer environments require task performance rather than the recall of facts, practice must follow a demonstration or worked example and require the application of the demonstrated procedure to complete a task and/or solve a problem.

Rosenshine\textsuperscript{24} has also described the research base supporting guided practice and have provided guidelines for constructing demonstration and practice exercises such as the following “17 Principles for Effective Instruction” (p. 19):

1. Begin a lesson with a short review of previous learning.
2. Present new material in small steps with student practice after each step.
3. Limit the amount of material students receive at one time.
4. Give clear and detailed instructions and explanations.
5. Ask a large number of questions and check for understanding.
6. Provide a high level of active practice for all students.
7. Guide students as they begin to practice.
8. Think aloud and model steps.
9. Provide models of worked out problems.
10. Ask students to explain what they have learned.
11. Check the responses of all students.
13. Use more time to provide explanations.
14. Provide many examples.
15. Reteach material when necessary.
16. Prepare students for independent practice.
17. Monitor students when they begin independent practice.”

With elements of Rosenshine’s 17 principles and Merrill knowledge types in place, a “blueprint” of the simulation is available to support all of the stages in the development of computer or live simulations. We turn next to a discussion of the ways that simulation software can incorporate this kind of design and support learning, practice, and transfer.

SOFTWARE TO SUPPORT THE DEVELOPMENT OF MEDICAL SIMULATIONS FOR TRAINING

Classically the development of simulations could be divided into five largely sequential phases: requirements analysis, design, implementation, testing, and maintenance. This model of development arose from common practice in systems analysis and software development, where it was sometimes called the systems development lifecycle. It was first formally described by Royce,\textsuperscript{25} who pointed out that the approach “is risky and invites failure” (p 2). More recent approaches have emphasized rapid prototyping, using methods of successive approximation and partial products, with the goal of integrating the customers into the development process throughout. Extreme Programming, or XP,\textsuperscript{26} is a prominent example of a rapid prototyping approach to software development. In practice, many large-scale development projects, including the development of simulations for military training, are driven by the procurement process to include elements of the classic systems development lifecycle approach. When a rapid prototyping approach is used, the phases are repeated, in varying degrees of formality, for each prototype. These phases are ordinarily all carried out (except in the case of single-use simulations, which would not have a maintenance phase), whether the simulations are computer-based or are conducted by human actors.

— Requirements Analysis. This phase of development should be guided by a CTA or by a similar characterization of targeted performance goals. The CTA determines what must be observable to the learner, what actions the learner can take, what actions the learner should take, what conditions may pertain to the medical task that is being learned, and what the consequences of possible actions would be under the various conditions.

— Design. The CTA guides the selection or construction of learning experiences based on the types of knowledge or skill that are required to perform a task. In simulation contexts, practice opportunities can be provided that appropriately exercises the learner’s decision-making opportunities.

— Implementation. In the case of actor-delivered simulations or simulated patients, implementation requires the development of scripts that may include conditional
branching, so that the simulated patient (and, in some cases, the simulated laboratory) understands how to appropriately respond to most of the possible actions that the learner can take in a simulated medical episode. The actors must then learn the script. In computer-based simulations, the CTA-based design guides the development of an interactive simulation that provides ways to carry out all the necessary observations during a simulated medical episode, together with ways to carry out all the actions (medical interventions, etc.) that could take place. It also guides the development of appropriate causal representations of the effects of those actions. Implementation of computer-based simulations can be accomplished using general-purpose programming languages, but there are also specialized tools for building interactive graphical simulations for training.27–36

— Testing. The implementation phase of a computer-based simulation typically includes repeated cycles of detailed design, coding, and in-house testing. Once such a simulation is determined to be “running,” it is time for testing with representative learners. This phase typically reveals aspects of the user interface or the behavior of the simulation that are not well understood by some portion of the learner population. This can also happen with actor-based simulations, and the implementation phase may have to be revisited to make modifications based on experience with the actual learner population.

— Maintenance. If simulations are to be used more than once, they will require maintenance. Medical practices and procedures evolve, and medical simulations must evolve with them. Tools that support the maintenance of simulations can contribute to a cost-effective simulation life cycle.

Specialized software tools can support each of the four areas of simulation development (design, implementation, testing, and maintenance). For some of these phases of development, useful tools are already available.34.36 For other phases, new tools are called for.

SOFTWARE TOOLS FOR SIMULATION DESIGN

In an ideal world, software tools for simulation design would be tightly integrated with an effective system for CTA. The analysts producing a CTA for a medical procedure, for example, would enter information about the procedure into a database for CTAs. The CTA data would include the goals and subgoals of the procedure, the steps required to achieve each subgoal, the conditions under which each step can be taken, the alternative steps to be used when a condition is not met, and the observations that need to be made to check each condition and each step outcome. A simulation design system would tap into this database to create lists of observables that the simulation must have to support the procedure, lists of basic actions that must be possible to carry out the steps of the procedure, and the combinations of observable values that define the conditions that are important for the procedure. At a minimum, the implementers of the simulation would then have comprehensive lists of essential objects and states required to realistically implement the behaviors of the simulation that would be essential to carrying out the procedure that is to be learned.

It would be even better, of course, if these automated outputs of the simulation design tool could be fed directly into the next phase tool, the implementation system, to directly support the production of the simulation from the CTA-based design. We are not aware of any tools that offer this feature at this time. A tool, such as “iRides Author”34,36 would be significantly enhanced if it automatically used elements of the work product from a CTA in the simulation development process, rather than requiring the simulation developer to manually confirm that elements identified as necessary by the CTA are in fact present in the simulation.

SOFTWARE TOOLS FOR THE IMPLEMENTATION OF SIMULATIONS

There are a wide variety of tools that support the implementation of simulations for training. One way of characterizing these tools is in terms of their specificity of purpose. The least special-purpose implementation tools are general purpose programming languages, such as C, LISP, Smalltalk, or Java. Specialized programming languages for developing interactive simulations are more targeted. Some languages are particularly oriented to event-based simulation models, rather than interactive training simulations. These include Simula GPSS and SIMSCRIPT, which are chronicled by Nance.37 There are also authoring tools that permit the specification of simulation behaviors without conventional computer programming, such as programming by example systems, also known as programming-by-demonstration systems.38,39 There are also authoring tools that are designed to support the development of a special domain of simulations, such as surface warfare tactics simulations.40

iRides Author36,41 is a tool for creating training simulations that is a tool of intermediate specialization. It supports the development of interactive graphical simulations for a variety of purposes. Example products have included simulations for air traffic ground control, C-30 missile attack countermeasures, and Navy surface warfare tactics planning.42

iRides Author is a tool that is not specific to a particular simulation domain, so it requires that developers who make use of it are able to think abstractly and code relational expressions that express the values of attributes as functions of the values of other attributes. Some attribute values are automatically set when users take actions, such as pressing a mouse button while pointing to a simulation object, and all the expressions that refer to those attributes are automatically evaluated when such actions are taken. The iRides simulation language is an example of a constraint-based language.43
This makes “programming” a simulation in iRides Author something like writing the formulas of a spreadsheet. Just as a spreadsheet author does not have to worry under what condition a formula will be evaluated, neither does a developer using iRides Author have to be concerned with the flow of control in a simulation. When a user takes an action, the flow of evaluation of relational expressions is automatic.

There are several motivations for special-purpose implementation tools for simulation development. The more adapted the tool is to a subject-matter domain, the more likely it is that development, or some significant subset of development, can be carried out by SMEs, rather than only by a professional cadre of development experts, who may understand little of the target domain. This development approach has several advantages. First, the risk of communications errors is reduced, because there is no problem of programmer misunderstandings about the behavior of targeted systems. Second, required changes to the simulation may be carried out much more quickly and less expensively; instead of having to find the programmer who coded the simulation in the first place (or, worse, finding another programmer to puzzle through the first programmer’s code) to make a change, an SME can directly access the relevant elements of the simulation and make the necessary changes directly.

A tool that is much more special-purpose than iRides Author is the TAO Sandbox. This tool lets Navy tactics SMEs build tactical problems that can be played out simply by choosing maps or charts, dragging units such as ships, subs, missile bases, and airfields into position, selecting the behavior of hostile units, and, optionally, scheduling future events, such as a change in the hostile posture of an opposing unit or the appearance of a new unit in the scenario at a specified time. The domain specificity of the TAO Sandbox has made it possible for SMEs to build moderately complex tactical problems in less than an hour. Frequently, the most time-consuming part of such scenario construction is writing the mission briefing for the problem, rather than any of the steps that actually determine how the simulated scenario will play.

Even more important than the advantages of SME implementation is the potential that specialized implementation tools offer for automatically supporting elements of instruction in the context of the simulation. Simulations for training, unlike other software applications, such as word processors, spreadsheets, or drawing programs, are intended primarily to teach about processes and procedures in simulated worlds. There are many types of user interactions specifically to support instruction that such simulations can use, such as “highlight” an object to direct the student’s attention, “demonstrate” an action, “require” a specific action, “detect” a simulation state of pedagogical interest, and “replay” a recorded procedure. An implementation tool that is designed to produce simulations that provide these types of instructional interactions has an advantage over more generic tools, such as conventional programming languages, which require that such services be created “de novo” for every new simulation project. Although efficiencies in development are one advantage to such tool-supplied capabilities, a more important advantage is that their presence will increase the probability that instruction will actually be used in the simulation context. Without such capabilities in a simulation, it is likely that instructor feedback will be the only possible source of correction. Automated instruction not only offers cost savings, it also helps to ensure that instruction is consistent and based on evidence about the kind of feedback that supports learner performance in the simulation, rather than on the feedback decisions that often vary with different instructors. Reviews and meta-analysis of studies have clearly indicated that a majority of feedback strategies used by instructors either have no impact or make performance worse.

In many cases, simulation-authoring tools are designed to work in conjunction with run-time tools for simulation training, as described below in the section “Software to Support the Run-time Delivery of Medical Simulations.” For example, the iRides Author system produces simulation specifications that are interpreted by a run-time component, called iRides. Every simulation built with iRides Author is actually delivered by iRides.

SOFTWARE TOOLS FOR TESTING SIMULATIONS

The testing phase of simulation development can also benefit from a close relationship with the results of CTA. Testing should ensure that all the observables specified by the CTA, and all the action types, are available in the simulation. Correct portrayal of the effects of undesirable or incorrect actions in context must be ascertained, as well, of course, as the accurate portrayal of the effects of correct actions. In a truly integrated CTA-based development process, one of the products of the CTA would be a test plan for determining the correctness (i.e., validity) and utility of the simulation. In the case of computer-based simulations, an integrated suite of tools could actually construct an automatically executable test of the simulation produced, based on the CTA.

SOFTWARE TOOLS TO SUPPORT MAINTENANCE OF SIMULATIONS

Medicine is not a closed knowledge corpus. New diseases are discovered. New techniques of medical intervention are also developed. New instructional or assessment strategies are also discovered. The conditions under which certain medical procedures should be carried out are refined. As medical procedures evolve, simulations and simulation-based training must evolve in parallel, ideally very quickly, so that there is no significant gap between approved new practice standards and the suitability of a simulation environment for accurately representing and assessing adherence to those standards.

Again, a CTA-based approach offers hope for cost-effective simulation maintenance. When a medical procedure or process is revised, the previous task analysis can most often be modified to reflect the revisions quite easily. If there is a CTA-based simulation design tool, it can provide lists of observables,
actions, and relevant conditions. These lists can be compared with those produced by the previous analysis of the procedure. New observables and new actions can be added to the simulation, and new or revised conditions can be added or edited. If the analysis tool produces a test plan, that plan can be used to drive the testing/validation phase of the new implementation of the simulation.

SOFTWARE TO SUPPORT THE RUN-TIME DELIVERY OF MEDICAL SIMULATIONS

Implementation of a computer-based simulation for training can be carried out in either of 2 ways: One, it could produce a stand-alone executable “program,” compiled in the machine language of the processor. This is the product of a conventional computer programming process. Alternatively, it could result in a “specification” that is interpreted and delivered by a software tool that can deliver a variety of training simulations. Clearly, in the second case, the run-time simulation interpreter constitutes a software tool to support run-time delivery. However, simulations produced using the conventional programming language approach can contain software components that support simulation delivery. Most modern computer programs make copious use of “libraries” of pre-existing code. Conventional code libraries exist to support interapplication communication, printing, specialized graphics packages, and so on. Possible libraries to support simulation training could include:

— Journaling packages, which write out records of student actions and other significant simulation events.
— Replay packages, which support the reading in of journal records to replay student sessions, using the simulation.
— Performance measurement packages, which report actions and events that can be evaluated during simulation sessions.

Of course, these capabilities can all be included in the second type of run-time software, the run-time interpreter. In addition, a run-time simulation interpreter can provide instructional services, such as highlighting objects and carrying out actions in a demonstration, as described above. If all training simulations supported a standard set of services for instruction, they could provide those services to instructional applications. Munro34 has proposed a universal set of such simulation services for instruction.

SOFTWARE TO SUPPORT ANALYSIS OF PERFORMANCE

To automatically assess learner performance in the context of a game or simulation, there must be an assessment component that observes user performance in that context and makes judgments. In traditional simulation-based training, the assessor is a human being with expertise in the required behaviors. If the evaluator were instead an automated software component, assessments could be made at lower cost and with greater consistency. Assessment and instructional interactions can take place either during a simulation session or post hoc. To avoid the negative effects of excessive cognitive load,22 simulations that require rapid reactions to an unfolding scenario sometimes provide assessments in after-action reviews or other postsimulation activities. In many cases, assessments cannot be made based on immediate in-simulation activities, but must be the result of cumulated observations. Software can support such analysis of performance for learner assessment.

The TAO Sandbox reports user actions and other performance-related events (e.g., the destruction of a friendly surface unit by a hostile submarine’s torpedo) to a separate application, the CRESST Assessment Application (CAA),46 using socket-based communications. The CAA is an application developed at the Center for Research on Evaluation Standards and Student Testing at UCLA. The CAA contains a detailed model of user knowledge at a higher level of representation than actions and events. When a performance announcement is made to the CAA, it informs a Bayesian network that represents the student’s exhibited competencies. Each such announcement results in modifications to the Bayes net model of student knowledge about the domain and about how to carry out procedures effectively. A third software application, the “CAA Monitor,” tracks the CAA’s changes to the Bayes net and relates those changes to performance events. This application can chart changes in student knowledge as students perform in one or a whole series of simulation training sessions.

FUTURE DIRECTIONS

It is critical that we implement a collaborative team approach to the future development of health care simulations. The application of evidence-based training and focused CTA methods with the most successful health care SMEs will produce more complete and accurate representations of expert cognitive and behavioral strategies used in complex and challenging health care settings. This information will guide the development of simulations that provide efficient and effective practice of medical skills acquired during training. There is also scope for research on tools that will better integrate the CTA process with simulation development. With appropriate software support, CTA-based strategies will provide vital data that could be imported into the simulation authoring environment to guide the production of simulations that will support effective learning experiences. Simulators provide an environment for health care professionals to repeatedly and deliberately practice new skills while receiving immediate and corrective feedback, based on desired expert performance captured by the CTA and included in the instructional design embedded in a simulation. Current medical simulators largely focus on incomplete descriptions of technical skills since about 70% of critical decisions are missing. Future research should focus on the use of CTA to expose critical decision-making skills necessary to perform complex health care tasks. Incorporating “gold standard” CTAs as the basis for teaching and
assessing technical and decision-making skills should also be examined.

ACKNOWLEDGMENTS

The Office of Naval Research supported the development of iRides, iRides Author, and the TAO Sandbox (N00014-08-1-0179, N00014-06-1-0711, N00014-08-1-0126, N00014-08-C-0563, N00014-09-C-0813, N00014-10-C-0265). Research on the use of Cognitive Task Analysis and Simulators was supported, in part, by the U.S. Army Medical Research and Materiel Command (W81XWH-04-C-0093) and the U.S. Army Training and Doctrine Command (W911NF-04-D-0005). The work reported herein was also partially supported by a grant from the Office of Naval Research, Award Number N00014-10-1-0978.

REFERENCES


Using Cognitive Task Analysis to Develop Simulation-Based Training for Medical Tasks

Jan Cannon-Bowers, PhD*; Clint Bowers, PhD†; Renee Stout, PhD‡; Katrina Ricci, PhD§; COL Annette Hildabrand, AVC USA∥

ABSTRACT  Pressures to increase the efficacy and effectiveness of medical training are causing the Department of Defense to investigate the use of simulation technologies. This article describes a comprehensive cognitive task analysis technique that can be used to simultaneously generate training requirements, performance metrics, scenario requirements, and simulator/simulation requirements for medical tasks. On the basis of a variety of existing techniques, we developed a scenario-based approach that asks experts to perform the targeted task multiple times, with each pass probing a different dimension of the training development process. In contrast to many cognitive task analysis approaches, we argue that our technique can be highly cost effective because it is designed to accomplish multiple goals. The technique was pilot tested with expert instructors from a large military medical training command. These instructors were employed to generate requirements for two selected combat casualty care tasks—cricothyroidotomy and hemorrhage control. Results indicated that the technique is feasible to use and generates usable data to inform simulation-based training system design.

INTRODUCTION
The importance to the Department of Defense of a well-trained and prepared military medical capability is unquestionable. However, compared to training civilian health care professionals (which is itself highly complex and costly), meeting the training needs imposed by military medicine presents several significant challenges. First, military medical training occurs across a wider spectrum of learners, with heavy reliance on non-physician paraprofessionals as the first-line caregiver. Second, these caregivers are often required to perform life-saving surgical procedures in uncertain, stressful environments. Third, the length of the available training period is often not optimal—given the military’s need for skilled caregivers, training must often occur in a very compressed time frame. This places a high priority on training approaches that can efficiently prepare personnel to be effective in the field.

To respond to these challenges, military medical training has relied on skill-based training of critical procedures. Such training emphasizes the ability to learn and perform a series of specific procedures that comprise a critical life-saving task. This is often accomplished through hands-on practice in increasingly realistic simulation settings. Attempts to prioritize potentially high impact interventions are needed. That is, “which tasks are most likely to benefit from simulation-based training?” and “what needs to be done to advance simulations so that they are most effective?” are the questions that need to be addressed.

There is also a need to develop and validate surrogate measures of performance that can serve as criterion measures for determining whether effective training can be accomplished. Toward that end, the present study used cognitive task analysis (CTA) techniques to identify important task cues associated with two critical combat medical procedures. By specifying these cues, we believe we can create a set of standards against that simulations can be compared, allowing estimates of their potential effectiveness for use in training and assessing critical medical procedures. Moreover, the CTA results (once validated) provide a set of critical task cues that should be included in any simulation-based training system. Finally, we sought to determine what the likely trainee errors are on major task steps so that initial training requirements could be identified.

The following sections lay out our rationale for using CTA to elucidate critical cues, errors, and simulation requirements. We then describe the specific protocol and methods we used. Finally, we present a summary of what we found for two important combat casualty care tasks (cricothyroidotomy and hemorrhage control) and discuss the implications of this work for future efforts.

BACKGROUND: CTA
Traditionally, the most common approach to describing worker behavior has been the task analysis. Task analysis
decomposes jobs into tasks and component subtasks until each activity the operator must perform is documented. The resulting analysis is then distilled into a set of training needs and approaches.\(^1\)

However, as the nature of work has changed, traditional task analysis methods that focused primarily on observable behavior are inadequate for the modern workplace. The key shortcoming is that the use of traditional task analysis techniques resulted in no data about the thought processes that were involved in the work. As workers began to take on tasks that required not only physical activity but also data monitoring and analysis, decision making, problem solving, and other higher order tasks, there arose a requirement to better understand the cues and cognitive processes that were components of successful performance.\(^2\) This gave rise to the development of techniques designed to elicit knowledge from experts. CTA is a term that refers to the use of some, or all, of these techniques, in combination with traditional task analysis methods, to fully understand the behavioral and cognitive processes needed to perform a specific job or task.\(^2\)

CTA appears well suited to the study of specific combat casualty care tasks. Certain procedures performed by medical personnel, especially surgical procedures, clearly require the operator to perform a series of behavioral and cognitive tasks. In these cases, making good decisions is just as critical to performance as proper technique. Fully describing these tasks is an essential aspect of training development, and it appears that CTA can be helpful in this regard. For example, it has been shown that training needs analyses using CTA yield 35% more information than non-CTA based analyses.\(^3\) It has also been shown that surgical laboratory courses based on CTA results lead to better outcomes for surgical students.\(^4\)

In addition, CTA has been shown to be effective in identifying critical decision errors that may be targets for training.\(^5\)

CTA for training development has been slow to be adopted as a training development technique in medicine. This may be due to the high costs associated with the CTA process. A CTA of a single surgical procedure typically requires several hours of an expert’s time. Moreover, to ensure the reliability of the results, CTAs are often conducted with several experts. For example, it has been suggested that a minimum of four experts are required to fully articulate the subtasks of one surgical procedure.\(^1\) Finally, because CTAs are often designed for specific training purposes, they may focus on only one aspect of training, such as training needs analysis. It is often the case that other CTAs are repeated to accomplish other goals, such as simulator design, scenario creation, or performance measurement. Consequently, the expense and inconvenience of CTAs are often perceived as exceeding the value of the procedure.

In this article, we assert that a properly designed CTA can anticipate and satisfy the needs of the entire training spectrum. In this fashion, one can obtain the benefits of CTA without enduring the cost of multiple analyses. Specifically, we will describe four content areas that should be considered in designing a CTA for use in military medicine: identifying training needs, developing training scenarios, developing performance metrics, and identifying simulator requirements.

**CTA to Identify Training Needs**

The most typical application of CTA is to elicit information for use in establishing training needs. This approach seems particularly important for medical training, given its emphasis of the apprenticeship model. Learners often obtain much of their training by observing the performance of a subject matter expert. However, although this observation may be useful for acquiring procedural skills, there is a substantial risk that learners may misunderstand the cognitive processes that cue the initiation of a procedure, or decisions that are made during the procedure.\(^6,7\)

CTA for the identification of training needs relies on the observational approaches of traditional task analysis. The most common addition is to ask experts to “think aloud” while performing the task.\(^8\) The expert’s description yields information about their cognitive processes without intervention by the analyst. These protocols are often videotaped and transcribed for later analysis.

Given the complexity of many military medical situations, it may be beneficial to augment the traditional “think-aloud” protocol to more fully understand the range of cognitive activities required for successful performance. Specifically, there appears to be a benefit to having the expert perform the task on a Human Patient Simulator (HPS) rather than simply imagining performance.\(^9\)

**CTA to Create Training Scenarios**

Training scientists have advocated scenario-based training as a technique that is useful for acquiring complex skills and abilities.\(^10\) This type of training is particularly important for the acquisition of skills that require substantial practice to acquire mastery. This approach is being adopted in medical training, especially to aid the transition from novice to expert. However, the key determinant of success in this training approach is the scenario provided to the trainee. The scenario must be created so that it (a) is appropriate to the trainee’s current level of knowledge, (b) provides a reasonable level of challenge, (c) allows the practice of skills that have been taught previously, and (d) allows measurement of performance and opportunities for feedback.\(^11\)

A CTA can be a rich source of information for the analyst who anticipates the need to create training scenarios. One method of obtaining information for this purpose is to use the Critical Decision Method.\(^12\) This CTA approach uses probes to have experts describe past situations (or critical incidents) in which specific skills were critical. The analyst uses probes to identify the common patterns of cues the expert uses, conditions under which various strategies are useful, and expectations he/she may have for a situation; these are all important to the expert’s decision making process. This information can
then be used as a basis for developing training scenarios that mimic the actual performance environment.\textsuperscript{13}

Feedback is an important element in determining training outcomes in scenario-based training. In addition to creating the training scenarios, CTA can also be useful in the provision of feedback to trainees. For example, McCloskey et al\textsuperscript{14} trained instructors in some CTA procedures as a way to improve their feedback to trainees. Specifically, they trained the instructors to use probes to elicit information about the cues that were used, the interpretation of those cues, the decision-making processes that made use of them, and the manner in which they influenced the final decision. The results indicated that instructors were much more confident in their ability to provide meaningful feedback to the trainees.

**CTA to Create Performance Metrics**

One challenge for training professionals in this field is to assess the degree to which the trainee has acquired the knowledge, skills, and abilities targeted for training. This is especially difficult as the trainee progresses from performing observable skills to higher levels of expertise, when cognitive elements such as decision making come into play.

Professionals in this area have responded to this challenge by articulating a scenario-based approach to training assessment that is similar to that applied to training. In this approach, scenarios are designed in such a manner so as to elicit the targeted behaviors as evidence that the knowledge/skills have been mastered. However, there is often not a body of empirical evidence to serve as a foundation for these judgments. Therefore, we must rely on the information possessed by the task experts.

Eliciting assessment information from experts is not as easy as simply asking them. It has been shown that, in the process of becoming an expert, learners tend to re-organize information about the task so that it becomes “automatic.” This allows them to act very quickly, but makes it difficult to articulate the individual processes that may be part of that action.\textsuperscript{15} As such, CTA tools are often used in the process of “knowledge elicitation.” This refers to the application of these tools to assist experts in articulating otherwise inaccessible knowledge.\textsuperscript{16} Knowledge elicitation techniques are critical to the use of CTA for designing assessments.

Consistent with this position, Mislevy et al\textsuperscript{17} described a process for using CTA in developing an assessment of dental assistants. The key elements of this approach are to: select a set of cases that evoke the targeted skills; obtain think-aloud protocols from performers representing a continuum of expertise levels; analyze the protocols to identify those situations and behaviors that discriminate among performance levels; and construct assessment scenarios that use those situations that are most effective in discriminating among levels of expertise.

This approach differs from other CTA approaches in a few important ways. First, it places a responsibility on the examiner to select specific elements of the task for analysis. Rather than have the performers conduct an entire task, the analyst selects only specific parts of the larger task that are presumed to discriminate among expertise levels. Mislevy et al\textsuperscript{17} accomplished this by articulating a set of subtasks based on training manuals, licensure exams and so forth. They then had experts provide judgments about which tasks were likely to discriminate among levels of expertise.

A second important difference is in the creation of the assessment scenarios. The purpose of this step is to take the tasks selected above and to use them to create a reasonable, realistic, evaluation scenario. This scenario must contain all of the cues necessary to allow behaviors to occur that are likely to generalize to the actual performance situation. As Mislevy et al\textsuperscript{17} point out, the scenario must include not only the critical cues, but the “context, expectations, instructions, affordances, and constraints the employee will encounter” (p. 340). Further, the scenario must be “seeded” with events that provide an appropriate challenge to the examinee so that he/she exhibits the targeted behaviors.\textsuperscript{1} Hence, there is a need to create specific events that provide the opportunity for the examinee to display the targeted behaviors, if he/she has actually mastered them.

A final important difference in CTA approach for developing assessments is the manner in which the data are analyzed. Rather than attempting to catalog all of the necessary tasks and subtasks, the analysis phase for this purpose focuses only on those elements that discriminate between the behaviors of novices and experts. These elements are used to create an assessment instrument that will be used by trained observers. The items are limited to the key discriminating behaviors to make the scale manageable for raters. It may also be helpful to provide cues to the rater to focus their attention on important simulator events or behaviors that they should watch for.\textsuperscript{1}

**CTA to Identify Simulator Requirements**

CTA is often used in the process of developing new systems, particularly to understand the needs of the operator for tasks with significant cognitive demands.\textsuperscript{3} Interestingly, however, there is little guidance about how to use these methods in the design of training simulators. There is clearly an opportunity to optimize simulator design by using CTA methods as part of the requirements generation process.

Training simulators are designed to provide opportunities for learners to practice newly acquired knowledge, skills, and abilities in a realistic environment. To accomplish this goal, a simulator must satisfy several requirements in creating an effective practice environment. One such requirement is that the simulator provides the task cues required for trainees to select the appropriate course of action. Scientists in the area of expertise have described the importance of developing appropriate cue-pattern associations as a key element that discriminates between experts and novices\textsuperscript{3}; that is, connection between the critical task conditions (cues) with appropriate
responses. To achieve this, task cues must be presented at an adequate level of fidelity to allow transfer from the simulation to the actual task. Cues that are absent, or poorly presented, may not trigger appropriate behaviors when needed. Furthermore, if cues are improperly presented, there is a risk that trainees might develop incorrect associations, impeding their ability to transfer learning to the operational setting. This could lead to errors of varying magnitude in performance.

Given the above, it is tempting to include every possible cue in a simulation at the highest possible level of fidelity. However, matters of cost largely prohibit this. Therefore, there is a need to identify those cues that are required for experts to make their decisions, and to identify the minimum level of fidelity required to trigger realistic behavior. In that fashion, one can achieve the most cost-effective approach for training outcomes. Using a CTA is crucial here as research suggests that some simulated elements that were thought to have importance by design engineers were perceived to have little value by learners.18

Above and beyond issues of cue fidelity, CTA is also essential for establishing the needs for new simulation technologies and capabilities. For example, Low-Beer et al16 described a CTA-based approach to teaching digital rectal examination procedures. The authors determined that the existing simulation technologies did not give the instructors the proper cues to make a determination of expertise. Consequently, they created a modified version of the simulator that allowed instructors to better see the targeted behaviors. They concluded that the new simulator not only enabled more effective evaluation, but in subsequent simulator assisted CTAs, they identified several more behaviors that needed to be trained. The following section documents a study whose purpose was a CTA in a medical domain.

METHOD

Participants

The participants for this study were nine instructors from a military training command (for the sake of clarity, we will refer to them as “instructors” throughout the remainder of this article) where the targeted procedures are trained. These instructors all had several years of experience and substantial training in the two techniques of interest. It should be noted that, although there are some variations in the steps that specific training commands use for each procedure, we do not believe that the critical cues should be different for any particular set of steps.

Procedure

The Institutional Review Board–approved protocol lasted for 2 days, 1 day per procedure (e.g., cricothyroidectomy, hemorrhage control). Before starting the CTA, instructors completed an informed consent form. Then, for each procedure, instructors were asked to complete a think-aloud protocol while they performed the targeted procedure on an HPS. They were asked to imagine that they were doing the procedure in a combat situation. After their free response, they reviewed the procedures again while the interviewer provided specific probes to elicit the critical cues for performance. Next, instructors were asked specific questions about the existing simulator and its differences from live tissue. Instructors were then asked to perform the procedure with an emphasis on behaviors that discriminate between levels of performance in their students (e.g., novices vs. experts).

RESULTS

On the basis of our design, the categories for which we report responses and observations relevant to each of the major steps in each task were the following

(1) Cues used to perform the step
(2) What is different or lacking in terms of cues on the HPS relevant to performing the step and/or what instructors want the HPS to have
(3) Typical trainee or student errors in performing the step
(4) What the instructors can observe when evaluating trainees’ performance of the step and/or how they can observe this
(5) The decisions made in the process of performing each procedure

Cricothyroidotomy Task

Results from the cricothyroidotomy task are shown in Table I. Inspection of this table reveals first that we concluded that nine critical steps describe the task (shown in column 1). The second column indicates examples of the major cues that experts reported as crucial to each step. The next column delineates the current deficiencies in the simulator (which can also be considered requirements) used to train this task. Several of these relate to the simulated tissue (skin, membranes), absence of blood and lack of realism in several other aspects of the task. The fourth column indicates typical errors that novices are likely to make at each step. These errors can be used as a basis to assess trainee performance and to specify training content. The fifth column displays the observable behaviors at each step. This information can be used as the basis to develop a performance assessment tool or observational protocol (e.g., checklist) that can be used by instructors or raters. Finally, there is a description of the decision making demands.

For example, returning to the first step of cricothyroidotomy task (controlling the trachea), results indicated that it is important for trainees to understand the necessity for grasping the trachea between the thumb and middle finger of their non-dominant hand to free their index finger to palpate for landmarks, and to understand that if the trachea is not properly secured it will likely move and they can lose their landmark. Results also showed that simulations designed to train this step should allow trainees to actually grasp the trachea in this manner, include an average trachea width and variations to instruct proper grasping of the trachea, and allow
### TABLE I. CTA Results for the Cricothyroidotomy Task

<table>
<thead>
<tr>
<th>Major Step</th>
<th>Cues Used to Perform</th>
<th>Simulator Deficiencies or Requirements</th>
<th>Typical Errors</th>
<th>Observable Trainee Behaviors</th>
<th>Decision Making Demands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Trachea</td>
<td>Tactile cues from thyroid cartilage—grasp with non-dominant hand between thumb and middle finger and secure it from moving</td>
<td>Simulated skin is too rigid; thyroid rings are sometimes missing; cannot actually grasp trachea and trachea does not move; include trachea of varying widths</td>
<td>Insufficient control of trachea; not grasping it between thumb and middle finger of non-dominant hand to free index finger to palpate</td>
<td>Trachea movement because of releasing grip or inadequate grip; vertical location of grip</td>
<td>Trachea control adequate; release trachea</td>
</tr>
<tr>
<td>Palpate for Landmarks</td>
<td>Tactile cues of depression over cricothyroid membrane; feeling the groove below a notch (point of Adam’s apple or thyroid cartilage)</td>
<td>Simulator skin is too rigid; “membrane” opening is too large, causing the depression to feel too large; missing anatomical landmarks (e.g., trachea rings)</td>
<td>Missing landmarks; misinterpreting “false” landmarks; incorrect use of finger for palpation (e.g., using two hands to palpate); feeling too high or too low on trachea</td>
<td>Finger dip into depression over cricothyroid membrane; index finger usage and vertical location on trachea</td>
<td>Interpretation of physical cues</td>
</tr>
<tr>
<td>Make Vertical Incision</td>
<td>Visual cues from opening of skin layers; seeing underlying structures; tactile cues of underlying structures; tactile cues of scalpel blade and visual cues of blade’s depth</td>
<td>Simulator skin is difficult to incise and causes not too much resistance; skin does not open in proper layers; absence of underlying structures and blood</td>
<td>“Freezing up”; too timid or aggressive with incision; cutting too shallow or too deep; incorrect scalpel grip and use of tip of blade vs. belly of blade; cutting too high or too low</td>
<td>Location of incision; depth of incision; scalpel grip, angle, and whether or not tip or belly of blade is used</td>
<td>Location of incision; choice of scalpel technique</td>
</tr>
<tr>
<td>Reacquire Landmark—</td>
<td>Visual cue of seeing finger in incision; tactile cue of palpatating cricothyroid membrane, such as graininess</td>
<td>Simulator membrane does not look or feel like actual membrane</td>
<td>No results obtained</td>
<td>Whether or not landmark is reacquired; placement of finger in incision</td>
<td>Reacquisition of membrane successful</td>
</tr>
<tr>
<td>Cricothyroid Membrane</td>
<td>Tactile cue of “poking through” membrane; feeling of bone resistance if cutting horizontally vs. just poking through</td>
<td>Lack of supporting structures reduce realism; absence of redundant vasculature, vocal cords, and blood; absence of bone resistance</td>
<td>Severing surrounding blood vessels or vocal cords</td>
<td>Depth and location of “poke”; blade angle; width of horizontal movement</td>
<td>Location and depth of incision</td>
</tr>
<tr>
<td>Penetrate Cricothyroid Membrane</td>
<td>Visual cues from opening hemostat or scalpel; tactile cues of feeling the area is opening</td>
<td>No results obtained</td>
<td>Losing control of incision by removing scalpel before securing it with hemostats or a cric hook; tearing of ET tube balloon with tools</td>
<td>Proper use and location of tools; proper incision control</td>
<td>No results obtained</td>
</tr>
<tr>
<td>Dilate Area</td>
<td>Tactile and kinesthetic cues when inserting ET tube, including guiding it with tools</td>
<td>Simulator imposes much more resistance than live patient; lack of natural lubrication (e.g., saliva, mucous, blood)</td>
<td>Improper cric hook use, including orienting it toward the chin; excessive force on or twisting of ET tube; improper location of tube, including going subcutaneous</td>
<td>Proper use and orientation of hook; pulling hook to make anatomical structures more pronounced; rotation and position of ET tube</td>
<td>Tube successfully placed</td>
</tr>
<tr>
<td>Insert ET Tube</td>
<td>Visual and tactile connection of syringe and cuff; seeing that syringe is at proper cc line; seeing that it is removed</td>
<td>Simulator cuff leaks more than in human because of inadequate seal</td>
<td>Failing to make actual connection; leaving syringe attached, allowing leakage and deflation of cuff</td>
<td>Number of cc’s inserted; their checking of rigidity of cuff; detachment of syringe</td>
<td>Cuff appropriately rigid</td>
</tr>
</tbody>
</table>

(Continued)
the trachea to move if not properly controlled. Results further indicated that metrics could include determining whether the trainee had grasped the trachea between the thumb and middle finger of her non-dominant hand, had lost her grasp, or had allowed the trachea to move.

**Hemorrhage Control Task**

Results from the hemorrhage control task are shown in Table II. Consistent with Table I, the major steps in the hemorrhage control task appear in the first column, followed by cues needed to perform the step and implications for observable trainee behaviors.

<table>
<thead>
<tr>
<th>Major Step</th>
<th>Cues Used to Perform</th>
<th>Simulator Deficiencies or Requirements</th>
<th>Typical Errors</th>
<th>Observable Trainee Behaviors</th>
<th>Decision Making Demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locate Source of Bleeding</td>
<td>Visual appearance of wound, including depth and seeing severed artery along long bone; feel of wound, including depth and feeling severed artery and long bone; blood flow, including spurting vs. oozing, and rate of cavity filling; color and viscosity of blood</td>
<td>Absence of realistic artery and long bones; absence of realistic blood flow, including rate, spurting, and pooling; absence of realistic blood characteristics, including color and viscosity</td>
<td>Failure to locate source of bleeding; improper wound cleaning; inadequate application of pressure because of incorrect location applied or insufficient force</td>
<td>Location of pressure being applied; whether blood flow is stopped; wound cleaning and drying; direction of palpation</td>
<td>Interpretation of cues</td>
</tr>
<tr>
<td>Pack Wound</td>
<td>Tactile cue of finger against bleeding source; feeling of pressure as gauze fills cavity; feeling that intentionally packing with some force; feeling of wound as spongy or like raw steak</td>
<td>Absence of realistic cavity and blood vessels; absence of slowing of flow of blood as gauze contacts bleeder; absence of separation of muscle and skin if packed too aggressively</td>
<td>Failure to maintain pressure; failure to dry wound before packing; failure to make direct contact with wound; use of insufficient packing material; packing that is not aggressive enough or is too aggressive</td>
<td>Whether or not gauze is in direct contact with bleeder and whether blood flow is slowed; amount of gauze used; depth of gauze; force of packing and whether or not skin and muscle are separating</td>
<td>Interpretation of cues; correct force for packing; amount of packing</td>
</tr>
<tr>
<td>Apply Pressure Dressing</td>
<td>Visual cues for placement of dressing, including whether it is placed directly centered over wound</td>
<td>N/A</td>
<td>No results obtained</td>
<td>Placement of dressing and whether it is applied as a ball on top of wound</td>
<td>Interpretation of cues; dressing placement</td>
</tr>
<tr>
<td>Wrap Wound and Dressing With Elastic Bandage</td>
<td>Visual and kinesthetic cues indicating tightness of bandage</td>
<td>Simulator skin is unrealistically non-pliable</td>
<td>Failure to secure wound tightly enough because of using too long of throws when wrapping or insufficient force</td>
<td>Student exertion; length of throws used; tautness of bandage and whether it is shiny</td>
<td>Interpretation of cues; tightness of wrapping; amount of wrapping</td>
</tr>
<tr>
<td>Document</td>
<td>Documentation of type of gauze used and whether or not wound is bleeding through tape by writing on or marking tape</td>
<td>N/A</td>
<td>Improper or incomplete documentation</td>
<td>Improper or incomplete documentation</td>
<td>N/A</td>
</tr>
</tbody>
</table>
simulation, training, and measurement. These results can be used to determine training requirements, inform the design of simulations necessary to train the procedure, and develop assessment protocols, as can the results for the cricothyroidotomy task.

For example, regarding the first step of hemorrhage control task (locating the source of bleeding), results indicated that trainees need to understand the anatomy of the body and how an arterial bleed will tend to follow the long bones; that it will be deeper than a venous bleed; that it will flow more rapidly and more quickly fill the cavity; and that it will actually spurt from the severed artery, which can be palpated and sometimes seen. Results also showed that simulations designed to train this step should include long bones and the bleeding artery near it; a proper flow of blood; spurring blood; and a rapidly filling cavity. Results further indicated that metrics could include determining whether the trainee: had applied adequate pressure to stop the flow of blood, had cleaned and dried the wound to allow visual inspection, or had palpated in the correct direction to feel the source of the bleed.

CONCLUSIONS
Developing effective training and assessment for combat casualty care tasks is an imperative. Given the cognitive and behavioral demands of such tasks, this can be a complex and expensive undertaking. CTA techniques hold promise as a means to develop a comprehensive understanding of the major steps to be performed in a task, along with detailed information about how best to train those tasks, and important insights into how to develop sound metrics for them. On the basis of past work into CTA and our experience in conducting one for the combat casualty care tasks discussed above, we offer the following guidelines for conducting a comprehensive CTA that enables multiple goals to be accomplished.

(1) Select multiple subject matter experts (SMEs) (instructors are an excellent source) who can give detailed accounts of task/training requirements. Use an HPS as a reference point for SMEs to recount how the task is performed.
(2) Elicit the major tasks and subtasks (steps) needed to accomplish the task using a think-aloud protocol.
(3) Have SMEs perform the task again, using probe questions to elicit critical cues at each step.
(4) Ask specific questions pertaining to deficiencies in simulation and other training devices.
(5) Elicit specific errors likely to be made by novices at each step and determine what specific observable behaviors typify each step.

ACKNOWLEDGMENTS
This work was funded by the Naval Air Warfare Center Training Systems Division, Orlando, Florida, under contract PO SEA1333. The work reported herein was also partially supported by a grant from the Office of Naval Research, Award Number N00014-10-1-0978.

REFERENCES
Use of Cognitive Task Analysis to Guide the Development of Performance-Based Decision Making for IntraOperative Decision Making

Carla M. Pugh, MD, PhD*; Debra A. DaRosa, PhD†

ABSTRACT Background: There is a paucity of performance-based assessments that focus on intraoperative decision making. The purpose of this article is to review the performance outcomes and usefulness of two performance-based assessments that were developed using cognitive task analysis (CTA) frameworks. Methods: Assessment-A used CTA to create a “think aloud” oral examination that was administered while junior residents (PGY 1-2’s, N = 69) performed a porcine-based laparoscopic cholecystectomy. Assessment-B used CTA to create a simulation-based, formative assessment of senior residents’ (PGY 4-5’s, N = 29) decision making during a laparoscopic ventral hernia repair. In addition to survey-based assessments of usefulness, a multiconstruct evaluation was performed using eight variables. Results: When comparing performance outcomes, both approaches revealed major deficiencies in residents’ intraoperative decision-making skills. Multiconstruct evaluation of the two CTA approaches revealed assessment method advantages for five of the eight evaluation areas: (1) Cognitive Complexity, (2) Content Quality, (3) Content Coverage, (4) Meaningfulness, and (5) Transfer and Generalizability. Conclusions: The two CTA performance assessments were useful in identifying significant training needs. While there are pros and cons to each approach, the results serve as a useful blueprint for program directors seeking to develop performance-based assessments for intraoperative decision making.

INTRODUCTION

There is a need to assess intraoperative decision making. However, there is a paucity of training curricula and performance assessments addressing this important aspect of surgical skill. Previous work on intraoperative decision making has largely focused on surgical outcomes relating to procedure choices or technical approaches. In the article, “Intra-operative decision making in the treatment of shoulder instability,” the authors discuss surgical outcomes for patients who had an open arthrotomy versus a minimally invasive arthroscopic procedure as the treatment of choice. In another article, the authors discuss the use of a minimally invasive versus open approach to empyema. These studies address intraoperative decision making as it relates to surgical planning based on intraoperative findings.

Additional work focusing on intraoperative decision making has taken advantage of cognitive psychology script theory. In these studies, researchers developed, implemented, and evaluated script concordance tests (SCTs) as a means of assessing operative judgment. SCTs are a form of problem-case-based pencil-and-paper examinations. However, compared to standard multiple choice questions where there is one correct answer, SCT items are based on a Likert scale. As such, there is no single correct answer, and scoring is based on deviation from the expert range. SCTs have been used to assess clinical judgment in surgery, urology, and radiology. When used to assess intraoperative decision making, SCTs were noted to correlate with years in training but not the number of cases performed in the operating room.

The complexity of intraoperative judgment calls for a variety of research methods to better understand the range and types of decisions made in the operating room. In addition to innovative pencil-and-paper tests like the SCTs, performance-based assessments may provide critical insight. Performance-based assessments are considered to be a more direct measure of task performance compared to oral and pencil-and-paper examinations. Pencil-and-paper examinations are indirect indicators or correlates of performance in specific task domains. In general, performance-based assessments involve the observation of trainees performing tasks in various contexts. Because of the complexity of performance-based assessments, validity must be evaluated from several different constructs. The purpose of this research project was to perform a critical, multiconstruct evaluation of two, independently developed, performance-based assessments that focus on resident’s intraoperative decision-making skills.

METHODS

Methodology

This research project used a mixed-methods qualitative approach. The methods were based on case-oriented comparative research (COCR) and variable-oriented comparative research (VOCR). The primary goal was to compare two, newly developed, performance-based assessments that focus on intraoperative decision making. A secondary goal was to...
provide program directors and researchers a conceptual blue-
print that would facilitate the development, implementation,
and evaluation of performance-based assessments for intra-
operative decision making. To develop the blueprint, compar-
ison points (independent variables) were defined and used
to guide the case-oriented comparison and operationalize blue-
print components.

COCR has been used for over 40 years.\textsuperscript{14–16} This research
method often employs qualitative research techniques and
maintains the overarching goal of discovery. The COCR is
often compared with variable-oriented research, which is tra-
ditionally based in hypothesis-driven and quantitative realms.
Both approaches seek causal relationships; however, the
approach and communication of such relationships differs.\textsuperscript{14–16}
Although implementation of both methods varies depending
on context and investigator preference, a commonly stated
difference is that COCR is reserved for comparing small case
numbers\textsuperscript{2–4} and that conclusions drawn from a causal perspec-
tive should be done within cases and not across cases.\textsuperscript{17} Our
methods followed these guidelines.

\textbf{Study Cases}

Two independent research projects used cognitive task anal-
ysis (CTA) techniques to guide the development of assess-
ments geared toward intraoperative decision making. CTA
techniques are used to discover the cognitive activities and
knowledge that experts use to perform complex tasks.\textsuperscript{18,19}
The outputs are used as the gold standard by which non-
experts are compared. The first independent project used
CTA to create a skills lab curriculum for laparoscopic cholec-
ystectomy. The curriculum included a pencil-and-paper test
and a “think aloud” oral examination that took place while
junior residents, PGY-1 and PGY-2 ($N = 69$), performed
a porcine-based laparoscopic cholecystectomy. Faculty facili-
tators prompted think aloud moments using a set of standard-
ized questions that correlated to laparoscopic cholecystectomy
procedural steps. The questions, as shown in Table I, were
posed at the appropriate times throughout the simulated pro-
cedure and scored for accuracy and completeness on anchored
scales of 1 (low) to 5 (high). Participants were asked to discuss
their knowledge of the critical procedural steps (e.g., How will
you decide where to place the subxiphoid trocar?), error recog-
nition (e.g., Describe special risks of inserting a Veress needle
in thin or muscular patients), and error management (e.g., List
potential complications associated with port sites and describe
how to prevent them).\textsuperscript{20}

The second independent project used CTA to create a
simulation-based assessment. The assessment focused on
decision making during a laparoscopic ventral hernia repair.
This assessment was geared toward senior level residents
in the fourth and fifth postgraduate years, PGY-4 and PGY-5
($N = 29$). The residents were tasked with repairing a
10 cm $\times$ 10 cm ventral hernia using a newly developed,
decision-based simulator (Fig. 1). Residents were expected to independently make all of the decisions necessary in completing the hernia repair using laparoscopic techniques. Operative performance was video recorded and later scored using a checklist evaluation. Outcomes were based on task errors and task completion rates. In addition, their hernia repairs were graded based on mesh placement and coverage of the hernia defect.21

Data Collection and Analysis
For the case-oriented comparison, descriptive data were collected regarding project goals, performance assessment development, implementation, and outcome-related details.13 These data were used to guide case comparison based on commonalities and disparities. Additional data were collected using eight variables that have previously been shown to be important in assessing validity of performance-based assessments. These variables were used as part of the variable-oriented comparison and included (a) Fairness, (b) Cognitive Complexity, (c) Content Quality, (d) Content Coverage, (e) Meaningfulness, (f) Consequences, (g) Cost and Efficiency, and (h) Transferability and Generalizability.11

Data analysis was qualitative in nature. Between-case comparisons were used to provide insight on the pros and cons of each case regarding approach and outcomes.13 Within-case comparisons and assumptions were evaluated to gain a better understanding of individual project outcomes and what this means for trainees and program directors. Student surveys and subjective evaluations based on review of individual project methods and project outcomes were used to generate summative evaluations for each of the eight variables. Analysis goals were descriptive and qualitative in nature as opposed to causal related. In essence, the independent variables served as a framework for descriptive analysis.

RESULTS
Using the case-oriented comparative method, commonalities and disparities in process and outcomes were reviewed. Important commonalities included (1) use of performance-based assessments to assess intraoperative decision-making skills of surgical trainees; (2) use of CTA research methods as a means to guide development of the performance-based assessments; (3) use of high fidelity, hands-on tasks that allow the use of surgical instruments commonly and currently being used in today’s operating rooms; and (4) use of outcome measures that require completion of scoring rubrics/checklists based on the observation of experienced raters.

Important disparities included (1) experience level of the study participants, (2) content domain, (3) type of performance-based assessment, and (4) domain- and implementation-specific differences in assessment outcomes (Table II). Overall, the disparities did not outweigh the ability to perform an in-depth comparison. For the variable-guided assessment, case-related advantages included (1) Cognitive Complexity, (2) Content Quality, (3) Content Coverage, (4) Meaningfulness, and (5) Transfer and Generalizability. In contrast, both projects had some disadvantages in the 3 remaining evaluation areas (Table III).

When comparing outcomes, both approaches revealed major deficiencies in residents’ intraoperative decision-making skills. For study A, during the think aloud exercise, residents struggled to formulate and verbalize their decisions and were not able to talk and operate at the same time. For Study B, 89% of residents made critical decision-based errors that prevented them from successfully completing the hernia repair.

DISCUSSION
There is a paucity of training curricula and performance assessments that focus on intraoperative decision making. Intraoperative decisions, related to procedure choices or technical approaches, are known to effect surgical outcomes.1–3 As such, preoperative and intraoperative judgments are critical to patient welfare. Prior works relating to the assessment of intraoperative decision making have

<table>
<thead>
<tr>
<th>TABLE II. Summary of Case Commonalities and Disparities (bolded entries)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grounding Theory</td>
</tr>
<tr>
<td>Task Fidelity</td>
</tr>
<tr>
<td>Outcome Instrument</td>
</tr>
<tr>
<td>Task Analysis Development Goal</td>
</tr>
<tr>
<td>Content Focus</td>
</tr>
<tr>
<td>Learner Group</td>
</tr>
<tr>
<td>Decision-Making Assessment Method</td>
</tr>
<tr>
<td>Data Collection</td>
</tr>
<tr>
<td>Assessment-Related Outcome</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
used SCTs. Using this method, two studies have shown no correlation between number of procedures performed in the operating room and SCT ratings. There is, however, a strong correlation to training level. Despite the reported validity and reliability of SCTs for assessing intraoperative decision making, performance-based assessments are considered to be a more direct measure of task performance. However, because of the complexity of performance-based assessments, validity must be established based on several different constructs. The purpose of this research project was to perform a critical, multi-construct evaluation of 2 independently developed, performance-based assessments geared toward resident’s intraoperative decision-making skills.

The first performance-based assessment included a think aloud exercise during a porcine-based laparoscopic cholecystectomy. This assessment involved both PGY-1 and PGY-2 residents (N = 69). Faculty facilitators prompted think aloud moments using a set of standardized questions. The results of this study revealed that residents experienced difficulty formulating and verbalizing their intraoperative decisions and were unable to talk and operate at the same time. The second performance-based assessment required senior level residents, PGY-4 and PGY-5 (N = 29), to repair a 10 cm × 10 cm ventral hernia using a newly developed, decision-based simulator. Residents were expected to independently make all of the decisions necessary in completing the hernia repair using laparoscopic techniques. Results of this study showed that 89% of the residents made critical decision errors that prevented them from completing the hernia repair.

Key commonalities that allowed for case comparison included (1) use of performance-based assessments to assess intraoperative decision-making skills of surgical trainees and (2) use of high fidelity, hands-on tasks that allow the use of surgical instruments commonly and currently being used in today’s operating rooms. A preferred requirement of performance-based assessments is that they are conducted in a venue that is as close to the real environment as possible. It is this key factor that sets performance-based assessments apart from pencil-and-paper and oral examinations and allows for a more direct evaluation of task performance skills.

A variable-oriented approach was used to guide the case comparison. In 1991, Linn, Baker, and Dunbar published a review that summarized eight key factors that affect the validity of performance-based assessments: (a) Fairness, (b) Cognitive Complexity, (c) Content Quality, (d) Content Coverage, (e) Meaningfulness, (f) Consequences, (g) Cost and Efficiency, and (h) Transferability and Generalizability. These eight factors were used as independent variables to frame the data collection and discussion from a case comparison standpoint. These independent variables were not used to generate a hypothesis that sought effects on a common dependent variable, rather qualitative methods were used. Qualitative and descriptive analyses based on the previously listed variables showed that both approaches showed major advantages in five of the eight evaluation areas including (1) Cognitive Complexity, (2) Content Quality, (3) Content Coverage, (4) Meaningfulness, and (5) Transfer and Generalizability. In contrast, both projects showed some disadvantages in the remaining three evaluation areas (Table III).

A commonality in the disadvantages related to the novelty of performance-based assessments for intraoperative decision making. Currently, there are no standard curricula in wide use. As such, it may not be fair to use a performance-based assessment on this topic as a summative evaluation in residency training. However, a formative assessment may be appropriate. An additional disadvantage includes the potentially high-time requirements for implementation of performance-based assessments. This also affects cost and efficiency.

From a cognitive standpoint, both user groups clearly showed deficiencies at levels that were unexpected. Junior residents struggled to formulate and verbalize decisions while operating. However, when the performance tasks were
Cognitive Task Analysis for the Development of Performance-Based Assessments

separate (performing an operation vs. verbalizing intraoperative decisions) the residents performed well. Similarly, it was not expected that 89% of senior level residents would fail to complete an operation that they had performed before in the live operating room setting. Cognitive load theory offers an explanation for both of these observations. This theory is based on the assumption that the brain has limited working memory for dynamic processing combined with partly independent processing units for visual and auditory inputs. At any given time, these inputs are in competition with working memory. When experiencing novel situations, with a variety of new inputs, the capacity for working memory is less and may result in an inability to multitask (junior residents unable to talk and operate at the same time) or forgetfulness (junior residents forgetting concepts they had recently verbalized or senior residents forgetting the steps of a surgical procedure when allowed to be independent for the first time). These findings underscore a key weakness in using pencil-and-paper tests to assess intraoperative decision making: the inability to dynamically assess unforeseen knowledge and performance deficits that affect cognitive load, hence task completion.

For both cases, it must be underscored that there were no trick questions or rarely experienced scenarios. Table I shows the standardized questions used in the think aloud study. These questions represent basic surgical knowledge for laparoscopic cholecystectomy. However, this is new knowledge for the participants in this study (PGY 1-2), hence they are not as experienced with recalling and verbally communicating the facts. For the second study with the senior residents (PGY4-5), there were several points in the procedure where errors were committed including (1) poor laparoscopic port planning, (2) failure to repurpose a port to facilitate visualization or retraction, (3) failure to match mesh size to hernia defect, and (4) improper mesh prep on the back table. Each of these steps is critical in successful completion of a laparoscopic hernia repair. Forgetting or making errors in any one of these steps will significantly affect successful and timely completion of the repair. Although many of the residents had performed this procedure several times before, it appears that 50 cases (the highest case experience for this group of residents) over a 5-year period is not enough to reach mastery or independent performance.

A major limitation of this study was access to only two performance-based assessments that focused on intraoperative decision making. Although the case-oriented comparative methods are designed to study two to four cases, limitations in causality and outcome generalizability are duly noted. In addition, when using a mixed-method approach, there will always be some difficulty in data fitting, analyzing, and reporting. Study strengths include topic importance and the need to reliably assess intraoperative decision making. Although more traditional assessments such as pencil-and-paper tests lend themselves to robust item analysis, performance-based assessments require a multiconstruct approach. Our findings serve as a useful blueprint for developing and evaluating performance-based assessments that focus on intraoperative decision making.

ACKNOWLEDGMENTS

This research was supported in part by the Department of Defense, United States Army Medical Research and Material Command, USAMRMC-W81XWH0710190, and the Association for Surgical Education—Center for Excellence in Surgical Education, Research and Training (CESERT) Grant Award. The work reported herein was also partially supported by a grant from the Office of Naval Research. Award Number N00014-10-1-0978.

REFERENCES

Balancing Physiology, Anatomy and Immersion: How Much Biological Fidelity Is Necessary in a Medical Simulation?

Thomas B. Talbot, MD, MS, FAAP*†

ABSTRACT
Physiology and anatomy can be depicted at varying levels of fidelity in a medical simulation or training encounter. Another factor in a medical simulation concerns design features intended to engage the learner through a sense of immersion. Physiology can be simulated by various means including physiology engines, complex state machines, simple state machines, kinetic models, and static readouts. Each approach has advantages in terms of complexity of development and impact on the learner. Such factors are detailed within the article. Various other biological, hardware-based, and virtual models are used in medical training with varying levels of fidelity. For many medical simulation-based educational experiences, low-fidelity approaches are often adequate if not preferable.

INTRODUCTION
There has been a tremendous shift in medical training over the last decade. The venerable approach of passive observation, trial and error, slavish working hours, and lengthy “rounds” is moving toward medical simulation as a mainstay of training for physicians, nurses, and medics. Animal-based laboratories are moving to computer-driven training experiments that replicate the physiology of humans in an environment that encourages experimentation and repetition over one-off opportunities. As computer, material science, and electromechanical technologies advance, there is a concurrent desire to create simulation experiences with ever higher fidelity. Given this, it is important to ask how much fidelity is optimal. This article explores various aspects of biological fidelity with emphasis on physiology and anatomy. The intent is to contrast approaches based on type of training, development effort, and impact on the learner.

PHYSIOLOGICAL FIDELITY
Many training scenarios involve demonstrations of physiological action with an expectation that the learner diagnose a condition based upon the demonstrated physiology, make interventions and view a realistic physiological response as would be seen in a patient encounter. A variety of mechanisms exist that can do this with varying levels of fidelity, dynamism, and effort involved in their creation.

COMPLEX FIDELITY: ANIMAL PHYSIOLOGY
Complex human physiology can be simulated with animal surrogates. Live animals possess analogous physiology that is highly reliable, responds to just about any medication or physical intervention, and allows for realistic interaction with learners. The drawbacks to using live animals include the high cost and effort to maintain the animals, the need to avoid suffering, infectious disease considerations, dissimilar anatomy, lack of repeatability, and the possible or likely loss of the animal.1

COMPLEX FIDELITY: PHYSIOLOGY ENGINES
A replicable experience with complex fidelity can be achieved through physiology engines. Physiology engines are computer-coded mathematical models that simulate body systems. Basic physiology engines replicate the cardiovascular system and the effects of hemorrhage, fluids, and medications on the model. Some manikins include such engines.2 More complex physiology engines are multisystem with large pharmacology libraries and multidrug interactions. An example of a multisystem model is HumMod, created by the University of Mississippi Medical Center.3 HumMod can readily simulate a wide variety of conditions such as hemorrhage, heart failure, ketoacidosis, or hyperaldosteronism. The results of HumMod outputs are in the form of graphs or data that will closely match results from physiology research studies and textbooks. Finally, biological modeling can simulate biological processes down to the molecular level, though this requires intense computing power and is used for research purposes rather than for education.

Physiology models are a high-end solution that can run in real or accelerated time. They have the capability to mimic realistic physiological activity and can gracefully manage unexpected user inputs. They can cope with the effects of multiple interventions even if those interventions are antagonistic to each other. One problem is that realistic changes in physiology may be too gradual or subtle for the learner to notice unless on-screen indicators readily depict historical trends. Some physiology processes, such as sepsis or chemistry changes, unfold too slowly to be observed during an educational scenario. In these cases, the simulation will appear

*Institute for Creative Technologies, University of Southern California, 12015 Waterfront Drive, Playa Vista, CA 90094-2536.
†Telemedicine and Advanced Technology Research Center, 1054 Patchel Street, Fort Detrick, MD 21702-5012.
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-13-00212
insufficiently responsive and fail to engage the learner. Ironically, gradual responses to user inputs can reduce the user’s impression of biological fidelity. The need to closely observe monitor displays while tracking ongoing changes distracts the learner from observing the patient. It is often difficult for a simulation to correlate virtual patient verbal behavior or appearance with the state of the physiology engine. Because manikins are constructed mostly out of rigid plastic and few are motorized, they have little capability to change their general appearance. Efforts to show changes in patient appearance in Virtual Reality (VR) simulations based on physiological parameters have been attempted and are maturing (Fig. 1).

Physiology engines are often present in high-end medical simulations. Higher end manikins such as the METI iStan and Laerdal SimMan 3G use physiology models that focus on the respiratory and cardiovascular systems. With these systems, changes in pulse, blood pressure, and respiratory rate are concretely accessible from the physical examination of the manikin as well as on a monitoring display. These engines will respond appropriately to artificial ventilation, chest compressions, and cardiac medications, for example. The high interactivity and close linkage of the physiological response closely replicates an actual critical care encounter. A shortcoming of manikins is the limited behavioral repertoire and the presence of monitoring displays that are often watched over by learners who neglect the physical examination. Manikin features rated by medical students to be most useful include chest rise, palpable pulses, interactive voice, and the vital signs display.

Anesthesia simulation is a common use of physiology engines and it is conducted with manikins, virtual patient avatars, or with a simulated patient monitor (Fig. 2). The most sophisticated manikins can simulate gas exchange on real anesthesia equipment. Because anesthesia training is heavily biased toward pharmacological effects and subtle trends, physiology engines are ideal for this application.

Physiology engines are also strongly suited to detailed exploratory activities, especially for advanced learners. They are unique in that they permit repetitive explorations while attempting different approaches. This empowers the learner to discover relationships between interventions experimentally. Because of the expense and formidable effort to create physiology engines, the Telemedicine and Advanced Technology Research Center (TATRC), Armed Forces Simulation Institute for Medicine (AFSIM), and the Defense Medical Research Program are sponsoring the creation of an open-source physiology engine as a public resource for all to use freely. It is hoped that the public physiology engine will increase the adoption of this technology and the available corpus of high-fidelity medical simulation content.

COMPLEX FIDELITY: KINETIC MODELS

A different approach to simulated physiology is the SIMapse Nerve Agent Laboratory. SIMapse visually shows cholinergic neurotransmission, the effects of nerve agents, and actions of various nerve agent antidotes on different body systems. The approach of SIMapse is less mathematical and more kinematic; neurotransmitter molecules are graphically depicted in three-dimensional (3D) space and interact with receptors, destructive enzymes, and other agents. The physiological process is graphically showed as actions and behaviors. The results are a very close approximation to known science even though the simulation does not use actual

FIGURE 1. Patient with maxillofacial trauma shows the graphical quality of typical high-end VR training systems. Newer models in development can separate out physiological data such as pulse, respiration, temperature, and capillary refill and tie it to relevant animations for pallor, flushing, sweating, anxiety, and distress. Image courtesy of TruSim, a division of Blitz Game Studios.

FIGURE 2. A VR high-fidelity anesthesia trainer uses a physiology engine for precise determination of medication effects. Physiology engines permit combinations of user inputs that may be unexpected by the developer yet still produce reliably lifelike results. HumanSim image courtesy of the Virtual Heroes division of Applied Research Associates.
physiology data, yet it does deliver a high-impact learning experience (Fig. 3).12

MODERATE FIDELITY: COMPLEX STATE MACHINES

In practice, physiology engines are neither necessary nor desirable in many educational situations because simpler methods for depicting physiology states are often more practical. Moderate fidelity approaches to depicting physiology are often managed by complex state machines (CSMs) also known as hierarchical finite state machines.13 CSMs are computer programs consisting of logical rules and decision-tree-based logic that responds to user activity. User actions, simulation timers, and other events trigger different states that change the patient presentation and vital signs.

CSMs are designed for scenario-based training and respond to well-defined, known possible interactions. A major advantage of state machines is that expression of different states is often a marked change that is readily noticed by the learner. It is also easy to indicate changes in patient behavior, vital signs, or appearance because these changes can be concretely tied to a change in the state machine. The CSM approach is ideally suited for training scenarios with limited depth and scope. Simulations appear more responsive to the learner because they provide immediate and visible responses to learner interaction.

The major disadvantage of state machines is that they do not respond well to unexpected, complex, or combinatorial inputs. Undesirable inputs and program complexity are addressed by limiting the variety of possible interventions. Recovery from user errors once a scenario has moved down a decision-tree is difficult to program. Another disadvantage is that each branch point must be individually coded. Adding additional branch points or levels to the state machine can exponentially increase the development work necessary to build the scenario.14 Medium and high-end manikins often feature this type of state machine and may be packaged with a scenario builder toolkit. In the practice

FIGURE 3. The SIMapse Nerve Agent Laboratory v3 provides a realistic portrayal of nerve agent pharmacology through motion and sound for educational purposes. The program simulates physiological behavior without referring to actual physiological data. It teaches accurate physiology by showing physiology mechanisms, relationships, and pharmacology behavior. Trends over time are depicted as colored sparklines at the bottom of the display.
of visiting dozens of simulation centers in the United States, it is the author’s observation that few centers go through the effort to create custom scenarios with a toolkit. Instead, they tend to rely on scenarios provided by the manufacturer. VR scenario and game-based training often use the CSM model.

Another medium-fidelity approach is to couple a state machine to some sort of physiology model. When the U.S. Army Field Management of Chemical and Biological Casualties Course required a nerve agent pharmacology trainer targeting less advanced learners than would be appropriate for the SIMapse Nerve Agent Laboratory, the SIMapse engine was adapted for a multimedia application called Nerve Academy.15 Nerve Academy uses an on-screen lecturer who delivers mini-lessons coupled with preprogrammed use of the SIMapse engine. It also includes numerous interactive activities at the end of mini-lessons that trigger events in the engine. The result is a responsive learning experience with high fidelity that is very easy for the learner to use.

The advantage of combining a limited set of inputs or a state machine to a physiology model is that it is possible to have high biological fidelity and visible trends in data although allowing for simplicity of use and well-defined, responsive changes in the appearance of the model. Another advantage to coupling a limited interaction set to a physiology model is that doing so greatly simplifies the complexity and sophistication required of the model by limiting the parameters the model has to account for.

LOW FIDELITY: SIMPLE STATE MACHINES
Most educational scenarios and simulations in medicine use low physiological fidelity to good effect. Low-fidelity approaches require less technology, effort, and sophistication to pull off and, therefore, are easier to author.

Patient simulations can use simple state machines (SSMs) to great effect. SSMs consist of 3 or more fixed states that alter the appearance, communication, and physiology data of the simulated patient. They lack branching and conditional...

FIGURE 4. The Cyanide Exposure Simulator is an SSM consisting of 7 frames. It uses simple animation effects to create the impression of live physiology. Numerous physiology displays are hand-drawn graphics that are progressively revealed, offering the illusion of real-time monitoring. Except for button controls, the simulation required absolutely no programming. This example shows that low-fidelity approaches can convincingly convey a dynamic process to learners.
features of the more complex CSMs yet still offer many of their benefits. They are especially useful in adding a dynamic appearance to a simple case presentation when creation of an in-depth simulation is too resource intensive. They can be created with simple technology, such as web pages, interactive animations or PowerPoint (Microsoft Corporation, Redmond, Washington).

An example familiar to the author is a simple animation-based application called the “Cyanide Exposure Simulator.”15 The simulation is usually deployed during large classroom sessions. It consists of 7 possible states that show an inhalational exposure to cyanide and the effects on human physiology. States are selected with two buttons to traverse back and forth through a timeline. Each step down the timeline plays a 10- to 15-second video clip, changes signs, and alters graphs (Fig. 4). The overall effect is that of a dynamic simulation that seems to portray a wealth of biological data. The video with the actor conveys the clinical picture. The rolling displays for the pneumograph, Electrocardiography, and Electroencephalography are fixed lines that are progressively uncovered, producing the illusion of live monitoring. Vital signs of respiratory rate, heart rate, cardiac output, and blood pH are represented by triangle indicators on linear gauges. The transition between states is a simple animation that moves the indicators over 3 seconds. The impression of this gradual transition is that of physiology that is changing with the timeline. In truth, the vital signs presented are generic as everything else. The cyanide simulator effectively portrays cyanide effects even though no detailed physiology data are ever conveyed to the learner.

The simulation is usually deployed during large classroom sessions. Each step down the timeline plays a 10- to 15-second video clip, changes signs, and alters graphs (Fig. 4). The overall effect is that of a dynamic simulation that seems to portray a wealth of biological data. The video with the actor conveys the clinical picture. The rolling displays for the pneumograph, Electrocardiography, and Electroencephalography are fixed lines that are progressively uncovered, producing the illusion of live monitoring. Vital signs of respiratory rate, heart rate, cardiac output, and blood pH are represented by triangle indicators on linear gauges. The transition between states is a simple animation that moves the indicators over 3 seconds. The impression of this gradual transition is that of physiology that is changing with the timeline. In truth, the vital signs presented are generic as everything else. The cyanide simulator effectively portrays cyanide effects even though no detailed physiology data are ever conveyed to the learners. In fact, providing such level of fidelity that includes more detail would diminish educational effectiveness.16 Based on user feedback as a developer, the author finds that most learners are unlikely to detect a difference between this simple state model and an expensive version that uses a physiology engine if the physiology changes seem logical.

SSMs can also be used for intervention-based training. Virtual Nerve Agent Casualty (VNAC) pushes the envelope for a simple state model to simulate treatment of a severe nerve agent exposure.17 Video of a patient is played although the learner interacts by dropping antidotes and interventions onto the patient display. The state model plays a new video with each intervention and counts up correct interventions until a required quantity is reached. At this point, the state is changed that alters data on the screen (Table I). The simulation requires persistence on part of the learner but results in a satisfactory simulated experience despite the fact that the biological fidelity is low. A drawback of this simple application is that it is very unforgiving of learner errors and does not allow for a corrective pathway if treatment errors are made.

**LOW FIDELITY: GAME-BASED APPROACHES**

A game-based approach to depicting physiology is the health score. Health scores are simple numeric or bar graph representations of health. They consist of a point-based score or 100-point scale and are often called hit points. This very simple and low-fidelity representation of health status is ubiquitous and is readily understood by game players. In games, the player’s avatar loses hit points upon receiving damage. Losing all hit points results in death of the player. Hit points tend to gradually regenerate or are restored by activating plus-ups.

Medical simulations can also exploit health scores. One approach to health scores uses trend zones. For example, one can split the range of hit points with low, stable, and high zones (Fig. 5). High health scores will improve until health is full. Scores in the stable zone will improve very slowly and scores in the low health zone will automatically decrease. Simply setting the health score based on a virtual patient’s disease or improvement after intervention will now be followed by an automatic, ongoing action. This action can be

### TABLE I

<table>
<thead>
<tr>
<th>Conditions</th>
<th>State 2</th>
<th>State 3</th>
<th>State 6</th>
<th>State 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial State</td>
<td>Activating Conditions</td>
<td>Activating Conditions</td>
<td>Activating Conditions</td>
<td>Activating Conditions</td>
</tr>
<tr>
<td>Default</td>
<td>3 Atropine</td>
<td>4 Atropine</td>
<td>15 Atropine</td>
<td>5-Minute Delay in Treatment</td>
</tr>
<tr>
<td>Vitals:</td>
<td>3 Oxime</td>
<td>3 Oxime</td>
<td>4 Oxime</td>
<td>Seizures Not Controlled in</td>
</tr>
<tr>
<td>Heart Rate 140</td>
<td>1 Diazepam</td>
<td>3 Diazepam</td>
<td>4 Diazepam</td>
<td>10 Minutes, An Error is Made</td>
</tr>
<tr>
<td>Respiration: 30</td>
<td>Vitals:</td>
<td>Vitals:</td>
<td>Vitals:</td>
<td>Vitals:</td>
</tr>
<tr>
<td>Blood Pressure: 140</td>
<td>Heart Rate 120</td>
<td>Heart Rate 130</td>
<td>Heart Rate 160</td>
<td>Dead</td>
</tr>
<tr>
<td>40 Systolic</td>
<td>Respiration: 24</td>
<td>Respiration: 20</td>
<td>Respiration: 13</td>
<td>“You just didn’t treat her</td>
</tr>
<tr>
<td>Secretion 4+</td>
<td>Blood Pressure: 60</td>
<td>Blood Pressure: 100</td>
<td>Blood Pressure: 140</td>
<td>well enough”</td>
</tr>
<tr>
<td>Bronchospasm 4+</td>
<td>systolic</td>
<td>100 systolic</td>
<td>140 systolic</td>
<td></td>
</tr>
<tr>
<td>Twitch +</td>
<td>Secretion 3+</td>
<td>Secretion 2+</td>
<td>Secretion—</td>
<td></td>
</tr>
<tr>
<td>Seizure +</td>
<td>Bronchospasm 3+</td>
<td>Bronchospasm 2+</td>
<td>Bronchospasm—</td>
<td></td>
</tr>
<tr>
<td>“The patient is twitching and having difficulty breathing…”</td>
<td>Twitch +</td>
<td>Twitch —</td>
<td>Twitch—</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Seizure +</td>
<td>Seizure —</td>
<td>Seizure—</td>
<td></td>
</tr>
</tbody>
</table>

**ANNOUNCEMENT**

An in-depth simulation is too resource intensive. They can be created with simple technology, such as web pages, interactive animations or PowerPoint (Microsoft Corporation, Redmond, Washington).

The simulation is usually deployed during large classroom sessions. Each step down the timeline plays a 10- to 15-second video clip, changes signs, and alters graphs (Fig. 4). The overall effect is that of a dynamic simulation that seems to portray a wealth of biological data. The video with the actor conveys the clinical picture. The rolling displays for the pneumograph, Electrocardiography, and Electroencephalography are fixed lines that are progressively uncovered, producing the illusion of live monitoring. Vital signs of respiratory rate, heart rate, cardiac output, and blood pH are represented by triangle indicators on linear gauges. The transition between states is a simple animation that moves the indicators over 3 seconds. The impression of this gradual transition is that of physiology that is changing with the timeline. In truth, the vital signs presented are generic as everything else. The cyanide simulator effectively portrays cyanide effects even though no detailed physiology data are ever conveyed to the learners. In fact, providing such level of fidelity that includes more detail would diminish educational effectiveness. Based on user feedback as a developer, the author finds that most learners are unlikely to detect a difference between this simple state model and an expensive version that uses a physiology engine if the physiology changes seem logical.

**LOW FIDELITY: GAME-BASED APPROACHES**

A game-based approach to depicting physiology is the health score. Health scores are simple numeric or bar graph representations of health. They consist of a point-based score or 100-point scale and are often called hit points. This very simple and low-fidelity representation of health status is ubiquitous and is readily understood by game players. In games, the player’s avatar loses hit points upon receiving damage. Losing all hit points results in death of the player. Hit points tend to gradually regenerate or are restored by activating plus-ups.

Medical simulations can also exploit health scores. One approach to health scores uses trend zones. For example, one can split the range of hit points with low, stable, and high zones (Fig. 5). High health scores will improve until health is full. Scores in the stable zone will improve very slowly and scores in the low health zone will automatically decrease. Simply setting the health score based on a virtual patient’s disease or improvement after intervention will now be followed by an automatic, ongoing action. This action can be

**TABLE I.** Simple State Model of Virtual Nerve Agent Casualty (VNAC). (States 4 and 5 are Not Depicted for Clarity.) VNAC is a Product of the U.S. Army Medical Research Institute of Chemical Defense Chemical Casualty Care Division

<table>
<thead>
<tr>
<th>Conditions</th>
<th>State 2</th>
<th>State 3</th>
<th>State 6</th>
<th>State 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial State</td>
<td>Activating Conditions</td>
<td>Activating Conditions</td>
<td>Activating Conditions</td>
<td>Activating Conditions</td>
</tr>
<tr>
<td>Default</td>
<td>3 Atropine</td>
<td>4 Atropine</td>
<td>15 Atropine</td>
<td>5-Minute Delay in Treatment</td>
</tr>
<tr>
<td>Vitals:</td>
<td>3 Oxime</td>
<td>3 Oxime</td>
<td>4 Oxime</td>
<td>Seizures Not Controlled in</td>
</tr>
<tr>
<td>Heart Rate 140</td>
<td>1 Diazepam</td>
<td>3 Diazepam</td>
<td>4 Diazepam</td>
<td>10 Minutes, An Error is Made</td>
</tr>
<tr>
<td>Respiration: 30</td>
<td>Vitals:</td>
<td>Vitals:</td>
<td>Vitals:</td>
<td>Vitals:</td>
</tr>
<tr>
<td>Blood Pressure: 140</td>
<td>Heart Rate 120</td>
<td>Heart Rate 130</td>
<td>Heart Rate 160</td>
<td>Dead</td>
</tr>
<tr>
<td>40 Systolic</td>
<td>Respiration: 24</td>
<td>Respiration: 20</td>
<td>Respiration: 13</td>
<td>“You just didn’t treat her</td>
</tr>
<tr>
<td>Secretion 4+</td>
<td>Blood Pressure: 60</td>
<td>Blood Pressure: 100</td>
<td>Blood Pressure: 140</td>
<td>well enough”</td>
</tr>
<tr>
<td>Bronchospasm 4+</td>
<td>systolic</td>
<td>100 systolic</td>
<td>140 systolic</td>
<td></td>
</tr>
<tr>
<td>Twitch +</td>
<td>Secretion 3+</td>
<td>Secretion 2+</td>
<td>Secretion—</td>
<td></td>
</tr>
<tr>
<td>Seizure +</td>
<td>Bronchospasm 3+</td>
<td>Bronchospasm 2+</td>
<td>Bronchospasm—</td>
<td></td>
</tr>
<tr>
<td>“The patient is twitching and having difficulty breathing…”</td>
<td>Twitch +</td>
<td>Twitch —</td>
<td>Twitch—</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Seizure +</td>
<td>Seizure —</td>
<td>Seizure—</td>
<td></td>
</tr>
</tbody>
</table>
enhanced further by adding nonlinear response curves.\textsuperscript{18} With trend zones, the score will always be moving, encouraging the learner to intervene. This approach requires very little development effort.

A more sophisticated approach from the gaming world uses score modifiers. Score modifiers come in four basic forms: instant damage, instant healing, damage over time (DoT), and healing over time (HoT). Instant modifiers are a one-time reduction or improvement to the health score. DoT is a continuous score reduction over a specified amount of time. In the example of a bleeding virtual patient, the health score will progressively decline on the screen as bleeding is a DoT effect. If the learner gives the patient a blood transfusion, perhaps some points will be added to the health score, but the score will continue to decline. If the learner intervenes and stops the bleeding (i.e., tourniquet), then the DoT effect is cancelled out. Applying a blood transfusion (instant healing) will now permanently increase the score because the DoT effect is no longer present. Careful application of different score modifiers can mimic responsive and sophisticated physiology. The drawback is that health is being represented by a single moving bar; such a display typically does not show historical trends. Learners who grew up with video games are accustomed to this health scores even though it is not widely used for medical simulations. Further research is needed to assess the impact of health scores on learner perception.

TATRC is currently working with BreakAway to develop a multiplayer online hospital-based mass casualty simulation for coordinating a response to a very large number of casualties (Communication with Jennifer McNamara, CBRNE-GAME Leader). The clinical role in the game involves triage and treatment for hundreds of casualties. The patients will use a blended model, including a health score, score modifiers, and a simple state model with only three descriptions of patient presentation. Use of these shortcuts allows for patients that change noticeable over the course of the exercise and act responsively to player intervention. The learner will be busy prioritizing and selecting treatments and will likely not take notice that the fidelity is shallow. In fact, the simulation depends on the fact that fidelity is shallow so learners focus on “the big picture.” The question of how the learner perceives fidelity in the simulation will be the subject of game test research.

**LOW FIDELITY: STATIC PRESENTATIONS**
The most common method of depicting physiology is through a static presentation. Static presentations state vital signs and provide a case in written or verbal form. Static presentations can have as little or as much detail as desired with little effort required on the part of the author. This format is commonly used in written tests, magazines, multimedia, and web pages to effect and is well known to medical learners. While the advantages involve ease of authorship and distribution, the disadvantage is the lack of interaction. This format does a poor job of showing progression and the effects of intervention. Static presentations have strictly right and wrong answers without consideration of creative possibilities. This format is the most common type of educational patient case.

**PHYSIOLOGICAL FIDELITY: APPLICATIONS**
Each approach to depicting physiology has its own best use (Table II). Advanced learner simulations and exploratory activities are well suited to physiology engines. Interactive case scenarios can be conducted with a physiology engine, but CSMs and SSMs are usually preferred because of the relative simplicity of development and lower computing resource requirements. Because state machines consist of a few dozen calculations to perform versus thousands for a physiology engine, state machines require fewer computing resources to run. Although computers are always becoming more powerful, the resource requirements to run a physiology engine in real time become significant when attempting to run 100 or 1,000 simultaneous instances within a serious game. Another area where the logical simplicity of state machines is preferable includes low power and low computing resource environments such as mobile devices and tablets. Game-based training is well suited to CSMs and health scores. Presentations and mini-activities with few options are well suited to SSMs. Case studies are most easily written as a static presentation, though adding a state machine can increase the interactive possibilities. The available number of quality medical training scenarios is limited by the effort required to develop them. Fortunately, the majority of medical scenarios do not require high physiological fidelity. Clever developers use a number of tricks to create the impression of physiological fidelity. These techniques include visible responsiveness to user input and use of animation.

**ANATOMICAL FIDELITY**
Medical simulations use anatomy in both virtual and physical environments. Anatomic features are most important for training invasive procedures though they are also useful in practicing diagnostic skills. Various anatomical approaches are identified here along with their strengths and weaknesses.

Live models such as human standardized patients have extremely good fidelity and are useful for physical diagnosis.
skills. Major limitations include the limited availability of pathologies that are stable enough to use this technique. The expense of hiring standardized patients is enormous and humans are not suitable for rehearsing invasive procedures.19 Cadavers offer many advantages for anatomical realism and some can even be partially reanimated if fresh, but they are expensive, limited in their availability and administratively burdensome to possess.20

Live animals have their own limitations. They can be used for training purposes for both examination and invasive procedure training, though there are important ethical restrictions and precautions involved in their use. Their superior tissue properties are not yet achievable by artificial means. They can be operated on and do all the physical things that humans do. They do not require a team of developers to bestow their capabilities. Differences between the animal model and true human anatomy are a disadvantage to this approach.21

Task trainers are simulators that consist of discrete body parts or regions for training a specific task. Task trainers have been developed for intravenous access, central line placements, intravascular access, lumbar puncture, colonscopy, tracheal intubation, and many other procedures. They allow for repetitive practice of physical procedures at low cost. They are usually constructed of molded plastic and rubber-like compounds. They tend to have suitable anatomic landmarks but can be lacking in suitable tissue properties.22 Ideal tissues will feel like flesh and bone and include appropriate compliance, texture, moisture, bleed and traction properties. Most task trainers fall far short of this ideal. Crude materials also limit fidelity for training for physical diagnostic skills such as palpation of blood vessels and organs.23 Task trainers also require maintenance and replacement of consumable parts. Despite limitations, even unsophisticated task trainers have been proven effective in medical training.24

Manikins known to the author fall short on anatomical fidelity because they lack internal anatomy such as tissues, muscles, bones, organs, and vessels. They are accurate at a gross level and may include palpable pulses, chest rise, palpable ribs, and other features. Current manikins do not articulate naturally, lack muscle tone, and cannot be operated on. The skin is rubber-like or plastic and they lack realistic tissue properties. Nevertheless, they are useful in many training situations and numerous procedures can be rehearsed on them. Current manikins do tend to have sophisticated airways. They are often best for scenario-based training and demonstration of decision making and proficiency for a sequence of events than specific skill rehearsal that is often better performed with a dedicated task trainer. Depending on the use, even very basic models may be suited to medical training if the scenarios are designed properly.

A less well-known, but growing approach is the high-fidelity physical model. Physical models are task trainers for surgical intervention. They combine soft plastics, gelatinous tissue substitutes, composites, and cloth to represent the internal anatomy of an organ or wound and are often moulaged to represent a specific pathology. These often disposable trainers include
simulated skin, bone, nerves, muscles, and tissue planes. Common physical models include limbs, the inguinal canal, and other areas of the body. They allow for surgical rehearsal at low cost and are very practical technology.

VR systems use a 3D human representation on a computer display. For surgical purposes, they are viewed stereoscopically on a 3D screen or with display goggles. They usually include realistic appearing surgical manipulators or a 3Dpen-like device with haptic feedback. Enormous advances in graphical computing power now allow for realistic 3D portrayals of both external and internal human anatomy with detailed surface textures and appearance. These recent graphical advances are so dramatic that VR simulations from even a few years ago now appear antiquated. Graphical realism is no longer limited by the computer. It is limited the effort and expense of creating the 3D models and artwork. The technology behind the haptics (sensation of feeling), although awkward to get used to, can provide uncanny sensations of moving tissue, a scalpel, or hard points underneath the skin.

VR surgical systems are extremely difficult to develop material for and remain very expensive. Because of this, the available library of procedures and scenarios using this technology remains small. Worse yet, content developed on one system is not transferable to that of another manufacturer. The AF SIM and the National Capitol Area Simulation Center are attempting to ameliorate this through Tri-Service Open Platform for simulation (TOPS) by specifying a common interface between software and hardware systems (Communication with TOPS Principal Investigator, Alan Liu). AF SIM is also working to develop an open-source haptic-enabled surgery toolkit to promote additional content development and advancement of this technology.

INTERACTIVITY AND NARRATIVE
Fancy graphics, accurate anatomy, and physiological fidelity are not the only, nor the most important features of a successful medical simulation. The quality of a simulation-based training experience depends on successful engagement with the learner. Achieving engagement depends on a sense of immersion, successfully executing visuals, responsiveness, and a good narrative. Creating the sense of immersion is more important than 3D or the level of visual detail in the simulation. It depends on having a consistent simulation world with things to do or see that are interesting to the learner. Simulations can achieve engagement with the learner more successfully if actions they perform in the scenario are followed by a visible or audible response. Responsiveness connects the learner to the scenario. In the case where virtual humans are encountered, responsiveness establishes likeability and rapport with the learner. Nonverbal cues and gestures, even random ones, increase this sensation of rapport with the scenario and virtual patient. For these reasons, factors such as “response to user input,” “perception of biological fidelity,” and other factors are listed in Table II, which is a comparison of approaches to virtual patient physiology. This table is based on the author’s experience as a developer and interactions with hundreds of learners using various medical simulations. In addition to design choices such as the level of fidelity required and the amount of development effort one wishes to expend, one should not neglect the impact of old-fashioned storytelling; research strongly shows that narrative is a successful tool to engage people.

CONCLUSION
When it comes to methods to depict physiology in a medical simulation, some generalizations can be made (Fig. 2). Physiology engines excel at advanced simulations and exploratory learning where sophisticated learners want to try unexpected things and see accurate responses. Examples of this include anesthesia trainers and toxicology simulations. Complex state models are useful for interactive case scenarios and game-based training because they provide predictable behavior, excellent responsiveness, and can be reasonably complex while appearing to have high fidelity. Simple state models are well suited to interactive case scenarios, lecture presentations, and mini-activities because they are very easy to author yet are interactive and responsive to user input. Health scores are ubiquitously used in entertainment titles but not so much in medical simulations. They have the advantage of easy implementation, being dynamic and that they are a familiar format. Further research is needed to determine their suitability and acceptance of health scores in medical simulations. In fact, the author recommends that those intending to create medical simulations use the simplest technology possible that achieves the learning objectives. Either way, further research is needed to obtain data regarding learner perceptions of simulations that use these technological approaches for physiology as the extant literature is sparse compared to research on graphical realism, immersion, and narrative.

Biological fidelity in medical simulation is not an end in and of itself. Physiology, anatomy, interaction, narrative, and the technology behind them represent an array of tools. The choice of these tools must be determined by educational objectives. The fact that more complex systems require more time, effort, and money to create means that unnecessary use of high fidelity results in less available training content overall. Although some applications rightly require exacting fidelity, many do not. Medical educators should choose the most appropriate level of technology that achieves the goal. When doing so, they may often find that the level of fidelity required is lower than what they initially expected.

ACKNOWLEDGMENT
This study was partially supported by a grant from the Office of Naval Research, Award Number N00014-10-1-0978.
REFERENCES


ABSTRACT    This article reviews simulation used for medical training, techniques for assessing simulation-based training, and cost analyses that can be included in such assessments. Simulation in medical training appears to take four general forms: human actors who are taught to simulate illnesses and ailments in standardized ways; virtual patients who are generally presented via computer-controlled, multimedia displays; full-body manikins that simulate patients using electronic sensors, responders, and controls; and part-task anatomical simulations of various body parts and systems. Techniques for assessing costs include benefit–cost analysis, return on investment, and cost-effectiveness analysis. Techniques for assessing the effectiveness of simulation-based medical training include the use of transfer effectiveness ratios and incremental transfer effectiveness ratios to measure transfer of knowledge and skill provided by simulation to the performance of medical procedures. Assessment of costs and simulation effectiveness can be combined with measures of transfer using techniques such as isoperformance analysis to identify ways of minimizing costs without reducing performance effectiveness or maximizing performance without increasing costs. In sum, economic analysis must be considered in training assessments if training budgets are to compete successfully with other requirements for funding.

INTRODUCTION    The advantages of using simulation in training may be summarized as follows:

—Safety: Simulated lives and health can be jeopardized to any extent required for learning.
—Economy: Simulated materiel, equipment, and other resources—physical or fiduciary—can be used, misused, and expended as needed.
—Visibility: Simulation can provide visibility in at least two ways. It can (1) make the invisible visible and (2) control the visibility of details allowing the learner to discern the forest from the trees or the trees from the forest as needed.
—Time control: Simulated time can be sped up, slowed down, or stopped. It can also be completely reversed, allowing learners to replicate specific problems, events, or operational environments as often as needed.

These advantages seem to be applicable in medical training and education as elsewhere. Overall, simulation can provide massive amounts of practice with feedback, exposing individuals or teams to realistic situations that in real-world settings would range from the impracticable to the unthinkable.

All these advantages are relevant and interrelated. This article focuses on the economic value of simulation. It suggests ways to assess the use of simulation in medical education and training through objective economic and cost analyses. The article begins with brief reviews of simulation in medical education and training, and economic and cost analyses. These reviews are followed by a discussion of ways in which measures of simulation training effectiveness can be combined with cost analysis to yield assessments of costs, cost-effectiveness, and return on investment in medical training and Education.

SIMULATION IN MEDICAL TRAINING AND EDUCATION    As abstracted from reviews and comments (e.g., Bradley and Rosen), at least 4 forms of simulation appear to be used in medical training.

Standardized Patients    Applying formal procedures developed in 1964 by Barrows and Abrahamson and continued into the present, actors, real patients, or lay people can be trained as "standardized patients" to participate in role-playing exercises for assessing and improving a learner’s ability to carry out medical procedures, such as taking medical histories, performing physical examinations, ordering tests, providing counseling, and prescribing treatment. These patients can be available when and where they are needed, trained to respond consistently to examination questions, and used when training with a real patient would be inappropriate, as in counseling cancer patients. They are, however, expensive to recruit and train and cannot present cues normally provided by physiological examinations.

Examples:

—Gerner et al found that consultation quality ratings by standardized patients after their visits with 67 general practitioners predicted later ratings by parents concerning improvements in the weight control behavior of their children. 95% of the general practitioners reported that they found training with standardized patients to be useful.

*Science and Technology Division, Institute for Defense Analyses, 4850 Mark Center Dr., Alexandria, VA 22311.
†University at Albany, State University of New York, 1400 Washington Ave., Albany, NY 12222.

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

doi: 10.7205/MILMED-D-13-00258
A study by Betcher found that the use of standardized patients (graduate students in theater) in role-playing consultations, followed by debriefing, was effective in improving the communication skills and confidence of nurses and other caregivers by 5% to 37% in advising end-of-life patients and their families.

Safdieh et al. compared the long-term effects on quality of neurologic examinations performed by 58 medical students who were trained using standardized patients to those performed by 129 students who were trained without this experience. Two years after this training, the authors found that a statistically significant advantage in performing these examinations favored students who were trained using the standardized patients.

**Virtual (Computer-based) Patients**

Interactive software simulations of patients have been used in standard simulation exercises and in gaming simulations for training and assessing medical skills. These simulations are gradually taking the place of standardized patients, although the absence of strong artificial intelligence, which would allow full mixed-initiative dialogue, limits their applicability. However, growth in the use of virtual patients is likely to continue because of their ability to scale inexpensively to large numbers of physically dispersed learners, adapt quickly to prior knowledge and other individual characteristics of learners, and be available anytime and anywhere via the global information infrastructure.

Examples:

—Steadman et al. compared learning in a week-long acute care course by 31 fourth-year medical students who were randomly assigned to a group receiving a virtual patient with labored breathing versus a patient receiving interactive problem-based learning without simulation. The simulation group performed significantly better (71%—a 24-point gain over the pretest) on the final assessment than the problem-based learning group (51%—a 7-point gain over the pretest).

—Ten Eyck et al. used a crossover design involving 90 students to compare virtual patient simulation with group discussion in emergency medical instruction. The learners received one set of topics using one instructional treatment and then switched mid-rotation to receive the other set of topics using the other treatment. Material presented in simulation format produced significantly higher scores than material presented using group discussion methods.

—Botezatu et al. compared learning by 49 students studying hematology and cardiology topics using virtual patients with learning by students receiving more conventional instruction (lecture and small group discussion). They assessed learning immediately after instruction and its retention after 4 months. They found effect sizes ranging from 0.5 to 0.8 in favor of the virtual patients.

**Electronic Patients**

These types of patients are generally whole-body manikins that physically simulate patients, although some use of helmet-mounted virtual reality capability has also been used. Miliary and civilian simulation developers have collaborated to produce and assess electronic patients. An early effort by David Gaba (Stanford University) and CAE-Link developed a full-scale manikin system called Anesthesia Crisis Resource Management. It was designed for Air Force and National Aeronautics and Space Administration Crew Resource Management training. Examples:

—Alinier et al. compared the performance of intensive-care nursing students who were assigned at random to two groups. One group followed the usual course of training, and the other group used a “universal patient simulator.” The researchers found significantly greater performance improvement (twice the percentage increase) for the simulation-trained group than for the group receiving the usual course of training.

—Radhakrishnan et al. used electronically controlled manikins to compare the performance of manikin-trained and conventionally trained nursing students. They found that the manikin-trained students significantly outperformed the other students in patient safety and in assessing vital signs.

—Cendan and Johnson used a randomized, repeated-measures design to train 40 second-year medical students in treating neurogenic, hemorrhagic, septic, and cardiac shock. This study compared two instructional approaches such as web-based training text with a culminating simulation exercise and a manikin-based exercise with instructors who provided management and evaluation in response to student questions. All students were exposed to both approaches, with half completing the web-based exercises first and half completing the manikin-based exercises first. Learning from web-based and manikin-based instruction was similar; however, overall learning was greater when the web-based simulation was presented first.

**Part-Task Trainers**

Anatomical models of body parts are used to provide training. These “part-task” simulations are becoming more advanced to keep pace with medical treatments and technology. They are used in instruction that ranges from minimally invasive laparoscopy to major cardiologic surgery and delicate ophthalmological procedures. They have been enhanced considerably through the development and inclusion of haptic systems and interfaces.
Examples:
—Barsuk et al\(^{24}\) found significantly fewer needle passes, arterial punctures, and catheter adjustments and higher overall success rates among 76 simulator-trained residents using a haptic task trainer than among 27 residents trained without the simulation.
—Holzinger et al\(^{25}\) compared learning by 96 medical students who were randomly divided into 3 groups: a conventional text-based lesson group, a group learning from a blood dynamics simulator alone, and a group learning from the simulator blended with additional human instructor support. They found no difference between the first two conditions but significantly more learning in the third.

**ECONOMIC ANALYSIS: THE VALUE OF A POUND OF SIMULATION**

Administrative decision making largely consists of allocating resources among competing alternatives.\(^{26}\) Such decision making is not only a matter of adopting enhancements, but also includes determining what must be given up to do that. To some degree, this process is described by the “rational theory of choice,” which balances costs against factors that contribute to the achievement of a specific goal.\(^{27–29}\) This approach can be overdone by neglecting issues that are hard to measure (e.g., attitudes, culture, and trust), but it plays an essential role for cases in which benefits and costs can be clearly identified, such as business profit and loss, combat success or failure, and the achievement of objectives through instruction.

The objective of economic analysis is to inform decisions by assessing alternative courses of action and/or inaction. It does so by estimating the amount and probability of expected returns from each alternative and balancing these returns against projected costs, consequences, and constraints.\(^{30}\)

Economic analysts look for value to be gained and resources to be sacrificed for each alternative identified. Such deliberation assumes that the analyst or the decision maker has assembled and considered a comprehensive list of available alternatives. Introduction of an additional alternative can dramatically alter the decision space.

Economic analyses remain as subject to controversy as any other analyses. Underlying assumptions, inclusion and exclusion of data elements, proper data collection, sampling procedures, criterion levels, and similar issues contribute to controversy. An economic analysis can never be assuredly correct, but it can and should be explicit and should allow decision makers to determine how well and to what extent it informs their decisions.

The single most accessible and readily commensurable criterion for choosing among alternatives remains cost measured in fungible monetary units (e.g., dollars). In these cases, decisions key on units returned for units invested. Even with these data, the decisions do not make themselves. Other factors poorly suited to economic analysis also come into play. The effectiveness of military and medical training outweighs cost when human lives are at risk. Cost considerations—however necessary, objective, and well-conceived—should not be the sole concern in informing decisions.

**Assessing Costs**

Assessment of costs invested is a central factor in economic analysis. Costs may be categorized as one of the following: research and development, initial investment, operations and maintenance, and salvage and disposal.\(^{30}\) Salvage and disposal costs are omitted from many analyses because they are one-time only and difficult (usually impossible) to estimate accurately. Many research and development and initial investment costs are also one-time only, but they may be known and may be a matter of interest for some alternatives. When an alternative is being considered as a replacement for an existing program, both research and development costs and initial investment costs for the replacement can be included even though they are not included for the program in place. In these cases, analysts may decide that the costs for the current program are “sunk” (i.e., beyond recovery no matter what happens). These sunk costs do not factor into the decision.

A perennial and debilitating problem for cost analysis in education and training is the absence of generally accepted, standardized cost models. Such models would present unambiguously specified and well-defined cost elements that clearly identify what they do and do not include. Without these specifications, decision makers, among others, do not know clearly what the cost analysis and the cost analysts are telling them.

A variety of commentators have provided a basis for cost models to be used in instruction. For instance, Levin and McEwan\(^{31}\) suggested five classes of elements, or “ingredients,” to be considered in a cost model: personnel, facilities, equipment and materials, other program inputs, and client inputs.

Personnel costs include all the human resources needed by the approach. Levin and McEwan recommend that personnel be classified according to their roles (instructional, administration, clerical, and so forth), qualifications (training, experience, specialized skill), and time commitments (full time, part time). Facilities costs include all resources required to provide physical space for the approach. Equipment and materials include furnishings, instructional equipment, and supplies. Other inputs in this scheme include components that do not fit elsewhere (e.g., instructor training and insurance costs).

Other costs are especially relevant in military and industrial training, where student pay and allowances are funded by the same organization that provides the instruction, thereby increasing interest in the speed with which students reach objective thresholds of competency. Much of the rationale for applying technology in industrial and military training is keyed to its capacities for tutorial individualization, which allows the adjustments for prior learning and self-pacing to qualify...
students more quickly for duty or allows students to maximize their competencies—be all they can be—while holding instruction time constant.  

Kearsley developed a model much like Levin and McEwan’s but with an added dimension for the components, or categories, of instruction system development: analysis, design, development, implementation, and evaluation. These components can be combined with the typical cost categories of personnel, facilities, and equipment and materials. Integrating these two categories yields the cost framework shown in Figure 1, which presents an outline, not a fully developed cost model. Explicit discussion of what is included in, and/or excluded from, each cell of this framework will help analysts know what they are talking about and will help decision makers determine the extent to which an analysis can be applied to inform their decisions. It is rare, if not impossible, for a cost analysis in any area, including instruction, to be entirely correct. As discussed earlier, every such analysis requires assumptions and extrapolations, but it can and should be explicit. The framework shown in Figure 1 may contribute to this end.

Benefit–Cost Analysis

A benefit–cost analysis is used to determine whether the benefits returned by a candidate course of action outweigh the costs of investing in it. The calculation of a benefit-to-cost ratio is straightforward as described by Fitzpatrick et al.34 Phillips,35 and McDavid and Ingleson36 among others. It reduces all costs of an action to a single unit. It does the same for all benefits and then calculates the ratio of benefits to costs.

We can calculate a benefits-to-cost ratio using whatever metrics we choose, but the terms for input and output must be commensurable (i.e., both must be measured using the same units). Monetary units tend to be those most readily translated from whatever investment resources are required and whatever returns are produced. For that reason, these ratios are usually expressed in terms of dollars, pounds, euros, or whatever monetary unit communicates most easily and usefully to likely decision makers.

A benefit-to-cost ratio is calculated as follows (Phillips; McDavid and Ingleson):

\[
\frac{\text{Value of the result}}{\text{Cost of the investment}}
\]

It tells us how many units of value we get for every unit of cost.

For instance, Thompson reported that in 1667 public health officials in London found that expenditures to combat the plague would yield a benefit to cost ratio of 84:1.

Return on Investment

Return on investment is closely related to benefit-to-cost ratios. It is also a ratio, and calculating it is as straightforward as its name suggests. It is calculated as follows (Phillips; McDavid and Ingleson):

\[
\frac{\text{Value of the result} - \text{Cost of the investment}}{\text{Cost of the investment}}
\]

Return on investment must be calculated for some period of time, such as a year. As with monetary units, the length of time should be determined by analysts in consultation with decision makers who are likely to use the results of the analysis.

Example:

—Fletcher and Chatham studied returns from investing in several training innovations. They found ratios of 2.49 for the “Top Gun” investment in training Navy combat pilots, 3.37 for using technology-based, in-transit training to sustain and enhance the bombing skills of pilots, and 2.50 if technology-based training were used for 40% of Department of Defense specialized skill training.

Benefit–cost and return-on-investment analyses require value and cost to be commensurable. Of the two, return-on-investment analysis may be preferred because it indicates how many units of net benefits are returned, after investment.
costs have been subtracted, for each unit invested. Of course, spikes, dips, and diminishing returns have to be considered with differently timed units of investment, so averaging and curve smoothing may be required.

Return-on-investment analysis may be helpful for an ancillary reason. It treats costs for education and training explicitly as investments, not as infrastructure expenses. Treating these costs as infrastructure expenses is often their fate in training venues, including those of Department of Defense, where training is bundled with transit, hospitalization, and stockade costs.

Cost-effectiveness analysis

When commensurability is difficult, cost-effectiveness analysis can be used. Costs of investment can usually be expressed in monetary units, but the full return—the benefits—of instruction may not be amenable to monetary units. Cost-effectiveness analysis allows effectiveness (e.g., information retention, job knowledge and motivation of workers, supervisor ratings, and productivity) to be measured in its own units. In instruction, it accommodates a more comprehensive range-of-objective outcome than analyses requiring commensurability.

Cost-effectiveness is calculated as a direct ratio of cost to benefits or benefits to cost. In determining cost-effectiveness, the usual practice is to hold either costs or effectiveness constant across all alternatives being considered and observe variations in costs or effectiveness. Sometimes, either costs or effectiveness is simply assumed to be constant across the alternatives. One could argue that cost is implicitly assumed to be constant by its absence from many instructional evaluations. The assumption may be reasonable, but analysts should present data or information to validate it so that decision makers can decide for themselves if it is warranted.

The good news is that cost-effectiveness does not require commensurability. The bad news is that it is a relative term. Relevant decision alternatives must be specified in assessing it. The addition of an alternative for achieving the objective(s) after a cost-effectiveness analysis is done can change its conclusions and recommendations entirely. Despite common usage, we cannot properly say that an investment, by itself, is or is not cost-effective; however, no harm is done in calculating a cost-effectiveness ratio for it.

Example:

—Fletcher et al combined experimental data reported by Jamison et al and Levin et al with their own empirical findings to assess the costs of raising student scores on a standard test of mathematics comprehension by one standard deviation. They compared these costs for professional tutoring, peer tutoring, reducing class size, increasing instructional time, and using computer-based instruction. They found that the most cost-effective approaches among all these alternatives were peer tutoring and computer-based instruction.

Cost-effectiveness analyses have been used in health care since the 1960s to determine the relative value of specific interventions, such as a medication, surgical procedure, or counseling techniques.

Example:

—Tsai et al calculated cost-effectiveness ratios that compared hospital-based home care for patients who had mental illness with care based on traditional, outpatient therapy. They measured effectiveness in terms of disease maintenance behavior, psychotic symptoms, social function, and service satisfaction. Overall cost was the sum of costs for all direct mental health services. They found cost per unit of effectiveness to be $4.3 for home care and $13.5 for outpatient therapy.

One form of cost-effectiveness analysis is cost-utility analysis, where the return is assessed in terms of utility or value received by the beneficiaries of the investment. Cost-utility analysis is frequently recommended and promoted but rarely used in sectors other than health services, where decision makers often assess different quality-of-life alternatives for their patients. They must balance quality of life against additional years of life to help patients review the net benefit or utility provided by different treatments.

Assessment of Simulation

Decision making concerning potential improvements in training raises two basic questions. Compared to current practice, does it produce threshold levels of human performance capabilities at less cost, or does it increase human performance capabilities while holding costs constant? Both costs and effectiveness must be considered if assessments of training simulation are to inform decision making in a responsible manner.

Transfer effectiveness ratios

A key issue is the extent to which capabilities produced through simulation-based training transfer to “real-world” tasks. More specifically, does the human performance produced by simulation-based training either reduce costs without diminishing performance or improve performance without increasing cost? One approach to this issue is the use of transfer effectiveness ratios (TERs). TERs were developed by Roscoe and Williges for aircraft pilot training, but they apply to simulation-based training in general. A TER can be defined as follows:

\[
TER = \frac{T_c - T_x}{X},
\]

where TER is the transfer effectiveness ratio; \(T_c\) is the time or trials required for a control/baseline group to reach criterion performance; \(T_x\) is the time or trials required for an experimental group to reach criterion performance after \(X\) time or trials using simulation (or any other instructional approach
of interest); and \(X\) is the time or trials spent by the experimental group using the simulation.

Roughly, the TER indicates how many trials or units of time are needed to achieve criterion performance in the objective experience (e.g., flying an aircraft, repairing a radar repeater, and performing a medical procedure) are saved for every unit of simulation training invested.

Example:

—Taylor et al\(^{48}\) used TERs to compare times required to reach criterion performance in using specific aviation instruments with and without a Personal Computer Aviation Training Device (PCATD). One group was trained only during flight in the aircraft. A second group was trained first with the PCATD and later in the aircraft. Criterion performance was measured during flight. Taylor et al found that the PCATD group required about 4 hours less of in-flight training, suggesting a transfer effectiveness ratio of 0.15—or a savings of 1.5 flight hours for each 10 hours of PCATD time.

These findings suggest that the requisite levels of performance can be attained at lower cost using simulation—if the simulation costs less to operate than an airplane. If it does, then “the larger the TER, the better” is good news for the simulation.

Example:

—Orlansky et al\(^{49}\) compared the costs of flying military aircraft with the cost of operating (“fly”) simulators. They found that the cost of operating a flight simulator was about one-tenth the cost of operating military aircraft, so the use of a flight simulator was generally cost-effective if the TER for the simulator exceeded 0.10.

This finding is useful and significant. However, a few caveats are in order.

First, as Povenmire and Roscoe\(^{50}\) pointed out, not all simulation training hours are equal. Early trials or hours in a simulation may save more trials or time than later ones. A TER is likely to decrease monotonically and approach zero for large values of simulation training. This consideration leads to learning-curve differences between TERs and incremental TERs, or ITERs, with the inevitable diminishing returns captured best by the latter. An ITER can be defined as follows:

\[
ITER = \frac{T_{X-\Delta X} - T_X}{\Delta X},
\]

where ITER is the incremental transfer effectiveness ratio; \(T_{X-\Delta X}\) is the time or trials required to reach criterion performance with access to simulation after completing \(X-\Delta X\) units of time or trials; \(X\) is the time or trials spent by the experimental group using the simulation; \(T_X\) is the time or trials required to reach criterion performance, with access to simulation; and \(\Delta X\) are incremental units of time or trials after starting at unit \(X\).

Roughly, the ITER indicates the amount of transfer produced by successively greater increments of time or trials in the simulation. As Morrison and Holding\(^{51}\) pointed out, total time or trials to criterion begins to decrease as the use of effective simulation increases, but, sooner or later, they begin to increase. At some point, training total time or trials to criterion with simulation will exceed those without it and produce negative TERs.

Example:

—Taylor et al\(^{52}\) used ITERs to compare the number of trials to specific completion standards, time to complete a flight lesson, and time to a successful evaluation flight with and without a PCATD. One group trained only during flight in the aircraft, and three other groups trained first with the PCATD and later in the aircraft. Criterion performance was measured during flight. The number of trials to reach criterion was less for all three PCATD groups than for the aircraft-only group. The three experimental groups trained with the PCATD for 5, 10, and 15 hours, respectively. The 10-hour PCATD group required the fewest number of trials to reach criterion for five of the eight criterion tasks, the 5-hour PCATD group required the fewest number of trials for two of the criterion tasks, and the 15-hour PCATD group required the fewest number of trials on only one criterion task. Average ITERs were 0.662, 0.202, and 0.148, respectively, for the 5-hour, 10-hour, and 15-hour PCATD groups, indicating the diminishing returns from time in simulation training accounted for by ITERs.

Second, transfer effectiveness is tied to the specific skill, knowledge, or performance levels—the training objectives—being sought. This issue was illustrated in a study by Holman\(^{53}\) involving a CH-47 helicopter simulator. Holman found that if the knowledge and skills of interest were simply overall ability to fly the helicopter, the TER was 0.72. However, he also found that the 24 TERs for the specific skills he examined ranged from 2.8 to 0.0.

The TER that is relevant depends, as in all assessment, on the decision it is intended to inform, which includes the type of transfer sought. Holman required straightforward, “near” transfer, where many elements exercised by the simulator are similar, if not identical, to those required by the objective task performance. Near transfer echoes long-standing prescriptions for including identical elements that are shared by the learning (e.g., simulation) and the eventual task environments.\(^{54,55}\) Other applications may require “far” transfer from simulation to the objective task, where fewer elements are common to both simulation and task performance and higher level thought and analysis is required by the performer for transfer to occur.\(^{56}\)

Similarly, transfer may key on automated responses learned in simulation to the objective environment in a
straightforward “low-road” fashion. Alternatively, it may require less focus on automatic responses and greater abstraction of simulation performance to conceptual levels that transfer indirectly but broadly to many objective environments (i.e., “high-road” transfer, which requires purposeful attention in the learning environment to the development of learners’ transfer abilities57).

Given the physical and anatomical differences of human beings, both low- and high-road transfer seem particularly important in the development and assessment of medical training. These forms of transfer have received some attention in the development and assessment of simulations for medical training, but more may be in order.

Third, the operating costs of objective, targeted tasks differ markedly and can produce quite different tradeoffs in assessing the cost-effectiveness of simulation-based training. For instance, Povenmire and Roscoe50 considered flight simulation training for Piper Cherokee pilots, where the cost ratio of simulation to targeted performance was 0.73, thereby requiring a much higher TER for cost-effectiveness than Orlansky et al49 found for high-performance military aircraft.

Cost/effectiveness versus effectiveness/cost
TERs primarily concern ways to minimize costs while holding effectiveness constant. In this regard, the ratio between the fastest and slowest learners in typical classrooms appears to be at least 4:1.32 Learner ability remains a factor, but this ratio is most directly linked to prior knowledge. The variety and extent of prior knowledge increases with the age and experience of the learner, thereby making adjustments for it increasingly important for adult learners. Individualizing instruction by taking into account the knowledge and skill that each learner brings to training has been found to reduce time or trials to reach criterion levels of performance. Costs for specialized technical training in areas such as medicine might be reduced by as much as one-fourth if the capabilities currently available through computer technology were implemented to take advantage of these differences.38

On the other hand, maximizing effectiveness while holding costs constant may be more appropriate for military training. Personnel commands, which prepare orders to pass course graduates on to their next duty station, have found it prohibitively difficult to deal with individuals leaving training at arbitrary times. Fast learners who finish early are often detailed to necessary but undesirable duties and thereby have few incentives to save resources by shortening their time in training.

It appears to be more feasible and beneficial for military organizations to provide training that allows each learner to “be all they can be,” while holding graduation dates for all students constant. For instance, learners who have experience with a topic exercise in simulation might be presented a more difficult exercise on that topic to enhance their knowledge or skill while holding simulation time constant. This procedure could accommodate the needs of military personnel systems to synchronize the preparation of post-training orders and, through various personnel actions, provide incentives for learners to take full advantage of opportunities to train beyond threshold levels of performance—training that could best be made available using simulation.32,38

Cost savings under procedures to maximize performance while holding time constant have been shown to be considerable.38,58 Unfortunately, most of these savings are realized in duty commands and not in the training commands that must bear the costs of developing and providing the extra training for fast learners. These costs can be minimized using simulation, but, at present, local training commands have limited incentives to implement such procedures.

Further, return on investment appears to be relatively insensitive to development costs at military training scales.58 The Services could invest much more in the development of high-quality training and still receive strong monetary return on investment. The return to operational effectiveness for this investment is also likely to be substantial, but it is far more difficult to assess.

Isoperformance
TERs cover transfer issues, but we would like to cover costs along with transfer effectiveness in a single omnibus analysis so that allocations of training time or trials between, for example, simulation and “hands-on” exercises produce targeted levels of performance at minimal cost. Isoperformance provides one approach for solving this problem. The basic idea is to devise a function, usually depicted as an isoperformance curve, showing every point where different combinations of training inputs produce equivalent performance outputs.51,59,60 The solution, then, is to find the point on the curve where costs are minimized.

Isoperformance relates two or more training inputs to a training outcome held at some prescribed value or level. It is generally assumed that each input by itself could produce the desired-level outcome; however, some inputs may provide unique contributions to the outcome, necessitating their inclusion at least to some degree. Isoperformance identifies all combinations of the inputs needed to produce the objective performance.

Bickley61 pointed out that cost considerations in simulation-based training require at least 2 component functions. First, a function is needed to relate simulation trials or time to their costs. Second, a function is needed to relate costs to trials or time in simulation to performance on the “real” task or job.

The first consideration can usually be treated with a simple linear function to account for time or trials in simulation. The second consideration is more complicated. It is called an isoperformance curve because it trades off simulation time or trials with real task experience while holding performance on that task at some threshold level. It requires the analyst to specify a criterion level of performance and a level of confidence for achieving it. Given these considerations, the factors that determine criterion performance can be traded off against one another, as shown in Figure 2.
Performance—the output of the training—is expected to be the same everywhere on the total cost curve (the upper curve in Figure 2). Total costs initially decrease as simulation time or trials are substituted for those with the (presumably more expensive) real-world objective task or job. Costs then begin to increase as more and more simulation training is allocated and substituted in. Costs for nonsimulation training—initially, the middle curve in Figure 2—start in the same place as total costs but then decrease monotonically as more and more simulation training is substituted in. Notably, these costs rarely reach zero because sooner or later training will have to include time or trials in performing the objective real-world task or job. Costs for simulation—initially, the bottom curve in Figure 2—start at zero and rise monotonically with its increasing use.

Bickley recommended the following formulation for an isoperformance curve, which appears as the top curve in Figure 2:

\[ Y = ae^{-bx} + c \]

where \( Y \) is time or trials in the real task required to reach criterion performance; \( x \) is the time or trials in simulation; and \( a \), \( b \), and \( c \) are the parameters of the model.

Given a data set with reasonable variability in matching simulation time or trials to task proficiencies, the values for \( a \), \( b \), and \( c \) in this model can be calculated, and an appropriate isoperformance curve can be developed. Algorithms for doing so are available, as Bickley\textsuperscript{61} and de Weck and Jones\textsuperscript{60} point out.

The cost-effective solution under this formulation is given by the minimum on the upper, total cost curve. It can then be used to allocate training time or trials between simulation and the real equipment. In effect, it holds performance (or effectiveness) constant and suggests an allocation of inputs that minimizes costs. Carter and Trollip\textsuperscript{62} illustrated the other side of the coin. They used a mathematically equivalent approach to devise an optimal strategy for maximizing performance (or effectiveness again) given fixed costs.

The problem of collecting appropriate transfer data to use in TER or isoperformance analyses remains for some applications. Many of these analyses trade off simulation for training (e.g., aircraft piloting and tank gunnery) that otherwise would be exorbitantly expensive. Collecting adequate data to show all combinations of training inputs (e.g., simulation and aircraft piloting) that produce equivalent performance outputs can easily swamp a training developer’s budget.

Morrison and Holding\textsuperscript{51} suggested that a solution to this problem would be to use limited but valid empirical data accompanied by expert judgment to double-check findings and fill in gaps. They suggest pilot “dosage” experiments with no simulation training, a great deal of simulation training, and two to three different allocations of simulation training in between. Findings from such experiments could then be reviewed and supplemented by expert judgment to produce an approximate learning curve sufficient for either TER or isoperformance analysis. If times or trials to criterion in simulation are a matter of hours or days, if the training is for a critical task or job, and/or if the tasks to be learned are inexpensive relative to piloting military aircraft (as may be the case for many medical procedures), this approach seems reasonable and, in fact, prudent.

Morrison and Holding’s\textsuperscript{51} application of isoperformance analysis concentrated on gunnery training. The main idea

![FIGURE 2. Notional isoperformance curve drawn as a function of simulator and actual equipment costs.](image-url)
was to use simulation to save training ammunition. Other examples are available. For instance, Bickley61 focused on simulator versus flight time in the Army’s AH-1 helicopter. Jones and Kennedy59 discussed trading off personnel aptitude against training time. They also provide an appendix that shows step by step how to create an isoperformance curve. de Weck and Jones60 provide examples from spacecraft design and professional sports.

Isoperformance analysis can be applied to any trade-off issue, including the use of simulation in medical training. Basically, isoperformance curves are just cost curves. It may be time to invest more seriously in this approach.

SUMMARY AND DISCUSSION
Discussion in this article has ranged from generally applicable techniques (economic analysis) to those techniques specifically focused on simulation-based instruction (TERs and isoperformance). Issues of benefit cost, net benefit cost, and cost-effectiveness seem applicable in a straightforward fashion to any sort of medical training and education. However, commensurability is a problem: How are we to capture fully in monetary terms the value of a patient’s life, quality of life, and overall health? It is solved to an appreciable degree by cost-effectiveness analysis, provided that we identify a comprehensive set of realistic alternatives.

In contrast to cost-effectiveness, return on investment does not require the identification of all likely alternatives. Different returns from different investments in education and training can be compared later as they arise. However, return on investment focuses on investment costs, which may be unknown or sunk compared to existing alternatives. The result is that the research and development costs and initial investment costs of a new approach may need to be included. The new approach may then be at a disadvantage when considered and compared with return from existing approaches, where such costs are unknown, sunk, and omitted from the analysis.

Applications of TERs and isoperformance to provide economic analyses for simulation used in medical training and education seem both feasible and worthwhile if our analyses are to treat costs in training seriously. Adequate policies and procedures for the cost and effectiveness of training programs might be developed without the expenditure of time, effort, and cost required for optimization, but they will require generally accepted cost models with well-defined cost elements, including those associated with simulation for medical training. These approaches may earn their keep by advancing the field beyond guesswork and/or administrative fiat in the competitive allocation of increasingly scarce resources to medical education and training.

ACKNOWLEDGMENT
Funding for this article was provided by the Office of the Deputy Assistant Secretary of Defense (Readiness), Training Readiness, and Strategy Directorate.

REFERENCES
Cost Considerations in Using Simulations for Medical Training


Assessment Methodology for Computer-Based Instructional Simulations

Alan Koenig, PhD; Markus Iseli, PhD; Richard Wainess, PhD; John J. Lee, PhD

ABSTRACT Computer-based instructional simulations are becoming more and more ubiquitous, particularly in military and medical domains. As the technology that drives these simulations grows ever more sophisticated, the underlying pedagogical models for how instruction, assessment, and feedback are implemented within these systems must evolve accordingly. In this article, we review some of the existing educational approaches to medical simulations, and present pedagogical methodologies that have been used in the design and development of games and simulations at the University of California, Los Angeles, Center for Research on Evaluation, Standards, and Student Testing. In particular, we present a methodology for how automated assessments of computer-based simulations can be implemented using ontologies and Bayesian networks, and discuss their advantages and design considerations for pedagogical use.

INTRODUCTION
Medical simulations exist in many forms, from stand-alone multimedia computer programs to interactive tactile simulators, to high fidelity virtual reality experiences. They can be used to teach a range of skills from partial tasks to full procedures, and can be valuable in offering practice with patient interactions. Furthermore, simulations can be used to teach and test competencies not only related to procedures but also related to higher order thinking skills as well, such as problem solving and decision making.

However to be effective, medical simulations must be designed with careful attention paid to blending instructional and assessment design goals with appropriate types of interactivity and technological constraints. This article will explore some of the characteristics that have made medical simulations effective and will present a methodology developed by University of California, Los Angeles’, Center for Research on Evaluation, Standards, and Student Testing (CRESST), which highlights how instructional games and simulations can be designed and used to assess student performance.

MEDICAL SIMULATION DESIGN CRITERIA
Simulations can be designed for a variety of purposes, including individual instruction, team training, selection, and assessment. Despite this variability, most (if not all) medical simulation designs could benefit from deliberate consideration of the following: (1) goals of the simulation, (2) learning objectives, cognitive demands, and assessment, (3) affordances of the simulation, (4) the instructional strategies used, and (5) motivation considerations for users.

The goals of a simulation can range from general task goals (i.e., effectively treating a patient) to more specific goals (such as resource management of teams or equipment). From those, the knowledge, skills, and attitudes that a student should attain should be specified as learning objectives, and rooted in specific cognitive demands germane to their mastery. Objectives should include behaviors (what the student should be able to do), conditions (context in which the behaviors are shown), and standards (to what level of proficiency).

Simulation affordances address the actions and decisions students can perform/make inside the simulation and how well they map to the real world. It relates to task and cognitive fidelity, and involves determining the appropriate mechanisms or objects provided to students to facilitate their actions. Careful consideration must be paid to ensure that the affordances appropriately match the instructional strategies used.

The motivation for engagement with the simulation should be taken into account as well. While it is true that in many cases the student may be required to use the simulation, there is value in making simulations that are engaging, intuitive, and satisfying to keep students actively involved in the content.

SIMULATION FEATURES AND BEST PRACTICES
McGaghie et al in their meta-analyses of recent medical research from 2003 to 2009, enumerate twelve medical simulation features and best practices: (1) providing feedback, (2) deliberate practice, (3) curriculum integration, (4) outcome measurement, (5) simulation fidelity, (6) skill acquisition and maintenance, (7) mastery learning, (8) transfer to practice, (9) team training, (10) high-stakes testing, (11) instructor training, and (12) educational and professional context. Since describing each in detail is beyond the scope of this article, we briefly highlight three key ones, namely feedback, deliberate practice, and assessment.

Feedback was the most commonly studied feature of simulator training and considered the most important feature for...
student learning by Issenberg et al.\textsuperscript{15} One type of feedback is debriefing, also known as after-action review in military contexts. Debriefing has long been held as the cornerstone of simulator training.\textsuperscript{16} A critical piece of learning is not experience alone, but reflecting on the experience and rehearsing through simulating the scenarios, environments, etc., are particularly well suited to this.

Repetitive practice leads to faster automaticity and was considered key to transferring skills from simulators to real patients (43 studies).\textsuperscript{15} McGaghie et al\textsuperscript{17} further examined the role of practice in simulation-based medical education and found that it accounted for almost 50% of the variance in the average weighted effect size variable ($\eta^2 = 0.46$) from 32 studies, with the most benefit coming from the over 8.1 hours of practice category. Deliberate practice includes repetitive practice and informative feedback\textsuperscript{17} and is needed to reduce skill decay or degradation.\textsuperscript{18}

There are several strategies used in the assessment of simulations. They range from observations using checklists\textsuperscript{19,20} to very specific measurements taken by the simulation to more sophisticated statistical modeling. Qualitative methods such as journaling or case studies are also used. Depending on the type of medical simulation used, assessments in different areas are possible,\textsuperscript{21} including from their meta-analyses of 32 studies: improvement in knowledge (pooled effect size of 1.20), time skills (1.14), process skills (1.09), product skills (1.18), other behaviors (0.81), and effects on patient care (0.50). Other behaviors assessed included instructor ratings of competence, completion of procedures, and procedural errors if present.

Large effect sizes over 0.80 like almost all of these\textsuperscript{22} can be seen when simulations are designed and integrated properly into the curriculum so that the benefits of the simulation can be realized. In addition, newer analysis methods are now available to look at changes in proficiency over time.

\section*{GAME AND SIMULATION DESIGN, DEVELOPMENT, AND EVALUATION}

Simulation design and development is a process of defining learning and assessment goals, determining what is needed to achieve those goals, developing an appropriate simulation, and validating its effectiveness. In this article, our discussion applies to both games and simulations. For ease of reading, the term “simulation” will be used synonymously with “game.”

\section*{CRESST’s Instructional Simulation Design and Development Methodology}

CRESST defined and showed a methodology for designing and developing learning simulations, which includes the creation of simulations as assessments.\textsuperscript{23} The methodology (Fig. 1) delineates both processes and role. The methodology supports the ADDIE (Analysis, Design, Development, Implement, Evaluate) instructional design model,\textsuperscript{24} as well as other instructional design models. Two boxes surround the various components. The upper boxed area surrounds components that pertain to learning goals and objectives, cognitive demands associated with achieving the learning goals, explication of relevant ontologies, and the instructional and assessment parcels necessary for teaching and assessing the to-be-learned materials. The lower boxed area surrounds components that pertain to simulation design and development, including genre selection (i.e., first vs. third person perspective) and platform (e.g., game console, PC, or mobile device), modes of interaction with instructional content, a game play model specifying the elements and relationships of a game, and development of the simulation.

As can be seen by the overlap of the two boxed areas in Figure 1, educators and simulation developers address four of the same components: assessment and instructional parcels, genre and platform, player interaction framework, and game play model. These are the components where a balance between media and learning occur. For example, there are four ways a learner can interact with instruction.\textsuperscript{23} One is to present information directly to the learner. Another is to let the learner find it on his/her own, which means it may not be discovered or the timing of discovery may be earlier or later than desired. Based on the importance of the information (a decision by educators), the simulation developer may be required to use a particular player interaction method.

The arrows in Figure 1 indicate the flow of the design and development process. Portions of the methodology may require a compromise between educators and developers because of instruction and assessment requirements that may dictate selection of a particular genre or platform and because genres and platforms may restrict the types of instructional or assessment strategies available. There is a similar give and
take regarding player interactions, assessment, and instructional requirements, with genre and platform.

The game play model describes the components of a game, their relationships, and the directionality of the relationships. This model provides the architecture for ensuring the game adheres to the requirements necessary for being a game. If the goal of the medium is to be a simulation with little or no game play elements, the game play model can be bypassed.

The last component, development of the simulation, is an iterative process that continuously revisits the way learners interact with instruction and assessments, and the ways content is represented. For a detailed description of the design and development methodology, refer to Wainess and Koenig.

Blending Simulation With Instruction and Assessment

With all the proposed potential of simulations to support learning, hopes for learning outcomes have not been achieved because not all educational simulations are equally effective. The simple presence of educational content in a simulation does not guarantee its efficacy. There is a strong consensus in the research community that learning outcomes are affected by the “instructional strategies” used in simulations, not by the simulations themselves.

Recently, researchers have begun to argue that learning outcomes from simulations also depend on how well the domain instruction is integrated into the simulation. This argument is related to “cognitive load theory,” which is concerned with the learners’ cognitive architecture, including a limited working memory that can have an adverse effect on learning. “Cognitive load” is the total amount of mental activity imposed on working memory at an instance in time. There are three types of cognitive load: (1) “Intrinsic cognitive load,” which is the load involved in the process of learning; (2) “Germane cognitive load,” which is the cognitive load imposed by schema formation; and (3) “Extraneous cognitive load,” which is the load caused by any unnecessary stimuli. Using one set of mechanics for both the game and the instruction and assessment can potentially reduce extraneous load. That is, blending learning the simulation with learning the content can lead to efficient and effective instruction by reducing the amount of switching the learner must do between learning the simulation and learning “from” the simulation, thus, lowering extraneous load.

The remainder of this article will focus on the components within the Instructional Developer section of the methodology, with a particular emphasis on the CRESST Assessment Methodology.

CRESST ASSESSMENT METHODOLOGY FOR COMPUTER-BASED SIMULATIONS

CRESST has devised a methodology for automating assessments of student performance in computer-based simulations. Although the foundation of this methodology was developed primarily with military applications, the approach is generalizable to any domain of interest, including medicine. Figure 2 below shows the primary components of the methodology and how they are sequenced together.

Ontology Development for Assessment and Simulation Design

Ontologies are useful to the design of assessments because of their utility in organizing and storing knowledge “about” things, and about relations “between” things. They can be used to define and represent the current learning status and the educational goals in terms of knowledge (e.g., factual, conceptual, procedural, and metacognitive), skills (e.g., motor, problem solving, and leadership), and attitudes.

Ontologies can be considered knowledge bases that contain the definitions, groupings, and relations of all entities of a domain. More specifically, information stored in ontologies can be graphically rendered as networks or maps, with nodes representing entities and links between the nodes showing the relationships and dependencies between the entities. For a more detailed description of ontologies refer to Guarino et al.
Usually entities involved in the assessment process are split into domain-independent entities, such as cognitive demands and higher-level cognitive constructs, and domain-dependent entities, such as domain-specific knowledge. This split will help with the design of the initial ontology but will not necessarily be feasible for higher-level constructs that require a combination of cognitive demands and domain knowledge. We distinguish between higher-level entities that usually involve other entities and lower-level entities that represent the basic building blocks of the ontology.

The CRESST ontology development process draws on preexisting research in the field.8,42–45 Below, this process is described, with emphasis on linking educational and/or assessment goals with simulation observations.

**High-Level Ontology Development**
The development of ontology is driven by the educational goals, which include the assessment and learning goals of the simulation. The goals contain the domain specifications, the required knowledge and skills, as well as the cognitive demands required for successful assessment completion. High-level ontology development delineates and defines the high-level entities and their relationships. Examples of high-level entities are situation awareness,46 decision making, communication, metacognitive skills, strategic skills, reasoning skills, problem-solving skills,47–49 teamwork, standards for practice and content, standard operating procedures, and “big ideas.” If needed, high-level entities are split up further. For example problem-solving skills might require problem identification, problem representation/modeling, solution planning, and solution evaluation.

A simplified example of the high-level part of a CRESST-developed Navy shipboard damage control ontology is shown in Figure 3. (Note: damage control refers to operations pertaining to addressing the outbreak of fires and floods aboard ships.) It shows how overall damage control knowledge comprises knowledge and skills of other domains or subdomains.

**Low-Level Ontology Development**
The development of a low-level ontology requires the specification of entities on a more detailed level. These entities usually represent the “basic” knowledge and skills, which cannot be split up any further. Depending on the desired level of granularity of the assessment outcomes, these entities can be very detailed or more general. Examples of low-level entities include conceptual/procedural knowledge of a simple concept, a skill that requires no other skills or knowledge or a skill that requires other subskills, but whose subskills cannot be measured in the current simulation (and therefore are not added to the ontology).

To connect high-level and low-level entities, intermediate entities can be added. In Figure 4, a low-level part of the Navy shipboard damage control ontology is depicted with the names of the leaf nodes being “Firefighting Equipment,” “Safety Checks,” “PPE Gear,” and “Reliefs Standing By,” with the intermediate node being “Equipment Selection.” The node “Equipment Knowledge” connects with the corresponding top-level ontology, as depicted in Figure 3. (Note: “PPE” stands for personal protective equipment.)

**BAYESIAN NETWORK DEVELOPMENT**
The ontology becomes the basis for devising an assessment system capable of evaluating performance and making predictions about the student’s mastery of the domain. At the root of this assessment system is a Bayesian network, which is a directed acyclic graphical model for representing probabilistic dependency relationships between variables.50

The Bayesian network, in essence, exists as an operationialized representation of the ontology. It relates the low-level entities of the ontology (i.e., observable actions or events from the simulation) with latent, nonobservable variables that comprise the high-level ontology entities (such as problem-solving ability). By observing low-level entities, and knowing how (and with what influence) they affect higher-level entities, the Bayesian network can make probabilistic inferences of the student’s mastery of each node represented.

The goal is to design the Bayesian network structure such that it reflects an expert human rater’s thought processes when assessing student performance.51 To do this, the system must not only consider the discrete player actions but also consider them in the context in which they occur. Indeed, a single action can be good in one context and bad in another.
Skill mastery is inferred based on the worth of taking one series of actions weighed against the value of taking alternative actions, given the “state of the world.” By having a state-based system in which the dynamic Bayesian network monitors actions relative to previous actions and relative to the state of the world, we increase the validity of the inferences made, and have greater confidence in the predictions of student mastery.

Formative vs. Summative Assessments

Formative Assessments

Formative assessments involve the gathering and interpretation of evidence about student achievement to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of this evidence. In our formative assessment work, the simulation sends observable student actions and game states to the Bayesian network, which scores them using predefined rubrics that encapsulate both the knowledge base of the subject-matter experts and the contextual state of the world when the action was taken. From this, the Bayesian network infers probabilities of mastery and sends appropriate feedback back to the simulation. This feedback could take many forms, including: (a) directing the simulation to serve up a certain type of situation for additional practice; (b) sending instructions to alter the current situation; (c) serving up remedial information; or (d) sending a textual message to guide future performance. Implementing formative assessment facilitates real-time and adaptive responses back to the student, which can boost performance while in session with the simulator.

Summative Assessments

Summative assessments encapsulate all evidence gathered as of a certain time, typically at the conclusion of the simulation. Indeed, there are situations where it is either not needed or not desired to have real-time assessments. In these cases, the simulation would capture and log all meaningful actions and events, and then feed this data into the Bayesian network post hoc, where it would get scored using pre-established rubrics.

Reporting and Debriefing (After-Action Review)

To make meaningful sense of assessment data, various reporting and debriefing tools are used to summarize the student’s performance—both during the simulation, as well

FIGURE 5. Criterion-referenced report used in summarizing student performance in a shipboard damage control simulation developed for the Navy.
as between simulations. We generally consider two classes of reporting tools: criterion-referenced metrics, and growth modeling reports.

Criterion-Referenced Metrics

The automated assessment systems we attach to instructional simulations incorporate criterion-referenced assessments, which are designed to provide a measure of performance that is interpretable in terms of a clearly defined and delimited domain of learning or performance tasks.

In practice, these criterion-referenced reports consist of various charts, graphs, descriptive summaries, and statistics that indicate how well the student performed and/or their mastery/proficiency within a domain according to preestablished criteria. These types of reports are used both for summative feedback to students, as well as for after-action reviews conducted by instructors.

Figure 5 below is an example of a criterion-referenced report used in scoring student performance in a shipboard damage control simulation developed for the Navy. The three pie charts correspond to the phases of fighting a fire casualty aboard ship—namely, Red = “Size it up,” Yellow = “Fight it,” and Green = “Monitor it.” Each phase involves carrying out specific tasks. (Note: The colors red, yellow, and green here “do not” correspond to achievement; they are merely Navy-used terminology to describe the phases of firefighting.)

Scoring for how well these tasks were performed is shown in the bar charts on the right, with each task being scored using a rubric with a 0 to 100 scale. For example, in the simulation, the student outfitted his team with the correct PPE, so the PPE score was 100%. But for Electrical Isolation (i.e., cutting electrical power to the room that’s on fire), the action was taken too late and endangered the ship, so the score for this task was only 50%. From these individual scores, the Bayesian network then infers the student’s proficiency with that phase of operations (overall), and is expressed as a percent in the pie chart. In this case, the Bayesian network infers only a 60% chance that the student has full proficiency in carrying out the operations that comprise the Red Phase of firefighting operations aboard ship.

Growth Modeling Reports

Growth modeling looks at how individuals change over time and whether there are differences in patterns of change. With growth modeling, the goal is to track student performance over time across a consistent set of measures. This can occur both within a single simulation, and across multiple instances of the student engaging with that same simulation. When coupled with criterion-referenced performance reports, growth modeling helps to form a complete picture of the student’s capabilities within the domain being assessed.

Figure 6 illustrates this with student performance data taken from the Navy damage control simulation. It shows the student’s achievement in keeping overall fire size (i.e., Casualty Strength) at or below a predefined safety threshold when combating class “bravo” fires (such as a grease or petroleum fire) across six different game play sessions (each 30–60 minutes). Here, the lower the Casualty Strength, the better, as it means the fire is smaller in size. The goal for the student is to ensure the Casualty Strength does not exceed the safety threshold. In the fourth simulation run, the fire reached strength of 90 before the student was able to extinguish it. But in the fifth run, he improved his tactics and was able to extinguish the fire before it exceeded strength of 50.

Here, Runs 2, 5, and 6 are best because in each, the fire strength remained below the safety threshold.

SUMMARY

Computer-based instructional simulations are becoming more and more ubiquitous, particularly in military and medical domains. As the technology that drives these simulations grows ever more sophisticated, the underlying pedagogical models for how they are designed must evolve accordingly.

In this article, we have highlighted some of these key design considerations, namely learning goals and objectives, cognitive demands, simulation affordances, and instructional and assessment strategies. We then presented a CRESST Assessment Methodology that delineates how these elements can be represented in a domain ontology, and how this completed ontology can facilitate the development of a Bayesian network-based assessment engine.

The goal of the Bayesian network is to model expert thinking when assessing student performance in the simulator. It can be connected to the simulation to perform real-time formative assessments, or it can be used post hoc to carry out summative assessments of performance. The output of such assessments can be exhibited in various reporting and debriefing instruments, including criterion-referenced and growth-modeling summary reports.

Although the methodologies outlined here have proven useful in CRESST’s automated assessment work, there is still much research to be done. Our hope is that the approaches...
outlined here will inspire new ideas for how instruction, simulation, and assessment can be blended, and will foster an ongoing dialog among the educational and simulation development communities.

ACKNOWLEDGMENTS

The work reported herein was supported by grant number N00014-08-C-0563 with funding to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) by the Office of Naval Research. The work reported herein was also partially supported by a grant from the Office of Naval Research, Award Number N00014-10-1-0978.

REFERENCES

Application of National Testing Standards to Simulation-Based Assessments of Clinical Palpation Skills

Carla M. Pugh, MD, PhD

ABSTRACT With the advent of simulation technology, several types of data acquisition methods have been used to capture hands-on clinical performance. Motion sensors, pressure sensors, and tool-tip interaction software are a few of the broad categories of approaches that have been used in simulation-based assessments. One major distinction between some of the technologies used to capture hands-on performance is the location of the sensors, hence the type of data that is collected. Some of the simulation systems have the sensors on a tool, surgical instrument, or data glove. In this instance, the information captured helps to quantify hand or instrument positions during task execution. For example, when using a data glove, the associated metrics provide quantitative measures of various hand and individual finger positions throughout the task. Similarly, instrumented surgical tools allow information capture regarding the position of the surgical instrument during a procedure. In both instances, the data captured is clinician centered and focuses on the motor movements and positioning of the clinician.

In this article, we will focus on sensor-enabled training tools in which the sensors are not on the clinician’s hand or instrument but on the patient. In this instance, the information captured using this type of technology helps to quantify human interaction with specific anatomical structures. For example, when using a simulated patient with sensor-enabled organs, the associated metrics provide quantitative measures regarding patient contact including anatomical location and quality of clinical palpation during a physical examination. Moreover, using a simulated patient with sensor-enabled organs during a surgical procedure will allow data collection regarding instrument interaction with the patient’s anatomical structures. The resulting data in both scenarios are patient centered and provide detailed information regarding direct (hands) or indirect (instruments) patient contact.

As most procedures and physical examinations require hands-on contact with specific anatomical locations, patient-centered devices allow for assessment of a variety of errors including incomplete examinations, missed lesions, or use of excessive force. The purpose of this article is to present a focused review of three sensor-enabled simulations that are currently being developed for patient-centered assessments of clinical palpation skills. The Standards for Educational and Psychological Testing will be used as a framework to structure our review.

METHODS

Testing Standards

The Standards for Educational and Psychological Testing is a set of testing standards developed jointly by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. The standards address professional and technical issues of test development and use in education, psychology, and employment. The intent is to promote the sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices. While evaluation of the appropriateness of a test or testing application should depend heavily on professional judgment, the standards provide a frame of reference to assure that relevant issues are addressed. The standards apply equally to standardized multiple-choice tests and performance assessments. For the purpose of this article, our sensor-enabled, simulation-based assessments are considered to be a type of performance-based assessment: evaluation of performance during tasks that are valued in their own right.

Sensor-Enabled Simulation Technology

The focus of this article is on performance assessments using sensor-enabled mannequin technologies. Development and
implementation of this technology began in 1998. The initial clinical focus was physical examination of the female pelvis, female breast, and male digital rectal examination. All three systems include a partial mannequin; umbilicus to mid-thigh for the pelvic and digital rectal examinations; and left or right chest wall for the breast models. In addition, the mannequins have interchangeable parts that enable simulation of different clinical presentations, both normal and abnormal scenarios. For example, the pelvic examination inserts range from normal-anteverted and normal-retroverted uterine positions as well as enlargements of the uterus or ovary.

Paper-thin 2 mm force sensing resistors (sensors) are connected to important anatomical structures on the mannequin and imbedded organs. Data acquisition systems enable the sensors to be sampled at a specified sampling rate. For these clinical examinations, the sampling rate is set at 30 Hz. Figure 1 shows the pelvic exam simulator including computer interface and mannequin. The computer interface shows that the user is touching the cervical os at six pressure units (1 PU = 0.125 psi). Consequently, the corresponding register bar rises to a level of six, the indicator button in the cartoon diagram turns blue, and a check mark appears in the “Exam Checklist” window. During a simulated pelvic exam, this interface enables students and instructors to see where the examiner is touching and how much pressure is being used. The prostate and breast models have similar computer interfaces.

Figures 2A and 2B show line graph representations of a pelvic examination plotted as pressure over time. Each line represents a different anatomical area. When reviewing performance using the line graphs, the examiner’s touch can be tracked by anatomical location and several palpation characteristics including the level and type of pressure. In Figure 2A, the examining medical student applies several bursts of pressure to the fundus of the uterus. This is represented by a series of narrow spikes (see black arrow) in the 8–10 PU range. Simultaneously, there is constant pressure on the left posterior cervix (L-post) at a level of 6 PUs. These data quantify bimanual examination of the uterine fundus: one hand explores the fundus while the other hand applies counterpressure to the cervix to lift the uterus toward the abdominal wall and facilitate palpation. The combination of pressures used (spikes plus constant pressure) provide detail on the examiner’s approach including anatomical locations explored (i.e., sensor locations) and palpation characteristics used (i.e., pressure levels and spikes plus constant pressure). To emphasize a range in performance, the medical student in Figure 2B applies a combination of pressures to the cervix in the 6–8 PU range. The first spike (~600 time units) represents pressure applied to the cervical os. The second waveform (~750–1166 time units) combines a series of spikes and constant pressure on the left posterior cervix. Starting around 1200 time units, the examination continues with several manipulations of the cervix in the 5–8 PU range and ends with two low pressure (4 PU) spikes on the uterine sensors. Compared to the student in Figure 2A, the student in Figure 2B spends a lot of time applying pressure to several areas on the cervix; takes twice as long to accomplish bimanual examination of the uterine fundus; and uses much lower pressure and a smaller number of peaks when palpating the uterine fundus. While the second student was eventually able to accomplish the goal of bimanual examination, there are noticeable differences in performance when comparing the two students. These graphs show that specific physical examination events can be captured and quantified. Use of this data for performance-related decisions will be discussed using the standards as a guiding framework.

**Metrics**

While the line graph representations of the data provide a high-level, qualitative view of performance differences, to quantify individual differences, the data must be converted to measurable variables. The most commonly used performance assessment variables extracted from the sensor data include (1) examination time, (2) number of sensors touched, (3) maximum pressure, and (4) frequency. The operational definition of these variables is as follows:

**Time Variable**

The time variable is equivalent to the length of time necessary for an examiner to perform a complete examination. Mathematically, exam completion time was defined as the time at which the last sensor was touched minus the time at which the first sensor was touched. The exam was considered to have begun when the pressure on any given sensor reached 1 full pressure unit above baseline.

**Critical Areas Variable**

The critical areas variable represents the number of sensors touched during the simulated clinical examinations. For the pelvic exam there were seven sensors; four on the cervix, one on the uterus, and one on each ovary. For the digital rectal exam there were also seven sensors on the prostate, three on the right lateral lobe of the prostate, three on the left lateral lobe of the prostate, and one in the median raphe. For the breast exam there were eleven sensors; approximately two-three sensors in each of the four quadrants of the breast.

**Maximum Pressure Variable**

The maximum pressure variable represents the highest pressure reading recorded for a sensor during the simulated examination. For example, if the highest pressure readings were recorded for sensor no. 1 as seven PU and sensor no. 2 as ten PU for mathematical purposes, average maximum pressure would be the average of those two sensor variables; however, maximum pressure would be the individual highest pressure for each sensor.
Frequency Variable

The frequency variable represents the number of times an individual sensor was touched near the maximum pressure during the examination. The mathematical formulation for creating this variable involved counting the number of times the given sensor was sampled within 0.5 PU of the maximum for that sensor.21–27

In addition to extraction of these variables using MATLAB code (MathWorks, Natick, MA), specific data mining techniques have been applied to the sensor data to gain...
an understanding of quality and usefulness, including Markov models\textsuperscript{19,23–25} and visual analytics.\textsuperscript{26}

**Application of National Testing Standards**

Use of the sensor-generated data for performance-related decisions requires a standardized approach to ensure validity, reliability, fairness, and appropriate use.\textsuperscript{20} The Standards provide a three part guide for evaluating a wide variety of assessments. The following sections present a detailed review of our past work. The Standards are used as a guide to structure our discussion. All of the studies reviewed in this article were performed after approval by the local Institution Review Board.

**Standards Part I—Test Construction, Evaluation, and Documentation**

When designing the simulation-based assessments, test construction required a review of all elements that may be used to gather data for evaluation purposes. The two main elements included (1) the written clinical assessment form and (2) the computer-generated sensor data. Correct answers for the written clinical assessment vary according to the clinical scenario (normal or pathologic variation) represented by the simulation. Correct answers for the computer-generated variables are also closely linked to the clinical presentation. While there is ongoing work to determine the best construction and administration for our simulation-based assessments, for the purpose of this article, each clinical scenario represents a specific content domain, hence a major section within a test. Most administrations using the clinical simulations involve at least two different clinical scenarios. Opportunities for partial credit exist for both the written and the sensor data components. The following sections provide an overview of the various studies we have performed to guide the process of test construction and administration.

Validity, reliability, and measurement error are important to consider during test construction. Validity was initially assessed by evaluating a basic content construct—does this technology capture any performance measures of interest? To answer this question we used a sensor-enabled pelvic examination simulator with second-year medical students ($N = 73$). In addition to collecting sensor-generated performance data, we assessed diagnostic accuracy rates using participant’s written clinical assessments of two, clinically different, pelvic models.\textsuperscript{27} Using a 2-tailed Pearson’s correlation, we found that three of the four sensor variables were significantly associated with participants’ ability to generate an accurate clinical assessment of the simulator after performing an examination. The highest correlation was for the “number of critical areas touched” during an examination ($r = 0.311, p = 0.007$). The second highest correlation was for the “mean maximum pressure” used during the examination ($r = 0.279, p = 0.017$). Finally, the last variable with a significant correlation was

![FIGURE 2. Line graph representation of the sensor-generated performance data collected during a pelvic examination. (A) The arrow points to several, high-pressure palpations of the uterus. (B) The arrows point to high pressure-palpation of the right and left cervix.](image)
“mean frequency” used during the examination ($r = 0.267, p = 0.022$). While the correlation values for these three variables are moderate to low, they indicate that there is a relationship between the simulator variables; accuracy on the written assessments; and the use of direct, hands-on contact during an examination. In essence, these three variables capture aspects of the clinical pelvic examination that are important in achieving an accurate diagnosis. Reliability of the three variables in this setting was as follows: (1) time, $r = 0.72$; (2) critical areas, $r = 0.63$; (3) mean maximum pressure, $r = 0.77$; and (4) mean frequency, $r = 0.50$. While the time variable did not have a significant correlation with diagnostic accuracy in this setting, it appeared to have moderate to high reliability and we continued to evaluate this variable in other settings (i.e., different participants and clinical scenarios).

Validity was also assessed by evaluating the ability to use the simulator data to differentiate between experience levels. When assessing the experience level using the pelvic simulators, we compared medical student ($N = 43$) performance with that of experienced clinicians ($N = 20$). For the written assessment, mean examination scores showed medical students were less accurate than clinicians (students = 10.18/18, clinicians = 15.60/18.0, $p < 0.001$). There were also differences noted in examination techniques. Students were noted to spend more time on the exam (students = 82.01 seconds, clinicians = 31.07 seconds, $p < 0.001$) and used greater palpation frequencies when examining each area (students = 42.45 Hz, clinicians = 20.30 Hz, $p < 0.005$). When evaluating experience level using the digital rectal examination simulator we conducted a study involving surgical residents ($N = 24$) and medical students ($N = 30$). Participants were grouped according to the number of prior digital rectal examinations performed. Group 1 ($N = 27$) was the less experienced group having performed five or fewer previous examinations. Group 2 ($N = 27$) had performed six or more rectal examinations. Each participant examined two different simulators: Simulator A (easy diagnosis—normal rectum and a firm 2 mm prostate nodule) and Simulator B (difficult diagnosis—enlarged prostate + a subtle 3 cm rectal mass). When comparing technical performance on Simulator A, the more experienced group (G2) was noted to spend more time on the examination (G2 = 12.34 seconds, G1 = 7.22 seconds, $p < 0.01$). For this simulator (Simulator A—easy diagnosis), there were no significant differences in accuracy. The less experienced group had an accuracy rate of 80% and the more experienced group had an accuracy rate of 85%. In contrast, for the more difficult simulator (Simulator B) there were significant differences in technical performance and accuracy. When comparing performance on Simulator B, the more experienced group G2 spent more time on the exam (G2 = 17.52 seconds, G1 = 11.94 seconds, $p < 0.05$) and were more accurate in their assessment and documentation of the prostate findings (G2 = 64% accurate, G1 = 33% accurate, $p < 0.05$).

Validity was also assessed by evaluating the ability to use the simulator data to differentiate between clinical specialties and gender. When assessing clinical specialty using the breast examination simulators, we compared the performance of four clinical groups: (1) surgeons ($N = 37$), (2) nonsurgical MD’s ($N = 36$), (3) nurses ($N = 12$), and (4) medical assistants ($N = 15$) on three simulators: (1) Simulator A—dense breast, 2 cm hard mass; (2) Simulator B—fatty breast, no masses; and (3) Simulator C—dense breast with right upper quadrant thickening. When assessing overall approach to the breast examination, nurses were noted, on average, to palpate more anatomical areas during the examination (nurses = 9.4/11 areas, surgeons = 7.38/11 areas, nonsurgeons = 6.22/11 areas, and medical assistants = 6.71/11 areas, $p < 0.01$). In addition, there was a trend for the nurses to spend more time on the breast examination. However, this finding was only significant for one of the three models—Simulator A (nurses = 58.57 seconds, surgeons = 40.94 seconds, nonsurgeons = 32.33 seconds, and medical assistants = 39.65 seconds, $p < 0.05$).

Despite these observed differences in technical approach, and differences in mean accuracy rates for each clinical scenario ($A = 87\%$, $B = 76\%$, and $C = 68\%$), there were no significant differences in accuracy when comparing the four specialties on the three breast models. Although there were differences in hands-on contact and approach among the specialties in this study, there was no difference in accuracy. As such, the three clinical scenarios and simulator variables are not a valid tool for assessing important differences in surgical, nonsurgical, and nursing specialties.

When assessing gender, we compared male ($N = 38$) and female ($N = 57$) performance on three breast simulators: (1) Simulator A—dense breast, 2 cm hard mass; (2) Simulator B—fatty breast, no masses; (3) Simulator C—dense breast with right upper quadrant thickening. The results showed that females, on average, spent more time (M = 42.09 seconds, F = 56.66 seconds, $p < 0.05$), touched more anatomical areas (M = 6.30/11 areas, F = 7.97/11 areas, $p < 0.05$), and used greater pressures (M = 4.82 mmHg, F = 5.21 mmHg, $p < 0.05$) when compared to male clinicians. While there was a trend toward females being less accurate (M = 83.7% correct, F = 72.4% correct) the difference was not statistically significant. Although there were differences in hands-on contact and approach, when comparing males and females, there was no difference in accuracy. As such, Simulators A–C and the computer-generated simulator variables are not a valid tool for assessing gender-related accuracy differences. Moreover, it is possible that there are no gender-related differences.

In summary, the simulator variables appear to capture data that correlates with hands-on performance during simulated clinical examinations. In addition, the variables show promising results in discriminating experience levels. When evaluating specialty and gender, there appears to be differences in technical approach but not overall accuracy. As such, the variables may not be useful in discriminating between specialties and gender. Moreover, there may not
be any important differences between specialty and gender except approach.

*Standards Part II and III*

Standards Part II and III deal with fairness and testing applications. Fairness issues include lack of bias, equitable treatment in the testing process, and equal opportunities to learn. As the simulations are physical models and the sensors quantify hands-on touch, biases in language and linguistics are limited. Test taker rights such as access to test results and rights when reviewing testing irregularities are issues that will need to be addressed before formal use of the simulations for performance related decisions. Issues relating to testing applications largely address general responsibilities of those who administer, interpret, and use test results. When using previously validated tests in different venues, test users must ensure that there is continued test validity and reliability in the new setting (Standard 11.19). Use of tests for psychological evaluation, educational assessment, employment-related decisions, and program evaluation are other areas that must be considered as part of test applications. These areas will be considered as part of our future work in evaluating the simulation-based assessments.

**SPECIAL CONSIDERATIONS**

**Content Domain and Creation of Clinical Scenarios**

There are inherent difficulties in simulating human body parts. These difficulties present several challenges to the use of simulation as an assessment tool. From a manufacturing standpoint, the developer’s goal is to find the right combination of materials and molds that, once fabricated, are the most realistic representation of human tissue possible. In essence, the goal is the best match for context and functionality. When building a breast model for example, it may be possible to perfect the mold and achieve an extremely realistic look; the shape, the color, the skin detail may all be perfect. However, after achieving this perfection, the materials may not feel or behave like real breast tissue. Likewise, a breast model may feel realistic to touch but be found deficient in achieving a realistic look. An additional challenge in using simulation as an assessment tool relates to human perception. For example, when manufacturing a breast model that represents a patient with an obvious breast mass, the expectation would be that most manufacturing a breast model that represents a patient with a clear indication of a breast mass. However, diagnostic perception may be affected if the palpation is not obvious. While the overall goal, from an assessment perspective, is to provide a range of easy and difficult test questions (simulations), lower accuracy rates on the more difficult clinical scenarios must be evaluated from a validity perspective in a similar fashion and with the same rigor as multiple-choice questions. Defining this process for simulation is imperative to the success and applicability of simulation-based assessments.

**The Science of Touch**

In health care, despite the many technological advances, human touch remains important for many diagnostic and therapeutic interventions. Unfortunately, without performance measures, objective and formative feedback to health care trainees and practitioners is nearly impossible. As such, there are no standards and health care professionals continue to graduate and become credentialed to practice medicine without any real measure of their ability to perform hands-on procedures or make sound clinical judgments using palpation. Part of the problem lies in the complexity of touch (palpation). Despite its importance, the human sense of touch is poorly understood and understudied. Touch is extremely difficult to convey verbally and there are no objective means of explaining one’s own experience or perception based on the sense of touch. Over 20 years of extensive research on the sense of touch reveals a set of specific hand maneuvers that humans use to detect object characteristics. The hand maneuvers, called exploratory procedures, are stereotyped movement patterns consisting of certain characteristics, which are largely subconscious and reproducible in a variety of settings. Key findings from this work show (1) human beings are very good at recognizing common objects on the basis of touch alone; (2) object recognition is strongly based on specific object characteristics including texture, hardness, shape, temperature, and weight; (3) during object recognition, specific hand maneuvers are used to detect object characteristics.

Our work using sensor-enabled, patient-centered simulations to assess palpation skills in clinical medicine has shown promising results regarding the relationship between sensor outputs and specific exploratory maneuvers used during palpation. When using force sensing resistors (FSRs) on anatomical models that simulate common medical examinations (breast, pelvic, and digital rectal examinations), we found that specific palpation maneuvers were detectable in our data. Figures 3A–3F show laboratory-generated waveforms for specific palpation maneuvers. Laboratory participants were asked to perform the following maneuvers for a minimum of 4 seconds on a sensored plate: (1) balloting (multiple, vertical bursts of firm pressure), (2) circular motion (rubbing counterclockwise), (3) constant pressure, and (4) rubbing (firm back and forth pressure across a vertical line). While the resulting waveforms show similar patterns for rubbing and circular maneuvers, balloting in multiple areas is distinguishable from circular motion in multiple areas and constant pressure. Linking the waveforms to specific exploratory maneuvers will facilitate our understanding of palpation characteristics.
during physical examination. For example, as shown in Figure 2A, we now understand that the examinee was using the balloting maneuver when examining the fundus. Our future work will continue to explore the sensor data using Klatzky and Lederman’s classification of palpation maneuvers. We believe this may help to generate a better understanding of how the sensor-generated data can be used in performance assessments.

**FIGURE 3.** Waveforms extracted from the sensor data during specific palpation maneuvers: (A, B) Balloting; (C, D) Circular pressure movements; (E) Constant pressure; and (F) Rubbing.
Simulation-Based Assessments of Clinical Palpation Skills

SUMMARY

The purpose of this article was to present a focused review of sensor-based assessments of palpation skills using patient-centered simulations. The technology components, capabilities, and metrics were reviewed. In our review of validity, we found that there were significant correlations between the sensor-generated performance variables and accuracy when reporting and documenting clinical findings. While the correlation values were moderate to low ($r = 0.267–0.311$), they warrant additional investigation of the relationship between hands-on performance and rate of accuracy during a clinical assessment.

In our review of validity evidence, we focused on three constructs: (1) experience level, (2) clinical specialty, and (3) gender. The results for experience level revealed promising results in discriminating between groups based on technical approach and accuracy.\(^\text{16,28}\) The results for specialty and gender showed differences in technical approach but not overall accuracy. As such, the current variables and clinical scenarios are not valid discriminatory variables for these constructs.

The use of palpation remains important for many diagnostic and therapeutic interventions in the health care environment. Development of performance metrics and assessments to ensure minimum performance standards is an important endeavor that should be closely guided by national standards.

ACKNOWLEDGMENTS

The following researchers have made significant contributions to this body of work: Jacob Rosen, Lawrence Salud, Jonathan Salud, Alec Peniche, Abby Kaye, and Brandon Andrew. This body of work has been funded by the following foundations and agencies: National Board of Medical Examiners (NBME) Stemmler Fund; Media X Grant, Stanford University; Augusta Webster Educational Innovation Grant, Northwestern University; Eleanor Wood-Prince Grants Initiative, Northwestern Memorial Hospital; National Cancer Institute-Supplement Grant-3U01CA116875-03S1; The Baum Family Wood-Prince Grants Initiative, Northwestern Memorial Hospital; National Institutes of Health R01EB011524. The work reported herein was also partially supported by a grant from the Office of Naval Medical Research, Award Number N00014-10-1-0978.

REFERENCES


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. Defined broadly as a “person, device, or set of conditions which attempts to present evaluation problems authentically,” medical simulations emulate patients, anatomical areas, or clinical tasks. They include standardized patients, part-task trainers (e.g., pelvic replicas), virtual reality systems, computer simulations, and games, mannequins, and even multiple-choice questions presenting information on a case to be evaluated. Simulations can be used for instruction or assessment, and are currently used by many medical schools for end-of-course comprehensive examinations, by the Medical Council of Canada as part of the licensure process and as part of the United States Medical Licensing Examination, among many others.

Simulation-based training has become popular because it is usually less costly, and it provides experiences without risk to patients. In addition to the benefits of cost and risk avoidance, there are also benefits to learning. 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. 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. 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 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, 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.
Evaluation of Medical Simulations

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. 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.

These methods are based on traditional psychometrics or classical test theory, 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—think standardized tests, such as the Scholastic Aptitude Test. 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 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.

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. 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.

Validity is not a property of the evaluation; it is a property of the inferences made based on the results. 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. This argument is very different from early conceptions of validity 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, 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.

–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, 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. 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 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 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
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. 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. 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. provide an excellent treatment of assessing the Accreditation Council...
for Graduate Medical Education competencies in psychiatric programs. Brünen et al.\textsuperscript{54} provide indicators for measuring cognitive load, and Hays\textsuperscript{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.\textsuperscript{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\textsuperscript{57} van der Vleuten and Swanson\textsuperscript{58}) with written and computer-based clinical simulations or computer-based case simulations, also called patient management problems,\textsuperscript{1} and with mannequins.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{62,63}

**AUTOMATED SCORING**

There have been multiple frameworks for evaluation and use of automated scoring (see Williamson et al.\textsuperscript{64} and Shermis and Burstein\textsuperscript{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, not actual expert performance. This scoring approach has been used successfully to model expert and rater judgments in a variety of applications including essays\textsuperscript{69} and patient management skills.\textsuperscript{30}

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\textsuperscript{69} and patient management skills.\textsuperscript{30}

One of the major issues with expert-based scoring is the selection of the expert.\textsuperscript{70,71} Problems include experts’ biases, the influences of the experts’ content and world knowledge, linguistic competency, expectations of student competency, and instructional beliefs.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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. 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.); to Tatsuoka’s rule-space methodology, to the use of Bayes nets to model student understanding in domains such as Web searching, to the use of Bayes nets to model student understanding in domains such as Web searching, to the use of Bayes nets to model student understanding in domains such as Web searching, circuit analyses. 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. and Williamson et al.

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. 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.

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.

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., Boulet and Swanson, and Robinson et al. 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., 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 and Lenth. Lenth 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. 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 Giuliani et al.

**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. 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. 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.

**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.
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 instrumental 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, 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.
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. 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
Evaluation of Medical Simulations


Evaluation of Medical Simulations


Prevention of Surgical Skill Decay

Ray S. Perez, PhD*; Anna Skinner, MA†; Peter Weyhrauch, PhD‡; James Niehaus, PhD‡; Corinna Lathan, PhD, PE†; Steven D. Schweitzberg, MD, FACS§; Caroline G. L. Cao, PhD∥

ABSTRACT The U.S. military medical community spends a great deal of time and resources training its personnel to provide them with the knowledge and skills necessary to perform life-saving tasks, both on the battlefield and at home. However, personnel may fail to retain specialized knowledge and skills if they are not applied during the typical periods of nonuse within the military deployment cycle, and retention of critical knowledge and skills is crucial to the successful care of warfighters. For example, we researched the skill and knowledge loss associated with specialized surgical skills such as those required to perform laparoscopic surgery (LS) procedures. These skills are subject to decay when military surgeons perform combat casualty care during their deployment instead of LS. This article describes our preliminary research identifying critical LS skills, as well as their acquisition and decay rates. It introduces models that identify critical skills related to laparoscopy, and proposes objective metrics for measuring these critical skills. This research will provide insight into best practices for (1) training skills that are durable and resistant to skill decay, (2) assessing these skills over time, and (3) introducing effective refresher training at appropriate intervals to maintain skill proficiency.

INTRODUCTION

Prevention of skill decay is and should be a high priority for any training organization. Skill decay is the partial or full loss of trained or acquired skills and knowledge following periods of nonuse.1 Loss of critical skills and knowledge is of great concern, and is problematic in situations in which medical personnel receive initial training but may not have an opportunity to use these skills or knowledge for extended periods of time. Arthur et al1 in a meta-analysis examining the results of 53 articles within the skill decay and retention literature, found that after 365 days of nonuse or nonpractice, the average participant’s performance was reduced by almost a full standard deviation (d = 0.92). There is evidence that declarative knowledge (e.g., facts, principles, and concepts) decays at a slower rate than procedural knowledge (e.g., multiple steps that must be performed in a specified order to solve a problem or complete a task).2 Psychomotor skills also appear to be more resistant to skill decay than cognitive tasks, in the absence of mental rehearsal.1 Although the research literature on skill decay is comprehensive and extensive, this is not true of the literature specifically pertaining to laparoscopic surgical (LS) skill decay. One study that did focus on these skills was conducted by Brunner and Korndorffer3; they examined skill decay of LS skills, echoing the findings by Arthur and his colleagues, demonstrating that LS skills could be trained in a virtual reality learning environment and that nonuse exhibited the same pattern of skill decay reported in the meta-analysis conducted by Arthur et al.1 Despite this specific conclusion, a primary limitation within this area is the lack of research examining the acquisition and retention of the cognitive skills of LS (e.g., decision making and problem solving). Within the domain of general surgery, Jacklin et al4 studied the effects of providing cognitive feedback to surgical trainees in an attempt to improve their risk assessment of surgery (i.e., judgment of postoperative mortality risk). They found that the use of cognitive feedback improved the accuracy of trainees’ estimates of postoperative mortality risk. However, very little research has been conducted to examine the skill decay of these critical cognitive skills, and similar research has not been conducted within the LS domain.

Section 1 of this article reviews skill acquisition and retention research related to general skills, military skills, and LS skills. Section 2 introduces models that identify critical skills related to laparoscopy, and Section 3 proposes objective metrics for measuring these critical skills. Section 4 addresses the development of skill decay curves, and Section 5 proposes a model that examines refresher training of laparoscopic skills. The article concludes with a summary and conclusions that include future areas of research. This article includes two perspectives and approaches to addressing the issues surrounding LS skill assessment, training, and sustainment: (1) the development and use of a Laparoscopic Surgery Training System (LASTS) prototype for learning, refreshing, and assessing LS skills and (2) the design, development, and validation of a portable, open architecture Surgical Skills Training and Assessment Instrument (SUSTAIN) to support acquisition and retention of fundamental psychomotor, perceptual, and cognitive laparoscopic skills. Both systems seek to leverage the development and validation of an empirical model of skill acquisition and decay.

*Office of Naval Research, 875 North Randolph Street, Arlington, VA 22203.
†AnthroTronix, Inc., 8737 Colesville Road, L203 Silver Spring, MD 20910.
‡Charles River Analytics, Inc., 625 Mount Auburn Street, Cambridge, MA 02138.
§Cambridge Health Alliance, 1493 Cambridge Street, Cambridge, MA 02139.
∥Biomedical, Industrial and Human Factors Engineering, Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH 45435.
SECTION 1: BACKGROUND

Skill Acquisition and Decay

To analyze skill decay, we must first examine the process by which knowledge and skills are acquired. The rate of skill acquisition depends on many factors; however, there is evidence that rates of learning tend to follow a logarithmic or exponential function, rather than a linear relationship between the time to perform a task and the number of practice attempts.\(^5\)\(^6\) There are typically three stages of skill development: a cognitive stage, associative stage, and an autonomous stage; it is during the autonomous stage that expertise is achieved.\(^7\) Research has shown that the major factors influencing skill decay and retention include the length of retention (nonuse) interval, degree of overlearning, task characteristics (e.g., closed loop versus open loop, physical versus cognitive), methods of testing for initial learning and retention, conditions of retrieval (recognition versus recall), instructional strategies or training methods, and individual differences in trainees.\(^2\)

Skill Acquisition and Decay Within Military Tasks

The rate of skill decay within military tasks has been researched extensively in the past. For example, Wisher et al\(^2\) investigated the decay of skills and knowledge with 20,000 reservists, and found that gross motor skills decayed after approximately 10 months, whereas cognitive skills, such as recall of procedures, decayed within approximately 6 months. Thus, while both motor and cognitive skills are subject to decay with nonuse, the respective rates of decay differ. The Naval Education and Training Command developed a categorization of navy tasks on a scale of proneness to decay; tasks most prone to decay included recalling procedures and voice communications tasks, whereas those least prone to decay included gross motor skills and attitude learning.

Wisher et al\(^2\) also conducted an extensive review of general skill acquisition and retention/decay literature. Based on this review, they categorized military tasks into 3 components: (1) knowledge, (2) decision, and (3) execution. The knowledge category is based on the recall of domain-specific information; the decision category depends on cognitive processing of the domain-specific information; and the execution category refers to both the perceptual and motor requirements of a task. Wisher et al\(^2\) also identified specific task factors that affect skill acquisition and decay, such as task complexity and task demands. Other factors identified as affecting decay included task time pressure, whether or not job aids were used (job aids decreased decay), and the quality of job aids used (higher quality job aids decreased decay).

Although past research efforts have provided a greater understanding of the mechanisms underlying skill acquisition and decay, including the relative rates of decay of various types of skills, there are no detailed models and skill decay curves of military medical tasks. A need exists for validated models and skill decay curves within the context of specialized military medical skills, such as LS skills, because these specialized procedures are not usually performed during deployments, and are therefore susceptible to decay. Validated LS skill decay models would support the development of guidelines for accurately timed refresher training to prevent LS skills decay.

Skill Acquisition and Decay Within Laparoscopic Surgical Tasks

Over the past decade, attempts have been made to establish standards for laparoscopic surgical skills training and evaluation. The McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS) has been shown to be reliable and valid as an educational tool.\(^8\) This method of assessment has been incorporated into the manual skills training practicum portion of the Society of American Gastrointestinal and Endoscopic Surgeons’ (SAGES) Fundamentals of Laparoscopic Surgery (FLS) training program, and includes a portable video trainer box for rehearsal of five basic manual skills tasks: peg transfer, pattern cutting, endo-loop placement, extracorporeal suturing, and intracorporeal suturing (see Fig. 1).

FLS trainer scores have been shown to be predictive of intraoperative laparoscopic performance as measured by the Global Operative Assessment of Laparoscopic Skills (GOALS) manual assessment framework,\(^9\) making the FLS training program the current “gold standard” in laparoscopic training, and resulting in its rapid adoption as a primary component of many general surgery residency programs.\(^10\)\(^11\)

FLS manual skills training results in laparoscopic manual skills that have shown to be durable for up to 11 months,\(^12\) and skill retention has been demonstrated for three tasks similar to FLS tasks for up to one year.\(^13\) However, these studies relied only on time to complete the specified manual skills tasks, and retention of the cognitive components of LS training was not considered. Also, although maintenance of laparoscopic manual skills through rehearsal and retraining has been shown to prevent decay,\(^12\) no standards currently exist for retraining, and few deployable systems exist that can be used where they are most needed—in far forward military
medical facilities—to provide critical skills refresher training during long deployments.

Stefanidis et al\textsuperscript{14} provided preliminary evidence that a visual–spatial secondary task assessing spare attentional capacity may help distinguish among individuals of variable laparoscopic expertise when standard FLS performance measures fail to do so, and that automaticity metrics may improve current simulator training and assessment methods. Stefanidis et al\textsuperscript{15} tested this theory further, demonstrating that after approximately 10 hours of training (average of 84 trials) on the FLS intracorporeal suturing task with concurrent performance of a visual–spatial secondary task, novices demonstrated improvements in both suturing and secondary-task performance compared with baseline scores; however, none of the subjects achieved expert level secondary task proficiency. Recently, Stefanidis et al\textsuperscript{11} demonstrated that training to automaticity (overlearning) on the FLS intracorporeal suturing task while concurrently performing a visual–spatial secondary task resulted in improved operating room performance on a porcine model as compared to training to proficiency without the secondary task. This level of training required, on average, 163 trials, and nearly half of the participants assigned to the automaticity training group were unable to achieve expert performance levels on the secondary task. Therefore, while superior training and skill transfer was demonstrated, the training costs are high in terms of time and resources. Also, the secondary task used was unrelated to surgical tasks. It is possible that by incorporating a secondary task that addresses interoperative skills, including cognitive skills, training would be enhanced further, providing justification for the extended time associated with training to automaticity. This article not only explores the impact of automaticity on the acquisition and durability of an LS skill but also raises a number of questions, such as: (1) How much training is actually necessary to competently perform LS?; (2) Do fundamental LS skills that become automated generalize to other minimally invasive surgery tasks and specific procedures?; and (3) What is the impact of automaticity on the training of cognitive LS skills?

Recent research has sought to develop a deeper understanding of the cognitive factors involved in training. For example, Park et al\textsuperscript{16} have attempted to understand the role of declarative memory processes during psychomotor skills learning using a cognitive simulation based on the Adaptive Control of Thought—Rational model. However, they did not directly assess the role of decision making involved in surgical procedures. Palter,\textsuperscript{17} in a review of training curriculum for Minimally Invasive Surgery (MIS), suggests that MIS training should include cognitive teaching to compensate for knowledge gaps.

Finally, given that LS tasks require surgeons to rely heavily on 2-dimensional visual cues and haptic cues to perform complex tasks within a 3-dimensional space, visual–spatial skills are likely to play an integral role in the performance of these tasks. Visual–spatial skills involve the perception and processing of spatial relationships, such as mental transformation (e.g., mental rotation) and relative distance perception. Hassan et al\textsuperscript{18} found that among novices, visual–spatial perception is associated with manual skills performed on a laparoscopic skills virtual reality (VR) simulator; novice participants with a high degree of spatial perception performed laparoscopic VR tasks faster than those who had a low degree of spatial perception and also scored better for economy of motion, tissue damage, and total error. Ritter et al\textsuperscript{19} demonstrated that spatial perceptual abilities correlated well with duration of the learning curve (i.e., number of trials required to meet the specified proficiency criterion) on a validated VR flexible endoscopy simulator. Visual–spatial ability assessments included determination of the 3-dimensional orientation of a 2-dimensional grayscale cube via the pictorial surface orientation (PicSOR) test and the Cube Comparison test, which involves mental rotation of an object about its center. The PicSOR test was also shown in three studies by Gallagher et al\textsuperscript{20} to consistently predict performance on the FLS circle cutting task, as well as significantly predicting laparoscopic surgeons’ performance. Keehner et al\textsuperscript{21} also examined changes in performance as novices learned to use an angled laparoscope within a virtual environment; initial performance showed considerable variability among novices, with performance related to both general, nonverbal reasoning ability (assessed via the abstract reasoning task of the Differential Aptitude Test battery) and spatial abilities (assessed using the Mental Rotation Test and Visualization of Views Test). As learning progressed, the correlation of performance with general reasoning ability diminished after the first few sessions, whereas by contrast, the significant correlation with spatial ability persisted even after the group variance had diminished. This finding provided further support for a previous study by Keehner et al,\textsuperscript{22} which found a significant correlation between spatial ability and interoperative videoscopic skills performed on animals for a novice group; however, no significant correlation was found within a group of experienced surgeons. Thus, the importance of spatial ability in performance of laparoscopic skills seems to diminish with experience.

**SECTION 2: CRITICAL SKILLS**

The first step in creating a system to train and refresh LS skills is to identify and model the critical underlying skills. Currently, no standard method exists for identifying critical skills related to laparoscopy. Current understanding is also limited on how best to refresh decayed skills, and no standard refresher training currently exists. Previous task analyses of (MIS) procedures have focused on psychomotor skills.\textsuperscript{23} However, the psychomotor ability to operate the laparoscopic tools represents only a portion of the skills needed to successfully perform MIS. Visual–spatial skills are needed to orient within the abdomen and recognize structures such as tissue, organs, ducts, and blood vessels. Cognitive skill is needed to maintain proper situational awareness, understand the ramifications of observations, make the correct decisions, and act in...
the best interest of the patient and the operation. In addition, different surgeons perform procedures with different techniques, based on varying degrees of expertise and experience, as well as training background. For example, establishing initial laparoscope and instrument access during LS cholecystectomy can be performed using the Hassan technique or using the Veress needle technique, and expert opinions differ on which is optimal. Therefore, the model of skills must account for both normative errors and actions that are not correct in any of these techniques, as well as quasi-normative errors, which consist of actions that are correct for some techniques but not others. The following details two complementary approaches to identifying and modeling the critical skills underlying LS proficiency, with the goal of developing improved training, assessment, and maintenance of these skills.

**LASTS Critical Skills**

The LASTS uses models of surgical skill acquisition and decay, objective methods of surgical skill assessment, and individual skill and training models to maximize training effectiveness and minimize skill decay by creating training curricula customized to each surgeon’s knowledge and skills. The LASTS skill model design goals are detailed enough to represent the skills needed for surgical procedures, moment-by-moment, and are sufficiently powerful to support an assessment of the skills that will differentiate experts from novices. The model accounts for variations in the skills required by different procedures by expanding on previous hierarchical models of LS skills and representing the skill hierarchy of each procedure. Surgical procedures are decomposed into a tree of steps, tasks, subtasks, and where appropriate, into motions. The tree describes a choreographed sequence of motions, views, information elements, and decisions, as well as deviations for nonstandard anatomy, emergencies, or repairs for adverse events. The skill model includes psychomotor, visual–spatial, and cognitive skills, and uses annotations to represent the goals, dangers, techniques, expectations, information requirements, and situational awareness needed by the step, task, or subtask. In addition to the base actions of the procedure, there may be a need to perform optional and repair actions, which are not necessary in every case. For example, a small amount of bleeding often occurs when detaching the gallbladder from the liver in LS cholecystectomy, and this may require cauterization or other repair at any point during the removal step. These action specifications also include annotations that indicate when the action should be performed.

Figure 2 shows a portion of the skill tree for the “Remove Gallbladder” step of a cholecystectomy procedure. Our skill tree graphics include the set of steps for the procedure (each page is one step, so this figure shows a single step), as well as the tasks (the black boxes) and subtasks (the white boxes) of each step. Tasks and subtasks are linked via decomposition links (solid lines) and ordering links (dotted lines with arrows). Not depicted in these graphics are the possible deviations and repair steps. Because of the diagram’s space constraints and because these visualizations are meant as an overview of the procedure only, we do not show the annotations in this graphic. However, the model could be presented with an overlay of the cognitive complexity, motor complexity, or danger ratings to show how specific conditions or complications can affect the skills needed by the procedure.

**SUSTAIN Critical Skills**

In our ongoing research and development effort, we are supplementing the LASTS cognitive task analysis with empirical skill acquisition and retention data to inform the LASTS skill models, as well as developing and validating a SUSTAIN that includes novel, objective assessment metrics and training modules within a modular design to support training within a variety of environments, including deployed settings.

We focused initially on identifying assessment metrics that could assess LS skill acquisition, proficiency, and decay. Our initial investigations have focused on assessment of the current tasks and metrics used to train and assess LS psychomotor skills. Later investigations will leverage the outcomes of the LASTS cognitive skills identification.

The manual skills testing component of the FLS training curriculum is intended to train and measure psychomotor skill performance during basic laparoscopic surgical maneuvers; however, some visual–spatial skills are involved as well. Figure 3 provides a task description and decomposition of the constituent psychomotor and visual–spatial skills involved in each of the FLS manual skills tasks. Additionally, the most challenging aspect of each task and common strategies used were identified based on expert interviews and are included in Figure 3. Currently, the cognitive skills associated with laparoscopic surgery are trained separately within the FLS didactic program, and are assessed via a written test. Based on skill decay literature, it is cognitive skills that decay most rapidly and therefore are most susceptible to decay during periods of nonuse, whereas psychomotor skills decay at slower rates. Therefore, it is not surprising that the FLS manual skills retention studies have shown little decay over long periods of retention. However, the current metrics for assessing the FLS psychomotor skills are limited primarily to time to complete each task, and secondarily to overt errors. Some research within the domain of robot-assisted laparoscopic surgery has begun to examine novel and objective metrics for assessing similar skills. We suggest that more sensitive metrics may identify early indications of skill decay, and provide a means for assessing and refreshing training during and following deployments.

We propose that additional research is needed to examine the retention of the relevant perceptual and cognitive skills,
and also to explore potential novel objective metrics related to psychomotor skill acquisition and decay.

SECTION 3: OBJECTIVE METRICS

To develop, validate, and effectively use models of surgical skill for training and refreshing, objective metrics are needed that can determine a surgeon’s level of expertise with respect to the identified critical surgical skills. The surgical skill assessment must reliably differentiate expert surgeons from novices. An assessment that cannot make this distinction is not valid as a tool for understanding, training, and certifying surgical skills. The skill assessment metrics must provide measures of the trainee’s skill level on the spectrum from novice to expert to enable the selection of training materials and practice tasks for an individual profile. Ideally, these materials and tasks will meet the surgical trainee at his or her current skill level and provide an opportunity to improve. Metrics must also effectively measure the underlying psychomotor, visual–spatial, and cognitive skills associated with LS proficiency, both independently and in conjunction with one another. Finally, the established metrics must be objective and empirically validated.

LASTS Objective Metrics

LASTS’s design builds on the FLS skill measurement approach and leverages measures recorded automatically in simulation, such as speed, accuracy, or smoothness of motion based on tracking instruments and task outcome. For example, to assess psychomotor skills, the surgeon performs...
a physical task, such as suturing, using a virtual trainer. To provide a comprehensive view of surgical skills, LASTS is designed to assess not only psychomotor skills but also visual–spatial and cognitive skills (yet to be determined) as well. Visual–spatial tasks include manipulating a virtual laparoscope to obtain the proper view on an organ; finding structures, such as a gallbladder, from video of LS; and recognizing abnormal conditions, such as cirrhosis of the liver, from video of LS. Cognitive knowledge can be assessed via test questionnaires, such as the multiple choice questions used by the FLS didactic exam. A number of decision tasks can be derived from the model of surgical skills for the procedure, such as where to place trocars for patients of varying body habitus (i.e., overweight, underweight, or normal weight) or whether to convert to open surgery after watching a video of a complication event. These measures of decision tasks can be formatted into multiple choice or select-on-a-diagram forms, or the measures may occur during a training procedure within the context of a simulated scenario.

In addition to assessing individual skill performance, LASTS’s design also assesses performance of combined skills through tasks that incorporate two or more skills simultaneously. An example of a combined psychomotor and cognitive task is dissecting the gallbladder from the liver during a simulated LS cholecystectomy procedure. LASTS specifies that the surgeon must understand the cognitive elements of

<table>
<thead>
<tr>
<th>Task 1: Peg Transfer</th>
<th>Psychomotor Skills Involved</th>
<th>Visual-spatial Skills Involved</th>
<th>Challenging Aspects</th>
<th>Strategies Employed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eye-hand coordination, Ambidexterity, Bimanual skills (transferring), Control precision, Manual dexterity, Arm-hand steadiness</td>
<td>Depth Perception, Spatial processing, Spatial relationship, Near vision, Color discrimination</td>
<td>Time, Ambidexterity</td>
<td>Strategically grasping and placing peg for security and efficiency of movement</td>
<td></td>
</tr>
</tbody>
</table>

| Task 2: Pattern Cut | Application of tension / counter-tension | Depth Perception, Near vision, Shape processing | Precision when cutting at difficult angles | Approaching from angle, Switching cutting hand or direction part-way through, Distance and angle to apply tension |

| Task 3: Endoloop | Eye-hand coordination, Bimanual skills, Control precision, Manual dexterity, Arm-hand steadiness, Approximation and not stranulation | Depth perception, Spatial processing, Spatial relationships | Preventing loop from tightening prior to placement on the targeted line | Tightening loop with one hand, Pressing down on targeted line with applicator while tightening, Knowing when to apply one endoloop or two |

| Task 4: Extracorporeal Knot | Eye-hand coordination, Bimanual skills, Control precision, Manual and finger dexterity, Arm-hand steadiness, Approximation and not stranulation | Depth perception, Near vision, Spatial processing, Combining multiple perspectives | Proper needle grasping, Precise suture placement | Maintaining tension in both ends of thread without avulsing penrose drain, Understanding and pattern recognition of various types of knots |

| Task 5: Intracorporeal Knot | Eye-hand coordination, Bimanual skills, Control precision, Manual and finger dexterity, Arm-hand steadiness, Approximation and not stranulation | Depth perception, Near vision, Spatial processing, Spatial relationships | Proper needle grasping and alignment, Precise suture placement, Use of short stitch | Beginning with suture in non-dominant hand, Creating and maintaining a short tail, Aligning heel of needle with driver the stitch is wrapping around |

FIGURE 3. FLS manual skills task decomposition.
the task—the goals, risks, techniques, expectations, and situational awareness—and use those elements to modulate the psychomotor movements of the task. For example, in some cases during a specific step, the surgeon must be careful not to unintentionally puncture the gallbladder or the liver causing excessive bleeding. Finally, LASTS’s design assesses combined psychomotor, visual–spatial, and cognitive skill by simulating multiple steps within a procedure. During the simulation, LASTS uses the skill tree to assess the performance of each step, task, and subtask independently. Specific scoring mechanisms are in development, which will take into account correlations between both individual and combined skills with task performance outcomes.

**SUSTAIN Objective Metrics**

Several objective metrics have been identified that may support effective assessment of LS skills acquisition, proficiency, and decay, including simulator-based metrics, motion- and vision-tracking metrics, and cognitive assessment metrics. For example, laparoscopic virtual environment training systems have incorporated a number of automated performance assessment metrics, such as task completion time, economy of motion, instrument collisions, peak instrument force, and peak strain. While further validation of such metrics is needed, including demonstration of predictive validity within live tissue and cadaveric models, these metrics may provide useful objective assessment and tracking tools if integrated into the FLS video trainer and/or curriculum. Laparoscopic simulation trainers can detect various motions as part of their mechanical input and visual feedback systems; however, the current FLS trainer boxes are relatively low-tech and do not include instrumentation tracking capabilities. Although the low-tech nature of the current FLS training system supports low-cost training, the goal of the SUSTAIN research and development effort is to investigate potential objective metrics that might provide insight into FLS skills acquisition, proficiency, and decay, including more sensitive metrics of psychomotor skills to enable detection of skill decay despite slower decay rates of these skills as compared to cognitive skills.

This effort seeks to develop metrics based on instrument tracking, as well as metrics that are novel within this domain, such as motion tracking via wearable instruments (i.e., gloves), and additional metrics to support assessment of perceptual and cognitive skills, both independently and in conjunction with the FLS psychomotor skills tasks.

A wide variety of motion-tracking technologies, such as instrumented gloves, currently exist for applications ranging from graphical animation to recognition of sign language. Data that may be extracted from these technologies include hand tremors, tool grip, and tool state (e.g., activating tool, rotating tool), as well as movement certainty. Instrumentation of the LS surgical tools may provide valuable data, such as time metrics, efficiency metrics, and task-specific metrics (e.g., object grasping, transfer, and placing times), as well as efficiency metrics, such as overall efficiency (left hand, right hand, average), pick-up efficiency (left hand, right hand, reverse, average), passing efficiency (left to right, right to left, average), and placing efficiency (left hand, right hand, average).

Vision tracking technologies are used in a variety of domains to provide insight into human–computer interaction, perceptual processing, and cognitive process. Vision tracking technologies have become increasingly affordable and unobtrusive, and their use has increased in both research and operational applications as a result. Vision tracking enables researchers to identify visual scan and fixation patterns during perceptual–cognitive and perceptual–motor tasks, providing insights into learning. In addition to enabling researchers to study learning patterns to develop more robust models of skill acquisition and decay, vision tracking can also enable surgical instructors and training proctors to see where trainees are focusing their visual attention during task completion in real time.

A variety of cognitive assessment technologies and strategies exist that incorporate indicators of cognitive workload and learning, such as concurrent performance of a secondary task. 

A subset of the metrics identified, such as hand and instrument tracking, are currently being empirically tested, and will be considered for inclusion in the SUSTAIN prototype system. In addition, we will consider cognitive assessment metrics such as assessing the ability to concurrently perform a secondary task more effectively as performance on the primary task of interest becomes more automated. While Stefanidis et al demonstrated that secondary task performance provides a means for differentiating between individuals of variable LS expertise when standard performance measures fail to do so, and that training to automaticity using a secondary task correlates to improved intraoperative performance, the secondary task used in these studies was not directly related to surgical skills. It is possible that by incorporating a secondary task that addresses intraoperative skills, including cognitive skills, training would be enhanced further, providing justification for the extended time associated with training to automaticity.

**SECTION 4: SKILL DECAY CURVES**

Reviews of the skill decay literature have identified a core set of factors that influence the decay of trained skills. These factors include (a) length of the nonuse interval, (b) degree of overlearning (training beyond mastery), (c) task characteristics (e.g., psychomotor versus cognitive, number of steps involved), (d) method of assessing original acquisition and retention (i.e., type of test), (e) condition of retrieval (e.g., recall versus recognition), (f) instructional strategies and training methods, (g) individual differences (e.g., spatial ability), and (h) motivation. Modeling these factors within skill decay curves helps to predict skill degradation over time and introduce refresher training at appropriate intervals.
Skill acquisition curves have been developed, for example, based on the relationship between the time to perform a task and the number of practice attempts. Ebbinghaus proposed the first formal decay curve, which he called the “forgetting curve.” This model, shown in Figure 4, demonstrates that the rate of forgetting depends on several factors, but that typically single trial learning results in exponential skill decay, with each additional learning trial resulting in increased retention. The solid line indicates the rate of forgetting for a single learning trial; each dashed line to the right represents the rate of forgetting for additional learning trials on subsequent days.

LASTS Skill Decay Model
The skill decay models used in LASTS predict how surgical skills change over time and practice sessions, in relation to specific learning and decay factors. These models have three related parts: (1) how skills are initially acquired through training, including factors such as the order, duration, and repetition of training, as well as individual trainee differences, such as skill level and experience; (2) how skills decay over time through nonuse; and (3) how skills are reacquired through refresher courses. It is important to note that reacquisition is different from the initial acquisition because of previous exposure to the skills. The model includes three skill curves for each skill assessed: (1) learning, (2) decay, and (3) relearning. The learning and relearning curves measure skill level against training attempts (i.e., number of times performing the training task). LASTS includes these factors because they have been shown by Arthur et al to modulate performance. A learning curve shows how a surgeon’s skill improves with each attempt. The decay curves measure skill level over time.

Learning curves for LS are typically defined in terms of the time it takes to complete a task versus how many sessions have been trained. LASTS extends the task completion time measure to all of the factors that comprise the assessment methodology to create a unique set of learning curves for each procedure, step, and task. A key contribution of LASTS is the ability to combine these curves across skills. We expect that initial learning rates (i.e., time to reach proficiency) will decrease and decay rates will increase when the surgeons are assessed on multiple skills simultaneously. Based on our initial understanding of the literature on multiple learning curves, our hypothesis is that the curves will all have a similar shape because proficiency at a given skill will increase with each trial until a plateau is reached. The exact values of the curves are dependent on the skill as well as the student’s experience level and past performance. On average, decay curves show decreasing proficiency over time, as the skill decays.

SUSTAIN Skill Decay Model
The skill decay model proposed for SUSTAIN is both theoretical and empirical, and will provide input to the LASTS computational decay models. To begin developing an empirical model of LS skills that includes novel metrics, we conducted a pilot study in which both traditional (i.e., time and error) and novel (i.e., instrumented gloved hand motion tracking) FLS assessment metrics were collected for 11 medical students with no prior exposure to the FLS training curriculum and one expert laparoscopic surgeon. FLS manual skills performance data were collected for the novice participants before training (pretest), following completion of training to proficiency (post-test), and following an 8- to 10-week retention period (follow-up). Novice spatial abilities were also measured via three validated computer tests that assessed egocentric spatial ability, allocentric spatial ability, and mental rotation. Despite a small sample size, based on traditional FLS scores, novice performance was shown to be significantly higher at post-test, as compared to pretest; FLS scores decreased at follow-up, but did not differ significantly from post-test scores. A significant difference was also found between the novice pretest and expert’s scores for the five tasks, but not between the novice post-test and expert’s scores, or between the novice follow-up and expert’s scores; however, the novice post-test and follow-up scores would likely be significantly differentiated from expert’s scores with a larger sample population. These pilot data are indicative of skill acquisition and decay trends, providing the basis for an initial empirical skill model. We also established initial technical feasibility for the use of instrumented glove motion tracking to assess smoothness of surgeon hand motions, and for the incorporation of these novel metrics within a skill decay model. Although limited by the small sample size, this pilot study provided a basis for an initial model of laparoscopic surgical skill acquisition and decay that incorporates a variety of metrics, which will be further refined and expanded to include cognitive, perceptual, and psychomotor aspects, and will be validated under ongoing empirical research efforts.
SECTION 5: DESIGN OF REFRESHER TRAINING

Stefanidis and Heniford\textsuperscript{33} suggest that a successful laparoscopic skills curriculum should “encompass goal-oriented training, sensitive and objective performance metrics, appropriate methods of instruction and feedback, deliberate, distributed, and variable practice, an amount of overtraining, maintenance training, and a cognitive component” (p. 77).

The LASTS approach includes a curriculum generator, which will use the LASTS and SUSTAIN assessment methods, learning curves, and decay curves to create a curriculum of refresher training for a specific surgeon based on that surgeon’s initial skill level across various tasks and the length of retention interval for various skills and tasks. At this time, we have designed the framework of the curriculum generator, which we describe here, but we have not yet implemented its components. The curriculum generator will leverage the skill decay curves developed, which predict the rate at which various skills will decay, and will be designed to use a surgeon’s training history, history of procedures performed, and skills assessments to determine which skills are most likely to require refreshing for a specific procedure. The curriculum generator then creates a curriculum to refresh these skills. Because the curriculum focuses on refreshing the skills that have decayed for the surgeon (and not all skills), it decreases retraining time and addresses potential problems proactively. This curriculum may then be used as one part of a surgical training, certification, and recertification program. By addressing the factors of the skill decay curves that most influence decay, the curriculum generator may be able to train new surgeons and retrain experts to build robust skills that decay slowly, if at all.

For each student, the curriculum generator creates a training regimen and keeps a profile of the trainee to assess progress. The trainee can be tracked through the training program to recommend when they are ready for operating room experience, and to test this recommendation at computed intervals in collaboration with attending surgeons’ review. The system first determines the trainee’s skill levels using our assessment methods. It then selects training elements, such as didactics and tasks, from a library and generates new simulation training scenarios as needed. For a deploying surgeon, the system takes a last predeployment reading on skill level, and then receives the record of procedures performed during deployment. Additional information such as a surgeon’s training and individual competency as rated by supervisors is collected and inputted into the curriculum generator. With this information, the curriculum generator develops a refresher course. For example, a surgeon that performed few procedures, had little access to refresher material, and had long periods of time without operations will require more extensive refresher training than a surgeon that performed many operations, refreshing their skills with related procedures.

Psychomotor and visual–spatial skills are modeled by continuous values along multiple skill dimensions, and training elements are selected to challenge the surgeon at his current estimated skill levels using item response theory. Cognitive skills are modeled as knowledge bases of correct and incorrect rules, and training elements are selected that extend these knowledge bases using intelligent tutoring systems techniques. Finally, the curriculum generator generates surgical scenarios for practice on surgical simulators. It uses narrative generation techniques\textsuperscript{34,36} to present realistic scenarios that also accomplish training goals. LASTS determines which skills need refreshing by using the decay curves and continuous skill assessments.

To augment the individual training elements, LASTS uses a scenario-based training methodology, where surgical situations are presented to the trainees instead of abstract tasks. This approach is designed to enable a higher level of skill acquisition and decrease decay. Techniques in search-based interactive narrative and narrative planning models are particularly suited for this type of generation, since they can be used to generate a sequence of realistic scenarios (the narrative content) while optimizing the ability for the student to learn the content. Narrative planning approaches provide this functionality.\textsuperscript{32} The role of the narrative planning model is to create realistic, scenario-based training sessions that cover the skills in the necessary quantity and order.

The LASTS computational models and curriculum generator will be incorporated into the SUSTAIN prototype, which includes a modular design to support training within a variety of environments, including deployed settings.

SUMMARY AND CONCLUSIONS

Although the research literature on skill acquisition and decay is comprehensive and extensive, this is not true of the literature that pertains specifically to LS skill decay. Based on the current research literature, it is evident that a wide variety of factors related to task and training characteristics, retention intervals, and transfer task environments impact the rate of decay of various skill sets, and that subcomponents of skill sets decay at different rates. However, relatively little is known about the nature of LS skill retention. Empirical research is needed to develop and validate predictive skill decay models related to these highly specialized skills. This research will help provide guidelines for efficient and effective refresher training that can be implemented within standardized medical training curricula, particularly for military surgeons, to prevent laparoscopic skill deterioration during long deployment cycles.

The research questions yet to be addressed in this area are many: Are there different decay rates for different types of learning? Which instructional methods/strategies are the most effective for skill/knowledge acquisition to prevent skill decay? What are the most effective strategies for refresher training (reacquisition)? How do individual factors, such as spatial ability and handedness, impact skill decay?

Studies assessing decay of LS skills often do not assess cognitive skills or knowledge, which predicts decay. Also, the studies that address skill acquisition and decay suffer
Prevention of Surgical Skill Decay

from a lack of objective, reliable, and valid measures of skill level and decay, as well as a lack of data on reacquisition of LS skills. When acquisition is studied, assessments are usually in relation to new training techniques compared to a traditional method, and it is difficult to interpret and draw conclusions from these studies without objective, reliable, and valid metrics. Further, many of the studies assessing acquisition and retention are based on psychomotor tasks, and do not address cognitive or perceptual skills. To remedy this situation, we suggest the following approach: (1) develop and identify assessment metrics that are objective, reliable, and valid to assess these skills over time; (2) acquire a comprehensive understanding of the nature of laparoscopic surgical skill acquisition and decay, including identification of the requisite critical skills; (3) report skill decay curves that enable prediction of which critical skills decay and at what rate; (4) develop training strategies for simulation-based training that support rapid acquisition and long-term retention; and (5) develop retraining for sustainment of these critical and perishable skills.

An ongoing research and development effort has been funded by the Office of Naval Research and the Telemedicine and Advanced Technology Research Center to further the development and empirical validation of novel objective metrics for laparoscopic surgical skill acquisition, decay, and reacquisition model; design, develop, and validate a prototype simulation-based assessment and training module; and develop curriculum recommendations for the prevention of LS skills decay. The goal of this effort is to conduct research on LS to develop generalized methods, procedures, and training methods that could be used to prevent decay of other critical medical skills and knowledge.

ACKNOWLEDGMENTS

This work is being conducted in collaboration with the Uniformed Services University of the Health Sciences, the National Capital Simulation Center, and Harvard Medical School. This research is funded by the Office of Naval Research and the Telemedicine and Advanced Technology Research Center under two Small Business Innovation Research contracts, AnthroTronix, Inc. (N0001411CO420) and Charles River Analytics, Inc. (N0001411CO426). The work reported herein was also partially supported by a grant from the Office of Naval Research, Award Number N00014-10-1-0978.

REFERENCES

Prevention of Surgical Skill Decay


Effects of Simulation-Based Practice on Focused Assessment With Sonography for Trauma (FAST) Window Identification, Acquisition, and Diagnosis

Gregory K. W. K. Chung, PhD*; Ruth G. Gyllenhammer, MA*; Eva L. Baker, EdD†; Eric Savitsky, MD‡

ABSTRACT  We compared the effects of simulator-based virtual ultrasound scanning practice with classroom-based ultrasound scanning practice on participants’ knowledge of focused assessment with sonography for trauma (FAST) window quadrants and interpretation, and on participants’ performance on live patient FAST examinations. Novices with little or no ultrasound training experience received simulation-based practice (n = 24) or classroom-based practice (n = 24). Participants who received simulation-based practice scored significantly higher on interpreting static images of FAST windows. On live patient examinations where participants scanned the right upper quadrant (LUQ), left upper quadrant (LUQ), and suprapubic quadrant of a normal patient and an ascites-positive patient, the classroom-based practice condition had a shorter scan time for the LUQ and a higher number of participants attaining high-quality window on the RUQ (normal patient only) and suprapubic quadrant (positive patient only) and correct window interpretation on the LUQ (normal patient only). Overall, classroom-based practice appeared to promote physical acquisition skills and simulator-based practice appeared to promote window interpretation skills. Accurate window interpretation is critical to identification of blunt abdominal trauma injuries. The simulator used (SonoSimulator) appears promising as a training tool to increase probe time and to increase exposure to FAST windows reflecting various anatomy and disease states.

INTRODUCTION
The use of portable ultrasound is increasing in military settings, such as combat support hospitals as a triage and an evaluation tool. Ultrasonography complements standard evaluation techniques and can improve the speed and accuracy of diagnosis of blunt abdominal trauma (e.g., bruising or laceration to the liver resulting in localized internal bleeding). Unfamiliarity with ultrasonography, the cost of training users on ultrasound-guided procedures, and the lack of training opportunities are limiting the use of this beneficial technology. One potentially cost-effective method for providing users with ultrasound-guided procedural training is the use of simulator-based training. Some potential advantages over traditional medical training include: (i) it presents no risk to trainees or patients during practice attempts; (ii) it is more cost-effective than current training methods; (iii) it provides multiple modes of sensory interaction to maximize learning; and (iv) it aids reduction in skill decay. In this study, we compared simulation-based practice of ultrasound scanning to classroom-based practice of ultrasound scanning on both knowledge and performance measures. We focused on one type of procedure, the focused assessment with sonography for trauma (FAST) examination, as the context for the comparison.

FAST Examination
The FAST examination is an emergency ultrasound procedure that focuses on the detection of free fluid in hemoperitoneum, hemopericardium, pneumothorax, and hemothorax. Unlike other trauma screening procedures such as the physical examination, diagnostic peritoneal lavage, and the CT scan, the FAST examination is noninvasive, bedside, and repeatable; requires 5 minutes to complete; and does not require stable patients. These properties make portable ultrasound scanning an invaluable evaluation tool in combat support hospitals, where there is often a sudden intake of a large number of patients who have life-threatening blunt trauma injuries.

The standard windows of the FAST examination are the right upper quadrant (RUQ), left upper quadrant (LUQ), subxiphoid (or subcostal), and suprapubic. The window quadrants are examined for fluid in the gutters between the organs. Free fluid is consistent with internal bleeding resulting from blunt trauma injuries. In the RUQ (also called Morrison’s pouch), evidence of fluid may be found between the liver and the kidney. Fluid may also exist between the kidney and the spleen in the LUQ view or in the pelvis in the suprapubic view.

Cognitive Demands of FAST Window Acquisition and Interpretation
Recognition of anatomical landmarks is fundamental to ultrasound window acquisition. Both diagnostic medical sonographers...
Effects of Simulation-Based Practice on FAST

and emergency care physicians require training and proficiency in abdominal anatomy as a component of medical training and professional certification.8

Ultrasound scanning involves manipulating a probe or transducer against the patient’s body, wherein the probe acquires a 2-dimensional slice through a 3-dimensional anatomical volume. Minute changes to the pitch, yaw, and roll of the probe cause changes in the window image, and facile probe manipulation is required to acquire clear anatomical landmarks. This skill is not easily taught and precise verbal instructions relating the trainer’s actions to the image on screen can be difficult. Experienced sonographers use nuanced hand movements instinctively during image acquisition.9 The FAST examinations and ultrasonography in general are “operator dependent.”10 To make a correct diagnosis, the sonographer must first acquire an adequate window by using anatomical landmarks and probe manipulation techniques such as fanning. Complicating the task is that the detection of fluid may depend on positioning of either the patient or the probe, with subtle changes in either required for precise window analysis.7,11

Use of Ultrasound Simulators for Training

The goals of FAST training are for the trainee to be able to acquire an adequate window and to identify both the free fluid and the type of window when given a single ultrasound scan. A typical classroom curriculum for the FAST examination includes both didactic and hands-on instructions. Didactic training presents instruction on the principles of ultrasonography, an introduction to ultrasound mechanics or knobology, and discussion of the purpose, method, and interpretation of the FAST examination. Because window acquisition is dependent on probe movement, physical probe time is thought to be absolutely necessary to training but often the hands-on practice component is constrained because of limitations of practicing with model patients.7,12

Ideally, a trainee would gain experience by performing the FAST examination on as many cases as possible under the guidance of someone competent in the FAST examination. However, extended training experience with model patients is rarely practical. For example, scanning the LUQ, RUQ, and suprapubic quadrants would take 5 to 8 minutes in a training context per trainee. Assuming a typical class size of 20 students, with each student given two opportunities to scan, the total time required would be between 3 and 5 hours. Thus, using a classroom setting to provide trainees with practice with varying anatomy and varying free-fluid states remains impractical.

One of the greatest benefits of simulators may be the capability of the simulator to provide interactive learning opportunities across a large number of cases. Previous research points to the effectiveness of the simulator training in preparing trainees to both perform and interpret the FAST examination. For example, in a 4-hour FAST training course for emergency medicine resident physicians, the UltraSim sonographic (mannequin) training model was found to be as effective in preparing trainees to detect the presence of intraperitoneal free fluid in various FAST windows as the training with a live patient model (using peritoneal dialysis patients).7

In another study, the use of the SonoTrainer sonography simulation system enabled physicians trained with the system (vs. theoretical training alone) to make diagnoses of second-trimester fetal abnormalities with a detection rate of 86% and specificity of 100%.13

Various systems have been developed with a range of capabilities (e.g., UltraSim, VirUS, EchoComJ, and SONOSim3D, SonoTrainer).13 For the FAST examination, the simulator used in the current study (SonoSimulator)14–16 focuses on providing trainees the capability to visualize free fluid in two ways: as a single static image or as a continuously changing image similar to what would appear on an actual ultrasound machine window. However, we could find no research that directly compared simulator-based practice to classroom-based practice with respect to knowledge of given FAST windows and on execution of FAST examination procedures with live patients.

Research Question

The main research question was to what extent does practice using an ultrasound simulator with virtual patients affect participants’ knowledge and subsequent performance on live FAST examinations, compared to practice in a typical classroom context with a portable ultrasound machine and a live model patient? This study, when conceptualized in Kirkpatrick’s framework,17 would be considered an investigation of Level One: Reaction and Level Two: Learning.

METHOD

Design

A pretest/intervention/posttest control group design was used to examine the effects of the two practice conditions on participants’ knowledge of and performance on a FAST examination. Participants in the control condition received guidance and feedback from an instructor and hands-on practice with a model patient. Participants in the experimental condition practiced scanning on virtual patients using an ultrasound simulator, and received quality of scan and accuracy of diagnosis feedback from an ultrasound expert.

Sample

Forty-nine participants were recruited through e-mail announcements to two local medical schools and received $100 for participating. All procedures were approved by our institutional review board for research involving human subjects. One participant was an undergraduate student who was subsequently dropped.
Effects of Simulation-Based Practice on FAST

Forty-eight participants were randomly assigned to either the group-based practice condition \((n = 24)\) or the simulation-based practice condition \((n = 24)\). The mean age was 25.48 years \((SD = 3.09)\). Forty-four participants were medical students, three were nursing students, and one was a medical resident. The mean Medical College Admission Test \((MCAT)\) score \((n = 41)\) was 33.63 \((SD = 3.22)\). Twenty-eight participants reported no prior training in ultrasound procedures, 16 participants reported receiving 0 to 2 hours of training, and four participants reported receiving 2 to 4 hours of training. Fifteen participants reported receiving ultrasound scanning procedure training in a lecture or classroom and nine participants reported receiving some hands-on practice. Overall, the sample represented novices with limited or no experience with ultrasound scanning.

Model Patients
Patients with ascites were recruited as model patients, for the purpose of simulating the abdominal condition of patients with blunt trauma resulting in hemorrhage. Ascites is a pathological condition resulting from liver disease wherein fluid accumulates in the abdominal cavity.\(^{18,19}\) Data collection occurred over three occasions. For the first two occasions, the same model patient (moderate degree of free fluid) was used. Scheduling conflicts required the use of a second model patient (severe degree of free fluid) for the third data collection. Both patients were positive for free fluid in the RUQ, LUQ, and suprapubic regions.

Tasks
The four major tasks participants engaged in were tests of knowledge of the FAST examination, FAST examination instruction, FAST examination practice, and conducting FAST examinations on model patients.

Tests of Knowledge of the FAST Examination
Participants were administered a pretest of their knowledge of the FAST examination and abdominal anatomy, and a posttest of their knowledge of the FAST examination. The tests included items depicting normal and positive conditions revealed by a FAST examination.

FAST Examination Instruction
In this study, the FAST examination was limited to the RUQ (Morrison’s Pouch), LUQ (spleen), and the suprapubic (bladder) windows, as the model patients were ascites positive with free fluid in only these quadrants.

Participants viewed two instructional video modules in sequence.\(^{5,20}\) The first video was on the physics of ultrasound and general principles of sonography. The second video focused specifically on the FAST examination, describing how to view and interpret each of the four window quadrants: RUQ, LUQ, suprapubic, and subxiphoid. The video demonstrated a 3-minute FAST examination, discussed the implications of the mirror image artifact, discussed major pitfalls, and provided case study examples of conditions wherein free fluid is present.

FAST Examination Practice
Participants in the control condition received classroom-based practice and those in the experimental condition received simulator-based practice.

Classroom-based practice
The instructor provided participants 5 minutes of instruction on the basics of machine operation followed by a demonstration of the FAST examination. Each participant then practiced the FAST examination on the normal model patient while the instructor provided guidance and feedback (RUQ, LUQ, suprapubic quadrant). The instructor demonstrated anatomic landmarks in each quadrant and other procedures consistent with typical training procedure.\(^{7,21}\) Each participant was given two opportunities to conduct the FAST examination. One hour was allotted and used for the classroom practice session.

Simulator-based practice
The simulator system we used, SonoSimulator, was developed by Pelagique (Santa Monica, California). The virtual patients used in SonoSimulator were modeled with real-patient ultrasound scans of both normal and pathologic cases with a generic ultrasound design. The simulator database contained a range of cases with varying amounts of free fluid: absence of fluid, and minimal, moderate, and severe amounts of fluid. The variety of patients used to populate the virtual-patient database also provided scans with a range of anatomy.

The simulator presented up to 10 cases with varying levels of normality and severity. Participants manipulated the probe and the probe movements were reflected in a scan window that was updated in real time in a real ultrasound machine. Figure 1 shows a screenshot of the user interface for the RUQ.

Each case in the simulation was subdivided into three views: RUQ, LUQ, and suprapubic. Participants were told to find the ideal diagnostic window and then “freeze” the scan and provide a diagnosis. A report with the participant’s response was then printed and given to the expert sonographer. The expert sonographer evaluated the participant’s diagnosis and window quality and determined whether the participant should advance to the next view. Criteria for advancement were an accurate diagnosis and an excellent or fair window. One hour was allotted and used for the simulator-based practice session.

FAST Performance Test
The performance test required participants to conduct a FAST examination on two model patients (one normal, one ascites positive). No help or feedback was given to the participants. Participants used a portable ultrasound machine (the M-Turbo,\(^{14}\) MicroMaxx,\(^{15}\) or S-FAST\(^{16}\)). Images of the ultrasound window were captured to disk and later evaluated.
by an expert sonographer. During the first data collection we observed a few patients who could not find the appropriate landmarks and took an excessively long time to complete a scan. Thus, on the second and third data collection occasions we imposed a time limit of 2 minutes each for the RUQ and suprapubic scans, and 4 minutes for the LUQ.

**Measures**

Two types of measures were developed for this study: (i) knowledge-based measures used to evaluate participants’ knowledge of FAST-examination-related concepts; and (ii) performance-based measures used to evaluate participants’ skill at executing a FAST examination.

**Knowledge-Based Measures**

Knowledge-based measures were derived from information covered in the instructional videos\(^5\)\(^{20}\) as well as FAST-examination-specific concepts. Three broad areas were sampled: prior knowledge of anatomy, basic FAST examination procedures, and window interpretation. The sources were consulted to develop or adapt items for the knowledge measures: instructional materials (e.g., textbooks, guidebooks, and prior research),\(^22\)\(^{–25}\) and the computer-based instruction\(^5\) used in the study. Items were reviewed by various experts (director of emergency ultrasound, director of ultrasound and breast imaging, director of an emergency ultrasound department, an experienced sonographer, and two emergency physicians). Table I shows the final distribution of the items by topic and the scale reliabilities.

The knowledge measures were embedded in a 69-item pretest, which included items related to the FAST examination (15 constructed response items and 46 selected response items) and participants’ knowledge of abdominal anatomy (8 selected response items). The anatomy items served as a check of prior knowledge; participants were assumed to have knowledge of abdominal anatomy and thus these items were not of interest on the posttest. The posttest contained the same FAST examination questions without the anatomy questions, in addition to questions about participants’ backgrounds.

As may be seen in Table I, the pretest \(\alpha\) reliabilities were acceptable \((\alpha > 0.70)\) except for identification of abdominal organs \((\alpha = 0.46)\). The posttest \(\alpha\) reliabilities, however, were much lower. The drop in item reliability can be explained by the decrease in variance from the pretest to posttest that is

**TABLE I.** Knowledge-Based Measures \((N = 48)\)

<table>
<thead>
<tr>
<th>Scales</th>
<th>No. of Items</th>
<th>Internal Consistency (Cronbach’s (\alpha))</th>
<th>Test–Retest Reliability</th>
<th>Equality of Variance Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pretest</td>
<td>Posttest</td>
<td></td>
</tr>
<tr>
<td>Prior Knowledge of Anatomy</td>
<td>8</td>
<td>0.45</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Basic FAST Scanning Procedures</td>
<td>17</td>
<td>0.70</td>
<td>0.44</td>
<td>0.47</td>
</tr>
<tr>
<td>Anatomical Interpretation of FAST Windows</td>
<td>16</td>
<td>0.79</td>
<td>0.53</td>
<td>0.22</td>
</tr>
<tr>
<td>Window Interpretation</td>
<td>14</td>
<td>0.91</td>
<td>0.57</td>
<td>0.46</td>
</tr>
<tr>
<td>Identification of FAST Window Quadrants</td>
<td>14</td>
<td>0.85</td>
<td>0.23</td>
<td>0.19</td>
</tr>
</tbody>
</table>
consistent with participants’ performance becoming more uniform after training and practice. The variance on each posttest scale was significantly lower than the pretest scale. For example, on 6 items in the posttest identification of FAST window quadrants scale, over 85% of the participants got the item correct compared with 16% of the participants on the pretest. The low test–retest reliability suggests that the rank ordering of participants changed from pretest to posttest.

Performance-Based Measures (Live Patient FAST Examination)

Window acquisition time
The window acquisition time was measured with a stopwatch and represented the period between first contact of the probe with the model patient’s body and when the participant said “stop” to indicate an adequate window or the participant’s judgment that he or she could not acquire the window. Quadrant scan time was computed for the RUQ, LUQ, and suprapubic quadrant, summing across the two model patients. Cronbach’s \( \alpha \) was 0.56 for the RUQ, 0.81 for the LUQ, and 0.86 for the suprapubic quadrant.

Window quality
For each acquired window, an expert evaluated the quality of the window. The window was rated as “excellent, fair, poor, or other.” “Other” captured situations where the window acquired was nondiagnostic. Window quality was dichotomized into two categories, “excellent or not excellent,” and subsequent analyses examined the number of participants that acquired “excellent” windows by quadrant (RUQ, LUQ, and suprapubic quadrant) and patient type (normal, positive).

Diagnostic accuracy
For each acquired window, the participant rendered a diagnosis of that window. An expert evaluated the quality of the diagnosis. The diagnosis was rated as “correct, incorrect, or other.” “Other” captured situations where the window acquired was nondiagnostic. For analysis purposes, diagnostic accuracy was dichotomized into two categories, “correct or not correct,” and subsequent analyses examined the number of participants that interpreted the window correctly by quadrant (RUQ, LUQ, and suprapubic quadrant) and patient type (normal, positive).

Perceived Utility of Training
Participants were asked three questions about how well they thought the ultrasound practice session (either with the simulator or in the group session) prepared them to perform a “real” ultrasound examination of a live patient with respect to (i) probe manipulation, (ii) window acquisition, and (iii) window diagnosis. For each question, participants were asked to rate the amount of practice received as being very inadequate, inadequate, adequate, or too much. Participants were also asked for written comments related to each question.

Background Information
Data regarding demographic information and prior experience with ultrasound training were gathered. Participants were asked their age, gender, current position, and MCAT scores. Participants were also asked what type of prior ultrasound training they had received and the number of hours spent training on ultrasound procedures.

Procedure
Participants were introduced to the study and then were given the pretest of FAST knowledge (30 minutes). A scheduling error resulted in four participants not being administered the pretest. Participants were then given 120 minutes to view instructional videos on the physics of sonography and the FAST examination procedure. Participants received practice in either a classroom setting or a simulation setting (65 minutes). Following the practice session, participants were then required to conduct a live FAST examination on two patients (40 minutes). After the live patient examination, participants were given the knowledge posttest, filled out the feedback form, and completed paperwork to receive payment (40 minutes). The entire protocol took 285 minutes (4 hours and 45 minutes).

RESULTS

Preliminary Analyses
Prior to the main analyses, the conditions were tested for differences on self-reported number of hours of ultrasound training, MCAT scores, and pretest measures; no differences were found, which suggests the conditions were similar on their experience and knowledge of ultrasound scanning procedures.

Table II shows descriptive statistics for the knowledge measures, Table III shows descriptive statistics for the performance measures, and Table IV shows correlations among and between the posttest knowledge measures and performance measures. In general, we expected posttest scores to be higher than pretest scores because our participants were novices and they received instruction and practice on ultrasound scanning procedures. We also expected the simulator-based practice condition to score higher on interpreting windows because of the greater amount of practice on interpreting windows available with the ultrasound simulator compared with the classroom-based practice condition. Finally, we expected the knowledge measures in general to relate to performance measures, particularly on window quality and window interpretation.

As shown in Table II, participants’ posttest scores are much lower than their posttest scores, especially on the scales related to FAST windows. Scores on the performance measures show high variability for the time-based measures, which suggests nonuniform training outcomes. As shown in Table IV, in general, knowledge measures did not correlate
with performance measures. The exception was diagnostic skill. The more correct interpretations of static (paper) windows, the more correct interpretation of live patient scans ($r(47) = 0.34, p < 0.05$). No other knowledge measure was related to any of the performance measures.

Finally, we checked whether participants benefited from the instructional videos and ultrasound scanning practice. Learning the critical ultrasound knowledge components was presumed to be essential to conducting live patient scans and window interpretation. Separate paired t-tests were conducted on each knowledge measure, as shown in Table V. Participants, independent of condition, learned the content very well as indicated by effect sizes over 1.0. Effect sizes over 1.0 are rare in training research. In general, if the pretest was at the 50th percentile, an effect size of 1.0 would increase the percentile score from 50 to 84.

### TABLE II. Descriptive Statistics of Knowledge Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Max. Possible</th>
<th>Control (Classroom-Based Practice)</th>
<th>Experimental (Simulator-Based Practice)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification of Abdominal Anatomy</td>
<td>8</td>
<td>M: 6.50, SD: 1.22, Min: 3.00, Max: 8.00</td>
<td>M: 6.65, SD: 1.09, Min: 5.00, Max: 8.00</td>
</tr>
<tr>
<td>Basic FAST Scanning Procedures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>17</td>
<td>M: 6.79, SD: 2.90, Min: 0.00, Max: 11.00</td>
<td>M: 6.45, SD: 2.96, Min: 3.00, Max: 13.00</td>
</tr>
<tr>
<td>Posttest</td>
<td>17</td>
<td>M: 14.25, SD: 1.51, Min: 11.00, Max: 17.00</td>
<td>M: 14.58, SD: 1.50, Min: 11.00, Max: 17.00</td>
</tr>
<tr>
<td>Anatomical Interpretation of FAST Window</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>16</td>
<td>M: 5.83, SD: 2.91, Min: 0.00, Max: 13.00</td>
<td>M: 5.10, SD: 2.13, Min: 0.00, Max: 8.00</td>
</tr>
<tr>
<td>Posttest</td>
<td>16</td>
<td>M: 12.83, SD: 1.79, Min: 10.00, Max: 16.00</td>
<td>M: 12.25, SD: 2.17, Min: 6.00, Max: 16.00</td>
</tr>
<tr>
<td>Identification of FAST Window Quadrant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>14</td>
<td>M: 1.75, SD: 2.92, Min: 0.00, Max: 9.00</td>
<td>M: 1.35, SD: 2.91, Min: 0.00, Max: 11.00</td>
</tr>
<tr>
<td>Posttest</td>
<td>14</td>
<td>M: 8.50, SD: 2.40, Min: 3.00, Max: 12.00</td>
<td>M: 9.25, SD: 1.80, Min: 7.00, Max: 13.00</td>
</tr>
<tr>
<td>Diagnostic Interpretation of FAST Window</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>14</td>
<td>M: 1.92, SD: 2.89, Min: 0.00, Max: 11.00</td>
<td>M: 0.90, SD: 1.62, Min: 0.00, Max: 5.00</td>
</tr>
<tr>
<td>Posttest</td>
<td>14</td>
<td>M: 8.46, SD: 1.41, Min: 6.00, Max: 11.00</td>
<td>M: 9.88, SD: 1.70, Min: 7.00, Max: 12.00</td>
</tr>
</tbody>
</table>

*Summed across RUQ, LUQ, and suprapubic quadrants and two patients. 

### TABLE III. Descriptive Statistics of Performance Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Max. Possible</th>
<th>Control (Classroom-Based Practice)</th>
<th>Experimental (Simulator-Based Practice)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic Interpretation of FAST Window</td>
<td>6</td>
<td>M: 4.13, SD: 1.19, Min: 2.00, Max: 6.00</td>
<td>M: 3.83, SD: 1.13, Min: 2.00, Max: 6.00</td>
</tr>
<tr>
<td>Acquisition of FAST Window</td>
<td>6</td>
<td>M: 3.33, SD: 1.58, Min: 0.00, Max: 6.00</td>
<td>M: 3.33, SD: 1.56, Min: 0.00, Max: 6.00</td>
</tr>
<tr>
<td>Window Scan Time</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUQ Scan Time (s)</td>
<td>240 (s)</td>
<td>M: 125.71, SD: 133.04, Min: 17.00, Max: 548.00</td>
<td>M: 142.75, SD: 62.13, Min: 33.00, Max: 257.00</td>
</tr>
<tr>
<td>LUQ Scan Time (s)</td>
<td>480 (s)</td>
<td>M: 155.00, SD: 132.17, Min: 25.00, Max: 548.00</td>
<td>M: 271.08, SD: 181.85, Min: 53.00, Max: 673.00</td>
</tr>
<tr>
<td>Suprapubic Scan Time (s)</td>
<td>240 (s)</td>
<td>M: 68.87, SD: 105.76, Min: 22.00, Max: 548.00</td>
<td>M: 88.38, SD: 48.82, Min: 9.00, Max: 180.00</td>
</tr>
</tbody>
</table>

*Summed across RUQ, LUQ, and suprapubic quadrants and two patients. A time limit was imposed on scan time after the first data collection, where we observed a few participants taking an excessively long time to complete a scan.

### TABLE IV. Intercorrelations (Pearson) Among Posttest Knowledge and Performance Measures (N = 48)

<table>
<thead>
<tr>
<th>Posttest Measure</th>
<th>Knowledge Measures</th>
<th>Performance Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4</td>
<td>5 6 7 8 9</td>
</tr>
<tr>
<td>Knowledge Measures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Basic FAST Scanning Procedures</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2. Anatomical Interpretation of FAST Window</td>
<td>0.32*</td>
<td>—</td>
</tr>
<tr>
<td>3. Identification of FAST Window Quadrant</td>
<td>0.16</td>
<td>0.20</td>
</tr>
<tr>
<td>4. Diagnostic Interpretation of FAST Window</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Performance Measures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. No. of Correct FAST Window Interpretations</td>
<td>0.13</td>
<td>0.11</td>
</tr>
<tr>
<td>6. No. of Excellent FAST Windows</td>
<td>-0.05</td>
<td>0.09</td>
</tr>
<tr>
<td>7. Total Scan Time (s)</td>
<td>0.16</td>
<td>0.07</td>
</tr>
<tr>
<td>8. RUQ Scan Time (s)</td>
<td>0.16</td>
<td>-0.02</td>
</tr>
<tr>
<td>9. LUQ Scan Time (s)</td>
<td>0.16</td>
<td>0.09</td>
</tr>
<tr>
<td>10. Suprapubic Scan Time (s)</td>
<td>0.14</td>
<td>0.12</td>
</tr>
</tbody>
</table>

*p < 0.05 (two-tailed); **p < 0.01 (two-tailed); ***p < 0.001 (two-tailed).
Was There an Effect of Type of Practice on Knowledge of FAST Examination Procedures and FAST Examination Performance?

To address this question, we examined whether there were treatment effects on knowledge and performance outcomes. We checked for differences between conditions on the various knowledge scales and performance was examined by checking for differences on time to scan, and the quality of window acquisition and window interpretation for the RUQ, LUQ, and suprapubic quadrants.

Effects of Type of Practice on Knowledge

Separate t-tests were conducted on knowledge of basic FAST scanning procedures, anatomical interpretation of FAST windows, identification of FAST window quadrants, and diagnostic interpretation of FAST windows (Table II). Participants in the experimental condition ($M = 9.88, SD = 1.70$) scored significantly higher on diagnostic interpretation of FAST windows than those in the control condition ($M = 8.46, SD = 1.41$), $t (46) = 3.14, p = 0.003, d = 0.91$. Participants who received simulator-based practice scored 17% higher on items requiring diagnostic interpretation than those who received classroom-based practice. No other significant differences were found.

Effects of Type of Practice on Performance

Performance during the live patient examination was evaluated using three measures: (i) time-to-scan; (ii) quality of the acquired window; and (iii) quality of the interpretation of the window. Data were analyzed by type of patient and quadrant.

Time-to-scan

Separate analyses were conducted by type of patient (normal, positive) and for each quadrant. For each quadrant, a repeated-measures analysis of variance was conducted, with type of patient (normal, abnormal) the within-subjects factor, and condition (control, experimental) the between-subjects factor.

For the RUQ, an effect of type of patient was also found, with participants in both conditions taking significantly longer to scan the positive patient ($M = 79.9 s, SD = 59.3 s$) compared to the normal patient ($M = 56.5 s, SD = 64.2 s$), $F (1, 45) = 5.44, p = 0.02, d = 0.38$. Participants in general took about 41% longer to scan the positive patient. No other significant differences were found.

Similarly, for the suprapubic quadrant, an effect of type of patient was found, with participants in both conditions taking significantly longer to scan the positive patient ($M = 45.4 s, SD = 49.5 s$) compared with the normal patient ($M = 33.5 s, SD = 37.6 s$), $F (1, 45) = 6.00, p = 0.02, d = 0.27$. Participants in general took about 36% longer to scan the positive patient. No other significant differences were found.

For the LUQ, an effect of type of patient was found, with participants in both conditions taking significantly longer to scan the positive patient ($M = 124.0 s, SD = 82.3 s$) compared with the normal patient ($M = 98.1 s, SD = 103.1 s$), $F (1, 42) = 4.85, p = 0.03, d = 0.28$. Participants in general took about 26% longer to scan the positive patient. A significant main effect was also found because of condition, with participants in the control condition taking less time to scan the patients ($M = 160.82 s, SD = 132.23 s$) than those in the experimental condition ($M = 283.36 s, SD = 183.49 s$), $F (1, 42) = 6.46, p = 0.015, d = 0.77$. Participants who received simulator-based practice took about 76% longer to scan patients compared to those who received classroom-based practice. No other significant differences were found.

In summary, these results suggest differences in the time-to-scan between the normal and positive patients in general. The only significant effect of practice was on the LUQ quadrant, with participants who received hands-on classroom-based practice performing the scan faster than those who received only simulator-based practice (which did not include practice finding the initial anatomical landmarks).

Window quality

Table VI shows the distribution of control and experimental participants who acquired high-quality windows. High quality was defined as a window quality rating of “excellent.” “Other” was defined as a window quality rating of “fair, poor, or other.”

Separate $\chi^2$-tests were conducted for each quadrant by type of patient. For the normal patient (RUQ), participants in the classroom-based practice condition had a higher rate of high-quality windows acquired ($n = 18$) than did participants in the simulator-based practice condition ($n = 11$). Conversely, participants in the simulator-based practice condition had a higher rate of non-high-quality windows acquired ($n = 13$) than did participants in the classroom-based practice condition ($n = 6$), $\chi^2 = 5.37, p = 0.02$. Seventy-five percent of participants who received classroom-based practice were...
able to acquire high-quality windows, compared to 46% of participants who received simulator-based practice.

Similarly, for the positive patient (suprapubic quadrant), participants in the classroom-based practice condition had a higher rate of high-quality windows acquired \((n = 16)\) than did participants in the simulator-based practice condition \((n = 4)\). Conversely, participants in the simulator-based practice condition had a higher rate of non-high-quality windows acquired \((n = 20)\) than did participants in the classroom-based practice condition \((n = 8)\), \(\chi^2 = 12.34, p < 0.001\). Sixty-seven percent of participants who received classroom-based practice were able to acquire high-quality windows, compared to 17% of participants who received simulator-based practice.

Window interpretation
Table VII shows the distribution of control and experimental participants who correctly interpreted windows. “Other” was defined as a rating of “incorrect or other.”

Separate \(\chi^2\)-tests were conducted for each quadrant by type of patient. For the normal patient (LUQ), participants in the classroom-based practice condition had a higher rate of high-quality windows acquired \((n = 17)\) than did participants in the simulator-based practice condition \((n = 9)\). Conversely, participants in the simulator-based practice condition had a higher rate of non-high-quality windows acquired \((n = 15)\) than did participants in the classroom-based practice condition \((n = 7)\), \(\chi^2 = 5.37, p = 0.02\). Seventy-one percent of the participants who received classroom-based practice were able to interpret windows correctly, compared to 38% of the participants who received simulator-based practice. Conditional analyses were also conducted to examine whether there was a difference in diagnosis quality given an adequate window acquisition (i.e., a window rating of excellent or fair). There were no differences between conditions by quadrant and patient, or by overall diagnosis quality.

**Participants’ Perceptions of the Effectiveness of Practice**
We examined participants’ perceptions of the utility and effectiveness of the practice received with respect to preparing them to conduct a FAST examination on a live model patient. Because nearly all participants responded with ratings of inadequate amount of practice or adequate amount of practice, the survey responses were collapsed into two categories, inadequate (comprising ratings of very inadequate and inadequate) and adequate (comprising ratings of adequate and too much practice) (see Table VIII). \(\chi^2\)-tests of independence were conducted for each question, to test for an association between condition and participants’ perceptions of the adequacy of the amount of practice.

For the acquisition aspect of scanning, there were no differences by condition. For the physical aspect of scanning, participants in the classroom-based practice condition reported a higher rate of adequate amount of practice \((n = 17)\) than did participants in the simulator-based practice condition.

---

**TABLE VI.** Comparison of Window Quality by Condition \((N = 48)\)

<table>
<thead>
<tr>
<th>Quadrant</th>
<th>Control (Classroom-Based Practice)</th>
<th>Experimental (Simulator-Based Practice)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of Participants With High-Quality Windows</td>
<td>Other</td>
</tr>
<tr>
<td>Normal Patient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUQ*</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>LUQ</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>Suprapubic</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td>Positive Patient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUQ</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>LUQ</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>Suprapubic***</td>
<td>16</td>
<td>8</td>
</tr>
</tbody>
</table>

\*\(p < 0.05\) (two-tailed); **\(p < 0.001\) (two-tailed).

---

**TABLE VII.** Comparison of Window Interpretation by Condition \((N = 48)\)

<table>
<thead>
<tr>
<th>Quadrant</th>
<th>Control (Classroom-Based Practice)</th>
<th>Experimental (Simulator-Based Practice)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of Participants with Correct Interpretations</td>
<td>Other</td>
</tr>
<tr>
<td>Normal Patient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUQ*</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>LUQ</td>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>Suprapubic</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>Positive Patient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUQ</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>LUQ</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>Suprapubic</td>
<td>11</td>
<td>13</td>
</tr>
</tbody>
</table>

\*\(p < 0.05\) (two-tailed).
TABLE VIII. Distribution of Participant Responses About the Amount of Practice

<table>
<thead>
<tr>
<th>Question</th>
<th>Inadequate Amount of Practice</th>
<th>Adequate Amount of Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Physical Aspect of Scanning (i.e., Manipulating the Probe)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classroom-Based Practice</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td>Simulator-Based Practice</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>2. Acquisition of a Scan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Window Aspect of Scanning (i.e., Being Able to Acquire a High-Quality Window)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classroom-Based Practice</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Simulator-Based Practice</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>3. Diagnostic Aspect of Scanning (i.e., Being Able to Identify Normal or Abnormal Conditions)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classroom-Based Practice</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>Simulator-Based Practice</td>
<td>11</td>
<td>13</td>
</tr>
</tbody>
</table>

condition \((n = 11)\). Conversely, participants in the simulator-based practice condition reported a higher rate of inadequate amount of practice \((n = 13)\) than did participants in the classroom-based practice condition \((n = 6)\). \(\chi^2 = 3.85, p = 0.05\).

For the diagnostic aspect of scanning, participants in the simulator-based practice condition reported a higher rate of adequate amount of practice \((n = 13)\) than did participants in the classroom-based practice condition \((n = 6)\). Conversely, participants in the classroom-based practice condition reported a higher rate of inadequate amount of practice \((n = 17)\) than did participants in the simulator-based practice condition \((n = 6)\). \(\chi^2 = 3.85, p = 0.05\).

Participants in both conditions appeared to perceive the adequacy of the practice sessions differently. Participants who received classroom-based practice reported adequate practice in general on acquiring a window compared to participants who received simulator-based practice, and inadequate practice on diagnosing window scans compared to participants who received simulator-based practice.

DISCUSSION
Participants learned from the training materials. There was clear evidence of the effectiveness of the training materials (the instructional videos and the practice). Participants more than doubled their scores on the posttest, with gains of 117%, 128%, 466%, and 532% on basic FAST scanning procedures, anatomical interpretation of FAST windows, identification of FAST window quadrants, and diagnostic interpretation of FAST windows, respectively. How much the instructional videos or the practice contributed to the posttest gains cannot be determined because measures were not administered between the instruction and practice components of the study.

There was clear evidence that participants in the simulator-based practice condition were able to interpret correctly more FAST windows than participants who received classroom-based practice (17% higher scores on the knowledge posttest). Participants in the simulator condition practiced on multiple normal and positive cases, and were able to compare an acquired window to reference windows that represented normal and positive states (minimal, mild, and severe). Participants in the simulator-based practice condition reported that the simulation practice aided them in interpreting scans. Participants in the simulator-based practice condition reported a higher rate of adequate amount of practice than did participants in the classroom-based practice condition. These findings point to a major benefit of the FAST simulator-based practice: the capability to offer practice with multiple cases and conditions synchronized with the probe manipulation.

Simulator-based practice appears to be similar to classroom-based practice in diagnostic accuracy of windows acquired during live patient examinations. There was no difference across conditions in the number of correct interpretations on five of the six quadrants across one normal and one positive. The exception was for the LUQ (normal patient), where more classroom-based practice participants were able to correctly diagnose the window.

Classroom-based practice may promote greater window acquisition skills compared to simulator-based practice. Statistical differences in window quality were found that favored the classroom-based practice condition for the RUQ (normal patient) and suprapubic quadrant (positive). In both cases, a higher number of participants in the classroom-based practice condition were able to acquire excellent windows compared to the simulator-based practice condition. These results may be pointing to one important difference between practice conditions. In the classroom-based practice condition, participants received individualized guidance from the instructor on probe positioning and rotation, and sometimes included the instructor physically guiding the participant’s hand to establish correct placement on the patient. In contrast, the simulator had the probe locked into the ideal initial positions, thereby providing participants in the simulator-based practice condition no opportunity to find the correct initial position or exposure to the various anatomical landmarks that would arise in a search for the target location. (This feature has been incorporated in more recent versions of the simulator.) This difference was a major issue that surfaced in participants’ self-reports. Participants in the classroom-based practice condition reported that they received an adequate amount of practice compared to participants in the simulator-based practice condition, who reported an inadequate amount of practice.

There appears to be no difference in the time to scan a quadrant between the two practice formats. For the LUQ, the simulation-based practice condition took significantly longer to scan the patients (76% longer). Although there were no statistically different scan times because of condition on two of the three quadrants, these quadrants had large standard deviations in general, with the control condition having significantly larger scan time standard deviations.
In summary, these findings suggest the general effectiveness of the simulator practice: (i) superior diagnostic interpretation on a knowledge-based test; and (ii) similar levels of performance on live patient examinations—despite having no prior hands-on practice (i.e., the first scan with a patient was the live patient test).

**Limitations**

There are two limitations to this study. First, the availability of only two ascites-positive model patients limited the range of cases and situations on which a participant was tested. Questions remain about sampling and generalizability—to what extent does performance on the live test in this study adequately represent performance on actual patients likely to be encountered? The second limitation is a potential training effect in the classroom-based practice condition. We do not have information on how representative the classroom instructor was of ultrasound trainers in general. The instructor used in this study reported over 20 years of experience teaching ultrasound concepts and procedures. Our observation of the classroom instruction suggested an instructor who was able to provide clear explanations and demonstrations of procedures, and provide effective feedback and guidance to students who had difficulty acquiring a window and initial probe position. In addition, the class size used (about eight participants) may have been smaller than a typical training class.

**Implications for FAST Training**

One of the most interesting findings of this study was that using the virtual trainer (SonoSimulator) for practice did not result in markedly inferior performance on the physical aspects of scanning. The simulator-based training was sufficiently effective to enable participants who received no hands-on practice to perform comparably to the participants who received hands-on practice on most of the performance measures across the normal and positive patients used in this study. However, self-reports of the participants who were given simulator-based practice point to the importance of being able to find the initial probe location for a quadrant.

In the broader training context, the findings of this study are consistent with those of design features of effective simulators. The simulator was designed around the cognitive demands of FAST window acquisition and interpretation. Through repeated exposure to cases, users are engaged in the review, identification, acquisition, and interpretation of a FAST window. High fidelity is used judiciously—only to link the real-time probe response to its corresponding window. Users can view interpolated windows of scans of actual patients resulting in exposure to varying free-fluid conditions and varying anatomy, and then compare an acquired window to windows of normal and various positive conditions. This latter capability is an important instructional feature as it helps users identify the window characteristics of the various free-fluid states.

One of the most important training benefits of simulators is extended training time. Having available virtual patients to scan with various disease conditions avoids the restrictions associated with model patients, such as availability and willingness to endure long training sessions with a number of trainees. In the case of ascites, patients are often too ill to even participate in a study. Another benefit of the simulator is documented pathology—any anatomical anomalies or a specific severity of condition can be included in the simulator for unlimited number of views.

Finally, the utility of the simulator as an anytime-anywhere refresher trainer is clear. Assuming basic competency at initial probe location and landmark identification, the use of the FAST simulator as a means to relearn procedures anytime-anywhere seems ideal as the simulator emphasizes the mapping between probe movement and the window quality. Perhaps the most powerful capability of the simulator is to provide users with practice recognizing various disease conditions or window anomalies (e.g., artifacts) that would be difficult to observe otherwise.

**ACKNOWLEDGMENTS**

SonoSimulator is protected by U.S. patent number 8,297,983. “Multimodal Ultrasound Training System” assigned to the Regents of the University of California. Lead Inventor: Eric Savitsky, MD. SonoSim retains exclusive licensing rights to the associated intellectual property associated with the patent.

Eric Savitsky is a Founding Partner of Pélagique, which funded this study.

The work reported herein was supported by a subcontract from Pélagique to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). The work reported herein was also partially supported by a grant from the Office of Naval Research, Award Number N00014-10-1-0978. Pélagique’s work on the SonoSimulator is supported by the U.S. Army Medical Research and Materiel Command under Contract No. W81XWH-11-C-0529.

**REFERENCES**

Effects of Simulation-Based Practice on FAST

Adaptive and Perceptual Learning Technologies in Medical Education and Training

Philip J. Kellman, PhD

ABSTRACT Recent advances in the learning sciences offer remarkable potential to improve medical education and maximize the benefits of emerging medical technologies. This article describes 2 major innovation areas in the learning sciences that apply to simulation and other aspects of medical learning: Perceptual learning (PL) and adaptive learning technologies. PL technology offers, for the first time, systematic, computer-based methods for teaching pattern recognition, structural intuition, transfer, and fluency. Synergistic with PL are new adaptive learning technologies that optimize learning for each individual, embed objective assessment, and implement mastery criteria. The author describes the Adaptive Response-Time-based Sequencing (ARTS) system, which uses each learner’s accuracy and speed in interactive learning to guide spacing, sequencing, and mastery. In recent efforts, these new technologies have been applied in medical learning contexts, including adaptive learning modules for initial medical diagnosis and perceptual/adaptive learning modules (PALMs) in dermatology, histology, and radiology. Results of all these efforts indicate the remarkable potential of perceptual and adaptive learning technologies, individually and in combination, to improve learning in a variety of medical domains.

INTRODUCTION Recent advances in the learning sciences offer remarkable potential to improve medical education. These advances are relevant to almost all domains of medicine, and they have direct application to maximizing the benefits of simulation and cutting-edge technologies. In this article, I describe two innovations in training technology that apply to simulation and other aspects of medical learning: perceptual learning (PL) and adaptive learning technologies. PL techniques teach pattern recognition, structural intuition, and fluency. Adaptive learning technologies can optimize learning for each individual, embed objective assessment throughout learning, and implement mastery criteria.

Understanding the role and value of these emerging technologies requires some discussion of traditional conceptions of learning and how these are changing, as well as elaboration of basic elements and benefits of each technology. We consider conceptions of learning, perceptual learning technology, and adaptive learning technology in the first three sections. Then, we describe recently developed medical learning applications of perceptual and adaptive learning technology, in the areas of clinical diagnosis, radiology, dermatology, and histopathology. In the final section, we consider synergies between these learning technologies and simulation tools and techniques in medicine.

REVISITING LEARNING In most instructional settings, learning is organized around two types of knowledge. This is not surprising, as these two types are often considered exhaustive, even in many cognitive psychology texts. Declarative knowledge includes facts and concepts that can be verbalized. Procedural knowledge includes sequences of steps that can be enacted. A conventional view of learning, shared by nonspecialists and researchers alike, is that learning consists of accumulating these facts, concepts, and procedures. The standard view has been called a “container” model of the mind: Learning consists of facts, concepts, and procedures that we place into the container (the mind), and for later performance, we retrieve these items.

Persistent problems in learning and instruction suggest that this learning worldview is defective. Students who have been faithfully taught and have diligently absorbed declarative and procedural inputs fail to recognize key structures and patterns in real-world tasks. Students may know procedures but fail to understand their conditions of application or which ones apply to new problems or situations. And, learners may understand but process slowly, with high cognitive load, making them impaired in demanding, complex, or time-limited tasks.

These characteristic problems can be observed in learning domains from mathematics to surgical training. They suggest that much is missing from the typical view of learning. What is it? Some answers are clearly available if one looks, not at the literatures on education or learning, but the literature on expertise. Studies of expertise—what people are like when they are really good at things—recurrently implicate a number of abilities that emerge from changes in the

Department of Psychology, University of California, Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90095-1563.

Any opinions, findings, and conclusions or recommendations expressed in this article are those of the author and do not necessarily reflect the views of the U.S. Department of Education, the National Science Foundation, or other agencies. The findings and opinions expressed here do not necessarily reflect the positions or policies of the Office of Naval Research. Systems that use learner speed and accuracy to sequence learning events, as well as some aspects of perceptual learning technology described here, are protected by U.S. patent 7052277 and patents pending, assigned to Insight Learning Technology, Inc. For information, please contact either the author or Info@insightlearningtech.com.

doi: 10.7205/MILMED-D-13-00218
way information is extracted: PL. Kellman suggests that PL effects fall into two broad categories, discovery and fluency effects. Table I summarizes a number of these in each category. Discovery effects refer to learners finding the information that is most relevant to a task. One important discovery effect is increased attentional selectivity. With practice on a given task, learners come to pick up the relevant information for relevant classifications while ignoring irrelevant variation. Practice also leads learners to discover invariant or characteristic relations that are not initially evident (cf., Chase and Simon) and to form and process higher level units (Goldstone; for reviews, see Bereiter and Scardamalia, Gibson, and Goldstone). Fluency effects refer to changes in the efficiency of information extraction. PL leads to fluent and sometimes automatic processing, with automaticity in PL defined as the ability to pick up information with little or no sensitivity to task load. As a consequence, perceptual expertise may lead to more parallel processing and faster pickup of information.

It is fair to say that studies of expertise have done more to describe these characteristics of experts than to reveal how these changes come about, except for the observation that expertise grows over long experience. More foundational work suggesting how these changes arise was done by Eleanor Gibson and her students several decades ago. Gibson defined PL as "changes in the pick up of information as a result of practice or experience" and argued that such changes tended to be domain-specific improvements, resulting from classification experience, involving the discovery of characteristic or invariant properties distinguishing objects or situations from one another.

Recently, PL has become a major focus of research in cognitive science and neuroscience (for reviews, see Kellman, Fahle and Poggio, and Kellman and Garrigan). For present purposes, 3 clear ideas are most relevant. First, PL is a pervasive process of learning that serves to optimize information extraction to improve task performance. Second, with appropriate procedures, all kinds of feature and pattern extraction can be improved by using PL. Third, these improvements are often dramatic, sometimes improving task performance by orders of magnitude.

One example of a complex task in which dramatic PL effects have been studied is chess. On a good day, the best human chess grandmaster can beat a chess-playing computer that examines upward of 200 million possible moves per second and incorporates methods for evaluating positions and strategies culled from grandmaster consultants. By comparison, human players do relatively little raw search in chess, examining perhaps as many as 4 possible moves and following these to a depth of several successive possible moves. Despite this huge discrepancy in search ability, humans can play chess at astonishingly high levels. Remarkably, the incredible abilities of skilled chess players, relative to novice players, turn out not to depend primarily on sophisticated reasoning or a greater storehouse of factual knowledge. They depend on perception of structure: learned pattern classification abilities of remarkable flexibility, complexity, and sophistication. Much of the relevant perception of structure is not verbally accessible. With appropriate learning experiences in a specific domain, PL allows humans to reach almost magical levels of expertise, but the relevant learning experiences are not those of traditional classrooms or tutorials. These observations about the origins of advanced expertise apply to many high-level domains of human competence; in medicine, they are crucial for understanding the skills of the expert radiologist, pathologist, and surgeon.

Likewise, PL appears to form the core of the notion of "situation awareness," which can be described as "being aware of what is happening around you to understand how information, events, and your own actions will affect your goals and objectives." Situation awareness is crucially important to many domains of military training and performance, as well as aviation and air traffic control, and many other complex tasks. The PL effects given in Table I summarize much of what is involved: selectively and automatically picking up task-relevant information, detecting important relationships, and being able to extract information with low-enough cognitive load to allow handling of complex and overlapping task demands.

In these and other domains, there is a common misconception about PL effects in expertise, related both to the oft-repeated maxim that becoming an expert requires 10,000 hours of practice and the typical view of learning as storing something in the mind. The misconception is that what happens in the transition from novice to expert has to do with committing to memory a great number of examples. A related idea is the suggestion that stored instances somehow become "mental models." In chess, for example, it may be asserted that the experts succeed because they have memorized many games. These ideas do not provide a workable account of the expertise furnished by PL. Although experiencing many instances can be an important input to PL, storage of instances does not...
produce much of the relevant expertise, nor is it a component of leading computational models of PL (see Kellman and Garrigan,\textsuperscript{10} for a recent review).

The reason involves what is needed to effectively use any facts, procedures, or models stored in memory (especially if there is a lot stored). Effective performance relies crucially on pattern recognition. When faced with a new situation, the question is: Which of the items, procedures, or models stored in the brain is relevant to this situation? This is a problem that requires classifying the new input. PL is the learning process that ultimately, through changes in the attunement, scope, and fluency of information extraction,\textsuperscript{3,10} distinguishes the expert from the novice who does not see what is relevant or who is blind to the distinguishing features that place the input into one category rather than another.

In domains that matter, this can hardly ever be done by use of memorized instances. The skilled radiologist, for example, must detect the pathology in a new image or set of images, where the tumor may be manifest in a different location, size, orientation, and contrast, and situated amidst novel and variable background anatomy and image noise, as compared with any images seen previously. The power of exposure, classification, and feedback involving a wide variety of cases is that information selection and pattern discovery mechanisms are honed, allowing the pickup of relevant structures, and equally important, that the information extraction mechanisms discard or ignore irrelevancies that do not drive important classifications. The expert emerges from PL experience with an attuned information extraction system, not a storehouse of memorized instances. The access to relevant stored information can work effectively in complex domains only after the input is rapidly and accurately classified.

It is paradoxical (or instructive) that one encounters the instance memorization account in reference to chess, as this is a domain in which the futility of memorizing can be shown by quantitative proof, based on the fact that the sheer number of possibilities dwarfs any capacity to remember and replay specific games. It has been calculated that after 40 moves of a game of chess, there are about $10^{120}$ different possible games. This exceeds by a considerable amount the number of atoms in the universe (about $10^{80}$)! Even chess-playing machines, whose memory capacity far exceeds humans, both in volume and accuracy, are not able to play chess primarily by looking up familiar games.

**PERCEPTUAL LEARNING TECHNOLOGY**

Most recent PL research has focused on low-level sensory discriminations.\textsuperscript{9} This focus derives from an interest on understanding plasticity in the brain, and from the fact that sensory coding is best understood in the early cortical levels of the brain. Considerable research, however, indicates that PL is equally applicable to high-level, complex tasks.\textsuperscript{3,4,13–15} Many of these research efforts in both high- and low-level PL have led to an improved understanding of the conditions that produce PL.

These developments are significant, because conventional instructional techniques do little to advance expert pattern recognition and fluency. In many domains, there has been a tacit assumption that we cannot teach this kind of knowing. In accord with this assumption, radiologists, surgeons, and pathologists, as well as chemists, pilots, and air traffic controllers, are told that expert intuitions will arise, not from “book learning,” but from “seasoning,” “experience,” or the passage of time.

From the standpoint of cognitive science, the passage of time is not a strong candidate for a learning mechanism. Instead it turns out that this kind of learning can be systematically addressed and accelerated using appropriate computer-based instantiations of principles of PL.\textsuperscript{13,15} We call these perceptual learning modules (PLMs).

A complete description of PL techniques is beyond the scope of this article, but a few basics will serve to characterize the approach. PLMs use interactive learning trials; learning advances through many short trials in which the learner performs some classification task and receives feedback. Classification episodes are the engine that drives PL processes to discover and process fluently key features and relationships relevant to the task. Equally crucial are specific kinds of variation in the display sets. Instances never or seldom repeat. Positive instances must vary in characteristics irrelevant to the classification, to allow learning of invariances. Negative instances must share with positive instances the values and dimensions of irrelevant properties. Research suggests a number of other important considerations about trial formats, spacing, and sequencing.\textsuperscript{16} The key to understanding PLMs, relative to traditional instructional modes, is that in PLMs one is seldom asked to solve an explicit problem or give a declarative answer; rather the tasks in PLMs call upon the learner to classify, locate, distinguish, or map structure across multiple representations.

Work with PLMs shows that relatively brief interventions can produce large learning gains in many domains. Some examples include aviation training,\textsuperscript{13} mathematics,\textsuperscript{14} and science learning.\textsuperscript{17} In some especially novel applications, PLMs are being used to improve intuitions about patterns that may lead to drug discovery in the pharmaceutical industry.

A number of studies indicate the role of perceptual structure in science, technology, engineering, and mathematics (STEM) learning domains,\textsuperscript{18} as well as the potential of PL interventions to accelerate expert information extraction and fluency in mathematics.\textsuperscript{10,14,16,17,19,20} PL interventions seem to be able to overcome pervasive obstacles in mathematics learning. In a recent series of PLMs targeting interrelated concepts in linear and area measurement, units, fractions, multiplication, and division, middle school students using PLMs in targeted interventions consistently showed strong and long-lasting learning gains on assessments including primarily transfer items, with effect sizes in the range from 0.84 to 2.69.\textsuperscript{14,16} PLM techniques systematically address aspects
of expertise for which direct instructional methods have not been previously available.

Both in terms of the breadth of applications and the possibility of radically improving learning in high-stakes domains, no area is more promising for PL technology than medical learning. Within radiology alone, there are a huge number of perceptual classifications relating to classification of pathology and normal variation, spanning not only a variety of disease conditions but also several different imaging modalities. Some of these involve a small number of fixed views; others involve 3D models capable of generating many views and requiring perceptual exploration to process fully, whereas in still others, such as ultrasound, the crucial information is often available only in animated sequences. It is well known that the speed and accuracy of the expert radiologist in exploring, seeing, and classifying develop over long and unsystematic experience (and may be highly variable across individuals). Likewise, pathologists must distinguish and classify different tissue conditions and pathogens, and dermatologists must classify skin conditions. Nor is PL confined to visual displays; heart sounds and breathing abnormalities are auditory examples, and we could enumerate haptic and tactile examples as well.

Equally important are the perceptual–procedural combinations required in surgical and interventional procedures. Although we are accustomed to thinking about the dexterous hands of the surgeon, the crucial role of perceptual expertise in guiding procedures, recognizing tissues and organs, and providing feedback from action illustrates Benjamin Franklin’s astute observation: “The eye of the master will do more work than both of his hands.”

We have begun to engineer PL technology into a number of domains of medical learning, and the potential appears limitless. Before describing these initial efforts, it will be useful to introduce the companion innovation that allows us to get the most from these efforts in PL interventions and also improves the efficiency of other types of learning: adaptive learning technology.

ADAPTIVE LEARNING TECHNOLOGY

In most instructional settings, student learning is limited by the failure of instruction to adapt to the individual. Students have different starting points and differ in aspects of lessons they learn well or poorly. Testing often arrives at the end, not in the midst, of learning, and it often involves global scoring rather than rich descriptions of what has and has not been learned. Moreover, testing usually targets accuracy alone, or perhaps speed for an entire test. Seldom are combined accuracy and fluency measures used to assess detailed aspects of learning; nor are assessments fed back continuously to optimize each individual’s learning. Lacking such links between continuous assessment and the flow of learning events, it is also rare for the learner to be guided to mastery criteria involving accuracy and fluency for all components of learning tasks. These limitations can potentially be overcome, and learning dramatically improved, by the use of adaptive learning technology.

**Adaptive Response-Time-Based Sequencing (ARTS) System**

Since the classic work of Atkinson in the 1960s, a variety of adaptive learning schemes have been proposed, with the goal of using the learner’s performance along with laws of learning and memory to make learning more efficient. These systems have usually been tested with the learning of discrete items, such as foreign language vocabulary words, and have been shown to outperform random presentation of items. Most systems adapt the presentation of items based on the learner’s accuracy on previous trials, and some guide learning by algorithms that derive estimates of probabilities of items becoming well-learned, based on models of learning.

The success of previous adaptive learning systems suggests the overall promise of adaptive approaches. Existing systems, however, have important limitations. One is that model-based systems require a prior experiment, using similar learners and random presentation of learning materials, to estimate parameters for implementing the adaptive scheme. Another is that reliance on accuracy omits important information that may be provided by response times (RTs).

We have developed a new adaptive learning system that uses both accuracy and speed to determine the spacing and sequencing in learning, as well as in implementing mastery criteria. We call it ARTS – Adaptive Response-Time-Based Sequencing. We describe some basics of the system and then describe its utility.

Consider a set of $n$ items (facts, patterns, concepts, procedures) to be learned. How can we optimize learning of the set for the individual learner? We assume an interactive learning system, in which learning consists primarily of learning trials. On each trial, some item, problem, or situation is presented, and the user must process and make a response. We optimize learning by applying principles of learning to a number of items simultaneously in a priority score system, in which all items (or categories in category sequencing) are assigned scores indicating the relative importance of that item appearing on the next learning trial. Priority scores for each item are updated after every trial, as a function of learner accuracy and RTs, trials elapsed, and in view of mastery criteria. Learning strength is assessed continuously and in some implementations, cumulatively, from performance data. In most applications, the sequencing algorithm chooses the highest priority item on each learning trial. Adjustable parameters allow flexible and concurrent implementation of principles of learning and memory, such as stretching the retention interval automatically for each item as learning strength grows.

Our system relies on a database that stores all categories in PL and all instances in factual learning contexts (e.g., multiplication facts, vocabulary, chemical symbols, etc.). Performance
data for every trial and every category or instance are acquired and used by a sequencing algorithm. For simplicity, we describe the system in terms of item sequencing, although it applies also to category learning, in which each presentation involves a novel instance. Another simplification is that even in basic factual learning, multiple formats may be used across trials to test a single item (to produce generalizable learning and enhance interest), but we omit further details.

We describe aspects of the system here omitting mathematical and technical detail. (See Mettler et al\textsuperscript{24} for more information.) Our framework has great flexibility and may use a variety of equations relating elapsed time or trials, accuracy, and RT to the priority for presentation. When any particular function of these variables is used, there are parameters that may be adjusted to suit particular learning contexts or even individual learners. Priority scores for items are dynamically updated after each trial. In many applications, initial priority scores are given to all items, and an item’s score does not change until after it is first selected for presentation. This establishes a baseline priority for feeding in new items that may be balanced against changing priorities for items already introduced. Preset orderings in learning can be accomplished by the assignment of initial priority scores that are higher for some items or categories than for others.

The full set of learning principles and objectives that may be embedded in ARTS is too extensive to describe here, but some important ones include:

**Rapid Item or Category Reappearance After Errors**

Errors result in assignment of a high-priority weighting. With ordinary settings, the error weighting will exceed all initial priority score assignments, as well as the highest priority that may result from a slow, correct answer. However, reappearance of missed items is still subject to enforced delay.

**Interleaving/Enforced Delay**

To prevent recurrence of an item while its answer remains in working memory, the system is normally configured to preclude the presentation of the same item on consecutive trials.

**Joint Optimization for the Entire Learning Set**

A priority score system allows joint satisfaction of a number of learning principles applied to an entire set of items, as all factors feed into a priority score for each item or category. Scores are dynamically updated after each trial, and items or categories compete for selection on each learning trial.

**Retirement and Mastery Criteria**

Adaptive learning focuses the learner’s effort where it is needed most. Commonly, learning effort and time are limited; therefore, it often makes sense to prioritize. We use the term retirement to describe removal of a learning item or category from the learning set, based on attainment of mastery criteria. Pyc and Rawson\textsuperscript{25} used the term “dropout” for this idea and found evidence that greater learning efficiency can be achieved with this feature, especially in highly demanding learning situations. RTs provide important clues to the type of processing the learner is using. When a learner answers a problem by calculating or reasoning, they will tend to be slower than when retrieving the answer from memory. A key effect of PL, for example, is becoming able to extract relevant structure with low attentional load, which is an important contributor to expertise in many domains. These are independent reasons for using RT in mastery criteria.

**Dynamic Spacing Based on RTs**

In our system, the priority for re-presentation of an item is a function of RT and accuracy. Even with an accurate answer, a long RT suggests relatively weak learning strength. The system can use various functions of RT but typically produces increasing priority for longer RTs. Use of RTs in adaptive learning offers a simple, direct framework for implementing important principles to produce efficiencies in learning. We hypothesize an internal variable of learning strength that may be influenced by the arrangement of learning events and inferred to some degree from performance. Learning strength is reflected in accuracy and speed in generating a factual answer or in making a classification in PL. Evidence supports response speed as an indicator of learning strength.\textsuperscript{25,26} Considerable research suggests that the value of a test trial (with successful retrieval) varies with an item’s learning strength.\textsuperscript{27,28} Thus, the best time to re-present an item is at the longest interval for which a correct retrieval can still be accomplished.\textsuperscript{29}

Controversy persists about whether and when expanding the retention interval is superior to schedules with equal spacing.\textsuperscript{27,30,31} Although these issues are subjects of continuing research, considerable evidence supports the idea that difficulty of successful retrieval is an important factor.\textsuperscript{27,28,32} Pyc and Rawson\textsuperscript{28} labeled this idea the “retrieval effort hypothesis”: more difficult, but successful, retrievals are more beneficial to learning. In recent work, they studied the relation of number of successful retrievals to later memory performance, while manipulating the difficulty of those retrievals in terms of number of intervening trials. Greater numbers of intervening trials predicted better retention. These investigators also provided evidence that, as had been suggested in other work, larger gaps produced longer average response latencies,\textsuperscript{28} a finding consistent both with the idea that a larger gap affects an item’s learning strength and that learning strength is reflected in RTs. Other recent research provides evidence for a substantial advantage of expanding the retrieval interval when material is highly susceptible to forgetting or when intervening material is processed between testing events,\textsuperscript{39} conditions that apply to many formal learning situations, including most medical learning applications. The flexibility of parameter adjustment in the ARTS system makes it possible to accommodate...
Adaptive and Perceptual Learning Technologies in Medical Education and Training

varied conditions of learning and even new findings regarding optimal spacing relations.

**Multipurpose, Multilevel Assessment**

ARTS offers not only new opportunities to improve learning but also wide-ranging possibilities for assessment. At the core of adaptive learning is performance tracking and adjustment based on embedded assessment. In our system, every concept and item in the database is tracked in terms of the learner’s accuracy and RTs on past trials. Both the raw data and derived measures are continuously available to gauge a learner’s progress. Aggregating across learners can show a class’s strengths and weaknesses for different categories of learning.

Recent research shows that ARTS outperforms random presentation and also outperforms a classic adaptive learning system in tasks involving learning of factual items. Other research indicates that ARTS improves learning in perceptual and category learning relative to other schemes.

**MEDICAL APPLICATIONS OF PERCEPTUAL AND ADAPTIVE LEARNING TECHNOLOGIES**

We have begun applying perceptual and adaptive learning technologies to medical learning, and the results are remarkably promising. We describe four of these efforts briefly.

**ARTS Technology for Optimal Sick Call Performance**

In a recent project funded by U.S. Army RDEC0, (a collaboration of UCLA, Insight Learning Technology, and Pelagique, Inc.), we used ARTS in prototype learning systems for learning of factual material and medical diagnosis. The focus was on initial clinical “sick call” diagnosis by corpsmen and medics, and the goal was to improve factual learning through adaptive factual learning modules and integration of probabilistic information in diagnosis in cognitive task modules.

In an efficacy study using premedical students carried out in the UCLA Human Perception Laboratory, the ARTS-based factual learning modules produced highly effective learning of medical material (such as signs and symptoms of meningitis, supraglottitis, etc.) and outperformed a control group using conventional study methods. Moreover, the cognitive task modules, which aimed at training information integration and higher level pattern recognition in diagnosis, added substantial benefits beyond mastery of the basic factual information.

**PL in Radiology**

Radiological diagnosis includes many domains in which subtle perceptual discriminations must be made, and radiological training could likely be radically improved by appropriate deployment of perceptual and adaptive learning technology. In a pilot project, we have begun to apply these methods to X-ray diagnosis of wrist injuries. Figure 1 shows a sample screenshot. A variety of trial types, including distinguishing normal from injured wrists and classification of single or multiple injuries in particular images, are used in the module to maximize PL. Studies are ongoing, but initial results suggest that this format for learning can produce strong advances in perceptual expertise from relatively short investments of learning time.

**Perceptual/Adaptive Learning Modules (PALMs) in Dermatology and Histopathology**

In collaboration with the David Geffen UCLA School of Medicine, we have recently developed and tested two computer-based PALMs in the pre-clerkship curriculum for first- and second-year medical students, one for recognizing pathologic processes in skin histology images (Histopathology PALM) and the other for identifying skin-lesion morphologies (Dermatology PALM). The goal was to assess their ability to develop pattern recognition and discrimination skills leading to accuracy and fluency in diagnosing new instances of disease-related patterns. We used pre- and post-test design, with each test consisting of the
presentation of a visual display along with possible answers for categorizing it. No feedback was given in the assessments. The PALM, given to UCLA medical students in between pre- and post-test, consisted of short interactive learning trials requiring the learner to classify images. The PL components in these modules included displaying a large display set, such that instances of categories did not repeat; moreover, as much as possible, irrelevant variables were balanced across categories (For instance, different dermatological conditions involved approximately the same range of body parts in the displays.) The initial PALMs in these domains were simple; they included only a single type of trial (display presentation with verbal category labels). More varied and complex trial types are known to facilitate PL, but these will be explored in subsequent work.

The adaptive learning components included use of category sequencing algorithms, which optimized spacing based on individual performance, as well as implementation of mastery criteria for each category, based on both sustained accuracy and fluency criteria.

The Dermatology PALM, designed to enhance the skin-lesion morphology curriculum presented in Year 2, consisted of 12 categories of lesion morphologies and was completed by 161 of the 162 second-year students. The Histopathology PALM was designed to complement the skin histopathology curriculum of Year 1 students by enhancing their ability to discriminate the different patterns of presentation observed for cell and tissue injury/repair, inflammation, neoplasia, and normal skin histology images, each at high-power and low-power magnifications. This module was completed by all 161 first-year students. The Histopathology PALM was also required of Year 2 students both to measure retention of the subject from Year 1 and to serve as review and enhanced learning of the material. The Dermatology PALM was offered to Year 1 students, as a control, on a voluntary basis and was completed by 78 students. These modules were completed quickly, with learning criteria typically reached in 15–35 minutes.

As shown in Table II, substantial improvements between pre- and post-test scores were observed, with large (mean effect sizes >0.7) and highly significant ($p < 0.0001$) increases in accuracy and speed in categorizing previously unseen images. Comparing performances for Years 1 and 2 on each of the modules, it can be seen that pre-test scores were much higher for dermatology lesion morphology in Year 2 than in Year 1, which is expected because the students in Year 2 had recently received lectures and an online learning experience. In contrast, this material was touched on only briefly for Year 1 students. Post-test scores, however, were highly similar. Histopathology pre- and post-test scores were similar for Year 1 and 2 students (Table II), showing strong learning gains for both groups. Finally, students reported that the PALMs increased their confidence and were useful, and they indicated that they would like more of these in other units.

### APPLICATIONS OF PERCEPTUAL AND ADAPTIVE LEARNING TECHNOLOGIES IN MEDICAL SIMULATION

Although efforts are in their infancy, the promise of perceptual and adaptive learning technologies for improving medical learning is already obvious. Not much work, however, has yet addressed procedural learning and simulation. These areas are ripe for development, as these new technologies are well suited to getting the most from simulation training. Simply having cutting-edge simulations does not solve the problem of how to improve learning. Perceptual–procedural learning technologies and adaptive methods using objective criteria of learning have much to offer in this regard. In this section we note some issues, benefits, and considerations in applying these new technologies to simulation.

#### Perception–Action Loops in Procedural Learning

We often think of skilled practitioners, such as pilots or surgeons, as having “good hands,” but the key to their skills

---

**TABLE II.** Results of Dermatology and Histopathology PALMs with First and Second Year Medical Students

<table>
<thead>
<tr>
<th></th>
<th>Pre-Test$^a$</th>
<th>Post-Test$^a$</th>
<th>$p$</th>
<th>$t(df)$</th>
<th>Effect Size</th>
<th>$N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 1 Histopath</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>54% (13%)</td>
<td>66% (12%)</td>
<td>&lt;0.0001</td>
<td>9.6</td>
<td>0.98</td>
<td>161</td>
</tr>
<tr>
<td>RT</td>
<td>15.82 (11.20)</td>
<td>6.16 (2.33)</td>
<td>&lt;0.0001</td>
<td>11.6</td>
<td>1.19</td>
<td>161</td>
</tr>
<tr>
<td>Year 1 Derm (optional)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>66% (11.5%)</td>
<td>84% (8.1%)</td>
<td>&lt;0.0001</td>
<td>12.5</td>
<td>1.55</td>
<td>79</td>
</tr>
<tr>
<td>RT</td>
<td>8.36 (6.11)</td>
<td>3.90 (1.03)</td>
<td>&lt;0.0001</td>
<td>7.0</td>
<td>1.00</td>
<td>79</td>
</tr>
<tr>
<td>Year 2 Histopath</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>51% (13.5%)</td>
<td>64% (14.4%)</td>
<td>&lt;0.0001</td>
<td>7.9</td>
<td>0.90</td>
<td>162</td>
</tr>
<tr>
<td>RT</td>
<td>8.87 (4.6)</td>
<td>5.30 (2.7)</td>
<td>&lt;0.0001</td>
<td>11.1</td>
<td>0.95</td>
<td>161$^b$</td>
</tr>
<tr>
<td>Year 2 Derm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>82% (8.9%)</td>
<td>88% (8.9%)</td>
<td>&lt;0.0001</td>
<td>7.62</td>
<td>0.76</td>
<td>161</td>
</tr>
<tr>
<td>RT</td>
<td>9.01 (4.77)</td>
<td>3.72 (0.94)</td>
<td>&lt;0.0001</td>
<td>14.5</td>
<td>1.54</td>
<td>161</td>
</tr>
</tbody>
</table>

Source: Krasne, Hillman, Rimoin, Burke, Kim, Drake, Craft and Kellman (unpublished data). $^a$Mean percent correct is shown in each assessment, with the standard deviation (SD) shown in parentheses. $^b$One outlier (pre-test value 182 seconds) was excluded from the analysis.
is often the expert pickup of the information that guides the procedure (cf., Kellman and Kaiser 15). A surgeon, for example, must recognize anatomy in novel cases, distinguish various tissues and structures, and sense the position, progress, and force of instruments. Surgery involves a host of delicate perceptual–procedural learning tasks. In such procedural tasks, much of the learning consists of improvements in the pickup of information. PL technology can be readily adapted to perception–action loops, in which perceptual discriminations and perceptually guided actions are objectively assessed and accelerated through adaptive spacing and sequencing methods.

Realism in Perceptual–Procedural Learning
High-quality simulations offer ideal synergies with perceptual–procedural learning technology. Because PL requires becoming attuned to subtle patterns of information, richer simulation improves the likelihood that training will be relevant to actual practice. Simulation also allows the variation in displays and scenarios that is crucial for allowing perceptual–procedural learning processes to distinguish crucial invariances from irrelevant variation and discover the best information for guiding action.

Objective Scoring, Assessment, and Certification
A truly revolutionary opportunity afforded by realistic simulation for procedure training, combined with adaptive methods, is the opportunity to objectively assess performance and use performance data to optimize the learning and implement mastery criteria. With suitable tracking technology, a learner’s movements in a virtual space may be tracked and scored for accuracy and speed in a variety of tasks. Use of the ARTS system offers particular advantages, in that fluency, as well as accuracy, is used both in optimizing the learning process and also as a goal of learning. It has been said that surgeons fall into three categories: fast and good, fast and bad, and slow and bad; there are no slow, good surgeons. Although speed is known to be crucial for effective execution of medical procedures, it has not typically been used systematically in training or assessment. Obtaining and using the objective data required to unlock the benefits of perceptual and adaptive learning techniques may require considerable investment, but doing so will produce accelerated learning, reliable certification, practice and refreshment of seldom-used skills in refresher training, and reduction of errors in subsequent medical practice.

CONCLUSION
For effective learning, cutting-edge simulation must be combined with cutting-edge learning techniques. Both in simulation and in many other areas of medical learning, two broad-based innovations offer remarkable potential to accelerate learning and enhance performance. Perceptual–procedural learning technology offers ways of bringing intuitive pattern recognition, interpretation of new cases, fluent processing, and procedural execution into the realm of systematic and objective instruction. Adaptive learning technology, such as the ARTS system, improves learning by attuning the level, spacing, and sequencing of learning events to each individual learner, allowing more efficient learning, better retention, and certification of mastery. These technologies apply to many medical domains, and it will be exciting to turn their potential into reality and see the benefits both in improved training and medical practice.

ACKNOWLEDGMENTS
I thank Cheryl Hein, Carmack Holmes, Erik Dutson, Robert Bjork, Christine Massey, K.P. Thai, and Everett Mettler for helpful discussions and contributions. The author gratefully acknowledges support from: the National Science Foundation (NSF) REESE Program award 1109228; the National Institutes of Health (NIH) award 5RC1HD063338; the U.S. Department of Education, Institute of Education Sciences (IES), Cognition and Student Learning Program (CASL) awards R305A120288 and R305H060070; the UCLA Center for Advanced Surgical and Interventional Technology (CASIT); and the Institute for Innovative Technology in Medical Education (iInTIME); as well as a U.S. Department of Education, IES SBIR award ED-IES-10-C-0024 to Insight Learning Technology, Inc. The work reported herein was also partially supported by a grant from the Office of Naval Research, Award Number N00014-10-1-0978.
Adaptive and Perceptual Learning Technologies in Medical Education and Training

Evidence-Centered Design for Simulation-Based Assessment

Robert J. Mislevy, PhD

ABSTRACT  Simulations provide opportunities for people to learn and to develop skills for situations that are expensive, time-consuming, or dangerous. Careful design can support their learning by tailoring the features of situations to their levels of skill, allowing repeated attempts, and providing timely feedback. The same environments provide opportunities for assessing people’s capabilities to act in these situations. This article describes an assessment design framework that can help projects develop effective simulation-based assessments. It reviews the rationale and terminology of the “evidence-centered” assessment design framework, discusses how it aligns with the principles of simulation design, and illustrates ideas with examples from engineering and medicine. Advice is offered for designing a new simulation-based assessment and for adapting an existing simulation system for assessment purposes.

INTRODUCTION
Simulation environments open the door to a radically new paradigm of learning, characterized by interaction and adaptation. Physicians can practice heart surgery, for example, and pilots can land planes with failed engines. Students can interact with simulated characters to develop experience with emotionally difficult situations. Given the increased acceptance and widespread use of simulations for learning, it is natural to consider using simulation environments for assessment as well. Several such simulations are currently in operation, such as the Primum computer-based simulated-patient problems in Step 3 of the United States Medical Licensing Examination. The step from a simulation to a simulation-based assessment is not an easy one, however. Despite considerable overlap in content and technology, a valid simulation-based assessment also requires principles and tools from assessment design and psychometrics. A shared design framework in which each team member can see how his expertise fits with the others is invaluable for producing a design that serves the purpose of the assessment.

This article describes such a framework, namely evidence-centered assessment design (ECD). ECD’s language and representations support assessment designers in projects across different domains, task types, and purposes. In particular, ECD has proved useful as the design framework for simulation-based assessments, including computer network engineering, science investigation in virtual worlds, and problem solving in dental hygiene. This article presents an overview of ECD and offers advice for developing a simulation-based assessment from an existing simulation capability or creating one de novo.

BACKGROUND
Novices in a wide variety of domains face similar difficulties. They do not know what information is relevant. They do not know how to integrate information. They do not know what to expect. They do not know what to do next. Even when they have an idea of what to do, they often cannot do it well enough or fast enough.

Nevertheless, humans routinely display some amazing capabilities. We carry on rapid conversation, for example, by processing sounds, grammar, domain knowledge, social norms, conversational conventions, and pragmatic moves—all at once, in milliseconds. This is possible because people are extremely good at working with patterns: patterns of perceiving, patterns of thinking, and patterns of acting, which we become attuned to and facile with, and can assemble flexibly in real time once we are sufficiently practiced. Expert performance is made possible through the continual interaction between the external patterns that structure peoples’ interactions in situations, such as language and professional practices, and a person’s internal neural patterns for recognizing, making meaning of, and acting through these patterns.

We develop our internal cognitive patterns through experience; i.e., by participating in activities that are structured around these patterns, discerning regularities, seeing what happens as we and others act, and acting ourselves in more complicated and more varied situations. More specifically, people become experts in domains such as sports, medicine, and culinary arts by taking part in the activities of that domain—learning to work on the problems, reading the literature, talking with the people, and acting in the situations. They learn to use tools and the strategies of the community. Through reflective practice—best with feedback, often starting from simplified situations, usually with feedback and support from others—they build their capabilities and overcome the pervasive limitations that plague novices.

Effective simulation environments highlight the key features of relevant situations—for practice, for feedback, for seeing what happens next, for honing skills, and for gaining facility with tools. They provide experiences that build up the patterns for knowing what is important, what it means,
what can be done, and how to interact with people and situations. They highlight key patterns, allow repetition and diverse practice, and provide opportunities for feedback.

A simulation emulates some features of actual situations but not others. It may speed up, slow down, change size, or simplify aspects of real-world situations. Within the limits of practicality, a designer must determine what aspects of real situations to include in a simulation environment and what aspects to modify or omit. These design choices cannot be made without knowing the intended purposes of the simulation. Higher fidelity to real-world situations does not necessarily make for better assessment. Fidelity with respect to targeted capabilities, at the targeted level, is more important, but even that is not the whole story.

Designing simulations for “learning” requires focusing on the features of situations that provoke the targeted knowledge and skills, at a level that is just beyond the capabilities of the student. This notion is akin to Vygotsky’s “zone of proximal development”: Given where a learner is currently, what aspects of situations will best solidify skills, add the next layer of complexity, or develop new variations of a familiar theme, with the support of a teacher or software, or appropriate structuring of the challenge? The features of situations that are critical for interaction must be in the simulation. That is, the means for the student to act on the system (its affordances) must be included, and the system must react to student’s actions in ways that reflect the underlying principles of the system. Features to omit are those that add irrelevant complexity, or require too much knowledge or skill that is not central. A high-fidelity system that models the way experts view problems may not be accessible to beginning students. Complexity can thus be introduced in stages to maximize effective learning. To help students learn optimally, a simulation system for learning can provide opportunities to slow down or stop action to reflect on what is important in a situation, what to do next, or why something happened.

Designing simulations for “assessment” requires a focus on information the assessment user needs about examinee’s knowledge and skill. Sometimes the user is the examinee, as in self-assessment and cognitive tutors. In most cases, though, the users are different from the examinee, and they need information to evaluate examinee’s capabilities, either as to overall proficiency or more specific aspects of knowledge and skill. The purpose may be low-stakes formative feedback or a high-stakes uses such as licensure or hiring decisions.

Much of the thinking for designing simulations for learning overlaps with designing simulations for assessment: The kinds of problems and situations in which one learns to think and act in a domain are the same kinds of problems and situations that provide evidence about these capabilities. Exactly how that evidence is to be evoked, captured, interpreted, summarized, and reported is the realm of assessment designers and psychometricians. An assessment design framework coordinates their contributions with those of domain experts and simulation designers.

OVERVIEW OF EVIDENCE-CENTERED DESIGN
An educational assessment embodies an evidentiary argument to reason from what students say or do in particular task situations, to claims about what they know or can do more broadly. Messick summarizes its form neatly.

[We] would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 16).

The ECD framework distinguishes layers at which activities and structures contribute to instantiating an assessment argument in operational processes (Fig. 1). Common language and representations across different forms of assessment help developers structure their work, both conceptually and operationally, in ways that encourage reusability (e.g., design patterns for generating tasks, adaptable scoring procedures).

— The first layer, “Domain Analysis” is marshaling substantive information about the domain. This helps us to understand the kinds of problems and situations people deal with, the knowledge and skills they draw upon, the representational forms they use, and the characteristics of good work. This is the substantive foundation for an assessment.

— The “Domain Modeling” layer organizes information from domain analysis along the lines of assessment arguments. Supporting tools, such as the design patterns discussed below help developers think through the assessment argument.

— The “Conceptual Assessment Framework (CAF)” concerns technical specifications for operational elements. An assessment argument is now expressed in terms of coordinated pieces of machinery, such as measurement models, scoring methods, and delivery requirements. The data structures and reusability of the CAF models help bring down the costs of task design.

— “Assessment Implementation” encompasses activities that prepare for operational administration, such as authoring tasks, calibrating psychometric models, piloting and finalizing evaluation procedures, and producing assessment materials and presentation environments.

— “Assessment Delivery” addresses the processes of presenting tasks to examinees, evaluating performances to assign scores, and reporting the results to provide feedback or support decision making. The architecture is viewed in terms of processes and messages, whose
meaning is grounded in Domain Modeling, whose structure is laid out in the CAF and whose pieces are built in Assessment Implementation.

WALKING THROUGH THE LAYERS
This section takes a closer look at the ECD layers, with an eye toward simulation-based assessment.

Domain Analysis
Domain analysis defines and documents the content or experiential domains to be assessed. This is the relevant content knowledge, the ways people use it, and the kinds of situations they use it in. What do we know about progressions of thought or skill, patterns of errors or common representations? What knowledge, skill, goals, tools, and representations are relevant? How do people interact with the physical environment, conceptual representations, and other people to accomplish goals? By what standards are these efforts judged in practice? Designers can consider the domain from a number of perspectives, including cognitive research, curricula, professional practice, ethnographic studies, expert input, standards, and current testing practices.

For performance-based assessments, designers must understand the kinds of situations that pose the recurring problems that Salthouse noted. People’s patterns of thinking and acting in such situations will be at the heart of the assessment, in forms and at levels that match its purposes. Practitioners’, researchers’, and instructors’ insights are invaluable. Task analysis has long been a staple of licensure testing. The critical incident technique has proven useful in domains such as medicine, engineering, and law enforcement.

More recently, cognitive task analysis (CTA) has been used to uncover the knowledge structures that people use, and how knowledge is related to performance. CTA is especially suited to developing learning and assessment environments, where what matters is not just what people do, but how and why they do it. For example, Mislevy et al identified nine broad classes of behaviors that tended to distinguish dental hygiene experts from novices, including gathering and using information, communication and language, and patient assessment. Novices might identify salient features in multiple representations, such as soft tissue charts and radiographs, but the experts produced a conception of a patient’s condition that would lead to the pattern of features across representations. These observations helped ground a design rationale for task design and scoring in simulation-based assessment for dental hygiene.

Domain Modeling
In domain modeling, designers organize information from domain analyses to articulate assessment arguments. ECD tools that support this work include “claims and evidence” worksheets, Toulmin diagrams for assessment arguments, and design patterns for constructing assessments around capabilities such as design under constraints and model-based reasoning.

Let us look more closely at Toulmin diagrams. Figure 2 formalizes the Messick quote in terms of philosopher...
Stephen Toulmin’s schema for arguments.23 At the top is a claim about an examinee’s knowledge and skill. At the bottom is the observation of the examinee acting in some situation. An assessor’s reasoning moves up from what is important in the examinee’s actions (which the examinee produces) and the features of the situation that are important in eliciting those actions (partly determined by the assessment designer, but also partly determined by the examinees as they interact with the task). These data support the claim through a warrant, or rationale, for why examinees with particular capabilities are likely to act in certain ways in the situation at hand. The backing for the warrants comes from the Domain Analysis.

The single boxes for “data concerning the student” and “data concerning the situation” are sufficient for self-contained small tasks. In simulations, though, performances evolve as the examinee interacts with them (Fig. 3). Some features of both the situation and the performance, such as the final solution can be evaluated without attending to the details of the interaction. However other potentially useful performance data arise along the way, and the evaluation must recognize them and take situational features into account to evaluate them. For example, did the examinee in a role play work herself into a situation where it became necessary administer adrenalin, and if so, was adrenalin administered? The circumstances that arise, and the opportunity to observe examinees’ reactions, will generally vary more as simulations are less constrained.

The “alternative explanations” shown in Figure 2 are particular importance for validity. Evaluations of performances can mislead us in two ways. First, the situation may not include features that it needs to evoke the targeted knowledge. For example, if a simulation is meant to assess students’ proficiency to obtain information in emotionally difficult patient history conversations, just selecting questions from a checklist misses the key interactional feature. Second, performing well on the task may require knowledge and skill beyond the targeted capabilities. For example, lack of

![Figure 2: An extended Toulmin argument diagram for assessment arguments.](image)

![Figure 3: Potential points to identify data in performance assessments.](image)
familiarity with a simulation interface can lead spuriously to poor performances. Messick refers to these two threats to validity as “construct underrepresentation” and “construct irrelevant demands.” Although medical practice often involves an interplay among multiple aspects of knowledge and skill, using simulation for learning and assessment affords the opportunity to focus on isolated capabilities, targeted combinations of knowledge and skill, or particular situations. Optimal design requires careful consideration of just which aspects are at issue for a given purpose. Messick’s article “The interplay of evidence and consequences in the validation of performance assessments” remains the best source on how to think about what features should and should not be represented in simulations.18

The CAF
In the CAF, the domain information is combined with information about goals, constraints, and logistics to create a blueprint for an assessment. The CAF comprises models containing objects and specifications for operational aspects of work, including (1) the creation of tasks, evaluation procedures, and statistical models, (2) delivery and operation of the assessment, and (3) analysis of data coming back from the field. While domain modeling emphasized the interconnections among aspects of peoples’ capabilities, situations, and behaviors, the CAF capitalizes on the separability of the objects that are used to instantiate an assessment. Because the models and their components can be rendered in digital form, they are amenable to assisted and automated methods of generation, manipulation, operation, and assembly.25,26

Figure 4 is a schematic of the three central models in the CAF and objects they contain.4,27 The Student Model contains variables for expressing claims about targeted aspects of students’ capabilities. Their number and character depend on the purpose of the assessment—a single student model variable to characterize students’ overall proficiency in a domain of tasks for a certification decision, for example, or a multidimensional student model to sort out patterns of proficiency from complex performances or to provide more detailed feedback.

A “Task Model” formally describes the environment in which students say, do, or make something to produce evidence.28 In simulation tasks, more information is required than in fixed tasks about the actions an examinee can take and the reactions of the simulation environment in response. One way to specify a simulation task is the initial status and transition rules of a finite state machine.29 A key decision is the form(s) in which students’ performances will be captured, i.e., the Work Product(s)—for example, a sequence of steps in an investigation, the locations of icons dragged into a diagram, or the final solution of a design problem. The dental hygiene CTA found that examinees’ interpretations of diagnostic tests were more informative than the sequence of tests they chose, so an additional work product was recommended: filling out an insurance form with key findings and interpretations. Task model variables indicate salient features of a situation and are used in evaluating performances—data concerning the situation in the assessment argument. The values of some of the task model variables are set a priori by the task designer, whereas in interactive assessments dynamic task model variables are determined as the performance unfolds.

An “Evidence Model” bridges the Student Model and the Task Model. The “evaluation component” says how one identifies and evaluates the salient aspects of the work products, expressed as values of Observable Variables. Evaluation procedures can be rubrics for human scoring or algorithms for automated scoring procedures. Margolis and Clauser30 describe the regression-based automated scoring procedures used in the National Board of Medical Examiners’ simulations, and Canché et al31 describe the rule-based approach used in a training simulator for CPR skills that provides formative assessment feedback. Information in the Observable Variables is synthesized in the “measurement model” component. The simplest measurement models are classical test theory models, in which scores for salient features are added. Modular measurement models assembles pieces of more complicated models such as those of item response theory or Bayesian networks.32,33 Much can be gained when the structures of evidentiary relationships in complex tasks and multivariate student models are expressed in reusable measurement model fragments. Task authors can then create unique complex tasks but know ahead of time “how to score them.”10

Assessment Implementation
The Assessment Implementation layer of ECD is about constructing and preparing the operational elements specified in the CAF. This includes authoring tasks, finalizing rubrics

![FIGURE 4. The Central models of the Conceptual Assessment Framework.](image-url)
or automated scoring rules, estimating the parameters in measurement models, and producing simulation states and transition rules. Using common and compatible data structures increases opportunities for reusability and interoperability, and helps bring down the costs of simulation-based assessment in the areas of task design, measurement models, authoring frameworks, and automated scoring.34–37

Assessment Delivery: The Four-Process Architecture

The Assessment Delivery layer is where students interact with tasks, their performances are evaluated, and feedback and reports are produced. In four-process delivery architecture by Almond et al, the processes pass messages among themselves in a pattern determined by the test’s purpose.27 The messages are either data objects specified in the CAF (e.g., parameters, stimulus materials) or are produced by the student or other processes in data structures that are specified in the CAF (e.g., work products, values of observable variables). Common language, data structures, and a partitioning of activities promote the reuse of objects and processes, and interoperability across projects and programs.

Figure 5 shows the four principal processes. The “activity selection process” selects a task or activity from the task library, or creates one in accordance in light of what is known about the student or the situation. The “presentation process” is responsible for presenting the task to the student, managing the interaction, and capturing work products. Work Products are passed to the “evidence identification process,” or task-level scoring. It evaluates work using the methods specified in the Evidence Model. It sends values of Observable Variables to the “evidence accumulation process”, or test-level scoring, which uses the Measurement Models to summarize evidence about the student model variables and produce score reports. A simulation-based task can involve many interactions among the processes in the course of a performance.27,37

FIGURE 5. Processes in the assessment cycle.

DISCUSSION

This article has touched on the subjects of designing simulations, simulation-based learning, and simulation-based assessment. The three topics overlap substantially. When the purpose of a simulation environment is learning, attention is on the cognitive patterns and activity patterns that people develop to be able to perform effectively in the relevant situations. Just what aspects of knowledge and skill are the targets of learning in the project at hand? What level of skill is the focus? What are the situations and action possibilities that these users need to develop their capabilities? How should practice opportunities and feedback be tuned to optimize their learning? Different answers to these questions can lead to very different choices for simulating aspects of the same domain. Accordingly, a wide range of simulation paradigms have been developed in medicine, from partial-task body-part models and realistic computer-controlled mannequins for practicing technique, to computer-based patient management problems and live standardized patients for developing cognitive and interpersonal skills.38

Moving further to supporting assessment introduces an additional layer of design considerations. This article provides an introduction to a framework that builds on assessment research to support the development of simulation-based assessments. Just because a simulation supports effective learning for some aspects of a system does not necessarily mean it will provide reliable or valid assessment of that learning. Further analysis is needed to optimally provide evidence to an external observer and to being able to characterize the properties of that evidence.39 Optimizing the evidentiary value of simulation-based performances may require going back to the design of the simulator and the situations with a psychometrician’s perspective. Fewer options and more constrained situations, for example, may be less effective for learning but more effective for focusing examinees’ actions on key aspects of cognitive structures or activity structures. Bringing floundering examinees back quickly
may not be as good for learning, but it is a better use of limited assessment time. Requiring an explicit work product may slow working through a simulation, but it will make some valuable evidence about unobservable thinking manifest. Simply having lots of data, in the form of gigabytes of time-stamped mouse clicks and key strokes, may not provide much evidence at all if it is not about salient actions in relevant situations.

This article has focused on the evidentiary argument considerations that go into these design decisions rather than psychometric methods for modeling the data. It highlights assessment-related issues that every member of the design team can understand and should be responsive to. To see exactly how different choices affect the evidentiary value of data or decision accuracy or instructional effectiveness, will require more esoteric methods from the psychometrician’s toolbox. A long-standing lesson applies: to design a rich simulation environment to collect data without consideration of how the data will be evaluated, and hoping psychometricians will somehow “figure out how to score it,” is a bad way to build assessments. Data mining techniques are limited by the strength of the relationship between what is in the data and its connection to the targeted inferences. Close collaboration and interaction from the very start of the design process is preferable—among users, who understand the purposes for which the assessment is intended; domain experts, who know about the nature of the knowledge and skills, the situations in which they are used, and what examinees do that provides evidence; psychometricians, who know about the range of situations in which they can model data and examine its evidentiary value; and software designers, who build the infrastructure to bring the assessment to life.

In particular, in using simulation as the basis of assessment in medical contexts, all of the complex considerations that go into designing simulations for learning are now joined by the equally challenging considerations that go into designing assessments. It is difficult enough to become an expert in either area, let alone both. The way forward builds on what has been learned from the rapid developments and proven successes of the various kinds of simulation in medical training. Grounding well-targeted exemplars, such as Primum simulations in an assessment design framework shows how to integrate the deep principles of medicine, learning, and assessment in a simulation-based assessment. The concepts and representations of evidence-centered design help by building the principles of assessment design into the work explicitly, enhancing at once its efficiency, validity, and transferability.

ACKNOWLEDGMENTS

I am grateful for insights and collaboration on simulation-based assessment with many colleagues from Cisco, CRESSST, the Dental Interactive Simulations Corporation, Educational Testing Service, SRI International, the University of Wisconsin, and the University of Maryland. The work reported here was supported in part by the Center for Advanced Technology in Schools (CATS), PR/Award Number R305C080015, as administered by the Institute of Education Sciences, U.S. Department of Education, and also partially supported by a grant from the Office of Naval Research, Award Number N00014-10-1-0978.

REFERENCES


Potential Applications of Latent Variable Modeling for the Psychometrics of Medical Simulation

Li Cai, PhD

ABSTRACT Use of simulation-based assessments and training has become increasingly widespread in medicine. It is recognized that simulations can yield a wealth of real-time information about the trainee or examinee’s performance, from which inferences about proficiency can potentially be drawn. However, for the inferences to be useful, psychometric evaluation should be conducted and validity evidence amassed. Traditionally, educational and psychological measurement has relied on psychometric models that are static, unidimensional, and based on observed scores. In this article, it is argued that modern psychometric models that are dynamic, multidimensional, and based on latent variables may be useful for evaluating medical simulations. It is also argued that modern computational methods based on Bayesian statistics may provide the technical foundation. Several examples are given and issues for further research are discussed.

INTRODUCTION

Technological developments have advanced how medical assessments and training material can be developed and delivered. For example, in the high-stakes U.S. medical licensure exam, computer-based simulations are already being used. General assessment design frameworks such as evidence-centered design have also been applied to simulation-based assessments. It is widely recognized that simulation-based training and assessment can generate a wealth of real-time data about the trainee or examinee’s performance as the simulations unfold, from which inferences about proficiency can potentially be drawn, and decisions such as certification or remediation can be made.

On the other hand, for the inferences from simulation-based assessments to be useful to the developers, users, and examinees, psychometric evaluation should be conducted and validity evidence pursuant to the joint testing standards should be amassed. Much of this evaluation effort aims at producing statements about the reliability, fairness, and validity of the assessment. To obtain such statements from empirically observed data, psychometric and statistical models are required. Psychometric models are mathematical instantiations of theoretical models of psychological and educational measurement. For instance, the classical true score theory implies one of the most well-known psychometric models \( Y = T + E \), where \( Y \) is the observed test score, \( T \) the true score, and \( E \) the random error of measurement. In applications, psychometric models are statistical models because they must be based on empirical data and contain parameters that summarize the data. For example, with test score data and the classical true score model, psychometric analysis can yield well-known summaries such as reliability and validity coefficients. As such, models and the modeling endeavor are the focus of this article.

The models considered here are special in that they all contain latent variables. A latent variable model is therefore a statistical model that contains unobserved variables. This is natural because the constructs being assessed, e.g., medical competency, are typically not directly observable. Spearman’s factor analysis model is arguably the first well-articulated latent variable model to witness widespread usage. From the perspective of statistical modeling, latent variables are nothing more than statistical devices that may help represent the theoretically driven constructs.

CHALLENGES

Traditionally, educational measurement has relied on psychometric models that are largely static, i.e., in contrast to models that explicitly incorporate a component for change from one time point to the next. Much emphasis is placed on scaling and internal consistency analysis based on correlational data collected from one test administration, e.g., in Nunnally and Bernstein’s classic text. This is both a reflection of the kinds of assessments that the field has dealt with in the past, e.g., end-of-grade tests for student outcomes on math or reading, as well as the technology of testing, e.g., paper-and-pencil forms. The same can be said of psychological assessments used in clinical or research settings. Paper forms are filled out, either by patient self-reports or through interviews, and are often sent back to owners of the psychological scales for score reporting. The timelag between the administration of assessments and the availability of scores along with the appropriate interpretations can be substantial.

In many operational, educational, and psychological assessments, the underlying models for test scores are based on classical true score theory. By some estimates, classical test theory treatises such as Gulliksen’s textbook indeed cover the vast majority of psychometric tools and procedures in widespread use. However, the classical test theory procedures are ill-suited to handling data that arise out of the dynamic interactions in simulations because of the complex
manner in which the data points are correlated, as well as the just-in-time scoring and feedback that are required features of simulation-based assessments. Moreover, simulation-based training also entails the need for accurate determination of change in levels of proficiency so that different tasks at the right levels can be chosen for different trainees. Thus, tracking individual level change is critically important, yet no coherent answers follow from the classical derivations that are focused entirely on observed test scores because inferring individual or group differences from observed differences in test scores requires measurement invariance.8

Unidimensionality has historically been a hallmark feature of classic psychometric models such as the Rasch model.9 The potential lack of unidimensionality of simulation-based assessments poses a further technical challenge to these classic models. A mathematical definition of unidimensionality is beyond the scope of this article,10 but it may be understood to imply that the test items measure one and only one construct. Multidimensionality can easily arise because a realistic performance task in a simulation-based assessment generally evokes responses that are the result of integrating more than one set of requisite competencies (understood as potentially binary dimensions indicating presence or absence of a skill) or proficiencies (continuously varying dimensions indicating the degree of proficiency). When simulation results from more than one time points are considered together (often in the form of a continuous stream of data collected in the background of computer-based environments), the longitudinal nature of the individual data points also create multidimensionality because of temporal dependencies in individual responses. Much needed to analyze or interpret these data are latent variable models that are dynamic and multidimensional in nature.

GENERAL LATENT VARIABLE MODELS

As mentioned earlier, a latent variable model is a statistical model that contains latent (i.e., unobserved) variables. Spearman’s seminal work4 on factor analysis resulted in the first latent variable model to witness popularity in educational and psychological measurement. Bartholomew et al11 proposed a more modern and inherently Bayesian perspective, in which observed and latent variables are not treated as qualitatively distinct entities. By adopting the Bayesian formulation, one can see that the existence of latent variables is consistent with the desire of the measurement enterprise, which is to make inferences about the unknown (latent) quantities of interest from the observed variables. In Bayesian inference, this is naturally handled by computing the posterior distribution of the latent variables given the observed variables and a model that connects the two. The posterior of a latent variable is simply the conditional probability of the latent variable given the observed variables. The pragmatic aspects of the Bayesian inferential machinery is made possible first by a well-known formula because of Thomas Bayes12 as well as the ensuing theoretical developments, and by the coupling of high-speed computing machines with the advent of remarkable stochastic algorithms (Markov chain Monte Carlo or MCMC) for performing posterior calculations.13

Commonly used latent variable models include factor analysis model, structural equation model (including latent growth curve model), latent class (or mixture) model, item response theory model, and latent profile model.11 This list is by no means complete, and there is in fact quite a bit of overlap across these seemingly different models. These models have each flourished in different disciplinary homes, which implied that different terminology has emerged for essentially the same concept. For example, educational testing has historically relied heavily on item response theory, whereas structural equation modeling is popular in psychological measurement. In item response theory, an observed variable is called an item, and a latent variable is called a latent trait, whereas in structural equation modeling, a latent variable is sometimes a disturbance term for an equation and an observed variable an indicator.

It is sometimes helpful to categorize the existing latent variable models according to whether the observed and the latent variables are discrete or continuous. One such typology was proposed in Bartholomew et al11. Table I is largely based on their system of classification. However, in a more generalized latent variable modeling framework, one may consider a latent variable model that is made up of features of from each of the four cells of Table I, e.g., having mixtures of continuous and discrete observed variables and latent variables. One may also add a time-dependent, dynamic aspect to the model that is not captured by any cell of Table I.

Using only elementary probability, a latent variable model consists of two fundamental parts. First, the model must specify a conditional distribution of observed variables given latent variables, i.e., \( P(\text{Observed}|\text{Latent}) \). Second, the distribution of latent variables should be specified as \( P(\text{Latent}) \). Multiplying the two parts together, the joint distribution of observed and latent variables is given by \( P(\text{Observed}, \text{Latent}) = P(\text{Observed}|\text{Latent}) \times P(\text{Latent}) \). The joint distribution is the latent variable model. It shows that the latent variables are of the same status as the observed variables. To find out about the posterior distribution of the latent variables, the Bayes formula is required:

\[
P(\text{Latent}|\text{Observed}) = \frac{P(\text{Observed}|\text{Latent})P(\text{Latent})}{P(\text{Observed})}
\]

where \( P(\text{Observed}) \) is the so-called marginal distribution of the observed variables. It is obtained from the joint distribution \( P(\text{Observed}, \text{Latent}) \) by integrating (or summing, if the latent variable is not continuous) across the latent variables \( P(\text{Observed}) = \int P(\text{Observed}, d \text{Latent}) \) = \( \int P(\text{Observed}|\text{Latent})P(d \text{Latent}) \). For example, if the latent variable is called \( X \), and it can only take on 2 values, 0 or 1, the marginal distribution is found by the law of total
probability: \( P(Observed) = P(Observed | X = 1) \ P(X = 1) \ + \ P(Observed | X = 0) \ P(X = 0) \). Predictions of the latent variable are based on the posterior \( P(Latent | Observed) \).

**CONCEPTUAL EXAMPLES**

**Generic Multidimensional Models**

The models that will be considered here may be represented using graphs instead of equations. In Figure 1, a generic multidimensional model is shown. The rectangles are observed variables and the latent variables are shown as circles. Directional relations are represented by single-headed arrows and correlations are represented by double-headed arrows. There are two correlated latent variables. Each latent variable exerts influences on a nonoverlapping set of observed variables. There are no additional single- or double-headed arrows connecting the observed variables themselves, which may be read from the graph to represent independence of the observed variables given latent variables.

This kind of path diagram is commonly found in the structural equation modeling literature.\(^{14}\) The structural equation models are linear simultaneous equation models derived from a successful merger of path analysis\(^{15}\) and factor analysis by Jöreskog.\(^{16}\)

At the first sight, the model depicted in Figure 1 is no more than a standard confirmatory factor analysis model. However, here the arrows do not necessarily imply a linear regression of the observed variable on a latent variable. The latent variables and observed variables are also not necessarily continuous and normally distributed, as the usual confirmatory factor analysis assumptions go. In fact, all the observed variables may be categorical item-level responses, e.g., as in a confirmatory item factor analysis model.\(^{17}\) Indeed, one of the latent variables could also be a latent classification variable (such as different solution strategies), whereas the other one remains a continuous latent trait (such as proficiency), which would turn the model into a mixture item response theory model.\(^{18}\) Or, the observed variables could be made up of a mixture of discrete and continuous variables. This kind of flexibility will allow the researcher to consider qualitatively different types of data in the same analysis, and this is illustrated in Figure 2 with a hypothetical example.

Consider each of the gray rectangles (to the left) to be observed variables that are indicators of a latent variable that describes the level of proficiency of the examinee (the larger circle to the bottom left). For instance, these may be the correct or incorrect responses to the tasks or problems that the examinee is trying to tackle, and one may use item response theory models to specify the relation between observed and latent variables. Consider at the same time, response time data can be collected. These are the rectangles to the right of the figure. Models for reaction time\(^{19,20}\) can be used to measure the differential speed at which examinees are responding. This is another latent variable, depicted as a larger circle to the lower right of the figure. The proficiency and speed latent variables may be correlated and depending on the study design, one may replace the correlation by another model indicating the degree to which there is speed accuracy trade-off\(^{20}\). Because for each task–examinee interaction, two observed data points are collected, one for accuracy and one for speed, there may still be residual correlation not fully accounted for. That is, the correlation between the accuracy and speed observations cannot be entirely explained by the latent accuracy and speed variables. Thus, additional latent variables are specified (the smaller circles at the top of the figure). These latent variables are effectively residual correlations,\(^{21}\) such that after controlling for their influences, the joint set of responses and reaction time observed variables become conditionally independent.

**TABLE I.** Classes of Latent Variable Models

<table>
<thead>
<tr>
<th>Latent Variable</th>
<th>Observed Variable</th>
<th>Continuous</th>
<th>Discrete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor Analysis/Structural Equation Modeling</td>
<td>Continuous</td>
<td>Discrete</td>
<td></td>
</tr>
<tr>
<td>Item Response Theory/Latent Trait Model</td>
<td>Discrete</td>
<td>Latent Profile Analysis/Mixture Modeling</td>
<td>Latent Class Analysis</td>
</tr>
</tbody>
</table>
The parameters of the model in Figure 2 would typically include item-level parameters such as the difficulty of the tasks, person parameters such as the predicted level of proficiency, and information on an additional time-related dimension such as the time intensity of one task versus another (i.e., some tasks require more time to solve than others in general). This model, together with the estimated parameters, may then be used in new simulation tasks and with new examinees or trainees to provide adaptive, online, real-time feedback and selection of tasks.

**Longitudinal Models**

Latent variable models are also useful for longitudinal data. Within the structural equation modeling framework, the so-called latent growth curve models have already been applied extensively to the analysis of repeated measures data. It is understood that latent curve models and hierarchical models for growth are closely connected or may even be equivalent under certain conditions. In these models, growth trajectories are estimated as functions of parameterized basis curves, linear or nonlinear, and the serial dependence are modeled by estimating additional covariance components for the random effects (individual-specific deviations from the average trajectories). For example, a standard linear latent growth curve model can provide estimates of the average initial status, average rate of change of the developmental trajectory, as well as variances and covariances of the individual differences in initial status and rate of change.

Here, another kind of latent variable model for longitudinal data is considered, that is, dynamical systems models with latent variables—some may call this a dynamical factor model. If the focus of growth trajectory models is on interindividual difference, then a dynamical systems model defined directly on the multivariate time series focuses on both intra-individual difference and inter-individual difference.

Figure 3 depicts a dynamic factor model for a set of observations from a single individual on four repeated occasions. Moving from left to right, one can see that at each time point, two potentially correlated latent variables are specified. These latent variables are each measured by a set of nonoverlapping observed variables. The latent variables at the next time point are related to the previous ones via a first-order dependence/regression relation, which is seen from the single-headed arrows that connect the circles representing time-specific latent variables. Browne and Nesselroade refer to this model as a process factor model.

The dynamical systems model depicted in Figure 3 can be separated into two parts (the state-space formulation). First, the sequence of latent variables evolving over time is known as the state sequence. Second, the connection between the latent and the observed variables is called the measurement model. This so-called state-space formulation is well known in time-series analysis and it has a natural correspondence with the general formulation of latent variable models given earlier in the section on “General Latent Variable Models.” The analysis is referred to as being on the time-domain because the model can be directly fitted to the multivariate time series data. The fitted model will provide model-based, smoothed (i.e., free of random irregularities) estimates of the evolution of the underlying unobserved states, as well as prediction and forecasting. Estimation of state-space models are often tied to various forms of Kalman filter techniques.

The model in Figure 3 is for a single individual, examining the intra-individual difference. Figure 4 simply joins together, in a hypothetical situation, observations from 5 individuals. For each individual, there is a multivariate time series of both latent variables and observed variables. The dynamic latent variable model provides a convenient approach to examine both intra- and interindividual difference in the latent state variables.
The seemingly complex model in Figure 4 is probably closest to the reality of data collected from computer-based medical simulations and games. Each time point may be thought of as a task or more generally an interaction between the examinee and the system. Markers of their proficiency, learning, may be collected from the observable responses, e.g., accuracy or strategies chosen. The latent variable states can be simple regression equations relating one continuous latent variable to another across two time points; or they could be discrete transitions from one state to another, with transition probabilities as parameters to be estimated. The parameters describing the latent variable states can be time-invariant or time-varying, depending on whether the latent time series can be thought of as a stationary or nonstationary process. The mathematical definition of a stationary stochastic process is well beyond the scope of this article, but it may be roughly understood as whether the distributions of the observed random variables remain invariant when they are shifted in time. When learning or practice effects are present, the changes in the latent variables may not be stationary. Time-varying parameters provide one possibility of accounting for the lack of stationarity in the observations. The measurement models at each time point may indeed be the flexible multidimensional models described in the section on “Generic Multidimensional Models.”

ADDITIONAL STATISTICAL CONSIDERATIONS

Parameter Estimation and Prediction

Models such as those shown in Figures 2 and 4 have high dimensionality in the latent variables. As shown in the section on “General Latent Variable Models,” a fundamental issue in latent variable modeling is that to obtain the key distribution of interest \( P(\text{Latent} | \text{Observed}) \), the marginal distribution of observed variables must be calculated by integrating (or summing) all latent variables out of the joint distribution. Often referred to as the curse of dimensionality, the integration problem is demanding because the amount of computation that is required is exponential in the number of latent variables, unless shortcuts are taken. Parameter estimation cannot proceed unless \( P(\text{Observed}) \) is found.

In general, Bayesian MCMC sampling-based estimation methods provide ready answers to the parameter estimation problem in latent variable modeling. The past two decades have witnessed dramatic increase in the use of MCMC in psychometric research. Models that were previously unimaginable because of the computational hurdle have now become well within reach (e.g., multidimensional item factor analysis, Edwards). For the dynamic latent variable modeling problem, however, a specific kind of Monte Carlo method may prove to be even more effective—the Sequential Monte Carlo method. In this method, the autoregressive nature of the unobserved latent variables is taken into account so that the sampling is more efficient than standard MCMC.

Checking Model Fit

Models shown in Figures 1–4, and other latent variable models not shown here are highly parameterized models. The inferences drawn from the models are contingent on the quality of the model. If the model does not fit the data adequately, the inferences cannot be trusted. Fortunately, the Bayesian estimation and computational framework has provided researchers with natural ingredients to make informed decisions on model selection and model fit checking. The so-called posterior predictive checking method is one such versatile procedure that effectively reuses output from the MCMC estimation to provide simple summaries of model fit.

DISCUSSION

This conceptual note is about the potential utility of general latent variable models for simulation-based medical assessments and training. These models are incredibly flexible, if
the researcher is willing to set aside the rigidity of previous generations of latent variable models. They can handle dependence (both intra- and interindividual) well. They are consistent with the overall message of evidence-centered design, and that the psychometric model should correspond well to the design of the assessment, showing the full complexities and probabilities dependencies of a system. The uncertainties about the latent variables and the parameters can be propagated by the use of the Bayes rule and conditioning, which provide ways to estimate system parameters and make forecasting and prediction possible. These models can inform the researcher of the characteristics of an “average” individual (e.g., a novice or an expert), but they also capture individual differences, which then may be related to known (or unknown) background characteristics. Finally, as new data accumulate, both by having more individuals or by having an individual spend more time within a simulation, parameter estimation can improve and model data fit can be tested empirically.

ACKNOWLEDGMENTS

Part of this research is supported by the Institute of Education Sciences (R305B080016 and R305D100039) and the National Institute on Drug Abuse (R01DA026943 and R01DA030466). The work reported herein was also partially supported by a grant from the Office of Naval Research, Award Number N00014-10-1-0978. The views expressed here belong to the author and do not reflect the views or policies of the funding agencies.

REFERENCES

Use of the Assessment–Diagnosis–Treatment–Outcomes Model to Improve Patient Care

Kevin F. Spratt, PhD

ABSTRACT

Objectives: The primary aim of this review was to compare and contrast the principles of the Assessment–Diagnosis–Treatment–Outcomes (ADTO) model with clinical prediction rules, shared decision making, and practice guidelines and provide a framework for validating medical simulations. Methods: Comparing and contrasting the various methods and tools used by the clinical community and patients to help make informed decisions regarding health care options. Results: The ADTO model provides an excellent theoretical framework to analyze and conceptualize the interrelationships between clinicians and patients as they go through the process of determining best practices. However, the ADTO model may be too comprehensive to implement within any given clinical prediction rule or hypothetical-construct-oriented medical simulation, at least at this point in the development of these fields. Conclusions: The way forward for developing substantive medical simulations that will tie together complex constructs is to start at the beginning and plan to make small steps as a basis for establishing the infrastructure that will lead to a giant leap for mankind.

INTRODUCTION

Since the 1980s, with the incorporation and growing importance of patient self-reported data into research and clinical practice, research methods in medicine have incorporated many of the methods developed in the social science, education, and agricultural disciples. Specifically, clinical prediction rules (CPRs), shared decision making (SDM), clinical or practice guidelines, and the application of medical simulations to a broad range of protocols associated with patient care have seen increased activity accompanied by rhetoric suggesting that these methods will lead us “out of the wilderness”—a place where it is difficult to find your way and understand what must be done to forge ahead to the promised land.

In 2002, Spratt introduced the Assessment–Diagnosis–Treatment–Outcome (ADTO) model as a framework for understanding: (1) why research, especially expensive randomized clinical trials (RCTs), has not driven improvement in medical care in general, and in orthopedics in particular—as evidenced by systematic reviews of research topics that typically conclude that more research must be done, or the emphasis in RCTs should be moved toward identifying patient subgroups since the aggregate differences between groups may not reflect the efficacy for all of the patients making up that group; and (2) what can be done to rectify these problems. This article has three aims: (1) compare and contrast the principles of the ADTO model with CPRs, SDM, and practice guidelines; (2) provide a framework for validating medical simulations; and (3) discuss the difficulties and barriers in changing the medical model in ways that will improve the quality of care and outcomes.

ADTO Model: Overview

The ADTO model simultaneously reflects the 4 major factors involved in patient care:

1. A—Assessing the patient’s symptoms, general health, health history, goals, and expectations to determine diagnosis
2. D—Specifying their diagnoses based on their assessment findings
3. T—Determining the “best” treatment based on the diagnosis
4. O—Evaluating the relevant outcomes associated with the treatment to assess efficacy and efficiency as well as costs

Although these 4 factors are well understood in medicine, surprisingly, the focus on establishing precise and accurate links between these factors assessment to diagnosis (A-D), diagnosis to treatment (D-T), and treatment to outcome (T-O) remains a challenge. Furthermore, it is important to understand that the links are hierarchical; meaning that specifying a reliable T-O link depends on precise and accurate A-D and D-T links.

In addition to these four building blocks and three links of the model, patient context, labeled as patient “X” factors or Px must also be considered. Px factors can include biological, and psychosocial factors that may mediate or moderate the final nature of any link (A-D, D-T, or T-O). For example, an assessment could provide information that leads to a diagnosis of infection. The usual “best” treatment for that infection might be to prescribe penicillin. However, a Px factor indicating that the patient is allergic to penicillin would contraindicate that treatment and an
alternative antibiotic that the patient is not allergic to would be linked to the diagnosis in this case.

When any of these links are poorly executed (lack of reliability) or poorly delineated (lack of validity), the chosen treatment and the relevant outcomes to assess treatment efficacy are based on false assumptions and, therefore, treatment efficacy evaluations are likely to be biased—meaning, at the most fundamental level, that the interpretation of the results may not reflect the true value of the treatment.

Within the ADTO model, evaluation of the treatment effects for a given patient and clinician may require different outcomes. For example, if quick recovery is the most important goal for patient A with Dx1, this might influence what treatment is the “best” fit and what outcomes are the “best” or most important factor for evaluating treatment benefits and risks. On the other hand, if pain relief is the most important goal for patient B with Dx1, again this might influence what treatment is the “best” fit and what outcomes are “best” or most important for evaluating treatment benefits and risks. Thus, under the ADTO model, a patient’s “X” factors may result in a common diagnosis (DX1) being linked to different “best” treatments that, in turn, link to different “best” outcomes.

From an ADTO perspective, an optimal study design and analysis would be to match the ADTO-based “best” treatment, which is dependent on the A-D and D-T links, with the ADTO-based “best” outcomes, which is dependent on the T-O link, and evaluate the variability of outcomes as a test of the validity of the A-D, D-T, and T-O links for maximizing treatment efficacy. Validity for the ADTO model is assessed by considering the magnitude and variance of the effect size associated with the change in outcome from baseline to follow-up adjusting for the magnitude of the baseline outcome variable. High validity is demonstrated when the effect size is large in the direction indicating improvement and the variance of the effect size is small since this pattern indicates not only the patients improved but also the improvement was consistent across the entire cohort.

When there is high variance in the effect size associated with a treatment, which is quite common in clinical RCTs, this argues that even when the group mean change is large, the effect size is large in the direction indicating improvement and the variance of the effect size is small since this pattern indicates not only the patients improved but also the improvement was consistent across the entire cohort.

Within the ADTO model, CPRs represent a methodological framework that provides a basis for establishing A-D, D-T, and T-O links. In the literature, CPRs are usually associated with the T-O link—seeking to predict the likelihood of success for a given treatment. As would be expected within the ADTO model, the success of these T-O-based CPRs tends to be low because of the lack of reliable or precise A-D and D-T links necessary to adequately define the “best” treatment of a given patient. In general, however, CPRs specify how known information can be used or structured, or modeled to predict unknown information. If the goal is to diagnose the patient, the known information probably includes the patient’s medical signs, symptoms, and results of specific physical examination tests, and the unknown information is the specific diagnosis. If the goal is to determine the “best” treatment for a given diagnosis based on the A-D link, the known information includes the diagnosis; patient-specific factors including their goals, expectations, preferences and values; and clinician-specific factors including their skill set and preferences; and the unknown information is specific treatment or treatment(s) that would maximize the likelihood of good outcomes, meaning high benefits (good efficacy and minimal costs) and low risks (side effects and complications). If the goal is to evaluate treatment success, then the known information is the factors that established the A-D and the D-T links, and the unknown information is the specific outcomes that would maximize the likelihood of accurately evaluating treatment consequences in terms of benefits and risks relative to the patient and clinician factors identified in the D-T link that informed the treatment decision. Thus, research done to establish CPRs to (1) identify diagnoses is consistent with the A-D link, (2) map a diagnosis to the “best” treatment is consistent with the D-T link, and (3) map the “best” outcomes—both benefits and risks—is consistent with the T-O link.

For many CPRs, the known information is based on expert opinion and, when available, evidence from the literature—often based on RCTs. It is generally held that experts in medicine “know what they know” and “know what they don’t know,” as opposed to nonexperts, who “think they know more than they know,” which is thought to be a result of their “not knowing what they don’t know.” The major problem with expert opinion is that experts, to paraphrase Mark Twain, “get into trouble
Use of the Assessment–Diagnosis–Treatment–Outcomes Model

because of ... what they know for sure that just ain’t so.”
A common consequence of this problem is that information systems designed to diagnose patients and establish baseline patient symptoms, health status, and physical examinations are limited to the “core” data elements known or thought to be useful for diagnosis, and, for the D-T link, for establishing what treatments would be best suited for the patient. This practice continues even though in many areas of medicine, results from RCTs often do not result in changes in practice. In part, these practices reflect one of Einstein’s more famous quotes: “The definition of insanity is doing the same thing over and over again and expecting a different result.” It also reflects the grain of truth that makes an old joke both funny and sad.

A Good Samaritan comes upon a man on his hands and knees under a lamp post apparently searching for something. The Samaritan asks the man what he is doing, and the man indicates that he is looking for his car keys. The Samaritan gets down on his hands and knees and helps the man search for the keys. After some time, the Samaritan indicates that the keys just do not seem to be here and asks the man how he came to drop the keys. The man indicated that he had dropped them when he tripped down the street a ways in a dark alley. Shocked by this information, the Samaritan asks why he is looking for his keys under the lamp post when he dropped them down the street in a dark alley. The man responded that the light is better here.

This joke is funny because of the absurdity of the reasoning the man used when deciding how he finds his keys. It is sad because this type of thinking seems to have generalized into many medical RCTs’ design and analysis protocols that recycle the methodology used in past studies that failed to produce results that affect practice. By not understanding that these methods have, in essence, mapped the lamp post when he dropped them down the street in a dark alley. The man responded that the light is better here.

This joke is funny because of the absurdity of the reasoning the man used when deciding how he finds his keys. It is sad because this type of thinking seems to have generalized into many medical RCTs’ design and analysis protocols that recycle the methodology used in past studies that failed to produce results that affect practice. By not understanding that these methods have, in essence, mapped the lamp post when he dropped them down the street in a dark alley. The man responded that the light is better here.

The general lack of progress in changing practice based on evidence-based medicine (EMB) relying on RCTs when the A-D, D-T, and T-O links have not been well established can be documented using the Cochrane Collaboration website. 29 As exemplified in a later section of this article, a common conclusion of meta-analyses available from the Cochrane Collaboration site is that there remains insufficient evidence to provide the clinician with guidelines to help inform their practice. Higher quality randomized controlled trials are needed. As predicted by the ADTO model, when there is

(1) Low reliability in the assessment tools used to establish diagnosis (a weak or invalid A-D link)
(2) Little clarity or consensus regarding what treatments are appropriate for a given diagnosis (a weak or invalid D-T link)
(3) No framework for determining what outcomes are most appropriate for evaluating treatment efficacy (a weak or invalid T-O link)
then randomized controlled trials designed to assess treatment outcomes are doomed to failure. 21 This is especially true when using traditional analyzes such as analysis of variance, which reflect aggregated differences at the treatment level rather than patient-level likelihoods of a good outcome.

What does failure mean in science? For randomized controlled trials, failure occurs when rival hypotheses are easily suggested and plausible for explaining the obtained results. Rival hypotheses shed doubt on the internal and/or external validity of a study. 21 Suppose that study researchers show that treatment A is superior to treatment B for patients with diagnosis X. Other clinicians, who may or may not be motivated by the desire to continue their practice of using treatment B, may argue that the reliability or validity of diagnosing the condition being treated was not sufficient and, therefore, it remains unclear if the patients being treated actually had disease X. The researchers counter that random assignment to treatment would likely balance the incorrectly diagnosed patients across treatment groups, to which the rival clinicians then
suggest that the efficacy and risks for both treatments are then unclear.

Since no study is perfect, there are almost always some rival hypotheses, and those motivated to discount any given study result will be able to suggest some, although the plausibility of the suggested rival hypotheses may be questionable. Thus, the goal of good science is to increase precision, replicate results, and modify methods to reduce the risks of possible and plausible rival hypotheses that have been championed by critics. Specifying these goals is often straightforward; obtaining the budgets necessary to pursue these goals is not.

In some medical circles, acknowledging lack of perfect methods is common and, in most cases, useful. For example, when treating a given patient, acknowledging that an exact diagnosis is not yet possible or remains controversial, but realizing, at the end of the day, a decision about how to treat the patient for the suspected diagnosis must be made. This is all well and good and—at least within the time frame—necessary; but why not also say, research to more precisely and accurately classify the diagnosis must become a priority? As long as the clinician can plead ignorance and proceed to treat without any consideration of conducting research to reduce that ignorance, the clinician is likely to follow the path of least resistance. As stated by Napoleon Hill,30 “The path of least resistance makes all rivers, and some men, crooked.”

Consider treatment of distal radius fractures in adults. These fractures of the wrist are common, are costly, occur across all age groups, and have problems with all three ADTO links:

(1) There is low intra- and inter-rater agreement in X-ray evaluation31–34 (a weak or unreliable and therefore invalid A-D link)

(2) The large number of RCTs and cohort studies done to develop consensus regarding what treatments are appropriate for a given diagnosis and patient with that diagnosis have been largely unsuccessful (a weak or unreliable and therefore invalid D-T link)

(3) There is no framework, or even any interest in developing a framework, for determining what outcomes are most appropriate for evaluating treatment efficacy (a weak or unreliable and therefore invalid T-O link).

The Cochrane Collaboration has used meta-analysis techniques to evaluate a number of aspects of care associated with distal radius fractures. Table I summarizes the various meta-analysis areas and their general conclusions. Note that for 7 of the 8 areas in which the Cochrane Collaboration conducted meta-analyses of randomized controlled trials investigating distal radius fractures in adults, either the evidence was insufficient to establish a “best” practice (1, 2, 3, 7, 8) or whatever possible “real” benefit that was observed came with limitations that made it difficult to recommend the treatment as a best practice (5, 6). Ironically, area 4, which indicated that there is some evidence for using external fixation in treating distal radius fractures, is a treatment modality that has become relatively

### Table I. Summary of Cochrane Collaboration Meta-Analyses and Conclusions Associated With Treatment of Distal Radius Fractures in Adults

<table>
<thead>
<tr>
<th>Area</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Closed reduction methods for treating distal radial fractures in adults72</td>
<td>Not enough evidence to decide whether there was any difference between the various methods tested.</td>
</tr>
<tr>
<td>2. Conservative interventions for treating distal radial fractures in adults73</td>
<td>Not enough evidence to tell what type of nonsurgical treatment is best for treating a broken wrist.</td>
</tr>
<tr>
<td>3. Different methods of external fixation for treating distal radial fractures in adults74</td>
<td>Insufficient robust evidence to determine the relative effects of the different methods of external fixation.</td>
</tr>
<tr>
<td>4. External fixation versus conservative treatment for distal radial fractures in adults75</td>
<td>Some evidence to support the use of external fixation for these fractures.</td>
</tr>
<tr>
<td>5. Bone grafts and bone substitutes for treating distal radial fractures in adults76</td>
<td>While bone scaffolding may improve anatomical outcome compared with plaster cast immobilization alone, there is insufficient evidence to conclude on function and safety; or on outcome for other comparisons.</td>
</tr>
<tr>
<td>6. Percutaneous pinning for treating distal radial fractures in adults77</td>
<td>Some evidence to support the use of percutaneous pinning. However, the precise role and methods of percutaneous pinning are not established. The higher rates of complications with Kapandji pinning and biodegradable materials cast some doubt on their general use.</td>
</tr>
<tr>
<td>7. Anesthesia for treating distal radial fracture in adults78</td>
<td>Insufficient robust evidence from randomized trials to establish the relative effectiveness of different methods of anesthesia, different associated physical techniques, or the use of drug adjuncts in the treatment of distal radial fractures.</td>
</tr>
<tr>
<td>8. Rehabilitation for distal radial fractures in adults79</td>
<td>Not enough evidence available to determine the best form of rehabilitation for people with wrist fractures.</td>
</tr>
</tbody>
</table>
unpopular in recent years. Instead, volar plating has become a much more frequent treatment, although the Cochrane Collaboration has yet to summarize the effectiveness of this treatment using meta-analysis.

**SDM**

Although CPR methods are appropriate for establishing the D-T link, SDM methods are becoming more and more popular as a means of helping patients and clinicians determine the “best” treatment for the patient. The SDM paradigm generally involves clinicians sharing their understandings of treatment options, including potential risks and benefits, and patients sharing their values, goals, and expectations for their treatment so that this information, in combination with the patients’ health state (e.g., comorbidities, frailty, specific treatment contraindications), can be used to determine the treatment with the best likelihood of maximizing efficacy relative to the patient’s goals, expectations, and values, and minimizing risks—side effects, complications, and high costs. Operationally, SDM typically involves treating and non-treating clinical personnel providing the patient with decision aids and decision support systems that provide information based on evidence-based practice regarding their diagnosis and the various treatment options available for treating the diagnosis. The goal is to help the patient establish a treatment preference based on their values, goals, and expectations. Typical measures used to evaluate the effectiveness of the SDM process include pre- to post-intervention improvement in (1) knowledge, (2) value clarification, (3) reducing decision conflict and regret, and (4) decision quality. Decision satisfaction and/or decision regret, often in relation to treatment outcomes, are also assessed at treatment follow-up.

A fundamental problem with SDM is common to all educational endeavors that attempt to teach people how to think and evaluate their options in complex situations: It is hard to do. A quote from Stoll and Schubert seems particularly relevant here:

- Data is not information;
- Information is not knowledge;
- Knowledge is not understanding;
- Understanding is not wisdom.

The truth of the first line is obvious to all who have been sent an Excel spreadsheet full of numbers, but with no variable names or data dictionary defining the variables. The truth of the second line becomes clear when, even armed with the variable names and what the values of each variable indicate, converting this information into knowledge requires additional effort, at a minimum generating some descriptive statistics. The truth of the third line becomes clear when one realizes that interpretation of the descriptive statistics or the inferential statistics if the data are comparing across groups is necessary to appreciate what value may lie in the knowledge. Finally, the truth of the fourth line becomes clear when having come to understand the meaning of this knowledge, there remain difficulties in determining how the implications of these understandings can be synthesized and contextualized so that actions result in making wise decisions or choices.

A new trend in SDM is to provide “coaches” to help patients assimilate the provided information or “facts” into knowledge, the knowledge into understanding, and the understandings into the wisdom that can aid the patient in developing their goals, and expectations as related to the various treatment options. The “facts” in quotes recognizes that facts, just as is the case with CPRs, are often based on expert opinion and randomized controlled trial results that typically represent the “average” patient, which is unfortunate given that the goal of SDM is to fit a particular patient and understand what treatment would be best for them. This points out the value of the Px factors that are part of the ADTO model.

The notion of using coaches, who are rarely treating-clinicians, but rather allied health personnel (e.g., nurses) trained to understand the pros and cons of the evidence and expert opinion made available to the patient in the decision aids and decision support system materials is an important methodological shift that may address the concern with asymmetrical knowledge in SDM protocols that are largely between the patient and the treating-clinician. Asymmetrical knowledge occurs when two people (sometimes referred to as the principal and the agent) wishing to strike some bargain do not have the same knowledge of the factors involved in making the bargain. A common stereotypic example would be when a woman (the principal) goes to an auto repair shop to see a mechanic (the agent) about having her car fixed. The mechanic, who perceives himself or herself having the upper hand regarding knowledge of automobiles, produces a work order that specifies a larger number of repairs than necessary and higher costs for those repairs because the woman client does not have sufficient knowledge to dispute or critically evaluate the work order.

This same asymmetrical knowledge situation is common in the patient—treating-clinician setting, where the patient (the principal) may be more knowledgeable about their symptoms, signs, and medical history, but the treating-clinician (the agent) is more knowledgeable regarding how to identify a diagnosis based on the patient’s information, what treatment options are possible, and what the potential risks and benefits of each treatment are to the patient, the clinician, the hospital or practice, and society. Just like the mechanic, the treating-clinician could run up the bill and, potentially, at the end of the treatment, provide the patient with a bill too large and a recovery too small.

Thus, asymmetrical knowledge often puts the more knowledgeable person, usually the agent (the treating-clinician), in a position of ethical or moral hazard. Regular news
Use of the Assessment–Diagnosis–Treatment–Outcomes Model

Clinical Guidelines

Clinical guidelines were developed to aid clinicians in figuring out how to best assess, diagnose, and treat patients. Traditionally, guidelines were developed by consensus that was largely based on clinical experience, expert opinion, and randomized controlled trials when available. With the rise of EBM,43–64 efforts to incorporate medical research into decision making and to train clinicians to become better consumers of medical research with the goals of translating the findings of this research into their clinical practice have proliferated. Expert opinion fills the gaps in guidelines when EMB is lacking. Not surprisingly, some clinicians remain skeptical of the role of guidelines in general and the role of EMB experts in establishing guidelines, and question the value of guidelines largely based on results of randomized controlled trials as the basis for determining a treatment plan at the patient level.

To some degree, clinicians are right to question the value of randomized controlled trials. The major problem with evidence based on randomized controlled trials is that this information is often based on aggregated data—the mean differences between treatment groups—that, by definition, is often not an accurate reflection of what the treatment effects are likely to be for the next patient. Although results gathered from randomized controlled trials often can be evaluated to provide information that would be useful for informing treatment choices at the patient level, these evaluations are not common in the literature. Spratt45 reported a patient-level evaluation of treatment outcomes in the Spine Patient Outcomes Research Trial (SPORT) for the intervertebral disc herniation (IDH) cohort of subjects where the aggregated analysis of the data65,66 showed superior results for patients treated surgically compared to patients treated nonoperatively. At 1 year, average change on the Short Form 36 version 1 physical function (PF) subscale from baseline was 44 points for patients treated surgically compared to 28 points for patients treated nonoperatively, p < 0.0001. Note that the PF subscale includes 10 items each describing various physical activities such as carrying groceries, walking a block, and engaging in strenuous activities. Each item is scored on a 3-point scale—no limitation, some limitations, severe limitations—and are scaled as 3, 2 and 1, respectively, with the sum of the items potentially ranging from 30 to 10. These scores are then linearly transformed to range from 0 to 100 with higher values reflecting greater physical functioning.

However, when analyzed at the patient level using Chi-Squared Automatic Interaction Detection67 (CHAID)—one of a number of recursive partitioning approaches—when improvement was defined as a positive change from baseline of more than 30%, an amount of change that reliably defines a minimally clinically important difference,68,69 at the aggregate level 90.75% of patients treated surgically reported improvement compared to 77% for patients treated nonoperatively, p < 0.0001. However, as illustrated in Table II, of the 10 distinct patient profiles identified by the CHAID-based decision tree results for the patients treated surgically, five of the profiles identified patients with improvement rates above 97% (1, 2, 4, 5, and 9),

<table>
<thead>
<tr>
<th>Selected Variables</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline PF</td>
<td>≤25</td>
<td>≤25</td>
<td>≤25</td>
<td>≤25</td>
<td>≤25</td>
<td>(25–40)</td>
<td>&gt;40</td>
<td>&gt;40</td>
<td>&gt;40</td>
<td>&gt;40</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Sensory Deficit</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>MCS ≤ 35</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>Female</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Herniation Type</td>
<td>Protrusion</td>
<td>E/S²</td>
<td>E/S²</td>
<td>E/S²</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline Sciatica</td>
<td>≤18</td>
<td>≤18</td>
<td>≤18</td>
<td>≤18</td>
<td>≤18</td>
<td>≤18</td>
<td>≤18</td>
<td>≤18</td>
<td>≤18</td>
<td>≤18</td>
</tr>
<tr>
<td>Percent Improved (%)</td>
<td>97.83</td>
<td>99.84</td>
<td>88.89</td>
<td>100.00</td>
<td>100.00</td>
<td>91.53</td>
<td>60.53</td>
<td>86.67</td>
<td>97.78</td>
<td>63.33</td>
</tr>
<tr>
<td>N</td>
<td>46</td>
<td>47</td>
<td>27</td>
<td>23</td>
<td>154</td>
<td>118</td>
<td>38</td>
<td>45</td>
<td>45</td>
<td>30</td>
</tr>
</tbody>
</table>

*E/S means that the disc herniation type was Extruded or Sequestered.*
three of the profiles identified patients with improvement rates slightly lower than the average 90.75% rate (3, 6, and 8), and 2 of the profiles identified patients whose improvement rates were lower than 64% (7 and 10), an aggregated success rate below the average success rate of 77% for patients treated nonoperatively. Note that for the Surgery profiles, 7 variables were involved in defining the 10 profiles but no one profile used more than 4 variables. For example, Surgery profile 1 included 46 patients with the profile defined by a baseline PF score ≤ 25, reporting some comorbidities but no sensory deficit; Surgery profile 7 had 38 patients, of which 60.53% had a good outcome with the profile defined as having a baseline PF score > 40 and a disc herniation classified as a protrusion.

In Table III, of the 8 distinct patient profiles identified by the CHAID-based decision tree results for patients treated nonoperatively, three of the profiles identified patients with improvement rates above 88% (5, 6, and 8), two of the profiles identified patients with improvement rates around the 77% aggregate rate (2 and 3), and three of the profiles identified patients whose improvement rates were lower than 63%, well below the 77% aggregate success rate for the nonoperative care (1, 4, and 7). Note that for the nonoperative profiles, 5 variables were involved in defining the 8 profiles but no one profile used more than 3 variables. For nonoperative profile 1, 38.89% of the 36 patients with this profile demonstrated improvement when they reported that baseline leg pain was ≤ 2 and symptom stability was not better. In contrast, for nonoperative profile 5, 100% of the 54 patients with this profile demonstrated improvement when they reported baseline leg pain as > 4 or when no leg pain score was provided (missing), and when the patient’s body mass index was less than 24.4.

At present, “best” evidence in the literature used in developing clinical guidelines, in developing decision aids and decision support systems, and in developing CPRs for comparing the relative efficacy of surgery or nonoperative care for IDH based on randomized controlled trials—where the group or aggregated differences in treatment are reported—would suggest that surgery has better efficacy. However, patient-level analysis of the SPORT IDH cohort, one of the largest and most well-known and discussed studies of this issue, provides a much different solution and approach for determining the “best” treatment for the next patient with IDH seeking care. Specific patient profiles showed substantially different probabilities of improvement. As a patient or a clinician, which information would you find more useful in helping to make a treatment choice?

(1) The average patient gain from baseline on PF at 1 year was 44 points for patients treated surgically and 28 points for patients treated nonoperatively.

(2) The overall rate of patients reporting improvement on PF at 1-year follow-up was 90.75% for patients treated surgically and 77% for nonoperatively treated patients.

(3) Patients with your surgical care profile (i.e., baseline PF > 40 with a herniation type classified as a protrusion) reported improvement from surgery about 60% of the time (see Table II), but patients with your nonoperative care profile (i.e., baseline leg pain of 4, and a body mass index of 23 as illustrated in Table III) reported improvement at 1 year 100% of the time.

“Perhaps both the patient and treating-clinician should consider all three sets of information when deciding on the “best” treatment for what the ADTO model described as patients of type “X,” meaning a patient with a given profile.”

### Medical Simulations

Simulations as a learning tool have a long history in medical domains involving hand–eye coordination and are especially useful when failure to correctly execute a particular action, often in response to an unanticipated set of circumstances, is associated with dire consequences. In these cases, learning by doing would often degenerate into one-trial learning, with the lesson learned is that your patient is dead, or at least much the worse for wear. Thus, simulators associated with piloting aircraft, functioning in outer space, and surgical procedures are common and some would argue indispensable in training personnel to a standard of excellence.

The use of simulations in medicine have now expanded to include what might be called hypothetical-construct coordination, since the simulations are designed to help train clinical

---

**TABLE III.** CHAID-Based Decision Tree Profiles Identifying Patient Subgroup Likelihoods of Greater Than 30% Improvement From Baseline for Intervertebral Disc Herniation Patients Treated Nonoperatively

<table>
<thead>
<tr>
<th>Selected Variables</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Leg Pain (0/6)</td>
<td>≤2</td>
<td>≤2</td>
<td>(2–4)</td>
<td>(2–4)</td>
<td>&gt;4, Missing</td>
<td>&gt;4, Missing</td>
<td>&gt;4, Missing</td>
<td>&gt;4, Missing</td>
</tr>
<tr>
<td>Symptom Stability</td>
<td>No</td>
<td>Better</td>
<td>Better</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Joint Symptoms</td>
<td>Body Mass Index</td>
<td>≤24.4</td>
<td>(24.4–26.6)</td>
<td>&gt;26.6</td>
<td>&gt;26.6</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Smoking History</td>
<td>Percent Improved</td>
<td>38.89%</td>
<td>84.00%</td>
<td>78.67%</td>
<td>52.00%</td>
<td>100.00%</td>
<td>90.00%</td>
<td>62.79%</td>
</tr>
<tr>
<td>N</td>
<td>36</td>
<td>25</td>
<td>75</td>
<td>25</td>
<td>54</td>
<td>31</td>
<td>43</td>
<td>54</td>
</tr>
</tbody>
</table>
personnel to integrate information from multiple sources, such as patient-reported symptoms, signs observed when examining the patient, other diagnostic tests such as radiographic findings (X-ray, magnetic resonance imaging, and Computed Tomography), and ultrasound with specific diagnoses and treatment options for the purpose of arriving at a particular decision or understanding of the system.

Current practices in medicine tend to develop, structure, and parameterize simulations by creating ontologies, where the knowledge presentation is used to capture information and knowledge about the subject at hand. In medical simulations focusing on hand–eye coordination, high fidelity regarding instrumentation, anatomy, and physiology is an important and perhaps even dominant aspect in the ontology development process. In contrast, in medical simulations focusing on hypothetical-construct coordination, high fidelity is, in essence, the ability of the actor to surface the necessary and sufficient information needed to achieve the simulation goal, which would be generically to identify the circumstances when and what: (1) supplemental assessments, beyond gathering typical intake information, are necessary to determine a diagnosis; (2) confirmation protocols for the provisional diagnosis are necessary to confirm the diagnosis; (3) factors are relevant for determining the most appropriate treatment; and (4) outcomes are relevant for documenting best treatment.

Thus, in hypothetical-construct-oriented medical simulations, as opposed to hand–eye coordination medical simulations, these ontologisms might well be based on information provided by CPRs and SDM materials. “As such, the cautionary tales associated with CPRs and SDM materials and protocols that rely on expert opinion and aggregated or group-level results from randomized controlled trials that are used to inform the A-D, D-T, and T-O links within the ADTO model apply.”

The ability to assess fidelity and validity in hand–eye coordination medical simulations is similar to the notion of representational measurement; it is relatively straightforward by visual inspection to evaluate the similarity of the simulation environment with the real-world environment and many of the metrics associated with success are also relatively straightforward including such things as reaction times, and adherence to protocols. For example (1) time to complete a task, (2) reaction time with the introduction of “adverse” events, and (3) adherence to set protocols, and accuracy and speed of completion of the maneuvers within each protocol. The ability to assess fidelity in hypothetical-construct medical simulations is more akin to nonrepresentational measurement, such as assessing achievement or intelligence by testing. Nonrepresentational measurement is necessary when one wants to quantify hypothetical constructs, and solutions for classic problems of reliability and validity associated with nonrepresentational measurement, along with the large body of measurement theory, are available to evaluate these constructs.

On the other hand, hypothetical-construct-oriented medical simulations represent a quantum leap in difficulty for a number of reasons. First, in some medical disciplines the lack of assessment sensitivity (the Probability of a test being positive $[T_e+]$ given that, in truth, the patient is positive $[T_r+]$ or $Pr[T_e+ | T_r+]$) and positive predictive value $Pr (T_r+ | T_e+)$ are not high-priority concerns; and second, and as documented throughout this article, the CPRs and shared decision making, and expert opinion and EMB clinical guidelines associated with many medical conditions do not provide information that is particularly useful for specifying high-quality ontologisms associated with determining at the patient level the A-D, D-T, and T-O links that the ADTO model suggests are necessary to establish a reliable and valid information system that can be used to further refine reliable and valid best practices for patient treatment. Reliability is the bedrock upon which diagnosis, treatment, and outcomes must stand.

One consequence of poor reliability in assessment is showed by Spratt and Koval (presentation at the Orthopaedic Trauma Association 24th Annual Meeting, October 2009). They evaluated inter- and intra-rater consistency associated with measuring distal radius wrist fractures from a series of three X-rays. This study codified aspects of six different distal radius classification systems that describe a distal radius fracture into a series of basic 2- and 3-option questions grouped within the four domains: Fracture (1) location, (2) displacement, (3) comminution, and (4) stability. The study design randomized the order of 24 sets of posterior-anterior (PA), Lateral, and Oblique X-ray sets for each of the 7 raters (4 orthopaedic residents, 1 orthopaedic hand surgeon, 1 orthopaedic trauma surgeon, and 1 musculoskeletal radiologist) and created a web-based application that combined Image J software (version 1.40g, NIH) to evaluate the X-rays with a conditional question algorithm where subsequent questions were based on the rater’s prior answer to characterize the fracture with the four domains. The purpose of this study was to begin the validation process of a new classification system by assessing both the inter- and intra-rater agreement. Thus, all raters were required to evaluate all 24 film sets on two occasions to establish intra-rater agreement and all 21 possible rater pairs (7 choose 2 = 21) were evaluated against each other for both the raters’ first and second readings of X-rays. Reliabilities were computed for each question in each of the four domains, the goal being to develop a classification system composed of the reliable components of each of the six various systems. Since the questions were generally 2- or 3-respose options only, Kappa statistics were used to assess consistency. Both Cohen’s and Brennan–Prediger Kappas were used since Cohen’s Kappa is the de facto standard but is also known to produce unstable results when the distribution of scores across the categories are highly skewed whereas the Brennan–Prediger Kappa statistic does not have this problem.
As shown in Table IV, there was 100% agreement amongst all raters on both rating occasions regarding the question of whether the fracture was of the left or right wrist. After that, consistency of response ratings was much less impressive. At assessment time 1 (first reading), the average inter-rater consistencies were moderate (\(K_c = 0.5130\)) to very poor (\(K_{BP} = 0.3522\)). At assessment time 2 (second reading) the average intra-rater consistencies were somewhat better: good (\(K_c = 0.6190\)) and poor (\(K_{BP} = 0.5714\)).

The average intra-rater consistency across the seven raters was moderate for both Cohen’s and Brennan–Prediger Kappas (\(K_c = 0.5342\) and \(K_{BP} = 77.68\)). Consistency or, more aptly, inconsistencies of ratings across the entire set of questions were similar. The poor to very poor Brennan–Prediger Kappas for the question associated with intra- versus extra-articular fracture, which is often a fundamental and primary factor associated with treatment options, suggested that the steps taken to reduce questions to 2 or 3 options and to use conditioning logic in the web-based application to force a standardized approach for evaluation did little to elevate the poor inter-rater consistency in clinicians trained in these assessments. Thus, RCTs designed to evaluate for differential efficacy of various treatments thought to be more beneficial for intra- or extra-articular distal radius fractures are likely to provide biased results given the A-D link—the ability to accurately classify the distal radius fracture as intra- or extra-articular—is not sufficiently reliable or precise. Ironically in some ways this lack of precision has positive consequences for the treating physician since the lack of precision provides the treating physician with few limitations in determining the “best” treatment for the patients since there is so little evidence to contraindicate any choice made.

**Closing Comments**

A major “take home” message from the concepts and consequences associated with the discussion in this article should be that CPRs, shared decision making, clinical guidelines, and medical simulations have a great deal in common and each of these methods would benefit by taking the lessons learned, both positive and negative, in each methodological discipline to heart. The ADTO model provides an over-arching point of view of the medical model and provides an excellent theoretical framework to analyze these medical issues, but the ADTO model may be too comprehensive to implement within any given CPR or hypothetical-construct-oriented medical simulation, at least at this point in the development of these fields.

However, the ADTO model approach of minimizing variance in outcomes by creating profiles or subgroups may represent an iterative modeling approach that can be used to simultaneously construct and subsequently build validity evidence for CPRs, SDM protocols, and medical simulations. Each of these methods is still in its infancy, perhaps farther along than the Wright brothers in their quest to build a flying machine but not yet to World War I aircraft. This metaphor is useful to consider when tempted to start by building too big. Consider what would have happened if success for the Wright brothers was defined for them as: the ability to fly 500 people from Kitty Hawk North Carolina to Paris, France in under 10 hours. Their 12-second first manned flyer would have been defined as a colossal failure and other groups exploring the possibility of flight whose definition of success in their early attempts was more reasonable would have made history. This success—building a Boeing 747 that can fly 500 people from Kitty Hawk to Paris in well under 10 hours—that occurred only 66 years after the 1903 flight at Kitty Hawk, was based on a large number of minor improvements and successive approximation, and a few breakthrough epiphanies, usually based on knowledge gained by carefully documenting both successes and failures. The important thing to remember is that the nature of medical science is much the same: long-term success comes by accumulating small gains across time.
REFERENCES

Harold F. O’Neil is a Professor of Educational Psychology and Technology at the University of Southern California’s Rossier School of Education. His research interests include the computer-based teaching and assessment of 21st century skills, e.g., adaptive problem solving; the teaching and assessment of self-regulation skills, e.g., metacognition and self-efficacy; and the effectiveness of computer simulations and games to solve education and training problems, e.g., medical. At USC, he has been active in the development of technology-based student assessment approaches involving public schools through a research relationship with UCLA/CRESST. His research is funded by the Institute of Education Sciences (IES), Office of Naval Research (ONR), Defense Advanced Research Projects Agency (DARPA), and the Telemedicine and Advanced Technology Research Center (TATRC). A prolific writer, Dr. O’Neil has recently co-edited four books on the following topics: What Works in Distance Learning; Assessment of Problem Solving Using Simulations; Computer Games and Team and Individual Learning; and Teaching and Measuring Cognitive Readiness (publication expected in 2013 by Springer Publishing). He is a Fellow of the American Psychological Association (APA), the American Educational Research Association (AERA), and the Association for Psychological Sciences (APS).

Karl E. Friedl retired from the US Army in January 2013 after thirty years of service and is currently a research mentor and consultant. His last assignment was as the Director, Telemedicine and Advanced Technology Research Center (TATRC) for the US Army Medical Research and Materiel Command in Frederick, Maryland. He received his PhD (Integrative Physiology) from the University of California at Santa Barbara, in the University’s Institute of Environmental Stress in 1984. As a research physiologist, he has published 150 papers, book chapters, and other reviews and technical reports. As a research director and laboratory commander, he has facilitated laboratory cooperation between government agencies including the DoD, VA, NIH, NASA, FDA, and USDA and organized and directed research initiatives such as military women’s health, Gulf War illnesses, bone health, and Parkinson’s research programs. His research focus areas are applications of metabolism and neurobiology technologies, assessment methods and metrics of research success, and strategies to accelerate research translation to practice, especially through useful models and simulations. He is a member of the American Society for Nutrition, AMSUS, Endocrine Society, and a fellow of American Institute for Medical and Biological Engineering (AIMBE). His awards and recognition include Legion of Merit (2d OLC), Morris K. Udall Award for Parkinson’s Research Advocacy, Ronald and Nancy Reagan Alzheimer’s Research Award, Founder’s Award for the Pennington Biomedical Research Center, Order of Military Medical Merit, and Diabetes Technology Society Research Leadership Award.

Ray S. Perez is a Program Officer at the Office of Naval Research (ONR), War Fighter Performance and Protection Division where he is the group leader for Training and Education. He is currently managing a research program on technology-based education and training, an area he has worked in for over 20 years. His research program scope includes investigations of the use of various artificial intelligence methods and techniques to enhance the design of web-based learning, intelligent tutors, simulations, and games for the application of emerging cognitive theories of learning and teaching. He also manages a program in neuroscience whose goal is to unravel the complex relationship between behavior and the underlying neural mechanisms responsible for learning and memory. He has received numerous awards for his work in these areas. He has co-edited four books on the use of technology in training and education, most recently a book entitled Computer Games and Team and Individual Learning. Prior to coming to ONR he served as Program Manager for the Presidential Technology Initiative Program at the Department of Defense Education Activity (DoDEA); Principal Scientist in Simulation and Advanced Instructional Systems at the US Army Research Institute for the Social and Behavioral Sciences (ARI); and Assistant Professor in the Department of Psychology, California State University Dominguez Hills. Dr. Perez received his master’s and doctorate in Educational Psychology with emphasis on Cognitive Psychology from the University of California, Los Angeles.

Kevin Kunkler currently serves as Chair for Medical Modeling and Simulation for the Joint Program Committee (JPC-1) at Ft. Detrick, Maryland. Dr. Kunkler received his medical degree from Indiana University School of Medicine and proceeded to Vanderbilt University Dept. of Surgery for his surgical residency. Dr. Kunkler served as VP and Medical Director at Immersion Medical prior to joining the University of Maryland School of Medicine as Visiting Instructor where he was loaned to the Telemedicine and Advanced Technology Research Center (TATRC) as Senior Subject Matter Expert for the medical modeling, simulation, and training technology portfolio. Dr. Kunkler is currently seeking a Masters in Regulatory Science at Johns Hopkins University.

ACKNOWLEDGMENTS

We thank Joanne Michiuye and Rob Sitter for their extraordinary efforts to bring this special issue together. Joanne Michiuye and Rob Sitter are members of the CRESST team at UCLA. They prepared this special issue, by coordinating and assembling the manuscripts through multiple drafts and edits, and interfacing with the authors and editors to achieve a coherent series of articles to address the theme of this issue. We also thank Francisco Flores of CRESST/UCLA for his exemplary assistance in planning and coordinating the workshop upon which this special issue was based.