Assessing the Skills of Surgical Residents Using Simulation

Mohsen Tavakol, PhD,* Mohammad Ali Mohagheghi, MD,† and Reg Dennick, PhD*

*Medical Education Unit, University of Nottingham, Nottingham, United Kingdom, and †Medical Sciences/University of Tehran, Cancer Institute Research Centre, Tehran, Iran

KEY WORDS: simulation-based assessment, surgical technical assessment, OSATS, virtual reality in surgery

COMPETENCY: Patient Care, Medical Knowledge, Practice Based Learning and Improvement

INTRODUCTION

Simulation-based assessment is a new paradigm in medical education. The history of simulation is rich and long, and much of it is derived from the military and the aviation industries since the 1930s.1 Simulation has been defined as “A situation in which a particular set of conditions is created artificially in order to study or experience something that could exist in reality.”2 Simulation provides a safe and supportive educational climate.3 Unlike patients, simulators do not become embarrassed or stressed; they have predictable behavior; are available at any time to fit curriculum needs; can be programmed to simulate selected findings, conditions, situations, and complications; allow standardized experience for all residents; can be used repeatedly with fidelity and reproducibility; and can be used to train for both clinical skills and high stakes examinations.4 Since the 1980s, the use of simulation technology in clinical education has become popular for the training and the assessment of medical students and residents.5 Increased attention to patient care and ethical issues, demands for innovation in clinical education, and accelerating advances in diagnostic and therapeutic procedures have all prompted a growing interest in the use of simulators for medical training. As skills become more complex, the challenge of assessment increases. Clinical simulators such as part task trainers, computer-based systems, virtual reality and haptic systems, simulated patients, simulated environment, and integrated simulators have been used effectively to assess technical skills.6 Such clinical simulators for assessment have 3 main purposes: to optimize the capabilities of all residents by providing motivation and direction for future learning, to protect the public by identifying incompetent residents, and to provide a basis for choosing applicants for advanced training.7

Primarily, medical skills simulators focus on training in specific tasks. The concept has now developed so that a simulator is not only a training device, but it is at the same time an assessment tool. It has been argued that training and assessment are simply 2 sides of the same coin.8 Therefore, simulators have the potential to be a valuable formative (proximate) and summative (terminal) assessment tool.

The purpose of this article is to provide an overview of assessment in simulation-based surgical skills training and to address the key features of the methods. We did not perform a systematic meta-analysis nor do we present an in-depth review of the topic. However, literature searches were performed in Medline and Educational database using the following keywords: simulation, simulation in surgery, virtual reality, simulation-based assessment, performance-based assessment, OSCE, OSATS, reliability and validity, haptic systems, objective assessment, competence, surgical performance, standards, residents (specialist registrar), hand motion analysis (HMA), and dexterity analysis. All abstracts of retrieved articles were reviewed for their relevance to the subject. Relative articles were reviewed, and their bibliographies were searched for associated articles not covered in computerized searches.

VALIDITY AND RELIABILITY OF CLINICAL SIMULATIONS

The 2 most important aspects of precision are reliability and validity. The general notion of validation incorporates 2 different processes: verification and validation.8 Verification is the process of assessing that a model is operating as intended. Validation is the process of assessing that the conclusions reached from a simulation are similar to those reached in the real-world system being modeled. In other words, “Validation is the process of determining that we have built the right model, whereas verification is designed to see if we have built the model right.”9

Some crucial questions that must be asked before making the choice of simulation-based assessment method include the following: (1) Is it valid? (Does it measure what it purports to
measure?), (2) Is it reliable? (Is the measure of performance consistent?), and (3) Is it feasible? (Is the method practical in a busy clinical environment? What is the cost?). The current enthusiasm for the validation of assessment tools, particularly simulators, is new in clinical education. However, scientific validation is a common practice in psychological tests, including objective tests. In 1974, the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education developed a set of standards for judging any assessment tools. These organizations provide rigorous guidance on validation, reliability, and error of measurement of tools.

Test validity is a concept that has evolved with the field of psychometrics but that textbooks still describe commonly as the degree to which a test measures what it was designed to measure. In other words, the answer to the question “Am I measuring what I am supposed to be measuring?” must be positive. Several validation benchmarks have been developed to assess the validity of a test or testing instrument. These benchmarks include face validity, content validity, construct validity, and criterion validity.

Face validity (realism) addresses the extent to which the examination resembles the situation in the “real world.” For example, does a suturing task in a bench top laparoscopic model resemble laparoscopic suturing in the real-world environment?

Content validity refers to the extent to which a measurement reflects the trait or domain it purports to measure. For example, an assessment of a resident who performed a laparoscopic cholecystectomy on an anaesthetized pig has higher content validity as a measure of surgical skill than does a multiple-choice examination on the anatomy of the gall bladder.

Construct validity describes the agreement between a theoretical concept and a specific assessment tool. For example, to demonstrate that a new simulator has construct validity as a measure of technical performance more senior surgeons should score higher on its assessment parameters than more juniors.

Criterion validity refers to the extent to which an assessment tool correlates with other measures of performance. Two types of criterion validity are as follows: Predictive validity refers to the ability of a tool to predict future performance [eg, the ability of performance on the Minimally Invasive Surgical Trainer—Virtual Reality (MIST-VR) to predict future performance as a surgeon]. Concurrent validity describes the correlation between an assessment tool and the perceived “gold standard” (eg, performance in the laboratory measured by global ratings compared with performance measured by academic staff ratings).

Reliability is a measure of a test to generate similar results when applied at 2 different points. In other words, an assessment tool is reliable if it yields consistent results of the same measure. It is unreliable if repeated measurements yield different results under similar conditions. The development of reliable behavioral assessment tools is relatively new to surgery. The behavioral approach to the assessment of 2 raters’ reliability takes a much more rigorous approach; as a statistical method, it is very straightforward and systematic. This approach simply asks whether the raters agreed or disagreed on the occurrence/nonoccurrence of an event. Usually, these events are scored from unambiguous event definitions. The metric used to assess the statistical quality of that methodology is known as inter-rater reliability (IRR). The IRR indicates the degree or proportion of times to which 2 or more independent observers agree in their rating of a test subject’s behavior. It is normally calculated as follows: observed event agreements/total number of observations. Standard error of measurement (SEM) is another method for estimates of agreement. The SEM is concluded from knowledge of the internal consistency (α) of the test and the SDs of the test scores. Thus, the formula for the standard error is as follows: 

$$SEM = SD \times \sqrt{1 - \alpha}.$$  

The larger the SEM, the higher the reliability of the data, and vice versa. This calculation gives a more accurate reflection of data in terms of absolute agreement.

The SPSS software program (SPSS Corporation, Chicago, Illinois) analyzes these statistical phenomena.

DEVELOPMENTS IN SKILL SIMULATORS

The last several years have seen the rapid growth of simulation centers in medical schools, community colleges, and other health-care training programs. The term “simulation center” now refers to a training and assessment laboratory that includes all simulation technologies such as warm simulators (eg, standardized patients or SPs) and cold simulators (full-size electromechanical human patient simulators, part task/body trainers, surgical trainers, and virtual reality programs and simulators).

Motivating this growth are the competency-based medical board SP examinations, the shift to competency-based curricula in medical schools, and the growth of technology companies that manufacture the simulation equipment (ie, simulators) and the supporting equipment (eg, digital video systems and data collection systems such as evaluation and management systems). In response to problems with traditional learning and assessment methods, developments in resuscitation and anesthetic training have seen the adoption of simulation in the training and assessment of residents. Within this movement, significant emphasis is placed on effective team-based working, with the drive to develop an interprofessional simulated ward environment and to modify criteria for high stakes assessment.

Simulators are not used just for training; they are used in the assessment of medical practice. The assessment of technical skills is a vital part of ensuring that we have trained high-quality residents before they are allowed to operate unsupervised. The push for simulation in the assessment of residents’ technical skills is because the assessment of technical skills in the hospital environment is difficult particularly in intensive care and in surgery settings. Furthermore, by using models for objective surgical skills assessment, residents can be allowed to operate completely independently, which provides an opportunity to make surgical decisions and to demonstrate judgment without encountering the obvious ethical dilemmas the same would arouse on a human patient. The advantages of using simulators in training and assessment include the following:
Risks to patients and learners are avoided, undesired interface is reduced, tasks can be created to demand, skills can be practiced repeatedly, training can be tailored to individuals, retention and accuracy are increased, and transfer of training from classroom to real situation is enhanced. These advantages have encouraged medical educators to focus on techniques that standardize the assessment process outside the therapeutic settings. One such technique is the Objective Structured Assessment of Technical Skills (OSATS), in which residents perform a series of standardized surgical tasks on inanimate models under the direct observation of an expert surgeon. However, although extensive data in the literature supports objective assessment of clinical skills, particularly the validation and reliability of assessment devices, developing simulations for the assessment of qualities such as teamwork and professionalism are still a challenge for medical educators.

**FRAMEWORK FOR CLINICAL ASSESSMENT**

Miller’s model for assessing declarative knowledge and procedural knowledge is very popular in clinical education. Miller’s pyramid presents 2 layers of declarative knowledge defined as “knows” and “knows how and two layers of procedural knowledge,” which are defined as “shows how” and “does.” Multiple-choice questions, essays, oral examinations, and patient management problems can assess declarative knowledge but are inadequate in assessing clinical performance that includes motor skills and attitudes and the interaction of an individual doctor with his/her colleagues and patients. Miller stated that “despite the significant advances in testing procedures that probe the quality of being functionally adequate, or of having sufficient knowledge, judgment, skill, or strength for a particular duty, skeptics continue to point out that such academic examinations fail to document what residents will do when faced with a patient, i.e., to demonstrate not only that they know and know how but can also show how they do it” (p. S63). It could be argued that without considering the assessment of the “shows how” (performance) layer of Miller’s pyramid, the assessment of residents (specialist registrars) in medicine would be a subjective process. It has been argued that greater emphasis needs to be placed on performance-based assessment. The use of standardized patients in simulated clinical encounters is the best thing to assess the clinical skills of students and residents.

Thirty-two years ago, Harden et al introduced the Objective Structured Clinical Examination (OSCE) as a means of assessing clinical behaviors that might be evaluated in a reasonable time frame using facilities and resources generally available at medical schools. The OSCE has been developed as a reliable method of assessing clinical performance. The OSCE model, which uses a series of standardized tasks assessed using structured sheets, could be used as a powerful assessment tool of the prototype performance-based assessment to evaluate the skills required of doctors. However, although objective structured clinical examinations, long and short cases, and simulated patients are used to evaluate clinical skills, they occur in an artificial examination setting and therefore lose some validity. They evaluate performance (shows how) rather than real clinical skills, which is the top layer of Miller’s pyramid (“does”) according to Miller. What concerns Miller is the action component of professional behavior, which is clearly the most difficult to measure accurately and reliably: what doctors do in a real-life environment. Miller argues that it is the assessment of “does” that is most relevant to a physician’s actual performance. For example, Rethans et al showed that scores given to general practitioners by simulated patients in an examination setting were significantly higher than the scores given for the same task in a real-life environment. Similar concerns come from other authors. Hodges, for example, questions the sociological assumptions within the OSCE and asks whether we can safely relate findings from an OSCE performance back to clinical practice. These findings draw attention to the need to evaluate what doctors do in practice (ie, move away from simulated situations to real practice settings) as well as the traditional examination that uses simulated patients in an artificial environment that cannot assess what the resident “does” in a real clinical world. Hence, we need to have new assessment tools to predict accurately what a doctor does when functioning independently in a clinical practice.

Some research studies have assessed as many different observations of real consultations as possible using global rating or judgments. Methods such as clinical work sampling and practice video recording have been developed to improve the existing system of in-training assessment. These methods have been criticized as being unreliable and invalid for assessing residents’ performance during clinical education. These methods require the observation of many different patient encounters as well as the judgment of an expert. Although the individual observations will be subjective, data collection together from different sources by a large number of experts will result in a sufficiently reliable score with acceptable resource efficiency. Incognito SP-based examinations are useful for measuring the “does” layer, in which incognito SPs are healthy persons who are trained to portray patients. They visit a doctor who does not know which patient is simulated. After the encounter, the SP completes a predefined checklist to score the doctor’s performance. Although the method has a high reliability and evaluates the top layer of Miller’s pyramid, some concerns exist with respect to the subjective opinions of the standardized patients. For instance, it is very expensive and time consuming; the SPs must be very well trained to portray a disease so that the doctor cannot recognize them as a “fake patient.” Furthermore, all simulated data and laboratory results must be produced to enhance the reality of doctor–patient encounters. Developing checklists or a rating is difficult because doctors with different clinical decision-making processes arrive at similar conclusions.

Based on these findings, in general one could argue that covering all layers of Miller’s pyramid requires a hybrid approach. As such, a need exists for more studies that uncover new methods for assessing the performance of practicing residents. We must also ensure the reliability, validity, feasibility and acceptability of new methods before using them in the real world. It is essential that
medical educators using simulation-based medical education gain basic declarative knowledge and understanding in assessment principles and methods, both formative and summative.

THE NEW APPROACH TO ASSESSING TECHNICAL SKILLS

The assessment of residents’ technical skills is an important part of ensuring that we have trained competent clinicians. It has been argued that most programs are based on subjective faculty assessment of residents’ technical skills. Very few programs use assessments that take place while the resident is performing a procedure, particularly a surgical procedure. Furthermore, studies have shown the surgical assessments that have happened in the past had poor reliability and validity. Basically, surgeons learn a procedure that is based on the model of apprenticeship. Although objective assessment is necessary because deficiencies in performance are difficult to improve without objective feedback, residents’ performance assessment is mainly subjective. High stakes examinations such as the fellowship of the Royal College of Surgeons, which is a professional qualification for practicing as a surgeon in the British Isles, focus mainly on the cognitive proficiency and clinical abilities of the residents and do not assess a resident’s technical skill or (operative performance) and judgment. Specialist registrars (residents) have log books in the United Kingdom, which lists the medical and surgical procedures performed by them. The log book is submitted at the time of examination, at the job interview, and during an exit clinical examination. It has been observed, however, that log books show merely procedural performance rather than a reflection of operative performance that results in low content validity.

Furthermore, it has been argued that the observational assessment of technical skills in the operating room is subjective and would have poor interobserver reliability. These issues, along with increasing public concern about medical competence, have led to the development of several different assessment tools in the hopes of providing more reliable and valid measurements of medical performance using low-tech and high-tech simulation modalities. They include the OSATS, HMA, virtual reality and haptic systems, and final product analysis.

OSATS

Clinical skills, especially surgical skills, are one of the most demanding of all performance tasks because life-threatening circumstances can be encountered. Objective performance assessment in the “real world” is problematic for obvious reasons. In response to this dilemma, ample literature now exists to recommend that the OSATS can be used to assess a resident’s surgical skills to provide an objective assessment that has both good validity and reliability. Pioneers in the 1990s such as Martin et al introduced the OSAT that measures performance precisely. The OSATS is an OSCE-like assessment in which residents perform standardized surgical procedures while being evaluated by an attending surgeon. The OSATS consists of 6 stations in which residents perform procedures on live animals or bench models. The assessment is made using a task-specific checklist in which every important step of a procedure is identified and a mark is assigned for every step the resident completes correctly (0 or 1). Furthermore, included is a global assessment score, which is a 7-item assessment of overall performance, such as respect for tissues and flow of procedure.

The OSATS assessment has the drawback of the resources and time involved in obtaining several attending surgeons to observe the performance of residents despite that it has high validity and the interstation reliability. Furthermore, animal rights movements have protested that animals should no longer be regarded as property or be treated as resources for human purposes, but they should instead be regarded as legal persons and members of the moral community. In addition to these issues, it has also been found that the OSATS have some challenges that require addressing before they becomes feasible for senior surgical resident, particularly exploring a link between technical skills and outcomes for patients and uncovering correlation between performance on simulations and real procedures. A danger exists that residents may become experts in using simulations rather than in the clinical context. More research studies are required to explore the assessment of surgical competence in realistic environments, such as a simulated operating theater for the assessment of communication, decision-making leadership, and crisis management.

HMA

HMA has been borrowed from the field of kinesiology, which is the study of human movement. HMA uses electromagnetic markers to assess and to follow clinical parts that relate to joint movement, such as purposeful hand direction, depth perception, finger-pressure coordination, finger visual tracking of moving objects, spatial visualization, finger strength, speed of movements, fine-control precision, and finger dexterity. Researchers at the Imperial College of Science, Technology and Medicine developed the use of manual dexterity examinations as a promising method of assessing technical skills. HMA is recorded using the Imperial College surgeon device (ICSAD), which is a commercial electromagnetic tracking system (Isotrak II; Polhemus Inc., Vermont). The device consists of an electromagnetic field generator and 2 sensors that are placed on the dorsal surfaces of the surgeons’ hands. Bespoke software (Bespoke Software, Inc., Clifton Park, New York) is used for converting the positional data generated by the sensors to dexterity measurers such as the number and speed of hand movements, the distance traveled by the hands, and the time taken for the task.

Another software program, the Blue DRAGON, was developed to visualize certain key steps of a surgical procedure and to measure hand movements of surgeons. It is a passive robotic arm that can measure precise hand motions in all directions and can record the amount of force generated. These measurements are analyzed to determine accurately the surgeon’s perfor-
mance. Such technology not only discriminates between residents and expert surgeons, but also provides a visual schedule as to how to improve the performance of residents.47-50 Another tool was also designed in 1997 in the Ninewells Hospital at the University of Dundee entitled “The Advanced Dundee Endoscopic Psychomotor Tester” (ADEP) to select residents for endoscopic surgery. The tool measures the ability of residents’ psychomotor skills in minimal access surgery.51 Despite the validity and reliability of the ADEP for measuring operative skills, some larger studies are needed to confirm that these devices can be used to assess an individual’s ability to learn or perform certain skills in surgery. Although several studies suggest aptitude tests might be used to predict an individual’s proficiency, much of the existing literature in this area has had conflicting results.52 The conflicting results could be because these tools ignore some characteristics of individuals, such as personality and how surgeons deal with stress and decision making, which are imperative within the profession.

VIRTUAL REALITY AND HAPTIC SYSTEMS

Virtual reality is defined as “a collection of technologies that allow people to interact efficiently with 3D computerized databases in real time using their natural senses and skills.” (p. 912).53 The ultimate goal of virtual reality is to visualize virtual objects or environments to all human senses in a way that is identical to their natural counterpart. The success of virtual reality is caused by improvements in computing technology and development of methods for collecting data such as medical imaging.6 The MIST-VR is a new generation of state-of-the-art computer-driven systems that was the first virtual reality laparoscopic simulator.53 The system is extremely effective to assess basic laparoscopic skills, especially psychomotor skills or procedures that must be learned.8 In the future, it might be argued that virtual reality systems will change the current status of surgery skills dramatically as these systems are low-fidelity models.

Haptic (force and touch) is a recent addition to virtual reality systems. In experimental psychology and physiology, the word “haptic” referred to the ability to experience the environment through active exploration, typically with our hands, as when palpating an organ to gauge its shape and size.54 Haptic technology has been applied recently to medical education. For instance, the Virtual Haptic Back project developed a realistic graphic model of the human back that has been used for teaching medical palpatory diagnosis.55 Although the importance of haptic technology in virtual and real environments has been focused on medical training, some studies have used it as an assessment device.56-58

FINAL PRODUCT ANALYSIS

As it is very difficult to estimate the complications after surgery, particularly in the long-term, some surgeons used bench models for measuring operative outcomes. For example, in 1 study surgeons assessed the quality of the final product after performance of 6 bench model representations of general surgery such as choledochojejunostomy, rectal anastomosis, femoral artery anastomosis, ileostomy, J-tube insertion, and inguinal hernia repair. They found that the method had good construct validity. They also showed that a poor correlation exists between product analysis scores and OSAT scores.59 This finding may imply additional research studies on both bench models and live models for assessing residents.

SUMMARY

Much is still to be learned about the assessment of simulation-based surgical skills training. However, assessing surgery skills through simulation is a new horizon in medical education. Providing a safe environment for surgical residents to assess their performance rigorously without placing patients in jeopardy is valuable. Using simulators (both warm and cold) as a means to assess trainees has been established. However, also problems concerning the validity and reliability of such simulation-based assessment tools exist, particularly in surgery, that may need to be investigated even more to decide whether to use them as a tool for assessing the performance of surgical residents.

REFERENCES


