

Psychometric Challenges of Using Simulations for High-Stakes Assessment

John R. Boulet, PhD; David B. Swanson, PhD

Objectives

- The reader will understand the use of different simulation methods to train and assess physicians.
- The reader will understand how evidence for the psychometric adequacy (reliability, validity) of simulation scores can be procured.
- The reader will understand the challenges of using simulations to assess physician competencies.

Key Words

reliability
validity
psychometric
patient simulation
clinical skill assessment

Over the past two decades, the use of simulations in medical education has become widespread (1-4). In the United States, most medical schools have well-established, standardized patient (SP) programs for teaching clinical skills. Standardized patients are laypeople who are trained to realistically and consistently portray actual patients for purposes of assessment and/or training. In addition, many medical schools have incorporated computer-based training modules, part-task-trainers (e.g., breast models, pelvic models), and full-scale integrated simulators to aid in the instruction of undergraduate medical students. The reasons for the increasing popularity of simulation-based training are clear: They can provide students with realistic clinical experiences without risks to patients; the tasks/scenarios can be designed to meet important needs, with increasing complexity introduced in a controlled way; skills can be practiced repeatedly, with tailoring to individual needs; and the likelihood of transfer from instruction to real-world situations is enhanced (5,6).

Similarly, simulations are increasingly used for assessment, both for low-stakes tests in medical schools and residency training programs and for high-stakes licensing and certification examinations. Dozens of medical schools in North America have implemented end-of-course, end-of-clerkship, and comprehensive examinations involving SPs (7). The Medical Council of Canada has used a multistation SP-based examination (Medical Council of Canada Qualifying Examination Part II) as part of the licensure process since 1993 (8,9); the Educational Commission for Foreign Medical Graduates (ECFMG) introduced the Clinical Skills Assessment in 1998 as part of the certification requirements for graduates of international medical schools (10); and the General Medical Council of the United Kingdom has conducted an SP-based assessment as one component of the Professional Linguistics and Assessment Board examination since roughly the same time (11). Beginning in the summer of 2004, a high-stakes, SP-based examination will be included in the United States Medical Licensing Examination (USMLE). For medical students graduating in 2005 or thereafter, passing this examination, known as USMLE Step 2 Clinical Skills (CS), will be required for licensure in all US jurisdictions. In addition, another simulation-based assessment method, computer-based case simulations, has been used in USMLE Step 3 since 1999.

On the horizon are a variety of other simulation-based assessment methods, including part-task-trainers (e.g., pelvic replicas) for teaching and testing psychomotor skills, virtual reality and haptic systems for teaching and testing surgical and other skills, and integrated simulators combining sophisticated life-size mannequins with computer programs allowing the mannequins to produce realistic physiological responses to pharmacologic and other interventions. These methods hold promise for testing the proficiency of individual members of healthcare teams and, for high-fidelity operating-room theater-type simulators, for the teams as a whole (12).

Although the fidelity, cost, and efficiency of simulation-based assessment methods vary extensively, their utility for high-stakes assessment depends crucially on their psychometric characteristics. In this chapter, we summarize a range of simulation-based assess-

ment methods, review their current use in USMLE, and then discuss some of the psychometric challenges that must be met for their promise to be fully realized. Following this, we provide some guidelines for developing simulation-based assessments, drawing heavily on the existing literature for SP assessments. Finally, given the emerging role of integrated simulators (mannequins) for both formative and summative assessment, we outline their potential use for credentialing and certification decisions.

Types of Simulations: The Fidelity Continuum

Simulations are designed to reproduce some aspect of the working environment. This may vary from the replication of a few aspects of a clinical task to the re-creation of an entire clinical environment, such as an operating theater. It is convenient to divide the environment into displays and surround (13). Displays are those environmental stimuli that are directly and intentionally influential in task performance, while the surround includes other stimuli that may influence performance indirectly or unintentionally. Generally, simulations replicate only some aspects of the displays and surround, and the fidelity of a simulation—its degree of realism or authenticity—ranges along a scale from completely artificial to the actual real-life situation. Fidelity is not unidimensional: any real-life situation is multifaceted, and only some characteristics of the environment are reproduced. Different simulation methods result from making different decisions about the aspects of the environment that are important to reproduce. Generally, these decisions are guided by the purpose of the assessment and the skills that the test is intended to measure.

One might view multiple-choice questions (MCQs) as simulations (at least those that begin with a brief patient description and require examinee to make a clinical decision) at the low end of the fidelity continuum, which provide an assessment of examinees' ability to apply their knowledge to written descriptions of case situations. Assessments using SPs lie at the other end of the fidelity continuum, providing a realistic context for measuring the skills involved in taking a history and performing a physical examination. Computer-based clinical simulations, part-task-trainers, and integrated simulators fall in between. An overview of each is provided in the next subsections; if the method is used in USMLE, the overview summarizes how it is used.

Multiple-Choice Questions

Patient-based MCQs have been used for decades; they have appeared in all three steps since the introduction of USMLE in 1992. In Step 1, these take the form of brief descriptions of patient care situations followed by questions challenging examinees to use their understanding of basic biomedical science to explain or predict patient findings (14). Roughly 60% of the items on Step 1 currently take this form. In Step 2, virtually all items begin with a description of a clinical situation; these are longer and less classic than patient descriptions on Step 1. Examinees must differentiate important findings from incidental ones and indicate a clinical decision, generally a diagnosis or the next step in

patient care (15). Step 3 test items also provide a robust description of a physician-patient encounter. Items are often presented in a format termed "case clusters," in which a series of MCQs address different facets of an unfolding clinical situation (16).

Across all three steps, the MCQs are best viewed as a form of low-fidelity patient simulations aimed at assessment of decision-making skills. The degree of fidelity depends, in part, upon the length of the patient description (17). Table 1 provides two versions of a sample MCQ that vary in detail and in the extent to which patient findings are provided in interpreted versus undigested format, illustrating how fidelity can vary across MCQs. In the future, the fidelity of MCQs on USMLE is likely to increase, as all three steps take advantage of computer-based test administration to incorporate multimedia into item "stems," thus enriching patient presentations.

Computer-Based Clinical Simulations

Computer-based clinical simulations (and before them, written clinical simulations, also termed patient-management problems) have been used in medical education since the 1960s and in assessment for almost that long (18). The Primum version of computer-based case simulations (CCS) was introduced in USMLE Step 3 in 1999; it was the first major change in the USMLE format since its inception. In CCS, the examinee is presented with a brief description of a patient, including a chief complaint and a brief history (16,19). From that point forward, the case unfolds as the examinee works up and manages the computer-simulated patient, obtaining diagnostic information, ordering therapeutic interventions, and monitoring patient progress. Any of several thousand diagnostic and therapeutic maneuvers can be requested by the examinee in free text on an "order sheet." As simulated time passes, the patient's condition changes based on the underlying medical problem and the examinee's interventions; results of tests are reported and the impact of interventions must be monitored. Examinees are scored on CCS using an algorithm that essentially compares their patient management strategies with policies obtained from experienced clinicians. Examinees must balance thoroughness, efficiency, timeliness, and avoidance of risk in responding to clinical situations with dangerous and unnecessary actions that lower scores. Though CCS cases have proven expensive to develop, administer, and score, psychometric analyses have indicated that CCS measures something somewhat differently than do MCQs, with a reasonable degree of precision (20,21).

Though not currently used in USMLE, it is clearly possible to further increase the fidelity of computer-based clinical simulations through the incorporation of multimedia. It is a reasonable expectation that this will occur within the next few years.

Part-Task-Trainers

Part-task-trainers are designed to replicate only a portion of the real clinical environment. Generally, these trainers resemble anatomical areas of the body (pelvic and breast trainers, models for soft tissue and joint injection, skin pads) and are used to provide training in the basic psychomotor skills involved in

performing physical examination maneuvers and minor procedures (e.g., venipuncture, catheterization, incision and drainage of abscesses).

Part-task-trainers are not currently used on USMLE, though they may be introduced into the Step 2 Clinical Skills examination in the near future (22), and they were used on the original National Board examination given in 1916 (23). The part-task-trainers are coming into common use by medical schools and residency programs for assessing psychomotor (particularly surgical) skills (24-27).

Integrated Simulators

Integrated simulators combine part-or-whole-body mannequins with computers that drive the findings "portrayed" by the simulator (28). Some mannequins can produce sounds that create the impression of the "patient" talking, physical signs including pulse and heart sounds, and even pupillary reactions and urine output. Computer-generated physiological signals can be fed to clinical monitoring equipment allowing results of simple (electrocardiogram, noninvasive blood pressure, oxygen saturation) and complex (central venous, pulmonary artery, and intracranial pressure) monitoring to be reported. Pharmacologic and other interventions (e.g., provision of oxygen and administration of fluid) can be automatically "sensed," with display of appropriate effects.

Although high-fidelity integrated simulators are currently quite costly, their use in clinical training is growing rapidly (3,4,12,29,30), both in undergraduate and graduate medical education. Their use in assessment, thus far, has tended to be for lower-stakes examinations intended to provide a basis for feedback on areas of examinee strength and weakness, but these methods hold promise for use in high-stakes examinations as well, in part because they have some notable advantages over SP-based assessments. First, they can be used to model rare events, especially those where errors are not reversible. Second, the types of conditions that can be modeled with SPs are limited; for example, it is difficult to use SPs for trauma-type scenarios. Most importantly, since real-time responses to therapeutic interventions can be modeled, the management of patient conditions requiring drug and other interventions can be assessed; SP-based assessments rarely include patient management as part of the interaction. For acute care scenarios such as anaphylaxis or myocardial infarction, where management decisions are critical, life-size mannequins seem likely to become the assessment method of choice.

Standardized Patient Assessments

SPs are nonphysicians taught to portray patients in a standardized and consistent fashion. SPs can be asymptomatic, can simulate a wide variety of abnormal physical findings, or may have a disease with stable physical findings. Examinees interact with SPs as though they were interviewing, examining, and counseling real patients (31,32).

Tests involving SPs are often called "objective structured clinical examinations" (33). These are administered by having

Table 1. Sample multiple-choice questions in nonvignette and vignette formats.

Which of the following renal abnormalities is most likely to be present in

a child with nephrotic syndrome and normal renal function?

- a. Acute poststreptococcal glomerulonephritis
- b. Hemolytic-uremic syndrome
- c. Minimal change nephrotic syndrome
- d. Nephrotic syndrome due to focal and segmental glomerulosclerosis
- e. Schonlein-Henoch purpura with nephritis

A 2-year-old African-American child is brought to the physician by his mother because he developed swelling of the eyes and ankles over the past week. Blood pressure is 100/60 mm Hg, pulse is 110/min, and respirations are 28/min. In addition to swelling of his eyes and 2+ pitting edema of his ankles, he has abdominal distention with a positive fluid wave. Serum concentrations are as follows: creatinine 0.4 mg/dL; albumin 1.4 g/dL; and cholesterol 569 mg/dL. Urinalysis shows 4+ protein and no blood. Which of the following is the most likely diagnosis?

- a. Acute poststreptococcal glomerulonephritis
- b. Hemolytic-uremic syndrome
- c. Minimal change nephrotic syndrome
- d. Nephrotic syndrome due to focal and segmental glomerulosclerosis
- e. Schonlein-Henoch purpura with nephritis.

examinees rotate around a circuit of "stations" at which they perform a variety of clinical tasks. Depending upon the purpose and format of the test, these may include activities such as taking a history from an SP, performing some portion of a physical examination, providing counselling or patient education, and ordering or interpreting diagnostic studies. Often, SPs are trained to complete checklists and rating forms at the end of encounters, recording the history findings obtained, the physical examination maneuvers performed, and the counselling provided, as well as rating communication skills of examinees. Alternatively, physician-raters may observe SP-examinee encounters and complete checklists and rating forms. The time allowed at each station can vary from a few minutes to an hour, depending upon the tasks to be performed; most commonly, stations last from 10 to 30 minutes.

SP-based examinations have been used in high-stakes examinations for more than a decade. Based on examinee volume, the SP-based ECFMG Clinical Skills Assessment is probably the largest hands-on, simulation-based examination in medicine (34). Since 1998, the ECFMG has tested over 43,000 graduates of international medical schools. In 2003 alone, over 11,500 test administrations were conducted, including more than 120,000 encounters between examinees and SPs. The Clinical Skills Assessment assesses clinical skills in four distinct domains: data gathering (history taking and physical examination), written communication, interpersonal skills, and spoken English proficiency (35).

The ECFMG and the National Board of Medical Examiners worked together for several years to introduce the USMLE Step 2

Clinical Skills examination in June of 2004. Step 2 Clinical Skills is a 1-day test that has a structure similar to a physician's workday in an ambulatory setting. Each examinee sees a series of 12 SPs and interacts with them as if they were real patients: establishing rapport, eliciting pertinent historical information, performing a focused physical examination, responding to questions, and providing counseling when appropriate. After each 15-minute encounter with an SP, the examinee has 10 minutes to complete a patient note, recording pertinent positive and negative history and physical findings, listing diagnostic impressions, and outlining plans for further evaluation (36). (These notes are subsequently scored holistically by trained physician raters.) While an examinee completes a patient note, the SP fills out a case-specific checklist documenting the history findings obtained and the physical examination maneuvers performed. The SP also completes a rating form assessing the examinee's communication, interpersonal, and English-speaking skills (37).

Reliability and Validity of Scores on Simulation-Based Assessments

According to the Standards for Educational and Psychological Testing (38,39), validity "refers to the appropriateness, meaningfulness, and usefulness of specific inferences made from test scores." Validity, thus, is not a property of a test, but a property of the inferences made from test scores, and the same test may produce scores that are valid for some purposes but not for others. Regardless of whether an assessment is based upon responses to a set of MCQs, behavior with real patients in actual practice, or performance on one or a series of simulations, there are at least three links in the chain of inferences required to interpret a score (40,41):

1. Evaluation (scoring) of the performance at hand—deciding whether the performance is good, poor, or somewhere in between
2. Generalization of results from the observed performance to other similarly structured, but not identical, tests
3. Extrapolation of the results from the assessment context to expected performance in actual practice

And the chain of inferences supporting the valid interpretation of test scores is only as strong as its weakest link.

Different assessment methods run into problems at different points in the inference chain. For example, suppose we are interested in examinees' ability to make appropriate patient care decisions in some domain of medical practice, and we develop an MCQ test consisting of 100 items, each of which describes a clinical situation in that domain and requires examinees to indicate the next step in patient care. The evaluation and generalization links for interpretation of scores on an MCQ test like this one tend to be very strong. Assuming that items are written and reviewed carefully, they should each have a clearly correct answer, so evaluation of the performance at hand should be straightforward. Because relatively large numbers of MCQs can be included on even short tests (e.g., a 100-item test might require two hours of testing time), it is likely that total scores on the test will be reproducible: they will be strongly related to scores on similar tests covering similar content but with differ-

ent items. The weak link for interpretation of scores on an MCQ test tends to be extrapolation, which is most plausible when the observed performance is very similar to the performance about which conclusions are wanted—and this clearly is not the case for our hypothetical MCQ exam. However, if the test is systematically constructed to consist of items describing commonly occurring patient situations posing "can't-afford-to-miss"-type questions, it may indeed be plausible to believe that examinees receiving low scores will not be able to provide safe and effective patient care, though the opposite may not be believable for those achieving high scores (41).

The pattern of strengths and weaknesses for observing patient care in actual practice is the opposite of the pattern for MCQs: evaluation and generalization tend to be problematic, while extrapolation tends to be a strength (40). Typically, there is little control over the cases that provide the basis for assessment in actual practice. Often, encounters in practice are either too simple (a patient with well-controlled blood pressure needing a hypertension recheck) or too complicated (a patient with a thick chart documenting multiple visits for long-standing, complex multisystem disease). For the latter, experts may disagree on the relative merits of the care that has been provided (or should have been provided), thus weakening the evaluation link. Even if there are agreed-upon practice guidelines for the condition(s) under study, generally some judgment is required in order to apply general guidelines to the care of an individual patient, introducing the potential for subjectivity and bias in evaluation. In addition, performance may justifiably be influenced by characteristics of the specific patient involved or by context variables beyond the control of both the evaluator and the practitioner. Generalization is likely to be poor because it is often difficult and inconvenient to obtain information about performance in practice, resulting in small, nonrepresentative samples of practice behavior from which inferences are to be drawn about performance more generally.

As discussed in the later subsection on generalization, the quality of care provided on one case tends to be a poor predictor of the quality of care provided on other cases (regardless of the assessment method used), so small convenience samples drawn from actual practice may not provide a very accurate basis for assessing general level of performance. However, if an assessment is based upon performance in actual practice, extrapolation should be less of a problem. Clearly, though, if an "examinee" knows that particular encounters are being "scored," this may cause better (or worse) performance. Ensuring that the measurement process is unobtrusive is probably particularly important if aspects of professional behavior are of interest in the assessment: an examinee is not likely to engage in unprofessional behavior if he/she knows that specific encounters will be used for assessment purposes (40).

Because of variability in the types of simulations and in how they may be configured for use in assessments, it is difficult to come up with justifiable, general conclusions about the strength of the evaluation, generalization, and extrapolation links for simulation-based methods. If simulation scenarios are carefully and deliberately developed with scoring in mind (33), scoring can be fairly objective and the evaluation link can be strong.

Similarly, if the test includes enough scenarios (though this is generally not true—[32,42]) and multiple raters participate in the assessment, the generalization link can be strong. And, to the extent that key elements of the real clinical setting are incorporated into the simulation, extrapolation can be strong as well. Unfortunately, depending upon the care used in developing the simulations, the test length, and other factors, all of these links can be weak as well. In the next subsections, we review each of these in more detail in the context of simulation-based assessments.

Evaluation: Approaches to and Problems in Scoring of Simulations

Regardless of the type or fidelity of a simulation-based assessment method, there are four types of criteria that are commonly used to generate scores: explicit process criteria, explicit outcome criteria, implicit process criteria, and implicit outcome criteria; combinations of these types of criteria can also be used. Table 2 provides examples of each type of criteria. The decision on which criteria to use generally depends on the skills to be measured, how those skills are manifested in examinee behavior, and to a degree, convenience. As a simple example, either implicit or explicit outcome criteria are generally appropriate for grading essays (e.g., a referral letter)—there is not much additional information to be gained by observing the examinee writing the essay. In contrast, for clinical situations where the linkage between clinical process and (simulated) patient outcome is weak, the use of process-oriented criteria is generally more appropriate (43). Often, process criteria are more relevant to determining if the care provided conforms to accepted standards.

With some variation, depending on the type of simulation and purpose of the examination, explicit process criteria are probably the most commonly used. These can vary widely in nature, from use of very detailed scoring criteria that assign a weight to every action taken by an examinee to very selective scoring criteria that focus in on key clinical decisions. The latter approach is more common in high-stakes examinations, though the former can be very useful in formative assessments because it tends to enable provision of more detailed feedback on performance.

From a research perspective, much of the research on scoring has taken place using data from SP assessments. Detailed information related to scoring of simulations generally and SP-based tests specifically can be found in Swanson (42) and van der Vleuten and Swanson (32). For history-taking and physical examination skills, checklists are often used to document examinee performance. The checklists are case-specific, reflecting the history and physical findings that should be obtained for a patient with a given presenting complaint. The checklists normally consist of explicit process criteria (e.g., questions asked or findings obtained) and can be weighted to take importance into account. It is generally difficult to incorporate data gathering sequence into score calculation, at least if the SP is responsible for completing the checklist. Alternately, provided that individuals with clinical expertise are available to serve as

Types of Criteria	Example
Explicit Process	Case-specific checklist used in a standardized patient chest-pain station to record the history findings obtained and physical examination maneuvers performed by an examinee
Implicit Process	Global judgment of a physician-rater observing an examinee's work with an integrated simulator in a trauma-type scenario
Explicit Outcome	Indicators of overall patient status (alive vs dead; complications; physiological indicators) at the conclusion of a computer-based clinical simulation
Implicit Outcome	Global judgment of a physician-rater inspecting the sutures made by an examinee on a skin pad
Combined Criteria	Task-specific checklist of explicit process and outcome criteria for observation and inspection of an end-to-end anastomosis of pig bowel

raters, holistic ratings can be obtained either in real time or by videotape review. Here, various rubrics can be used to obtain "expert" ratings of performance. These rubrics are normally based on implicit process criteria (e.g., interacts appropriately with the patient, asks relevant questions) and yield scores that reflect examinees' overall proficiency in interacting with the SP. However, depending on the purpose of the assessment, rater training, and the particular clinical skill being assessed, reasonably accurate scores can be obtained, regardless of whether explicit or implicit process criteria are employed (44,45).

Development of scoring checklists for SP-based assessments is fairly straightforward. For example, one could model an atypical pneumonia case where a 19-year-old woman comes to the clinic because of a cold for the past week. Symptoms such as fatigue, diffuse muscle aches, nonproductive cough, and sharp pain in the front of the chest could be simulated quite easily. If one were assessing basic interviewing and physical examination skills, there would be a number of history-taking checklist items that could be used (e.g., physician asks about muscle or body aches, fever, pain when taking a deep breath, medications, vomiting). Likewise, items on a physical examination checklist might include examination of the throat, flexing of the neck, palpation of the anterior cervical lymph nodes, etc. The SP being interviewed and examined, or some secondary observer, can easily document what the examinee did and did not do. The examinee's score is simply the percentage of history-taking questions asked (or findings obtained) and physical examination maneuvers performed. Individual checklist items can also be weighted to reflect their clinical importance. The use of checklists is advantageous for a number of reasons. First, documenting the ques-

tions asked and physical examination maneuvers performed are relatively objective tasks. Also, with proper training, raters can be reasonably accurate (37). Second, if the checklist is well constructed, the percentage of findings obtained is predictive of "expert" holistic ratings (46).

Explicit process criteria have also been used for written and computer-based clinical simulations, though problems with such scoring keys have been commonly encountered (18). In many clinical situations, a broad range of patient management strategies are possible, and it can be difficult to develop scoring keys that appropriately reward different strategies that are similar in quality and similar strategies that differ in quality, in part because of differences in examinee response style. Some examinees approach clinical simulations as if they are playing the game, *Twenty Questions*: they use an efficient approach, taking only those actions absolutely necessary for patient management. Other examinees use a more thorough approach, taking all actions that are not contraindicated. More detailed scoring keys typically reward the thorough examinee and penalize the efficient examinee, and it is possible to rack up points using the thorough approach, with scores tending to reflect the number of actions taken as well as the quality of those actions. In some research, compared with physicians-in-training, practicing physicians were more likely to adopt the efficient, *Twenty-Questions* approach, taking shortcuts based upon clinical experience. As a result, studies comparing performance of experienced clinicians with physicians-in-training have produced some anomalous results (18).

Because of problems in scoring, it is important to explore alternate approaches. A good example comes from a study of computer-based clinical simulations done by the National Board of Medical Examiners and American Board of Internal Medicine (1981). Two scoring methods were used. In one, every examinee selection was scored; this might be termed the "trees" method. For the other, only major patient care decisions were scored; this might be termed the "forest" method. Reliability analyses indicated that the trees-type scoring method yielded scores that were far more reproducible across cases, probably because this method differentiated examinees using an efficient "*Twenty-Questions*" test-taking strategy from those adopting a more thorough approach. In the study, expert physicians rated transcripts summarizing examinee selections on each case. Results indicated high correlations between expert ratings and forest-type scores but low correlations between ratings and trees-type scores. Thus, the more reliable scores were less-valid indicators of the quality of performance. This study, as well as results of research on clinical simulations more generally, was crucial in developing improved scoring systems for the new generation of computer-based clinical simulations (21,47).

Not surprisingly, checklist scoring has also been attempted for assessments that use mannequin-based cases (48,49). Similar to SP cases, checklists can be easily developed for acute-care scenarios typically modeled using mannequins. For example, one could easily model an anaphylaxis case where the physician is called to see a young woman who is in the recovery room following a tonsillectomy. She is not awake and is thrashing in bed with an occasional paroxysm of coughing. Depending on the

sophistication of the mannequin, vital signs (e.g., blood pressure=85/60) could also be modeled with reasonable fidelity. For this type of simulated patient, a checklist could also be constructed: establish complete neuromuscular recovery, auscultate chest, diagnose presence of bilateral wheeze, request bolus of intravenous fluids, intravenous epinephrine (correct dose), etc. However, unlike SP assessments, it is not possible for the patient to record the physician actions. Instead, scorers (e.g., nurses, physicians) can document examinee events in real time or via videotape review. This process can be aided by a simultaneous full display of patient vital signs (e.g., pulse oximetry, electrocardiogram, blood pressure). While research on the use of checklist scoring for mannequin-based scenarios is relatively sparse, some studies suggest that they can be used to discriminate between broad levels of performance (2).

The use of checklists to score mannequin-based cases, especially those that model acute care scenarios, can be problematic. First, in such situations, certain actions are typically much more important than others. While weighting these more heavily in the scoring rubric may alleviate the problem, there may be little consensus on what the weights should be. More importantly, if the weights are extremely large or negative weighting is incorporated (e.g., subtracting points for negative actions such as incorrect drug dosages), scoring can be complex, and the reliability of any decisions regarding competency can be compromised. Here, depending on the weighting schema, examinees could obtain a high or low score for a given scenario based on a single action. While this may be realistic, significant measurement error can be introduced if the score for an entire simulation is effectively based on a single action. Second, unlike SP-based assessments, the timing and sequence of specific actions is essential. Although there should be a reasonable ordering of data-gathering activities in a typical patient interview, asking about muscle aches before or after medication usage does not really matter much. In contrast, for acute care scenarios, the order of specific actions can be critical. Providing intravenous fluids prior to establishing an airway would definitely be inappropriate. For acute care mannequin-based scenarios, timing can be critical. For the anaphylaxis case mentioned previously, the time to make a correct diagnosis will affect management decisions (e.g., epinephrine dose), which in turn could seriously impact patient outcome.

Although detailed checklists have been constructed for mannequin-based simulations (48), other ways of quantifying examinee performance, including those based on implicit process criteria or patient outcomes (explicit outcome criteria), may be more useful. Global scoring, where an expert provides a holistic rating of the overall performance, has been commonly used. Although this type of scoring has been criticized because of subjectivity, with proper training experts can provide scores that are both reliable and valid (2,36,50). Unfortunately, for mannequin-based assessments with complex patient manifestations, the experts would need to be physicians, and this could substantially increase the cost of any assessment. Key actions could also be used to score performances. For many acute care scenarios, there are a relatively small number of key actions that one would expect of the physician. For a ventricular tachycardia case, one would expect the examinee to make a diagnosis,

initiate a correct therapy for arrhythmia, and shock the patient. If the scenario timing is relatively brief (say, 5 minutes), then one could simply score these explicit process actions. This strategy would be efficient and relatively objective. Likewise, one could also incorporate a factor related to speed of response by timing these actions and building this into the scoring rubric, with the examinees performing the key actions sooner receiving additional points. The scoring system could also be based on patient outcomes. For example, following respiratory failure and intubation, the desired outcome would be effective ventilation (explicit outcome). Here, physiological monitors could be used to develop scoring algorithms that rewarded examinees for actions that led to patient improvement and penalized them for actions that led to negative outcomes (e.g., death).

Generalization: Sources of Measurement Error in Simulation-Based Test Scores

The purpose of any assessment is to permit inferences to be drawn concerning the proficiency of examinees—more specifically, inferences that extend beyond the particular clinical situations included in the assessment to the larger domain from which the clinical situations in the assessment are but a sample. Performance on the sample provides a basis for estimating proficiency in the broader domain that is actually of interest.

Depending on the nature of the sample, those estimates can be more or less generalizable (reliable, reproducible): they may provide a good or a poor basis for predicting performance on other similar but not identical assessments. If the sample is too small, estimates of examinee proficiency are not reproducible from one assessment to the next. If the sample is not representative of the broader domain of interest (e.g., including only critical care situations in a test of competence in anesthesiology), test results will be biased and will not provide a good basis for estimating proficiency in the domain that is actually of interest.

Most tests of clinical competence, regardless of the specific assessment method used, can be viewed within the sampling framework illustrated in Figure. For each examinee, a test is composed of a sample of cases (simulated clinical scenarios in the context of this writing) from a domain of cases and a sample of judges/raters from a domain of judges/raters. Within this framework, inter-rater agreement (reliability) is the extent to which judges agree in assigning scores to an examinee on a particular case. In contrast, one might use the term “inter-case” reliability to refer to the extent to which an examinee obtains similar scores on different cases from the same judge. For an estimate of an examinee’s competence in a domain to be generalizable, an adequate number of both judges and cases must be sampled in the assessment. Thus, test design can be viewed, in part, as the development of a “sampling plan” for cases and judges, and not all sampling plans are equally effective. In the figure, the sampling plan illustrated by the row of cells labeled A (call it the “inquisition model”) includes an adequate sample of judges, but only a single case; test scores generated using this and similar approaches will not be generalizable because the number of cases included in the sample is too small. The sampling plan illustrated by the column of cells labeled B includes an adequate

sample of cases, but only a single judge is evaluating performance. As a consequence, the score received by a given examinee will be heavily influenced by the stringency of the judge: if an examinee is assigned to a “hawk,” he or she will tend to receive scores that are too low, and the reverse is true if an examinee is assigned to a “dove.” This has been a common problem in conduct of oral examinations: pass/fail outcomes tend to be more determined by the stringency/leniency of the judge than the proficiency of the examinee. The diagonal of cells labeled C in the Figure illustrates an efficient sampling plan: if a different judge is assigned to each case, the maximum numbers of judges and cases are included in the sample, producing the most generalizable score for a given test length and amount of judge-time.

Ironically, most assessments of clinical competence have focused on ensuring adequate inter-rater agreement, without attending to variation in performance across clinical situations. Yet, in virtually all studies, regardless of the assessment technique used, differences in an examinee’s performance from one clinical situation to another are as large, or larger, than the differences across judges for the same clinical situation. For example, for scores on an SP-based assessment of clinical skills, several hours of testing time and 10 or more cases are required to obtain a score that is sufficiently generalizable for use in making high-stakes decisions (32,51). The same is true for assessment of decision-making skills with computer-based clinical simulations (18,52,53). Comparable results have been obtained for assessments using oral examinations, chart audit, and ratings of performance provided by patients, nurses, and medical faculty (54).

These findings imply that it is important to include a large sample of clinical situations in simulation-based assessments, as well as large numbers of judges. Specific numbers will vary depending upon the skills to be tested, the breadth of the clinical domain of interest, the control that can be exercised over the testing situation, the training of the raters, the preparation of guides to scoring, and other factors. Statistical techniques (termed generalizability studies) have been developed specifically for the purpose of investigating sampling requirements for alternate assessment methods and for evaluating alternate sampling approaches (55-57).

Extrapolation: Inferring Real-World Behavior from Simulation-Based Test Scores

The desire to strengthen the extrapolation link while maintaining good control over test content and scoring provides (often tacitly) much of the motivation for use of simulation-based assessment methods. Developers of simulation-based assessments often seek to test professional judgment and problem-solving skills in realistic contexts, challenging examinees to set priorities and deal with the subtleties of important clinical situations in a safe environment. In adopting this approach, they, in effect, are attempting to short-circuit questions of validity by simulating the real-world clinical situations of ultimate interest as closely as possible (40). Use of high-fidelity simulations can, however, introduce the same extraneous factors that make it difficult to use behavior in the real practice setting as a basis

for assessment. This approach also tends to result in simulations that require more testing time, as well as introducing more random and systematic error in scoring and generalization. The solution is to carefully introduce enough realism and complexity to obtain a valid indication of examinees' ability to apply their knowledge and skills, while constraining the structure and length of simulations sufficiently that the evaluation and generalization links remain strong. Decisions about how to make such trade-offs are often difficult.

It is also important to recognize that, no matter how realistic a simulation is, it is still a simulation, and examinees do not necessarily behave as they would in real life. Surveys of examinees have consistently shown that, from an examinee's perspective, simulation-based methods provide a more realistic assessment of, for example, clinical skills than multiple-choice tests (18). However, studies comparing examinee behavior on simulations with behavior in the real world have shown striking differences. In part, the differences are due to cueing present in some simulation formats. In particular, actions are more likely to be taken if they are suggested by lists of options provided in some simulation formats (unless those lists are also explicitly present in real life—e.g., laboratory order sheets). This can result in more thorough data gathering, usually leading to different (and better) performance on simulations than in real life.

Since data-gathering activities with real patients are usually uncued and open-ended, it seems as if uncued simulation formats should be more appropriate. However, the use of uncued formats also has consequences that must be considered. For example, in a comparison of performance with SPs and natural-language-based computer simulations, Feightner and Norman (58) reported marked differences in the patient history information elicited: far fewer findings were obtained in the computer simulation format, probably because the software did not provide a particularly good simulation of interacting with a patient. Other studies report similar variations in performance for different simulation formats (18).

The vagueness of more realistic, uncued formats can also result in differences in examinees' perceptions of "what the question/task" is, which induces differences in examinee behavior that are unrelated to skills. Because simulations almost invariably omit some features of the real environment, behavioral artifacts can occur if those features have an important influence on behavior. For example, in the landmark study of medical problem solving reported in Elstein et al (59), instructions to participants working up SPs did not make it clear if the participants were only to take care of each SP's presenting problem at that visit or to assume responsibility for long-term care. As a consequence, some physicians performed short, focused workups of SPs' presenting problems; others did complete histories and physicals and initiated patient management activities addressing prevention of new problems. This variation was almost certainly due to differing perceptions of the intent of the simulation, rather than reflecting real-world differences in practice style. The same kinds of problems have been observed with other kinds of simulations: if expectations are unclear, the behavior of the examinees will be influenced by the varying perceptions of the tasks posed for them.

Validation of Scores on Simulation-Based Assessments

The new Standards for Educational and Psychological Testing defines validation as the development of evidence providing "...a sound scientific basis for the proposed score interpretations" (39, p. 9). Thus, validation involves the development of a coherent argument for the proposed interpretation of scores, as well as arguments against plausible alternate explanations (60). Because identification of poorly performing examinees is often of particular importance, it is especially important to investigate that low scores are due to lack of proficiency rather than impediments to performance unrelated to the knowledge and skills of interest (e.g., unclear instructions, unrealistically speeded testing conditions, other extraneous factors).

Much of the work on validation of scores on simulation-based assessments has looked at a) the correlations between those scores and other measures and/or b) the magnitude of differences in performance among groups varying in training or experience. Because we rarely have strong hypotheses about what the magnitude of these correlations or group differences should be, these studies tend to be informative only if results are counterintuitive (32,42,52). For example, if the mean score of acknowledged experts in a clinical domain is the same or lower than a group of junior medical students, it raises serious questions about validity. Similarly, if performance on a simulation is completely (or negatively) unrelated to performance on an MCQ-based assessment of knowledge relevant to the simulation, validity questions are again raised. But, because we do not generally know how large the correlations or group differences should be, many different outcomes of such analyses can be viewed as providing "support" for the validity of the inferences drawn from simulation-based test scores.

Though it is reasonable to conduct correlational studies and studies of group performance as a part of a validation effort, it is generally more productive to use an alternative approach in which a series of "threats" to the validity of score interpretation are identified. Some of these have already been discussed: scoring keys that do not appropriately award alternative approaches to patient management, poor inter-rater agreement, sampling plans for constructing tests that result in limited (or biased) samples of cases or judges, and failure to appropriately represent important elements of the real clinical environment influencing performance. Others depend upon the specific simulation method being used. For example, for computer-based clinical simulations, depending upon the complexity of the computer interface, examinee performance may be influenced by computer skills or by the opportunities that examinees have to familiarize themselves with it (61). For part-task-trainers, the time allotted to complete a task may have a significant influence on performance. For a high-fidelity trainer, this can be appropriate, valid measurement information for tasks that need to be done quickly. But, for a low-fidelity trainer, if it simply reflects the fact that the trainer is unfamiliar, it may be inappropriate and result in less valid scores.

The basic message, though, is to identify alternate explanations for good and poor performance and to investigate those

	Judge 1	Judge 2	Judge 3	...	Judge n
Case 1	A B C	A	A	A	A
Case 2	B				
Case 3	B				
...	B				
Case m	B				
					C

Figure. Assessment of clinical skills from a sampling perspective. A: Broad sample of judges; inadequate sample of cases. B: Broad sample of cases; inadequate sample of judges. C: Optimal sampling plan for given numbers of cases and judges.

carefully. Such work, which has become more prevalent for performance-based assessments (62-64), tends to provide very useful information for improving the assessment method, as well as a stronger basis for test validation.

Guidelines for Developing Simulation-Based Assessments

Although research pertaining to the reliability and validity of scores from commonly employed simulation-based assessments (e.g., MCQs, computer simulations, SPs) is widespread, relatively little work has been done with respect to integrated simulators. By drawing on the lessons learned in the more general literature concerning assessment of clinical competence, general test development guidelines applicable can be derived. Newble et al (65), is a particularly valuable, readable "how-to" guide for development of simulation-based assessments. Based on these guidelines, which are linked to fundamental psychometric requirements, the potential benefits and drawbacks of using integrated simulators for high-stakes assessment purposes can be surmised.

Specific Lessons from the Standardized Patient Literature

Unlike integrated simulators, SPs have been used successfully for high-stakes assessment decisions, and their use is supported by over 40 years of research and development. While integrated simulators may be used for different assessment purposes (e.g., patient management skills), under certain circumstances it may be feasible to use them for high-stakes assessment decisions. Thus, lessons learned from SP-based assessments will be invaluable.

As mentioned in the score generalization section above, content sampling is an important limitation of any performance-based assessment. For SP examinations and integrated

simulators, a finite number of patient conditions can be modeled. Therefore, great care must be taken in choosing which and how many scenarios an individual examinee must complete as part of the assessment (66). Content specificity will also be an issue, perhaps more so for tasks based on the integrated simulator. Unlike SP assessments that generally key on basic clinical skills, specific experience and knowledge in one domain (e.g., cardiology) could significantly impact performance. Also, if an examinee has little experience with a certain procedure (e.g., intubation) then one should expect him or her to do well on specific simulation tasks. In general, examinees with little direct experience in a given clinical area would be expected to perform poorly, regardless of their skill level. As a result, performance in one scenario will not necessarily be a good predictor of performance in another. Therefore, to get a reliable estimate of performance, an examinee will need to be assessed across numerous tasks.

Many of the other psychometric issues associated with performance-based examinations are similar, regardless of whether mannequin-based scenarios or SP-based methods are used for assessment. First, regardless of the rubric used, there will always be some measurement error. In order to minimize these errors, criteria for crediting examinees must be well defined, and scorers/raters must be sufficiently trained. Second, establishing the validity of simulation-based scores can be difficult and time consuming. For both SP and integrated simulation scenarios, the examinees must use their imaginations to overcome the artificiality of the assessment environment. As a result, extrapolating from performance in the simulated environment to real-world situations can be risky. Although there have been studies to show that skills developed and assessed in the simulated environment transfer to real patients (67,68), this transfer is far from perfect. Third, it is often difficult if not impossible to model some conditions well (e.g., swelling in SP assessments). For integrated simu-

lators, the simulation environment is also imperfect. Changes in skin color, muscle tone, etc., which may be key clinical clues, cannot at present be imitated. Since many common conditions cannot be modeled, the generalizability of the assessment results to the universe of patient complaints will be limited and highly dependent on the set of tasks included in the assessment. Finally, for testing large numbers of examinees, it is essential that scores are comparable across test forms. Here, score-equating procedures are essential (51). These mathematical techniques are used to adjust scores based on the psychometric properties (e.g., difficulty) of the test form. More importantly, because modeled scenarios are expensive to develop, they are often few in number, magnifying potential security concerns. Special care must be taken to ensure that examination materials are not vulnerable to duplication and theft.

For SP-based examinations, many of the psychometric issues noted above have been studied and addressed in operational testing programs. Follow-up studies, where the skills and proficiencies of examinees are assessed with real patients, have also been completed (68). These types of investigations, including operational feasibility and population-based outcome studies, would be extremely useful for mannequin-based assessments.

Using Integrated Simulators for High-Stakes Assessments

Many challenges are associated with using mannequins for high-stakes certification or credentialing decisions (e.g., costs). However, for particular skill sets, mannequins may hold some advantages over SP assessments. Although SPs are generally highly trained, for certain cases there can be some variability in the way they respond to physician queries or physical examination maneuvers. Therefore, examinees may get the same information, albeit with somewhat different queries and physical examination maneuvers. For integrated simulators, the patient will react in a physiologically consistent fashion. If two examinees administer the same drug in the same dosage, the mannequin will react in the same way. Unlike integrated simulators, SPs are extremely useful for evaluation of doctor-patient communication skills (69,70). While some of the high-tech mannequins can talk via a speaker in the mouth, this is most useful for pain-type responses, not establishing rapport between doctor and patient. Team communication skills could potentially be assessed with mannequin-based cases, but current scoring models may be inadequate.

One of the greatest strengths of mannequin-based assessment is the ability to manage patient conditions and deal with a host of physical findings. This is not possible for most standardized patient assessments, especially those that are used for high-stakes credentialing decisions. For many medical specialties (e.g., anesthesiology, critical care, emergency medicine), the ability to treat simulated acute-care patients is extremely beneficial and avoids potential negative "real" patient outcomes. Many of the modeled physical findings are quite realistic, enabling the assessor to get a reasonable picture of how the examinee may react in a real-life situation.

One of the major drawbacks of SP-based assessments is the time it takes to assess an individual examinee. Depending on the

chief complaint, it can take 10 to 20 minutes to take the history and perform the relevant physical examination maneuvers. More importantly, to get a reliable estimate of an examinee's skill level, it may take in excess of 10 individual SP encounters. Thus, the exam can be fairly long and expensive to administer. In contrast, scenarios developed for integrated simulators, especially those that involve acute-care concerns, can be relatively brief. This not only affords the opportunity to sample more extensively, increasing the generalizability (reliability) of scores, but can also minimize staff costs associated with examination administration. Likewise, scoring is more straightforward in that less material needs to be reviewed and evaluated.

Conclusion

Based on the available literature, there are a number of psychometric issues associated with the use of simulations for high-stakes assessment. While there are numerous types of simulations covering a broad fidelity continuum, the basic challenge remains the same—ensuring that the scores have meaning and are useful for making specific inferences regarding the proficiency of examinees. Whether we are concerned with the evaluation of the observed performance, generalization of scores to other similar tests, or the extrapolation of test results to expected performance in actual practice, gathering evidence to support the use of the test scores is a necessary task if the assessment results are to be used for high-stakes assessment decisions.

In medicine, a number of performance-based examinations are currently being used for certification and licensure decisions. These assessment methodologies, including computer-case simulations and SP-based methods, have undergone extensive testing to ensure that potential threats to the validity of the scores and associated pass/fail decisions are addressed. As more sophisticated assessment methods (e.g., integrated simulators, haptics systems) are earmarked for use in credentialing decisions, research efforts will need to be focused on determining what can and cannot be measured effectively, establishing the reliability and validity of the scoring systems, and ensuring that performance in the assessment environment translates to performance with real patients.

REFERENCES

1. Bennett RE. Using new technology to improve assessment. *Educ Meas Issues Pract.* 1999;8(3):5-12.
2. Boulet JR, Murray D, Kras J, Woodhouse J, McAllister J, Ziv A. Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. *Anesthesiology.* 2003;99:1270-1280.
3. Good ML. Patient simulation for training basic and advanced clinical skills. *Med Educ.* 2003;37(suppl 1):14-21.
4. Schuwirth LW, van der Vleuten CP. The use of clinical simulations in assessment. *Med Educ.* 2003;37(suppl 1):S65-S71.
5. Maran NJ, Glavin RJ. Low-to high-fidelity simulation—a continuum of medical education? *Med Educ.* 2003; 37(suppl 1): S22-S28.
6. Dent J. Current trends and future implications in the developing role of clinical skills centres. *Med Teach.* 2001;23:483-489.
7. Association of American Medical Colleges. Emerging trends

- in the use of standardized patients. *Contemp Issues Med Educ.* 1998;1(7).
8. 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(suppl 10):S4-S6.
 9. 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(suppl 1):S19-S21.
 10. Whelan G. High-stakes medical performance testing: the Clinical Skills Assessment program. *JAMA.* 2000;283:1748.
 11. Tomblason P, Fox RA, Dacre JA. Defining the content for the objective structured clinical examination component of the professional and linguistic assessments board examination: development of a blueprint. *Med Educ.* 2000;34:566-572.
 12. Byrne AJ, Greaves JD. Assessment instruments used during anaesthetic simulation: review of published studies. *Br J Anaesth.* 2001;86:445-450.
 13. Fitzpatrick R, Morrison EJ. Performance and product evaluation. In: Thorndike R, ed. *Educational Measurement.* Washington, DC: American Council on Education; 1971.
 14. Federation of State Medical Boards, National Board of Medical Examiners. 2004 USMLE Step 1 Content Description and Sample Test Materials. Philadelphia, Pa: Federation of State Medical Boards and National Board of Medical Examiners; 2003.
 15. Federation of State Medical Boards, National Board of Medical Examiners. 2004 USMLE Step 2 Content Description and Sample Test Materials. Philadelphia, Pa: Federation of State Medical Boards and National Board of Medical Examiners; 2003.
 16. Federation of State Medical Boards, National Board of Medical Examiners. 2004 USMLE Step 3 Content Description and Sample Test Materials. Philadelphia, Pa: Federation of State Medical Boards and National Board of Medical Examiners; 2003.
 17. Swanson D, Case S. Trends in written assessment: a strangely biased perspective. In: Harden R, Hart I, Mulholland H, eds. *Approaches to the Assessment of Clinical Competence: Part 1.* Norwich, England: Page Brothers; 1992:38-53.
 18. Swanson D, Norcini J, Grosso L. Assessment of clinical competence: written and computer-based simulations. *Assess Eval Higher Educ.* 1987;12:220-246.
 19. Clyman S, Melnick D, Clauser B. Computer-based case simulations from medicine: Assessing skills in patient management. In: Tekian A, McGuire C, McGaghie W, ed. *Innovative Simulations for Assessing Professional Competence.* Chicago, Ill: University of Illinois, Department of Medical Education; 1999:29-41.
 20. Clauser BE, Margolis MJ, Swanson DB. An examination of the contribution of computer-based case simulations to the USMLE Step 3 examination. *Acad Med.* 2002;77(suppl 10):S80-S82.
 21. Dillon GF, Clyman SG, Clauser BE, Margolis MJ. The introduction of computer-based case simulations into the United States medical licensing examination. *Acad Med.* 2002;77(suppl 10):S94-S96.
 22. Federation of State Medical Boards, National Board of Medical Examiners. 2004 USMLE Step 2 CS Content Description and General Information Booklet. Philadelphia, Pa: Federation of State Medical Boards and National Board of Medical Examiners; 2003.
 23. Hubbard J, Levit E. *The National Board of Medical Examiners: The First Seventy Years.* Philadelphia, Pa: National Board of Medical Examiners; 1985.
 24. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg.* 1997;173:226-230.
 25. Goff BA, Nielsen PE, Lentz GM, et al. Surgical skills assessment: a blinded examination of obstetrics and gynecology residents. *Am J Obstet Gynecol.* 2002;186:613-617.
 26. Dunkin BJ. Flexible endoscopy simulators. *Semin Laparosc Surg.* 2003;10(1):29-35.
 27. Pugh CM, Youngblood P. Development and validation of assessment measures for a newly developed physical examination simulator. *J Am Inf Assoc.* 2003;9:448-460.
 28. Schwid HA. Anesthesia simulators—technology and applications. *Isr Med Assoc J.* 2000;2:949-953.
 29. Issenberg SB, McGaghie WC, Hart IR, et al. Simulation technology for health care professional skills training and assessment. *JAMA.* 1999;282:861-866.
 30. Blackburn T, Sadler C. The role of human patient simulators in health-care training. *Hosp Med.* 2003;64:677-681.
 31. Vu NV, Barrows HS. Use of standardized patients in clinical assessments: recent developments and measurement findings. *Educ Res.* 1994;23(3):23-30.
 32. van der Vleuten C, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. *Teach Learn Med.* 1990;2(2):58-76.
 33. Harden RM, Stevenson M, Downie WW, Wilson GU. Assessment of clinical competence using objective structured examination. *BMJ.* 1975;1:447-451.
 34. Educational Commission for Foreign Medical Graduates. 2004 Clinical Skills Assessment Orientation Manual. Philadelphia, Pa: Educational Commission for Foreign Medical Graduates; 2004.
 35. Whelan GP. Educational commission for foreign medical graduates: clinical skills assessment prototype. *Med Teach.* 1999;21:156-160.
 36. Boulet JR, Rebecchi TA, Denton EC, McKinley DW, Whelan GP. Assessing the written communication skills of medical school graduates. *Adv Health Sci Educ Theory Pract.* 2004;9(1):47-60.
 37. Boulet JR, McKinley DW, Whelan GP, Hambleton RK. Quality assurance methods for performance-based assessments. *Adv Health Sci Educ Theory Pract.* 2003;8(1):27-47.
 38. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association; 1985.
 39. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association; 1999.
 40. Kane MT. The assessment of professional competence. *Eval Health Prof.* 1992;15:163-182.
 41. Kane MT. Validating interpretive arguments for licensure and certification examinations. *Eval Health Prof.* 1994;17:133-159.
 42. Swanson D. A measurement framework for performance-based tests. In: Hart I, Harden R, eds. *Further Developments in Assessing Clinical Competence.* Montreal, Canada: Can-Heal Publications; 1987:13-45.
 43. McAuliffe WE. Studies of process—outcome correlations in medical care evaluations: a critique. *Med Care.* 1978;16:907-930.
 44. Hodges B, McNaughton N, Regehr G, Tiberius R, Hanson M. The challenge of creating new OSCE measures to capture the characteristics of expertise. *Med Educ.* 2002;36:742-748.
 45. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. *Med Educ.* 2003;37:1012-1016.

46. 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(2):85-97.
47. Clauser BE, Subhiyah TE, Piemme LG, Clyman SG, Ripkey D, Nungester RJ. Using clinician ratings to model score weights for a computer-based clinical-simulation examination. *Acad Med*. 1993;68(suppl 10):S64-S66.
48. Morgan PJ, Cleave-Hogg D, DeSousa S, Tarshis J. High-fidelity patient simulation: validation of performance checklists. *Br J Anaesth*. 2004;92:388-392.
49. Murray D, Boulet J, Ziv A, Woodhouse J, Kras J, McAllister J. An acute care skills evaluation for graduating medical students: a pilot study using clinical simulation. *Med Educ*. 2002;36:833-841.
50. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med*. 1998;73:993-997.
51. 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(1):67-106.
52. Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons from the health professions. *Educ Res*. 1995;24(5):5-11.
53. Clauser B, Harik P, Clyman S. The generalizability of scores for a performance assessment scored with a computer-automated scoring system. *J Educ Meas*. 2000;37:245-262.
54. Woolliscroft JO, Howell JD, Patel BP, Swanson DB. Resident-patient interactions: the humanistic qualities of internal medicine residents assessed by patients, attending physicians, program supervisors, and nurses. *Acad Med*. 1994;69:216-224.
55. Brennan RL. *Generalizability Theory*. New York, NY: Verlag; 2001.
56. Brennan RL. Performance assessments from the perspective of generalizability theory. *Appl Psychol Meas*. 2000;24:339-353.
57. Brennan RL, Johnson EG. Generalizability of performance assessments. *Educ Meas Issues Pract*. 1995;14:9-12, 27.
58. Feightner J, Norman G. Computer based problems as a measure of the problem-solving process: some concerns about validity. *Proceedings of the 17th Annual Conference on Research in Medical Education*. 1978, 51-56.
59. Elstein AS, Shulman LS, Sprafka SA. Medical problem-solving. *J Med Educ*. 1981;56(1):75-76.
60. Cronbach L. Five perspectives on the validity argument. In: Wainer H, Braun H, eds. *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum; 1988.
61. Clauser B, Kane M, Swanson D. Validity issues for performance-based tests scored with computer-automated scoring systems. *Appl Meas Educ*. 2002;15:413-432.
62. Burger SE, Burger DL. Determining the validity of performance-based assessment. *Educ Meas Issues Pract*. 1994;13:9-15.
63. Haertel EH. Validity arguments for high-stakes testing: in search of the evidence. *Educ Meas Issues Pract*. 1999;18(4):5-9.
64. Hodges B. OSCE! Variations on a theme by Harden. *Med Educ*. 2003;37:1134-1140.
65. Newble D, Dauphinee WD, Macdonald M, et al. Guidelines for assessing clinical competence. *Teach Learn Med*. 1994;6:213-220.
66. Ziv A, Boulet JR, Burdick WP, Ben-David MF, Gary NE. The use of national medical care surveys to develop and validate test content for standardized patient examinations. Melnick D, ed. *Proceedings of the Eighth Ottawa Conference on Medical Education and Assessment*, 99-105. Philadelphia, Pa: National Board of Medical Examiners; 2000.
67. Kramer AW, Jansen JJ, Zuithoff P, Tan LH, Grol RP, van der Vleuten CP. Predictive validity of a written knowledge test of skills for an OSCE in postgraduate training for general practice. *Med Educ*. 2002;36:812-819.
68. 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.
69. Boulet JR, Ben-David MF, Ziv A, et al. Using standardized patients to assess the interpersonal skills of physicians. *Acad Med*. 1998;73(suppl 10):S94-S96.
70. Cohen DS, Colliver JA, Marcy MS, Fried ED, Swartz MH. Psychometric properties of a standardized-patient checklist and rating-scale form used to assess interpersonal and communication