Available online at www.sciencedirect.com
ScienceDirect

Teaching and Teacher Education 23 (2007) 239–250

TEACHING AND TEACHER EDUCATION

www.elsevier.com/locate/tate

Assessment methods in medical education

John J. Norcini*, Danette W. McKinley

Foundation for Advancement of International Medical, Education and Research (FAIMER®), 3624 Market Street, 4th Floor, Philadelphia, PA 19104, USA

Abstract

Since the 1950s, there has been rapid and extensive change in the way assessment is conducted in medical education. Several new methods of assessment have been developed and implemented over this time and they have focused on clinical skills (taking a history from a patient and performing a physical examination), communication skills, procedural skills, and professionalism. In this paper, we provide examples of performance-based assessments in medical education, detailing the benefits and challenges associated with different approaches. While advances in psychometric theory and technology have been paralleled by the development of assessment instruments that improve the evaluation of these skills, additional research is needed, particularly if the assessment is used to make high stake decisions (e.g., promotion and licensure).
© 2007 Elsevier Ltd. All rights reserved.

Keywords: Assessment; Clinical skills; Medical education; Psychometrics; Evaluation

1. Introduction

The use of a variety of different assessment methods has been characteristic of medical education, credentialing, and licensure since the 1950s. Prior to that time, the medical knowledge and clinical skills of doctors were often assessed using written and oral examinations. The written examinations were usually composed of open-ended questions of one type or another, which were graded by hand. The oral examinations (viva voce) typically required the student to go to a patient's bedside, gather information, and then present a diagnosis and treatment plan to assessors who asked questions and made judgments about the performance.

Since then, there has been rapid and extensive change in the way assessment is conducted in medical education based on a number of developments (Norcini, 2005). Historically, educators started with the available assessment methods and then used them for all of the competencies of a doctor, even when they were ill-suited to the task. For example, it is critical for a doctor to be able to communicate effectively with patients but an assessment of this aspect of competence is not tested well by written examinations or a viva in which the student–patient encounter is unobserved. To correct this problem, several new methods of assessment have been developed and implemented over the past 50 years. These new methods have focused on clinical skills (taking a history from a patient and performing a physical examination), communication skills, procedural skills, and professionalism.

*Corresponding author. Tel.: +1 215 823 2170.
E-mail address: JNorcini@Faimer.org (J.J. Norcini).

0742-051X/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tate.2006.12.021

In this paper, we will provide a framework for selecting methods of assessment and an overview of the assessment methods used in medical education, with a focus on recent developments in simulation and work-based strategies. The advantages and disadvantages associated with these assessments will be offered. Detailed examples of simulation and work-based assessment methods and the benefits and challenges related to their use will be provided.

2. Framework for selection of assessment methods

Historically, decisions about which method of assessment to use have rested primarily on validity and reliability. Validity is the degree to which the inferences made about medical competence based on assessment scores are correct (Messick, 1989). Reliability or generalizability is a measure of the relative magnitude of variability in scores due to error, with the aim of achieving a desired level of measurement precision (Shavelson & Webb, 1991). In addition to being critical, these two measures of the performance have the advantage of being quantifiable.

Recently, van der Vleuten and Schuwirth (2005) have expanded on these factors for purposes of assessment in medical education. To validity and reliability they added educational effect, feasibility, and acceptability. The educational effect of assessment capitalizes on students' motivation to do well and directs their study efforts in support of the curriculum. For example, if the goal of a particular educational intervention is increased knowledge then a written assessment will appropriately motivate students to study from books. Similarly, a goal of increased clinical skill is best supported by a clinical assessment that motivates students to interact with patients. Feasibility is the degree to which the assessment method selected is affordable and efficient for the testing purpose; assessments need to have reasonable costs. Acceptability is the extent to which stakeholders in the process (e.g., medical students and faculty, practicing physicians, patients) endorse the measure and the associated interpretation of scores.

The three factors that van der Vleuten and Schuwirth (2005) added are not readily quantifiable. Nonetheless, they argue correctly that the selection of an assessment device for a particular situation is a weighted combination of these factors. If any is completely unsatisfied, (e.g., the methods are not feasible or acceptable), then the method is unsuitable for purpose regardless of its validity or reliability. For the methods described below, these five factors will be highlighted where they are relevant to the discussion.

2.1. Simulation

Simulations are increasingly being used in medical education to ensure that examinees can demonstrate integration of prerequisite knowledge, skills, and affect in a realistic setting (Tekian, 1999). This reflects the subject matter experts' concerns that traditional and selected-response item formats measure knowledge, but not clinical performance. In medical education, licensure, and certification, both standardized patient simulations (e.g., Reznick, Blackmore, Dauphinee, Rothman, & Smee, 1996; Whelan, 1999) and computer-based simulations (Clyman, Melnick, & Clauser, 1999; Kneebone, 2003) are widely used to assess examinees' clinical skills and medical problem-solving. For the purpose of this review, simulations will be classified into two broad categories: standardized patients and computer-based simulations. Examples of assessment methods used in each of these categories will be presented.

2.1.1. Standardized patients

Standardized patient examinations are used to assess the ability of physicians and medical students to collect medical history and physical examination data and create a therapeutic relationship with the patient (Tamblyn & Barrows, 1999). A standardized patient is a person trained to accurately and consistently portray a patient with a particular medical condition. Based on an encounter between the standardized patient and a student, both the standardized patient and medical professionals can make judgments about the quality of the performance along a number of dimensions (e.g., history-taking, physical examination, interpersonal, and communication skills).

Scores for assessments composed of standardized patients are typically generated by applying criteria developed by subject matter experts (e.g., did the candidate listen to the heart or ask about a history of smoking) and/or collecting ratings of performance. Specific scoring criteria (or checklists) are developed for each patient scenario, and are generally focused on the examinee's ability to collect the relevant data from the patient and perform the necessary physical examination maneuvers (Tamblyn & Barrows, 1999).

Skills in summarizing and interpreting the information collected in the encounter with the standardized patient are often measured using post-encounter exercises consisting of open-ended questions or short-answers. For example, diagnostic and patient management skills are measured through the use of post-encounter notes, where the examinees may respond to open-ended questions about their findings and plans for treatment (Tamblyn & Barrows, 1999). Because of the considerable expertise needed to determine whether the correct information was utilized by the student and the listed diagnoses were probable, it is customary to have physicians or other knowledgeable health professionals provide the evaluation of these exercises (Tamblyn & Barrows, 1999). Usually, scores across these exercises are averaged to compute examination component (i.e., test-level) scores.

Doctor–patient communication skills are often assessed through the use of standardized scoring rubrics that are common to all cases presented (Boulet et al., 1998). Likewise, standardized patients are heavily used to assess interpersonal skills and two approaches are commonly employed. One approach is to evaluate examinees' interpersonal skills in every patient encounter. Another approach is to design patient encounters that focus on the assessment of interpersonal skills. In either case, the scoring criteria for such exercises should result in high correlations amongst the elements assessed.

One example of a standardized patient examination is the Educational Commission for Foreign Medical Graduates (ECFMG®) Clinical Skills Assessment (CSA). Until April 2004, ECFMG administered a ten-encounter CSA to graduates of medical schools outside the United States and Canada. The purpose of this assessment was to evaluate the clinical skills of graduates of international medical schools as part of a process that determines their readiness to enter postgraduate training in the US.

Research conducted to demonstrate the psychometric adequacy of scores from this assessment showed that with standardized test procedures and an adequate number of cases, reasonable levels of reliability and validity can be achieved. Depending on which aspect of competence is being assessed (e.g., the history and physical examination or interpersonal skills), the generalizability coefficients obtained met psychometric standards (Boulet et al., 1998; Boulet, Rebbecchi, Denton, McKinley, & Whelan, 2004; Boulet, van Zanten, McKinley, & Gary, 2001). Scores from CSA were shown to correlate with other markers of performance (e.g., Boulet, McKinley, Norcini, & Whelan, 2002). Moreover, correlations with written examinations were modest, adding empirical support to the notion that standardized patients assess aspects of competence not addressed directly by the traditional written measures.

One challenge associated with standardized patient examinations is the drift in rater stringency over time (McKinley & Boulet, 2004). Although acceptability is high, feasibility is an issue. Standardized patient examinations are expensive to develop and maintain. Regular review and update of case materials and changes in disease prevalence is necessary. Quality assurance procedures (Boulet et al., 2002) must be implemented to assure that scores and decisions produced are appropriate.

While standardized patient examinations have been successfully implemented in medical education, certification, and licensure, they cannot be used to assess all aspects of competence. For example, there is a need to both educate and assess the ability of doctors to perform procedures and manage life-threatening clinical situations. For these aspects of competence, the use of computer-based simulations is more appropriate. In the next section, we review some of the research conducted with these simulators.

2.1.2. Computer-based simulation

With the rapid development of technology over the past 50 years has come its application to assessment. As in other areas of education, the computer was first used to support the administration of large scale written examinations by facilitating scoring and reporting. In the late 1960s, however, the development of new psychometric models, such as Item Response Theory, allowed the intelligence of the computer to be put to use in selecting test items for individual examinees. This permitted tailored testing and/or other forms of assessment that maximized the precision of the assessment (Green, 1983; Hambleton & Swaminathan, 1985).

In medical education, this exploitation of the computer for psychometrics has been paralleled by its use in recreating, with high fidelity, various patient conditions and procedures. As in aviation, simulation enhances the training of doctors by allowing them to learn in safety and experience a variety of conditions that are relatively infrequent in practice. Moreover, it allows for assessment in a more realistic environment, enhancing the generalizability of scores.

There are hundreds of different simulators available and a review of the entire field is beyond the scope of this paper. However, three different types of simulators that employ computers will be described: computer-based programs, model-driven simulators, and virtual reality simulators.

2.1.3. Computer programs

There is considerable variability in this category of assessments but they are often interactive programs run on an ordinary computer that simulate some aspect of the doctor–patient interaction. The computer-based case simulation (CCS) of the National Board of Medical Examiners is an example of this type of method. Examinees are presented with a textual description of a patient (e.g., "A 42-year-old woman comes to your office complaining of lower back pain"). As simulated time advances, the examinee can manage the 'patient' in any number of ways including asking questions, ordering tests, and reviewing notes from other members of the health care team. The patient reacts as appropriate given the underlying medical condition and the actions of the examinee. Scores are generated based on an algorithm that compares their management strategies with those of expert clinicians (Dillon, Boulet, Hawkins, & Swanson, 2004).

Extensive research was conducted to determine whether this simulation would be feasible in the setting of a high stake assessment of a large number of examinees (i.e., whether to grant a license to practice medicine in the US). To develop scoring algorithms, expert clinicians reviewed and rated the actions of a small number of examinees. Those ratings were then used to derive case-based regression formulas that were applied to the large group of test-takers. Dillon, Clyman, Clauser, and Margolis (2002) showed that it was acceptable to use regression formulas to derive examinee scores for each of the cases.

In terms of validity, the CCS correlated modestly with the multiple-choice component of the examination (Dillon et al., 2002). The magnitude of the correlations indicated that while performance on the simulation was related to performance on the multiple-choice questions, only about half of the variation in scores could be considered common to both measures. This finding supported the use of the CCS as another indicator of clinical competence in one of the US medical licensure examinations.

Management of a patient simulated by the computer has the advantage of assessing problem-solving skills and providing examinees clinical challenges that cannot be posed by standardized patients. Further, the 'patients' can be more acutely ill, examinees can take a much wider range of actions, the settings of care (e.g., hospital, surgery) can be broader, and the patients' problems can unfold over a much longer period of time. These methods are also relatively inexpensive and this enhances their feasibility. However, the setting is not realistic and it is unclear exactly what these forms of assessment might contribute over and above traditional written measures.

2.1.4. Model-driven simulations

Advances in computer and materials technology has resulted in the development of physical simulators that model the human body with very high fidelity. Assessments based on these devices offer the opportunity for very realistic patient presentations and a focus on invasive procedures and clinical skills in acute situations.

An example of this type of device is the MedSim–EagleSim simulator, which is a computer-controlled mannequin. In one study of its measurement properties (Murray et al., 2002), examinees were asked to conduct a 5-min encounter with the 'patient' who had an acute and life-threatening condition. They were given assistance from a nurse and they were able to order suitable laboratory tests and begin therapy. A total of 64 medical students and postgraduate trainees participated in the study. Four expert raters who provided both holistic scores and completed case-specific checklists that had been developed by a panel of experts scored the performance. The checklists and holistic scores were moderately correlated, and rater consistency was higher for the checklist-based scores than for the holistic scores, although either score type would meet psychometric standards. Because raters were consistent in their scoring, results suggested that fewer raters could be used without increasing measurement error.

In another study using the same simulator, Boulet, McKinley, Whelan, and Hambleton (2003) found that moderately reliable six-case forms could be produced. Postgraduate trainees obtained higher scores than students, providing evidence of construct validity, since postgraduate trainees have more clinical experience. In addition, the use of specific guidelines for development and scoring resulted in high levels of agreement among the raters. Finally, their analyses indicated that increasing the number of scenarios, not the number of raters, could enhance reliability of scores.

This example demonstrates the use of technology to assess skills in critical care and management in a surgical setting. This form of simulation is very realistic and provides an excellent assessment of skills that are difficult to obtain in any other fashion. However, feasibility is an issue. The simulators themselves are very expensive, they require considerable space and staff support, and the development of cases and scoring requires significant expert input. Other types of simulators have been developed that provide a simulated environment where eye–hand coordination and psychomotor abilities can be assessed and research conducted with this type of simulation is discussed in the next section.

2.1.5. *Virtual reality/haptic devices*

These simulators are on the cutting edge of technology and they combine powerful imagery with other forms of sensory feedback. For example, they can permit the manipulation of three-dimensional organs or body systems as if they were real. Similar to those used in aviation, these simulators can also provide the user with a real time sense of touch. Because of the sophistication of the technology they require, these simulators often focus on a few related procedures (e.g., endoscopic procedures). They offer practice and assessment without risk to patients that replicate reality with high fidelity (Issenberg et al., 1999).

An example of this type of simulation can be found in work with the minimally invasive surgical trainer (MIST-VR) and the endoscopic sinus surgery simulator (ES3). The MIST-VR used spheres, cubes, and virtual surgical instruments to simulate laparoscopic surgery. The ES3 was developed for a specific surgical procedure, and requires navigation, accuracy, and ambidexterity. In an extensive study, Fried et al. (2004) developed a curriculum for the ES3, identified errors that could be classified as cognitive or technical, developed evaluation criteria, and studied three groups (practicing doctors, postgraduates, and medical students). Study participants were required to complete (1) six tasks on the MIST-VR to measure psychomotor ability, including grasping tissue and cauterizing three targets, (2) one novice level task on the ES3, (3) a test of pictorial surface orientation (PicSOr), and (4) measures of visual–spatial ability (card rotation, cube comparison, and map planning tests) developed by the Educational Testing Service.

As expected, practicing physicians performed best on the ES3 task, followed by postgraduates and medical students. The positive relationship between experience and performance provided some evidence of validity. To determine whether the scores derived from the ES3 task related in the expected fashion to other measures, scores were correlated with scores from the MIST-VR, PicSOr, and the visual-spatial ability measures. The total MIST-VR score and ES3 task score were highly correlated. Regression analyses showed the PicSOr perceptual task and visual–spatial scores were statistically significant predictors of performance on the ES3 task as well.

These forms of simulation are just being introduced so there is not yet a large body of research. Like the model-driven simulations, they are very realistic and provide an excellent assessment of skills that are difficult to obtain any other way. Moreover, as a group they tend to be somewhat less expensive than the model-driven simulators and to require less infrastructure for support. However, their use in assessment still requires an intensive content development effort and, since they are often focused on a single skill or procedure, several different simulators would be needed to cover even the narrower medical disciplines.

2.2. *Issues*

Simulations have been implemented in various testing programs based on the belief that fidelity is as important a feature of assessment as reliability and that the realistic challenges they pose enhance the validity of the examination and increase their acceptability to examinees and content experts. Research has shown that it is feasible to develop assessments that produce scores that meet reasonable psychometric criteria for validity and reliability. However, the use of simulations does raise a number of special issues that influence the validity, reliability, and feasibility of the methods (Norcini, 1999; Norcini & Boulet, 2003).

2.2.1. Fidelity

Simulation offers the opportunity to faithfully present the tasks or problems a doctor faces in practice. In response, there is a strong temptation to generate very elaborate recreations of individual tasks, each of which takes considerable time. However, much of what doctors do in responding to a problem is routine or repetitive and simulations faithfully reproduce this redundancy. Consequently, development of scoring criteria for each task requires a substantial amount of time. Given practical constraints, the result is a long test containing only a few problems and this narrow sampling limits the degree to which scores generalize to the domain of interest.

To address this issue, test developers often shorten the tasks and focus them around the testing point or critical incident. The work done in this area to date has been encouraging (Rothman, Cohen, & Bilan, 1996; Shatzer, DaRosa, Colliver, & Barkmeier, 1993). Ultimately, it will be important to balance fidelity and breadth of sampling since this affects all five factors in the selection of a testing method.

2.2.2. Equivalence

Most assessments are given over time, and sometimes place, so different versions of the same examination are needed. If they vary in difficulty it adversely affects the validity of scores. In multiple choice question (MCQ) examinations, this issue is addressed by ensuring that the content of the different versions is similar, their difficulty is comparable, and some form of equating is applied (Shea & Norcini, 1995). Simulations pose special challenges in this regard because a test is composed of relatively few tasks. This makes it hard to assure comparable content over different versions of the examination and to apply many of the statistical techniques used to equate scores. Equating has only rarely been done with simulations and the applications have been limited to the simpler equating methods (Norcini & Boulet, 2003; Swanson, Clauser, & Case, 1999). Considerable research in this area is needed.

2.2.3. Standardization

It is not uncommon in the assessments based on SPs to have more than one actor/actress portray the same case. Moreover, even when a case is played by the same actor/actress, over a long period of time, "performance drift" sometimes occurs (McKinley & Boulet, 2004). These issues adversely affect the validity of the assessment. Consequently, it is important to recruit standardized patients well, simplify their portrayal and scoring rubrics, train them extensively, and develop an ongoing quality assurance program that includes observing SPs, double scoring encounters, and thorough investigations of discrepancies (Boulet et al., 2002).

2.2.4. Reliability

As with most performance-based methods of assessment, the scores from simulations are affected by measurement error from a number of different sources. Consequently, scores on assessments composed of simulations tend to be less reliable per unit of testing time than other traditional formats like MCQs (Clauser, Margolis, & Swanson, 2002). These sources of error have been investigated in several studies and task variability is typically the major contributor (Boulet, McKinley et al., 2003; Boulet, Murray et al., 2003; Elstein, Shulman, & Sprafka, 1978; Norcini, 1999; Norcini & Boulet, 2003). The performance of doctors is patient or case specific, so to get a reliable estimate of ability it is necessary to sample broadly. Unfortunately, there is no other way to reduce this source of error other than including a number of different tasks in an assessment based on simulations. This clearly has implications for the feasibility of the methods, since this increase in the number of tasks would affect cost.

2.2.5. Case generation

The cases used as part of a simulation are much more difficult and expensive to generate than MCQs and this has implications for feasibility. The selection of topics is harder because each case must have identifiable correct and incorrect responses, be appropriately difficult and discriminating, and be important enough to justify the large amount of testing time devoted to it. In addition, the actual scripting of the case is harder because there is typically more than one acceptable way to carry out the task and the simulation must respond appropriately to both plausible and implausible actions. Finally, it is very time consuming to obtain knowledgeable reviews of the cases because the experts must find their way, or be led, through all of the possible response paths.

The primary response to these issues is to develop good tools to help authors write new cases. The tools should be computerized and guide the author through the writing exercise. Default information should be supplied by the software and it should identify internal conflicts and inconsistencies. Extensive tryout of the cases with doctors of various ability levels is essential. It would also be useful to extend the pool of material by disguising (changing the non-essential aspects of cases so they will not be recognized as similar) and modeling cases (systematic and substantive changes in content to generate a family of cases) (LaDuca, Templeton, Holzman, & Staples, 1986; Shea et al., 1992).

2.2.6. *Security*

For assessments where the stakes are high, security is a critical issue since prior knowledge of the specific content or correct course of action adversely affects the validity of the scores (Swanson, Norman, & Linn, 1995). Simulations may be less affected than MCQs by certain types of security breaches because no one examinee can become familiar with all of the pathways through a case. Moreover, it is more difficult for examinees to look up or fake the correct responses. However, memorization and sharing of test material remains a threat since the administrations of some simulations (like an SP examination) require rolling administration so examinees have considerable opportunity to share test content. This problem is exacerbated by the fact that an assessment built around simulations will typically be composed of only 8–12 cases or tasks. Even if the cases are more elaborate than MCQs, the fact that there are so few will increase the ease with which they are remembered. A large pool of test material is the primary defense against these security problems but that has implications for feasibility.

While simulations have the advantage of measuring skills that are different from those measured by MCQs, they can be costly to maintain and score. In addition, there is no guarantee that performance in the simulated environment will generalize to actual practice settings. In the next section of the paper, we will present methods of assessment that are used to assess performance based on management of real patients.

2.3. *Work-based assessment*

Unlike traditional didactic education, the training of doctors often occurs in the setting of patient care. This variation on the apprenticeship model offers a series of challenges and opportunities for assessment. Of most importance is the fact that the type and complexity of patient care problems that doctors face during training are the same as those encountered in practice. Trainees confront a broad array of health-care problems and just like doctors in practice they are required to integrate all of their skills in response. Simulation provides a means of beginning to assess these skills, but real patients often have more complex problems, are more acutely ill, and demand more skill than can be simulated through current technology. Consequently, assessment that occurs in this setting should more easily generalize to future performance.

At the same time, the educational mission dictates that the methods chosen for assessment protect the safety of patients and provide the opportunity for educational feedback to the trainee. To address these issues, several methods of assessment have been developed, many of which are variations on the traditional bedside viva. There is a stimulus with which the trainee interacts and this might be a patient or the patient's medical record. The encounter is observed, typically by a faculty member or a peer, who makes global judgments about certain dimensions (e.g., cognitive, interpersonal, technical) of the performance. In a sense, these methods are analogous to classroom testing for doctors.

One of the best examples of work-based assessment can be found in the Foundation Programme of the National Health Service (Beard et al., 2005). This is a 2-year educational program that forms the link between medical school and specialist training. It is composed of a series of placements in different specialties and settings. Learning objectives for each stage of training have been specified as have been the competencies expected by the end of training including clinical skills (e.g., the ability to take a history from a patient or to do a physical examination), communication skills, and teamwork. The methods of assessment used as part of this program include (1) case-based discussion, (2) the mini-clinical evaluation exercise, (3) the direct observation of procedural skills, and (4) the mini peer assessment tool.

2.4. *Methods*

2.4.1. *Case-based discussion (CbD)*

In CbD, the trainee picks two written patient records in which they have made entries and an assessor selects one for the session. A discussion between the assessor and trainee ensues, centered on this written record and designed to assess clinical decision-making and the application of medical knowledge. After the discussion, the assessor judges the quality of the performance and then provides constructive feedback. The trainee selects the timing, the records, and the assessor. It is expected that they be assessed six times during the year and that the patient records they choose be appropriately sampled from a list of patient problems.

CbD was originally called Chart Stimulated Recall and there is empirical research supporting its acceptability, reliability, and validity. A study by Maatsch, Huang, Downing, and Barker (1983) at the American Board of Emergency Medicine looked at the method as part of a recertifying examination for practicing doctors. They found reasonably reliable results with five to eight charts and assessors and score distributions and pass–fail results consistent with certification. Moreover, CbD was correlated with other measures of performance (e.g., chart audit) and it was considered the most valid of the various methods used by practicing doctors.

A later study by Norman et al. (1989) compared CbD scores for "referred" doctors (those in some type of difficulty) with a group of volunteers. CbD was able to distinguish these two groups and the scores were highly correlated with other methods of assessment such as an oral examination or an assessment composed of standardized patients. Similar results were found in a study of practicing doctors by Soloman, Reinhart, Bridgham, Munger, and Starnaman (1990). Scores on CbD again had a reasonable relationship with another measure of competence (structured viva) and with performance 10 years prior to both written and oral examinations.

2.4.2. Mini-clinical evaluation exercise (mCEX)

In the mCEX, a faculty member watches a trainee–patient encounter in any healthcare setting. The encounters are intended to be relatively short, about 15 min, and the trainee is expected to conduct a focused history and/or physical examination during this time. Afterwards, he or she provides the assessor with a diagnosis and treatment plan, the performance is scored using a structured form, and then educational feedback is provided. Trainees are expected to undertake six encounters during the year, with a different assessor for each encounter. Each of these encounters should represent a different clinical problem, appropriately sampled from the list of patient problems.

There have been a number of studies of the mCEX. In routine use, this method was found to produce reasonable confidence intervals with 4–8 encounters/assessors (Norcini, Blank, Duffy, & Fortna, 2003). The method applied well to a range of patient problems, settings, and types of visits. Ratings increased throughout the year of training and examiners were very satisfied with the format.

An application of mCEX in undergraduate training demonstrated that ratings had modest correlations with written exam scores, inpatient clerkship ratings, outpatient clerkship ratings, and final course grades (Kogan, Bellini, & Shea, 2003). In a postgraduate setting, Durning, Cation, Markert, and Pangaro (2002) found similar results; mCEX ratings were correlated with monthly evaluations in matched areas of competence.

Finally, Boulet et al. (2002) asked the faculty to evaluate videotapes of SP–student encounters using the mCEX rating form. The results of the SP checklists predicted faculty global ratings and their assessments of doctor–patient communication also correlated with faculty communication ratings. Similarly, Holmboe, Huot, Chung, Norcini, and Hawkins (2003) created videotapes of trainees that were unsatisfactory, satisfactory, and superior and faculty successfully discriminated among the three levels of performance.

2.4.3. Direct observation of procedural skills (DOPS)

DOPS is a variation on the mCEX in which the assessor observes the trainee while he or she is performing a procedure (e.g., giving an injection, drawing blood, inserting a tube), rates the performance, and then provides feedback. Trainees need to undertake six observed encounters during the year, each with a different assessor. The trainee chooses the timing, procedure, and assessor but the procedures need to be sampled from an approved list.

There is little research specific to DOPS but it is built on a large body of work on the global ratings of procedural skills. In a study of the Objective Structured Assessment of Technical Skills, Goff et al. (2002) had two faculty members complete checklists and rate the performance of OB/GYN trainees; ratings increased with amount of training. Similarly, Winckel, Reznick, Cohen, and Taylor

(1994) assessed performance during actual operations and found higher ratings with increased training. Finally, Grober et al. (2004) found that hands-on training in urologic microsurgery produced better global ratings of technical proficiency than didactic training. In general, these studies and previous work on the mCEX indicated that 4–8 encounters are sufficient to produce reasonable confidence in the final results.

2.4.4. Mini-peer assessment tool (mPAT)

In this method, the trainee nominates eight assessors from among those who are his/her supervisors and peers, including nurses and other health professionals to fill out a questionnaire concerning their technical and interpersonal skills. The trainees also complete a self-assessment using the same questionnaire. The assessment forms are sent directly to the assessors from a central office, rather than the trainee, to ensure confidentiality.

Feedback is collated centrally and is presented in a way that shows the self-ratings, the mean rating for the assessor, and the national mean ratings; comments are included. These data are shared with the trainee and the educational supervisor so that there can be agreement about strengths and weaknesses and a plan can be developed for improvement.

There is a body of research supporting the use of peer assessment both within higher education (Topping, 1998) and medical education (Norcini, 2003). For example, in a study of specialty certification by Ramsey et al. (1989), certified doctors had higher peer ratings and their assessments of technical skill were correlated with written exam performance. A later study by Ramsey et al. (1993) also showed that reasonable reliability was obtained with 8–12 peers and roughly five questions. Similar validity and reliability work has been done with internists and pediatricians in the UK and Canada (Archer, Norcini, & Davies, 2005; Lockyer, 2003).

2.5. Issues

Despite the many advantages of the work-based methods of assessment there are at least four challenges that these methods and this program face: standards, alternative assessments, selection of assessors, and equivalence.

2.5.1. Standards

On the basis of these methods of assessment, not many trainees will be considered unsatisfactory (Norcini, 2003). Given their long educational process and the selection criteria for entry to medicine, it is possible that virtually all trainees are indeed satisfactory. However, it is more likely that faculty in a training program are reluctant to give failing grades to their trainees; they have a conflict of interest. To address this problem, training of the assessors has proven helpful in increasing their stringency and discrimination (Holmboe, Hawkins, & Huot, 2004). Nonetheless, there remains a need for national assessment to address this problem.

2.5.2. Alternative assessments

For those few trainees who are considered unsatisfactory, another assessment process is needed. The purpose of these assessments is to rule out false negatives and to provide diagnostic feedback that would inform remediation. To accomplish these purposes, traditional measures such as knowledge tests would be useful. In addition, inclusion of an examination composed of standardized patients might be appropriate as a test of clinical and communication skills.

2.5.3. Selection of assessors

In the Foundation Programme, trainees have some control over who examines them and indirectly over the content of the assessment. At least in the case of instruments like the mPAT, this does not seem to be a biasing factor (Ramsey et al., 1993). Nonetheless, there remains the appearance, if not the reality, of unfairness. As a result, it is important to ensure that the assessments cover the spectrum of problems and that several faculty members are involved. Where the resources exist, it may also be preferable for the faculty to decide when trainees will be assessed, by whom, and with which patients.

2.5.4. Equivalence

Assessments will not be equivalent across sites and training programs because the patients will vary in difficulty and the faculty will vary in stringency. Sampling both faculty and patients broadly will reduce this effect to some degree but increasing the numbers alone will not remove bias. Consequently, the results of this type of assessment will not be useful for ranking all trainees nationally or for making decisions regarding certification or licensure. These assessments are more appropriate for identifying whether additional training or remediation is needed.

3. Summary

Since the 1950s, there has been rapid and extensive change in the way assessment is conducted in medical education. Several new methods of assessment have been developed and implemented over this time and they have focused on clinical skills (taking a history from a patient and performing a physical examination), communication skills, procedural skills, and professionalism. In this paper, we provided an overview of the assessment methods used in medical education, with a focus on recent developments in simulation and work-based strategies.

Standardized patients, computer programs, model-driven simulations, and virtual reality devices were described as representative of the work being done with simulation. As a group, these devices offer realistic clinical challenges that enable the assessment of a variety of skills that are inaccessible to traditional methods. Communication skills, the ability to take a history from a patient or perform a physical examination, and the management of patient scenarios both over time and in acute conditions are examples of the types of content to which these methods are well-adapted. The development of instructional guidelines and detailed scoring criteria are being researched, particularly for computer-based simulators. These efforts will most likely increase integration of computer-based simulators in medical education curricula. Issues of feasibility remain with many of these methods and there are underlying psychometric issues, such as how to equate scores, which need to be addressed in future research.

Among the work-based methods, the mCEX, the direct observation of procedural skills, case-based discussion, and the mPAT are described. These methods fit well into the apprenticeship-type training programs that are common in medicine, they provide realistic challenges, and they offer the opportunity for feedback that can have major educational effects. These work-based methods are the clinical analogs to classroom testing and consequently they are probably not appropriate for a high stakes setting where equivalence oversight is necessary.

Simulations and work-based assessments discussed here address medical educators' concerns about assessment of clinical performance (e.g., medical history-taking, physical examination skills, procedural skills, clinical judgment). Research conducted to date provides the basis for selection of each assessment method for specific purposes. The use of simulations in assessment will require additional research in ensuring that tasks are adequately sampled and of similar difficulty across administrations. While it would be difficult to implement similar task sampling and control over task difficulty for work-based assessments, these methods provide important information about trainees' abilities in actual practice with real patients. The work-based assessment methods presented may not be appropriate for high-stakes testing (e.g., licensure and certification) where equivalence over administration is required, but they can inform medical education and indicate where additional training or remediation is needed.

References

Archer, J. C., Norcini, J. J., & Davies, H. A. (2005). Use of SPRAT for peer review of paediatricians in training. *British Medical Journal*, *330*, 1251–1253.

Beard, J., Strachan, A., Davies, H., Patterson, F., Stark, P., Ball, S., et al. (2005). Developing an education and assessment framework for the Foundation Programme. *Medical Education*, *39*(8), 841–851.

Boulet, J. R., Ben-David, M. F., Ziv, A., Burdick, W. P., Curtis, M., Peitzman, S. J., et al. (1998). Using standardised patients to assess the interpersonal skills of physicians. *Academic Medicine*, *73*(Suppl. 10), S94–S96.

Boulet, J. R., McKinley, D. W., Norcini, J. J., & Whelan, G. P. (2002). Assessing the comparability of standardised patients and physician evaluations of clinical skills. *Advances in Health Sciences Education*, *7*, 85–87.

Boulet, J. R., McKinley, D. W., Whelan, G. P., & Hambleton, R. K. (2003). Quality assurance methods for performance-based assessments. *Advances in Health Sciences Education*, *8*, 27–47.

Boulet, J. R., Murray, D., Kras, J., Woodhouse, J., McAllister, J., & Ziv, A. (2003). Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. *Anesthesiology*, *99*, 1270–1280.

Boulet, J. R., Rebbecchi, T. R., Denton, E. C., McKinley, D. W., & Whelan, G. P. (2004). Assessing the written communication skills of medical school graduates. *Advances in Health Sciences Education*, *9*, 47–60.

Boulet, J. R., van Zanten, M., McKinley, D. W., & Gary, N. E. (2001). Evaluating the spoken English proficiency of graduates of foreign medical schools. *Medical Education*, *35*, 767–773.

Clauser, B. E., Margolis, M. J., & Swanson, D. B. (2002). An examination of the contribution of computer-based case simulations to the USMLE Step 3 examination. *Academic Medicine*, *77*(Suppl. 10), S80–S82.

Clyman, S. G., Melnick, D. E., & Clauser, B. E. (1999). Computer-based case simulations from medicine: Assessing skills in patient management. In A. Tekian, C. H. McGuire, & W. C. McGahie (Eds.), *Innovative simulations for assessing* *professional competence* (pp. 29–41). Chicago: University of Illinois, Department of Medical Education.

Dillon, G. F., Boulet, J. R., Hawkins, R. E., & Swanson, D. B. (2004). Simulations in the United States Medical Licensing Examination$^{\text{TM}}$ (USMLE$^{\text{TM}}$). *Quality and Safety in Health Care*, *13*, 41–45.

Dillon, G. F., Clyman, S. G., Clauser, B. E., & Margolis, M. J. (2002). The introduction of computer-based case simulation into the United States Medical Licensing Examination. *Academic Medicine*, *77*, S94–S96.

Durning, S. J., Cation, L. J., Markert, R. J., & Pangaro, L. N. (2002). Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. *Academic Medicine*, *77*(9), 900–904.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem-solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

Fried, M. P., Satava, R., Weghorst, S., Gallagher, A. G., Sasaki, C., Ross, D., et al. (2004). Identifying and reducing errors with surgical simulation. *Quality and Safety in Health Care*, *13*, 19–26.

Goff, B. A., Nielsen, P. E., Lentz, G. M., Chow, G. E., Chalmers, R. W., Fenner, D., et al. (2002). Surgical skills assessment: A blinded examination of obstetrics and gynecology residents. *American Journal of Obstetrics and Gynecology*, *186*(4), 613–617.

Green, B. F. (1983). Adaptive testing by computer. In R. B. Ekstrom (Ed.), *Principles of modern psychological measurement* (pp. 5–12). San Francisco, CA: Jossey-Bass.

Grober, E. D., Hamstra, S. J., Wanzel, K. R., Reznick, R. K., Matsumoto, E. D., Sidhu, R. S., et al. (2004). Laboratory based training in urological microsurgery with bench model simulators: A randomized controlled trial evaluating the durability of technical skill. *Journal of Urology*, *172*(1), 378–381.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Dordrecht: Kluwer Academic Publishers.

Holmboe, E. S., Huot, S., Chung, J., Norcini, J. J., & Hawkins, R. E. (2003). Construct validity of the mini clinical evaluation exercise (Mini-CEX). *Academic Medicine*, *78*, 826–830.

Holmboe, E. S., Hawkins, R. E., & Huot, S. J. (2004). Effects of training in direct observation of medical residents' clinical competence: A randomized trial. *Annals of Internal Medicine*, *140*(11), 874–881.

Issenberg, S. B., McGaghie, W. C., Hart, I. R., Mayer, J. W., Flener, J. M., Petrusa, E. R., et al. (1999). Simulation technology of health care professional skills training and assessment. *Journal of the American Medical Association*, *282*, 861–866.

Kneebone, R. (2003). Simulation in surgical training: Educational issues and practical implications. *Medical Education*, *37*, 267–277.

Kogan, J. R., Bellini, L. M., & Shea, J. A. (2003). Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. *Academic Medicine*, *78*(Suppl 10), S33–S35.

LaDuca, A., Templeton, B., Holzman, G. B., & Staples, W. I. (1986). Item-modeling procedure for constructing content-equivalent multiple choice questions. *Medical Education*, *20*, 53–56.

Lockyer, J. (2003). Multisource feedback in the assessment of physician competencies. *Journal of Continuing Education in the Health Professions*, *23*(1), 4–12.

Maatsch, J. L., Huang, R., Downing, S., & Barker, B. (1983). *Predictive validity of medical specialty examinations. Final report for Grant HS 02038-04, National Center of Health Services Research. East Lansing Michigan: Office of Medical Education and Research and development*. Michigan State University.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement, (3rd ed.)*. Washington DC: Oryx Press.

McKinley, D. W., & Boulet, J. R. (2004). Detecting score drift in a high-stakes performance-based assessment. *Advances in Health Sciences Education*, *9*, 29–38.

Murray, D., Boulet, J., Ziv, A., Woodhouse, J., Kras, J., & McAllister, J. (2002). An acute care skills evaluation for graduating medical students: A pilot study using clinical simulation. *Medical Education*, *36*(9), 833–841.

Norman, G. R., Davis, D., Painvin, A., Lindsay, E., Rath, D., & Ragbeer, M. (1989). Comprehensive assessment of clinical competence of family/general physicians using multiple measures. *Proceedings of the Research in Medical Education Conference*, *7*, 75–79.

Norcini, J. J. (1999). Measurement issues in the use of simulation for testing professionals: Test development, test scoring, standard setting. In A. Tekian, C. H. McGuire, & W. C. McGaghie (Eds.), *Innovative simulations for assessing professional competence*. Chicago, IL: University of Illinois.

Norcini, J. J. (2003). Peer assessment of competence. *Medical Education*, *37*, 539–543.

Norcini, J. J. (2005). Current perspectives in assessment: The assessment of performance at work. *Medical Education*, *39*, 880–889.

Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. (2003). The mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine*, *138*, 476–481.

Norcini, J. J., & Boulet, J. R. (2003). Methodological issues in the use of standardised patients for assessment. *Teaching and Learning in Medicine*, *15*(4), 293–297.

Ramsey, P. G., Carline, J. D., Inui, T. S., Larson, E. B., LoGerfo, J. P., & Wenrich, M. D. (1989). Predictive validity of certification by the American Board of Internal Medicine. *Annals of Internal Medicine*, *110*, 719–726.

Ramsey, P. G., Wenrich, M. D., Carline, J. D., Inui, T. S., Larson, E. B., & LoGerfo, J. P. (1993). Use of peer ratings to evaluate physician performance. *Journal of the American Medical Association*, *269*, 1655–1660.

Reznick, R. K., Blackmore, D., Dauphinee, W. D., Rothman, A. I., & Smee, S. (1996). Large-scale high-stakes testing with an OSCE: Report from the Medical Council of Canada. *Academic Medicine*, *71*, S19–S21.

Rothman, A. I., Cohen, R., & Bilan, S. (1996). A comparison of short- and long-case stations in a multiple-station test of clinical skills. *Academic Medicine*, *71*, s110–s112.

Shatzer, J. H., DaRosa, D., Colliver, J. A., & Barkmeier, L. (1993). Station length requirements for reliable performance-based examination scores. *Academic Medicine*, *68*, 224–229.

Shavelson, R. J., & Webb, N. M. (1991). *A Primer on generalisability theory*. Thousand Oaks: Sage Publications.

Shea, J. A., & Norcini, J. J. (1995). Equating. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 253–287). Lincoln: Buros Institute of Mental Measurements.

Shea, J. A., Poniatowski, P. A., Day, S. C., Langdon, L. O., LaDuca, A., & Norcini, J. J. (1992). An adaptation of item modeling for developing test-item banks. *International Journal of Teaching and Learning in Medicine*, *4*, 19–24.

Solomon, D. J., Reinhart, M. A., Bridgham, R. G., Munger, B. S., & Starnaman, S. (1990). An assessment of an oral examination format for evaluating clinical competence in emergency medicine. *Academic Medicine*, *65*(Suppl. 9), S43–S44.

Swanson, D. B., Clauser, B. E., & Case, S. M. (1999). Clinical skills assessment with standardised patients in high-stakes tests: A framework for thinking about score precision, equating, and security. *Advances in Health Sciences Education*, *4*, 67–106.

Swanson, D. B., Norman, G. R., & Linn, R. L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher*, *24*, 5–11.

Tamblyn, R., & Barrows, H. (1999). Data collection and interpersonal skills: The standardised patient encounter. In A. Tekian, C. H. McGuire, & W. C. McGaghie (Eds.), *Innovative simulations for assessing clinical competence* (pp. 77–104). Chicago: Department of Medical Education, University of Illinois at Chicago.

Tekian, A. (1999). Assessing communication, technical, and affective responses: Can they relate like a professional? In A. Tekian, C. H. McGuire, & W. C. McGaghie (Eds.), *Innovative simulations for assessing clinical competence* (pp. 105–112). Chicago: Department of Medical Education, University of Illinois at Chicago.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68*, 249–276.

van der Vleuten, C. P. M., & Schuwirth, L. W. (2005). Assessing professional competence: From methods to programmes. *Medical Education*, *39*(3), 309–317.

Whelan, G. P. (1999). Educational commission for foreign medical graduates: Clinical skills assessment prototype. *Medical Teacher*, *21*, 156–160.

Winckel, C. P., Reznick, R. K., Cohen, R., & Taylor, B. (1994). Reliability and construct validity of a structured technical skills assessment form. *American Journal of Surgery*, *167*, 423–427.