The Use of Standardized Patient Assessments for Certification and Licensure Decisions

John R. Boulet, PhD;
Sydney M. Smee, PhD;
Gerard F. Dillon, PhD;
John R. Gimpel, DO

Although standardized patients have been employed for formative assessment for over 40 years, their use in high-stakes medical licensure examinations has been a relatively recent phenomenon. As part of the medical licensure process in the United States and Canada, the clinical skills of medical students, medical school graduates, and residents are evaluated in a simulated clinical environment. All of the evaluations attempt to provide the public with some assurance that the person who achieves a passing score has the knowledge and/or requisite skills to provide safe and effective medical services. Although the various standardized patient-based licensure examinations differ somewhat in terms of purpose, content, and scope, they share many commonalities. More important, given the extensive research that was conducted to support these testing initiatives, combined with their success in promoting educational activities and in identifying individuals with clinical skills deficiencies, they provide a framework for validating new simulation modalities and extending simulation-based assessment into other areas.
(*Sim Healthcare* 4:35–42, 2009)

Key Words: Licensure, Certification, Simulation, Standardized patient, Simulated patient, OSCE

There are many types of simulations that are currently being used to assess healthcare professionals.[1–4] In both Canada and the United States (US), many of these simulation modalities, including multiple choice questions, part-task trainers, and computer-based case simulations, have been used as part of the examination process used to certify and license physicians.[1,5,6] These simulation-based examinations, which can vary somewhat in terms of purpose and focus, all attempt to provide the public with some assurance that the person who achieves a passing score has the knowledge and/or requisite skills to provide safe and effective medical services, either independently or under supervision. Here, as with any simulation-based assessment, the structure, content, fidelity, and difficulty of the modeled exercises, combined with the scores, will determine what inferences one can make about the individual test taker.

From a simulation perspective, the use of standardized patients (SPs) for certification and licensure decisions has been a relatively recent phenomenon.[7] Historically, SP-based assessments were implemented as part of formative evaluation activities.[8–10] Individuals were trained to portray specific patient conditions, allowing medical students to practice their clinical skills and receive immediate feedback concerning strengths and weaknesses. In the 1980s, with an increased emphasis on evaluating what medical trainees could do, as opposed to what they knew, various organizations started research programs aimed at determining how assessments employing SPs could be structured to make valid skills-based proficiency decisions. Over the next two decades, the end result of these research activities was the implementation of a number of high-stakes assessments all aimed at measuring abilities in key clinical skills domains. Although these research efforts required extensive resources, they were successful in identifying the specific conditions and structures that are needed to produce defensible scores and decisions for multistation, performance-based, simulation activities.[11–17]

The introduction of SP-based certification and licensure examinations in medicine was a monumental achievement. Although other high-stakes simulation-based assessments have been developed and used in other professions, the logistical, economical, and psychometric challenges associated with national multistation clinical skills assessments were staggering.[18,19] Organizations that built these assessments all had to address concerns regarding test content (eg, types of scenarios to model), test administration models (eg, fixed versus temporary sites; number, timing and sequencing stations), measurement rubrics (eg, holistic or analytic), eligibility requirements, scoring models (eg, compensatory or conjunctive), and the establishment of defensible standards, just to name a few. Nevertheless, even with these hurdles, and despite numerous objections concerning the need to measure

From the Foundation for Advancement of International Medical Education and Research (J.R.B.), Philadelphia, PA; Medical Council of Canada (S.M.S.), Ottawa, ON, Canada; National Board of Medical Examiners (G.F.D.), Philadelphia, PA; and National Board of Osteopathic Medical Examiners (J.R.G.), Conshohocken, PA.

Reprints: John R. Boulet, PhD, Foundation for Advancement of International Medical Education and Research, 3624 Market Street, Philadelphia, PA 19104 (e-mail: jboulet@faimer.org).

The authors have indicated that they have no conflict of interest to disclose.

DOI: 10.1097/SIH.0b013e318182fc6c

clinical skills as part of certification/licensure process,[20] each of these organizations was able to produce a high-quality simulation-based assessment that was appropriate for their particular needs. In doing so, many lessons were learned, the most important being that simulation-based summative assessment of clinical skills was viable, even with large examinee populations, differing testing purposes, and varying examination administration protocols.

PURPOSE

The purpose of this article was to describe and contrast the Clinical Skills Assessment (CSA) programs that are employed in Canada and the US as part of the certification and licensure process for physicians. These assessments include the Medical Council of Canada (MCC) Qualifying Examination Part II (MCCQE Part II),[21] the United States Medical Licensing Examination (USMLE) Step 2 Clinical Skills (USMLE Step 2 CS),[22] and the National Board of Osteopathic Medical Examiners (NBOME) Comprehensive Osteopathic Medical Licensing Examination Level 2-Performance Evaluation (COMLEX-USA Level 2-PE).[23] To better understand the USMLE Step 2 CS, a brief overview of the Educational Commission for Foreign Medical Graduates (ECFMG) CSA is also provided.[24] The CSA was used to assess the clinical skills of international medical graduates (IMGs) before the introduction of USMLE Step 2 CS. Following this overview, a brief synthesis of the similarities and differences in the assessments and assessment programs is provided. With these distinctions in mind, and knowing the success and scope of the individual testing programs, it is possible to envision where summative simulation-based assessment activities could be enhanced, applied in other areas, and used for the evaluation of nonphysician healthcare professionals.

ASSESSMENT OF CLINICAL SKILLS

In general terms, clinical skills refer to information gathering and communication skills, applied during the patient encounter, that help to establish an accurate diagnosis and support high-quality treatment. Within the medical education and practice community, these skills have long been recognized as essential to patient care. Several organizations, including those responsible for the accreditation of undergraduate and graduate medical education (GME) programs, have included clinical skills among the competencies deemed important to the education and assessment of practicing physicians.[25–27] As a result, it is not surprising that considerable efforts have been made to develop, and subsequently defend, testing methods than can be used to reliably and validly measure these skills.

STANDARDIZED PATIENTS

SPs, often referred to as simulated patients or programmed patients, are people who have been trained to accurately portray the role of a patient with a specific medical condition or conditions. The term "standardized" refers to the fact that the person is specifically trained to model the "real" patient's condition, including symptoms and emotional states, and to do so consistently over time. Examinees who interview the same SP with the same presenting complaint will receive, on questioning, the same patient history. The physical findings relevant to the case, either real or simulated, need to be stable and, for a given modeled scenario, they must not vary from one SP to another.

LARGE-SCALE SP EXAMINATIONS

Medical Council of Canada Qualifying Examination Part II

Since 1912, the MCC has been setting an examination that is a prerequisite for medical licensure in Canada; the Licentiate of the MCC is granted to those who successfully complete it. In 1992, the MCC added the Qualifying Examination Part II (MCCQE Part II) to the assessment sequence. Initially the MCCQE Part II was a 20-station Objective Structured Clinical Examination (OSCE).[7,28] Although the use of OSCEs is now commonplace throughout the world, implementing a national summative, performance-based, assessment based on a series of SP encounters was, at the time, unprecedented. The impetus for implementing the MCCQE Part II came largely from the licensing authorities. In the late 1980s, because of the number and nature of related complaints that they received each year, members of these authorities began calling for an assessment of clinical and communication skills. The existing paper-and-pencil test of medical knowledge and problem solving (MCC Qualifying Examination Part I—MCCQE Part I) was not sufficient to address the emergent belief that candidates for medical licensure should be assessed more broadly.

To qualify for the MCCQE Part II, candidates must have completed successfully 12 months of postgraduate clinical training and passed the MCCQE Part I, currently a computer-adaptive test of knowledge and clinical decision-making. The number of candidates who qualify for the MCCQE Part II continues to grow. In 1992, 401 candidates took the examination. In 2007, 3481 candidates completed this assessment, a more than eightfold increase.

As the measurement qualities of the MCCQE Part II became better understood, the number of stations was reduced from 20 to 14, and is now set at 12. This reduction in station length could be attributed to evolving test development processes, allowing for a more efficient and appropriate targeting of test content to examinee ability. Each station is based on a clinical problem presented by a SP; scoring is completed by physicians who observe from within the room. Checklists and rating scales are used to generate the station scores. At this time, the MCCQE Part II is comprised of eight 10-minute encounters with a SP and six couplet stations that include a 5-minute encounter with a SP followed by a 5-minute written component (Two of the stations in the assessment, including one of the couplets, are used for pilot testing purposes). Four domains are assessed based on common presenting problems: history-taking skills, physical examination skills, patient management, and doctor-patient interactions. Patient safety issues and professionalism are also evaluated.

Each scored station, while potentially measuring slightly different skill sets, counts equally in terms of generating a total score. Although station scores are compensatory, mean-

ing poor performance in one station may be compensated by superior performance in another, the overall pass/fail decision is based on a conjunctive standard; candidates must pass both by total score (the sum of their station scores) and by the number of stations passed.

Results from the MCCQE Part II are reported as a standard score (mean = 500, standard deviation = 100). The examination is criterion-referenced, with the individual station pass marks set using the borderline group method.[29] Candidates receive a bar graph indicating their performance in each of four domains relative to the mean score for their testing cohort. The four domains are data gathering (from history taking and physical examination tasks), patient interaction (from rating scale items across stations), problem-solving and decision-making (based on certain stations; eg, acute care of trauma and the written work from the couplet stations), and legal, ethical, and organizational issues (which comprises a minimum of 10% of the total score). More extensive feedback is provided to those candidates who are unsuccessful; specifically, they are told which stations they failed and are provided with a more extensive description of the four domains.

To balance accessibility and costs, a multisite, fixed test form model with two administrations per year is employed. In the spring, one test form is administered twice over 1 day at 10 university sites across Canada. At most sites, the examination runs in two or more parallel tracks. In the fall, there are two test forms, one for each of 2 days of testing, and the examination runs at 16 sites. In spring, over 500 SPs are trained to simulate the patient problems. Twice that many are recruited for the fall. Ensuring that the SPs present their problems consistently and with sufficient fidelity for valid testing is critical. Each site has its own trainers who recruit and prepare the SPs according to the protocols developed centrally. Training videos, meetings with MCC staff, consultation with supervising physicians, along with telephone support are all part of a process aimed at ensuring the SPs are ready for the examination.

Like all large-scale testing programs, there have been some administrative challenges. Developing feasible, psychometrically sound cases (simulated scenarios) is an ongoing task and takes considerable time and effort. Because the MCCQE Part II is a national examination, the scoring instruments and the supporting materials for SP training are developed centrally by a multidisciplinary test committee. Cases range from those requiring relatively little simulation (eg, history of diarrhea) to those where the SP must accurately simulate specific patient presentations (eg, shortness of breath, decreased consciousness, pain, anxiety).

The MCC is continuously assessing different aspects of the MCCQE Part II. Numerous research studies suggest that both valid and reliable competency decisions are being made.[30–32] Most recently, the predictive validity of the MCCQE Part II was investigated by looking at the relationship between MCCQE Part I and Part II scores and complaint records from two licensing jurisdictions.[33] The authors concluded that poor performance on the MCCQE Part II patient-physician communication component and the clinical decision-making component from the MCCQE Part I were predictors for complaints.

Educational Commission for Foreign Medical Graduates Clinical Skills Assessment

Based on several years of extensive research and consultation with the MCC, the ECFMG CSA was instituted in July 1998.[34,35] This 11 station clinical skills examination was developed to evaluate whether graduates of international medical schools (IMGs) possessed the skills necessary to enter supervised GME programs in the US. Successful completion of this examination became one of the required elements for ECFMG certification. Initially, the assessment was offered at one fixed site in Philadelphia, Pennsylvania. In 2002, in collaboration with the National Board of Medical Examiners, a second testing site was constructed in Atlanta, Georgia. Between 1998 and 2004, 43,624 IMGs were tested (37,930 first-time takers) in a total of 372,674 simulated clinical encounters. During this time, numerous studies were published, several providing evidence to support the validity of the assessment scores.[36–38] Of particular note, research was conducted to show that SP and physician evaluations of clinical skills were comparable.[39] In 2004, administration of the ECFMG CSA ceased. Instead, IMGs were required to take and pass USMLE Step 2 CS (described below), a similar simulation-based assessment that was developed to measure the clinical skills of American allopathic medical students and graduates. The USMLE Step 2 CS examination is part of the USMLE sequence (There are three "Steps" to the USMLE. Step 1 is intended to assess whether the examinee understands and can apply important concepts of the sciences basic to the practice of medicine. Step 2 focuses on the examinee's knowledge, skills, and understanding of clinical science essential for provision of patient care "under supervision"—typically the point that medical school graduates begin their postgraduate education and experience. Step 3 is intended to assess whether the examinee can apply medical knowledge and understanding of biomedical and clinical science essential for the unsupervised, independent practice of medicine.) To qualify for a medical license to practice in the US, graduates of MD-granting schools in the US and graduates of medical schools located outside the US must take and pass all components of USMLE.

United States Medical Licensing Examination Step 2 Clinical Skills

From the time that introduction of the USMLE program was first proposed in the late 1980s, it was the intent of the National Board of Medical Examiners and the Federation of State Medical Boards (the organizations that sponsor USMLE) to include clinical skills among the areas assessed as part of the examination program supporting the US medical licensing system. After many years of development, this goal became a reality in June 2004 when USMLE Step 2 CS was administered for the first time.[40] At this point, the previously existing Step 2 examination, a 1-day, computer-based multiple choice questionnaire test, was renamed the Step 2 Clinical Knowledge examination. The introduction of Step 2 CS in the USMLE sequence was informed by the research of many organizations interested in the assessment of these important

skills and by the operational experiences of organizations that brought this type of format to the arena of large-scale, high-stakes assessment, in particular, the MCC and the ECFMG.[41–43]

The USMLE Step 2 CS examination, which is delivered at each of five regional testing centers (Atlanta, Chicago, Houston, Los Angeles, and Philadelphia), requires test takers to move through a series of 12 simulated encounters (stations), interacting with SPs, individuals who are trained to portray real patients. Examinees are given up to 15 minutes to interact with each SP. During that time they are expected to take a history and to perform a physical examination that is focused on the chief complaint of the patient and on the information that is revealed during the encounter. After the simulated encounter, examinees are given 10 minutes to write a patient note that summarizes and synthesizes their findings, including possible diagnoses. The mix of cases seen by any one examinee is guided by a group of content experts who are charged with overall design and development of Step 2. Based on a test blueprint established by this committee, each test form contains a blend of patient presentations that would not be uncommon for clinical practice in the US. This same committee is involved in the process used to establish passing standards.[44,45] Because the Step 2 CS examination is offered daily across five sites, a variable test form administration model is used. The test form (mix of clinical presentations and SP characteristics) for any given administration, at any site, is individually constructed to meet blueprint specifications. Efforts are made to minimize case and SP exposure for previously failing examinees who are repeating the assessment.

USMLE Step 2 CS examinees are required to pass three subcomponents: the integrated clinical encounter, which includes demonstration of skills in history taking, physical examination, and documentation; communication and interpersonal skills, which includes skills in information gathering/sharing and establishing rapport; and spoken English proficiency, which requires clear communication with the patient. With the exception of the postencounter notes, which are scored by a group of physicians who are specially trained to the specifics of the case, all scoring is done by the SPs who are extensively trained and monitored in their use of a series of checklists and rating scales that were specifically designed for gathering reliable and valid measures of these components. To pass the USMLE Step 2 CS, an examinee must pass all of the three subcomponents (integrated clinical encounter, communication and interpersonal skills, and spoken English proficiency) in a single administration. Failing examinees are provided with feedback outlining relative strengths and weaknesses in the various clinical skills components that are measured.

The USMLE Step 2 CS program has been fully operational for almost 4 years, delivering, scoring, and reporting results year round. More than 120,000 examinations have been administered, representing more than 1.4 million examinee-SP encounters. Because of the complexities of an overall system that handles, at any one time, thousands of examinees, hundreds of SPs, and multiple testing centers, there are substantial quality assurance measures in place[46] and, as a result, for the most part, the examination process has been completed with relatively few problems. Similar to the other USMLE examinations, significant efforts are dedicated to all phases of testing, including content development and validation, examinee scheduling, administration, scoring, equating, standard setting, and score reporting.

Despite the technical and administrative challenges, the implementation of the USMLE Step 2 CS program has been successful. USMLE Step 2 CS identifies examinees with deficiencies in important practice skills who might not otherwise have been identified based on the other examinations in the USMLE sequence.[47] In this way, the examination has made a significant contribution to the medical licensing process in the US and, at the same time, has called special attention, within the education and practice community, to the role of clinical skills in patient care activities. In a recent study that was based on interviews of 25 leaders of medical school CSA programs, respondents noted that the new national examination validated the importance of clinical skills for medical students.[48] Also, of particular note, numerous schools have changed the objectives, content, and emphasis of their preclinical curriculum in response to the implementation of the Step 2 CS.[49]

Comprehensive Osteopathic Medical Licensing Examination Level 2-Performance Evaluation

In 1994, the NBOME started the process of developing a SP-based clinical skills examination for osteopathic physician licensure. After considerable research and several feasibility and pilot studies, the COMLEX-USA Level 2-PE was launched in 2004.[50] Similar to both the MCC and the USMLE, this new assessment complemented the other examinations that are part of the licensure process for osteopathic physicians (COMLEX-USA or Comprehensive Osteopathic Medical Licensing Examination is a series of three osteopathic medical licensing examinations administered by the NBOME. The examinations include Level 1, Level 2-CE, Level 2-PE, and Level 3. COMLEX-USA is the most common pathway by which osteopathic physicians (DOs) apply for licensure, and is accepted in all 50 states and numerous international jurisdictions.) The COMLEX-USA Level 2-PE, which is usually taken in the 4th year of osteopathic medical school, tests the clinical skills of graduating students of osteopathic medical schools in the US. As of 2008, the accreditation body for osteopathic medical schools in the US (Commission on Osteopathic College Accreditation of the American Osteopathic Association) requires that all students pass COMLEX-USA Level 2-PE before graduation, and examinees are not eligible to take the COMLEX-USA Level 3 examination, the final examination in the COMLEX-USA series, unless they have passed COMLEX-USA Level 2-PE. Through the end of the 2007 calendar year, there have been a total of 992 COMLEX-USA Level 2-PE test administrations, involving more than 11,800 examinees.

Based on the COMLEX-USA Level 2-PE assessment design, examinees encounter 12 SPs in a simulated ambulatory clinical medical environment. The assessment takes 7 hours and is administered at a single fixed site (NBOME National Center for Clinical Skills Testing) located in the Philadelphia,

Pennsylvania area. For each of the 12 simulated encounters, examinees have 14 minutes to evaluate and treat the SP based on the clinical presentation. Following the 14-minute encounter, the examinee has an additional 9 minutes to complete a written patient note. Content design for the examination, including test form specifications, was informed by analysis of national practitioner databanks and expert consensus.[51] The mix of cases for a given test form is balanced with respect to acute, chronic, and health promotion/disease prevention presentations. To enhance content validity, the mix of SPs is governed by specifications related to patient characteristics, including gender and age. The COMLEX-USA Level 2 PE is administered almost every day, and sometimes both in the morning and in the evening. Consequently, a variable test form administration model is employed.

The COMLEX-USA Level 2-PE assesses skills in four clinical skill areas: doctor-patient communication, interpersonal skills, and professionalism; data gathering, which includes medical history-taking and physical examination; documentation and synthesis of clinical findings (including treatment); and osteopathic principles and osteopathic manipulative treatment (OMT). Doctor-patient communication, interpersonal skills, and professionalism are evaluated by the SPs using behaviorally anchored holistic scales. Data gathering proficiency is derived from case-specific checklist items, documented by the SPs following the clinical encounter. Written notes are evaluated by physician examiners located throughout the US using a holistic rubric. Unique to COMLEX-USA Level 2-PE, osteopathic principles and OMT are evaluated by physician examiners via a distributed video review system. Here, the physician examiners, also located across the US, access assigned clinical encounters through a secure web link and then provide structured performance ratings.

The four skill area scores, summarized over the encounters, are combined into two domains. The Humanistic domain summary score is based solely on the SP ratings of doctor-patient communication, interpersonal skills and professionalism. The Biomedical/Biomechanical domain summary score is a weighted composite of an examinee's data gathering, written patient notes, and OMT scores. For both domains, the generation of a summary score, over encounters, is compensatory, meaning that an examinee can compensate for poor performance in one station with excellent performance in another. However, across the two domains, COMLEX-USA Level 2-PE uses a conjunctive scoring model; examinees must achieve passing scores in both domains to receive a passing score for the examination. Examinations standards were initially set in 2004–2005 and, based on widely accepted testing protocols, updated in 2007. Only candidates who fail the examination are given specific feedback on their skills performance in the two domains and four skills areas.

To ensure that decisions based on the COMLEX-USA Level 2-PE examination scores are fair, an extensive quality assurance program has been implemented. In addition to pilot testing cases prior live usage, double scoring a large percentage of the encounters, investigating the relationships among scores, and regularly checking physician and SP rater stringencies, the performances of failing candidates are systematically reviewed to ensure that the decisions are accurate and can be defended.

The introduction of COMLEX-USA Level 2-PE, although logistically challenging, helps to fulfill the public and licensing authority mandate for enhanced patient safety through the documentation of the clinical skills proficiency of graduates from osteopathic medical schools. As a consequence, it has effectively highlighted the importance of clinical skills training as part of the osteopathic medical school curriculum.[52–54] Moreover, there has been an associated increase in the use of simulation throughout the medical school curriculum. Based on a survey of the deans of the 23 fully accredited Colleges of Osteopathic Medicine and branch campuses, Gimpel et al.[55] concluded that the use of SPs and mechanical simulators at colleges of osteopathic medicine increased substantially from 2001 to 2005.

DISCUSSION

The clinical skills examinations described above (MCCQE Part II, USMLE Step 2 CS, NBOME COMLEX-USA Level 2-PE) share many commonalities. They all use a multistation format where candidates rotate through series of clinical encounters, alternating between patient interviews and some form of postencounter exercise. Here, the development and choice of clinical encounters (stations, cases) is governed by detailed test specifications. Multiple stations are used in an effort to broadly sample the practice domain and to ensure that the scores, and associated pass/fail decisions, are reliable. All of the examinations model typical patient settings and doctor-patient interactions. This high-fidelity simulated environment provides the means to measure fundamental clinical skills, including history taking, physical examination, doctor-patient communication, and interpretation of clinical data. In measuring these skills, some combination of rating scales and checklists is used to produce examinee scores. Given the high-stakes nature of these examinations (access to the medical profession), significant resources are allotted to development and validation of the simulated clinical scenarios. For all three examinations, unscored pilot stations are incorporated into live examinations before their active use in making decisions about clinical skills proficiencies. In this way, data can be gathered to establish the fidelity of the simulation, the appropriateness of the clinical content, and the ability of the resultant scores, both ratings and checklists, to discriminate between those who possess the skills and those who do not. Finally, and likely most important, they all employ highly structured training and quality assurance protocols, both for the SPs and physician evaluators. This helps to ensure that valid inferences (ie, pass/fail decisions) can be made from the available scores and ratings.

Although the assessments share a common structure, there are some important differences that, taken collectively, serve to broaden the potential assessment domain and provide potential test administration frameworks that could be useful to other health professions that wish to evaluate clinical skills. First, the USMLE Step 2 CS and NBOME COMLEX-USA Level 2-PE run at fixed sites, whereas the MCCQE Part II operates periodically on weekends at actual clinics

across Canada. Although choice of variable or fixed sites is dependent on candidate volume, political considerations, and economics, quality exams can be offered under either administrative model as long as steps are taken to ensure proper standardization and security. Second, because of the almost daily administration of the COMLEX-USA Level 2-PE and USMLE Step 2 CS exams, test forms are continuously changed and are rarely repeated. For the MCCQE Part II administrations, which take place at the same time across different sites, a fixed form model is appropriate (The actual examination does not take place at exactly the same time across Canadian sites. Examinees at sites in later time zones are sequestered so that examination information cannot be shared.) Third, unlike the MCCQE Part II, which is usually taken in the second year of residency, the US-based examinations (COMLEX-USA Level 2-PE, USMLE Step 2 CS, former ECFMG CSA) are targeted at individuals who are just entering GME programs. As a result, the content of the MCCQE Part II is somewhat more challenging, requiring more advanced management and clinical decision making abilities. Fourth, because of differences in the practice characteristics of allopathic and osteopathic medicine, the clinical content modeled in the various assessments is not identical. For example, on the COMLEX-USA Level 2-PE there are proportionally more encounters involving SPs with musculoskeletal complaints. Moreover, unlike any of the other assessments, the evaluation of osteopathic principles and OMT is a fundamental part of this examination.[56] Given the differing purposes of these assessments, it is not surprising that they diverge somewhat in terms of focus. Modeling clinical encounters that are important to the profession, combined with tailoring the examinations to the expected performance level of examinee, provides a basis for establishing the content and construct validity of the assessments. A similar strategy could easily be used for non SP-based simulation activities, including those employing mannequins or part-task trainers.

Although the skills that are measured in these performance-based assessments are similar, the measurement protocols vary. For both the USMLE Step 2 CS and COMLEX-USA Level 2-PE, a score equating strategy is employed.[42] Because the examination content, and associated SPs, can vary considerably from day to day, it is important to account for potential differences in the difficulty of the test forms administered. Unlike the other assessments, the MCCQE Part II employs physician examiners who sit in the room while the clinical interview takes place. These physicians are trained to score the encounters and also to make summary, holistic, judgments of the adequacy of the performance. These summary measures are then used, in combination with assessment scores, to derive performance standards.[29] In contrast, for both the COMLEX-USA Level 2-PE and USMLE Step 2 CS, where SPs complete history taking and physical examination checklists, separate standard setting exercises are conducted periodically. Interestingly, while all three assessments employ some form of assessment of doctor-patient communication skills, there are no common rubrics or training protocols. For both the COMLEX-USA Level 2-PE and USMLE Step 2 CS, the SPs provide ratings of interpersonal and communication skills; for the MCCQE Part II, the physician in the room evaluates these traits. Finally, although employed somewhat differently, all of the examinations have both compensatory and conjunctive scoring elements. Test-level scores are generated by averaging performance in specific domains over the series of modeled encounters. For the COMLEX-USA Level 2-PE and USMLE Step 2 CS, a candidate's pass fail status is determined by summary performance in multiple areas. For the MCC Part II examination, candidates must also demonstrate an acceptable level of performance across a minimum number of stations.

Overall, based on a fairly limited usage of mock-up settings and simulation modalities, the three SP-based examinations are successful in fulfilling their assessment goals. For the most part, the restricted use of simulation modalities can be attributed to the fundamental purposes of the assessments, the logistics and economics of large scale assessment, technological limitations, and psychometric issues pertaining to scoring. Nevertheless, going forward, one can envision the adoption of other simulation strategies to broaden the assessment domain. For example, if logistical and psychometric issues could be effectively addressed, incorporating paired SP-Part task trainer stations could be an effective way to measure procedural skills and clinical decision making.[57–59] Likewise, although stations involving one SP and one examinee are efficient, at least from a testing perspective, the measurement of communication skills in this context is restricted to the doctor (examinee) and the patient. To evaluate teamwork, and certain facets of professionalism and ethical behavior, it would be appropriate to include other simulated healthcare workers and even standardized family members.[60–63] The MCC has already integrated some stations of this nature into their clinical skills examination; for example, working with a nurse to care for a trauma patient in an acute care setting and advising another healthcare professional over the telephone. Finally, even though some physical findings can be simulated by SPs quite well, many cannot (eg, trauma, breathing difficulties). As a result, for an OSCE that only includes SP-based encounters, it can be difficult to fully evaluate physical examination skills. Here, provided financial and logistical concerns can be addressed, electromechanical mannequins could be employed in some stations.[64]

Although the incorporation SP-based performance assessments as part of licensure and certification has spurred substantial research, there remain several important areas where further investigations are warranted. With respect to scoring, the available checklist and rating scales used for SP-based assessments, although appropriate for measuring basic clinical skills including history taking and physical examination, may not yield valid and reliable measures when employed for acute care situations, especially those modeled with electromechanical mannequins or even part-task trainers. Here, other constructs (eg, timing, sequencing, accuracy) will need to be incorporated within the measurement framework. In terms of content sampling, additional research focusing on the choice and structure of the various forms of simulation exercises is needed. Knowing which types of simulated scenarios provide for the best assessment conditions, and most valid and reliable scores, is essential if one seeks

meaningful and generalizable measures of ability. Likewise, if new ability measures are constructed, additional psychometric work will be needed to delimit the score, or scores, that separate those who are proficient from those who are not. Finally, and arguably most important, there is still relatively little published research that shows that performance in the simulated environment translates to real-world patient care. Designing and completing outcome studies that provide support for the validity of the performance measures derived from simulation-based assessments is paramount.

Conducting large scale, high-stakes performance assessments for medical licensure has been extremely successful. Although the MCC, USMLE and NBOME clinical skills exams have somewhat different purposes, administration models, and scoring protocols, they are all effective in providing a fair and equitable assessment of the clinical skills of their test populations. All three assessments are supported by a substantial number of research studies aimed at establishing the validity and generalizability of the test scores. As medical simulation further expands into other areas (eg, specialty board certification, selection of residents, continuing medical education, maintenance of certification), the processes used to develop and administer these examinations, with some modification, can be used as a model for assessment design and delivery. Should simulation-based assessment be adopted more broadly, especially for high-stakes competency decisions, one ought to expect a fairly large consequential educational impact, including an enhanced curricular emphasis on any particular skills that are evaluated as part of new assessment strategies. As other health provider groups seek to evaluate their trainees and make defensible competency decisions, the lessons learned in developing high-stakes, SP-based assessments in medicine will certainly prove to be quite valuable.

REFERENCES

1. Melnick DE, Dillon GF, Swanson DB. Medical licensing examinations in the United States. *J Dent Educ* 2002;66:595–599.
2. Cosby JC Jr. The American Board of Dental Examiners clinical dental licensure examination: a strategy for evidence-based testing. *J Evid Based Dent Pract* 2006;6:130–137.
3. Austin Z, Gregory P, Tabak D. Simulated patients vs. standardized patients in objective structured clinical examinations. *Am J Pharm Educ* 2006;70:119.
4. Vu N, Baroffio A, Huber P, Layat C, Gerbase M, Nendaz M. Assessing clinical competence: a pilot project to evaluate the feasibility of a standardized patient—based practical examination as a component of the Swiss certification process. *Swiss Med Wkly* 2006;136:392–399.
5. Dillon GF, Boulet JR, Hawkins RE, Swanson DB. Simulations in the United States Medical Licensing Examination (USMLE). *Qual Saf Health Care* 2004;13(suppl 1):i41–i45.
6. Hallock JA, Melnick DE, Thompson JN. The step 2 clinical skills examination. *JAMA* 2006;295:1123–1124.
7. Reznick RK, Blackmore D, Dauphinee WD, Rothman AI, Smee S. Large-scale high-stakes testing with an OSCE: report from the Medical Council of Canada. *Acad Med* 1996;71:S19–S21.
8. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. *Acad Med* 1993;68:443–451.
9. Barrows HS, Abrahamson S. The programmed patient: a technique for appraising student performance in clinical neurology. *J Med Educ* 1964;39:802–805.
10. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ* 1979;13:41–54.
11. Cusimano MD, Cohen R, Tucker W, Murnaghan J, Kodama R, Reznick R. A comparative analysis of the costs of administration of an OSCE (objective structured clinical examination). *Acad Med* 1994;69:571–576.
12. Reznick RK, Smee S, Baumber JS, et al. Guidelines for estimating the real cost of an objective structured clinical examination. *Acad Med* 1993;68:513–517.
13. Cohen R, Rothman AI, Ross J, Poldre P. Validating an objective structured clinical examination (OSCE) as a method for selecting foreign medical graduates for a pre-internship program. *Acad Med* 1991;66:S67–S69.
14. King AM, Pohl H, Perkowski-Rogers LC. Planning standardized patient programs: case development, patient training, and costs. *Teach Learn Med* 1994;6:6–14.
15. Klass DJ. "High-stakes" testing of medical students using standardized patients. *Teach Learn Med* 1994;6:28–32.
16. Norcini JJ, Stillman PL, Sutnick AI, et al. Scoring and standard setting with standardized patients. *Eval Health Prof* 1993;16:322–332.
17. Ben-David MF, Klass DJ, Boulet J, et al. The performance of foreign medical graduates on the National Board of Medical Examiners (NBME) standardized patient examination prototype: a collaborative study of the NBME and the Educational Commission for Foreign Medical Graduates (ECFMG). *Med Educ* 1999;33:439–446.
18. Friedman M, Mennin SP. Rethinking critical issues in performance assessment. *Acad Med* 1991;66:390–395.
19. Boulet JR, Swanson DB. Psychometric challenges of using simulations for high-stakes assessment. In: Dunn W, ed. *Simulations in Critical Care Education and Beyond*. Des Plains, IL: Society of Critical Care Medicine; 2004:119–130.
20. Wettach GR. A standardized patient enrolled in medical school considers the national clinical skills examination. *Acad Med* 2003;78:1240–1242.
21. Medical Council of Canada. Medical Council of Canada Qualifying Examination Part II (MCCQE Part II). 2008. Medical Council of Canada.
22. Federation of State Medical Boards, Inc. and National Board of Medical Examiners. United States Medical Licensing Examination: Step 2 clinical skills (cs) content descriptions and general information. 2008. Federation of State Medical Boards, Inc. and National Board of Medical Examiners.
23. National Board of Osteopathic Medical Examiners. Bulletin of information. 2008. National Board of Osteopathic Medical Examiners.
24. Whelan G. High-stakes medical performance testing: the Clinical Skills Assessment program. *JAMA* 2000;283:1748.
25. Heffron MG, Simspon D, Kochar MS. Competency-based physician education, recertification, and licensure. *WMJ* 2007;106:215–218.
26. Accreditation Council for Graduate Medical Education and American Board of Medical Specialties. Toolbox of assessment methods. Version 1. 1. 2000. Accreditation Council for Graduate Medical Education and American Board of Medical Specialties.
27. Frank JR. *The CanMEDS 2005 Physician Competency Framework: Better Standards. Better physicians. Better Care*. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2005.
28. Reznick RK, Blackmore D, Cohen R, et al. An objective structured clinical examination for the licentiate of the Medical Council of Canada: from research to reality. *Acad Med* 1993;68:S4–S6.
29. Smee SM, Blackmore DE. Setting standards for an objective structured clinical examination: the borderline group method gains ground on Angoff. *Med Educ* 2001;35:1009–1010.
30. Wood TJ, Humphrey-Murto SM, Norman GR. Standard setting in a small scale OSCE: a comparison of the Modified Borderline-Group Method and the Borderline Regression Method. *Adv Health Sci Educ Theory Pract* 2006;11:115–122.

31. Smee SM, Dauphinee WD, Blackmore DE, Rothman AI, Reznick RK, Des MJ. A sequenced OSCE for licensure: administrative issues, results and myths. *Adv Health Sci Educ Theory Pract* 2003;8:223–236.

32. Humphrey-Murto S, Smee S, Touchie C, Wood TJ, Blackmore DE. A comparison of physician examiners and trained assessors in a high-stakes OSCE setting. *Acad Med* 2005;80:S59–S62.

33. Tamblyn R, Abrahamowicz M, Dauphinee D, et al. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA* 2007;298:993–1001.

34. Ziv A, Ben David MF, Sutnick AI, Gary NE. Lessons learned from six years of international administrations of the ECFMG's SP-based clinical skills assessment. *Acad Med* 1998;73:84–91.

35. Sutnick AI, Stillman PL, Norcini JJ, et al. ECFMG assessment of clinical competence of graduates of foreign medical schools. Educational Commission for Foreign Medical Graduates. *JAMA* 1993; 270:1041–1045.

36. Whelan GP, Boulet JR, McKinley DW, et al. Scoring standardized patient examinations: lessons learned from the development and administration of the ECFMG Clinical Skills Assessment (CSA). *Med Teach* 2005;27:200–206.

37. Whelan GP, McKinley DW, Boulet JR, Macrae J, Kamholz S. Validation of the doctor-patient communication component of the Educational Commission for Foreign Medical Graduates Clinical Skills Assessment. *Med Educ* 2001;35:757–761.

38. Boulet JR, Rebbecchi TA, Denton EC, McKinley DW, Whelan GP. Assessing the written communication skills of medical school graduates. *Adv Health Sci Educ Theory Pract* 2004;9:47–60.

39. Boulet JR, McKinley DW, Norcini JJ, Whelan GP. Assessing the comparability of standardized patient and physician evaluations of clinical skills. *Adv Health Sci Educ Theory Pract* 2002;7:85–97.

40. Hawkins RE. The introduction of the Clinical Skills Assessment into the United States Medical Licensing Examination (USMLE): a description of USMLE Step 2 Clinical Skills (CS). *J Med Licensure Discip* 2005;91:22–25.

41. De Champlain AF, Swygert K, Swanson DB, Boulet JR. Assessing the underlying structure of the United States Medical Licensing Examination Step 2 test of clinical skills using confirmatory factor analysis. *Acad Med* 2006;81:S17–S20.

42. Swanson DB, Clauser BE, Case SM. Clinical Skills Assessment with standardized patients in high-stakes tests: a framework for thinking about score precision, equating, and security. *Adv Health Sci Educ Theory Pract* 1999;4:67–106.

43. De Champlain AF, Macmillan MK, Margolis MJ, et al. Modeling the effects of security breaches on students' performances on a large-scale standardized patient examination. *Acad Med* 1999;74:S49–S51.

44. Margolis MJ, De Champlain AF, Klass DJ. Setting examination-level standards for a performance-based assessment of physicians' clinical skills. *Acad Med* 1998;73:S114–S116.

45. Boulet JR, De Champlain AF, McKinley DW. Setting defensible performance standards on OSCEs and standardized patient examinations. *Med Teach* 2003;25:245–249.

46. Boulet JR, McKinley DW, Whelan GP, Hambleton RK. Quality assurance methods for performance-based assessments. *Adv Health Sci Educ Theory Pract* 2003;8:27–47.

47. Harik P, Clauser BE, Grabovsky I, Margolis MJ, Dillon GF, Boulet JR. Relationships among subcomponents of the USMLE Step 2 Clinical Skills Examination, the Step 1, and the Step 2 Clinical Knowledge Examinations. *Acad Med* 2006;81:S21–S24.

48. Hauer KE, Hodgson CS, Kerr KM, Teherani A, Irby DM. A national study of medical student clinical skills assessment. *Acad Med* 2005;80: S25–S29.

49. Gilliland WR, La Rochelle J, Hawkins R, et al. Changes in clinical skills education resulting from the introduction of the USMLE step 2 clinical skills (CS) examination. *Med Teach* 2008;30:325–327.

50. Gimpel JR, Boulet JR, Errichetti AM. Evaluating the clinical skills of osteopathic medical students. *J Am Osteopath Assoc* 2003;103: 267–279.

51. Boulet JR, Gimpel JR, Errichetti AM, Meoli FG. Using National Medical Care Survey data to validate examination content on a performance-based clinical skills assessment for osteopathic physicians. *J Am Osteopath Assoc* 2003;103:225–231.

52. Errichetti AM, Gimpel JR, Boulet JR. State of the art in standardized patient programs: a survey of osteopathic medical schools. *J Am Osteopath Assoc* 2002;102:627–631.

53. Gimpel JR, Boulet JR, Weidner AC. Survey on the clinical skills of osteopathic medical students. *J Am Osteopath Assoc* 2006;l106:296–301.

54. Baker HH, Cope MK, Adelman MD, Schuler S, Foster RW, Gimpel JR. Relationships between scores on the COMLEX-USA Level 2-Performance Evaluation and selected school-based performance measures. *J Am Osteopath Assoc* 2006;106:290–295.

55. Gimpel JR, Weidner AC, Boulet JR, Wilson C, Errichetti AM. Standardized patients and mechanical simulators in teaching and assessment at colleges of osteopathic medicine. *J Am Osteopath Assoc* 2007;107:557–561.

56. Boulet JR, Gimpel JR, Dowling DJ, Finley M. Assessing the ability of medical students to perform osteopathic manipulative treatment techniques. *J Am Osteopath Assoc* 2004;104:203–211.

57. Nestel D, Kneebone R, Black S. Simulated patients and the development of procedural and operative skills. *Med Teach* 2006;28: 390–391.

58. Kneebone RL, Nestel D, Vincent C, Darzi A. Complexity, risk and simulation in learning procedural skills. *Med Educ* 2007;41:808–814.

59. Hatala R, Issenberg SB, Kassen B, Cole G, Bacchus CM, Scalese R. Assessing cardiac physical examination skills using simulation technology and real patients: a comparison study. *Med Educ* 2008;42: 628–636.

60. Kobayashi L, Shapiro MJ, Gutman DC, Jay G. Multiple encounter simulation for high-acuity multipatient environment training. *Acad Emerg Med* 2007;14:1141–1148.

61. Morgan PJ, Pittini R, Regehr G, Marrs C, Haley MF. Evaluating teamwork in a simulated obstetric environment. *Anesthesiology* 2007; 106:907–915.

62. Malec JF, Torsher LC, Dunn WF, et al. The Mayo high performance teamwork scale: reliability and validity for evaluating key crew resource management skills. *Simul Healthcare* 2007;2:4–10.

63. Rosen MA, Salas E, Wilson KA, et al. Measuring team performance in simulation-based training: adopting best practices for healthcare. *Simul Healthcare* 2008;3:33–41.

64. Huang GC, Gordon JA, Schwartztein RM. Millenium conference 2005 on medical simulation: a summary report. *Simul Healthcare* 2007;2: 88–95.