Medical Teacher, Vol. 27, No. 3, 2005, pp. 200–206

Taylor & Francis
Taylor & Francis Group

MIRIAM FRIEDMAN BEN-DAVID MEMORIAL ISSUE

Scoring standardized patient examinations: lessons learned from the development and administration of the ECFMG Clinical Skills Assessment (CSA®)

GERALD P. WHELAN[1], JOHN R. BOULET[1,2], DANETTE W. McKINLEY[1,2], JOHN J. NORCINI[2], MARTA VAN ZANTEN[1,2], RONALD K. HAMBLETON[3], WILLIAM P. BURDICK[1,2] & STEVEN J. PEITZMAN[1]

[1]Educational Commission for Foreign Medical Graduates, Philadelphia, USA; [2]Foundation for Advancement of International Medical Education and Research, Philadelphia, USA; [3]University of Massachusetts, USA

Miriam Friedman Ben-David was instrumental in developing and implementing the Educational Commission for Foreign Medical Graduates' Clinical Skills Assessment (ECFMG CSA); to date, one of the largest and most comprehensive high-stakes performance assessments of clinical skills to ever be undertaken. Her efforts in this area paved the way for the assessment of the clinical skills of all IMGs seeking graduate medical education training in the US. More broadly, she helped to develop and improve the methodology that was needed to take OSCEs and standardized patient assessments from the realm of student training and formative evaluation into the arena of high-stakes, high-volume, medical certification and licensure testing.

Dr. Friedman Ben-David not only helped assemble a team of clinicians, educators, psychometricians, and linguists, but also guided the design of numerous pilot studies, each bringing the final design of CSA closer to fruition. Her experience, insight and wealth of knowledge allowed her to expeditiously develop a high quality assessment while still applying appropriate rigor to the evaluation of all the measurement components. She articulated the need to keep the design for such a high-stakes, high-volume, assessment as simple as possible; paradoxically, those who worked with her in developing and implementing this necessarily complex instrument best remember her mantra, ''It's not so simple!'' Her contributions to performance assessment in clinical medicine will be the foundation of continued advances for generations of educators to come, and improve the quality of medical care for millions of persons around the world. Her leadership, both as a scientist and as an educator, has inspired us all.

SUMMARY *Throughout the 40 year history of standardized patient assessments and OSCEs, there have been numerous advancements, including many that involve scoring the simulated clinical encounters. While there is no clear agreement on how examinees' performance should be documented or scored in an encounter, there is a consensus that several well-chosen SP encounters are required to produce reliable examinee scores. There also continues to be some debate as to who should do the scoring on an SP-based assessment. While logistics and cost will certainly play a role, it is probably best to use the person who is most familiar with the domain being assessed. In some instances this will be the SP; in others, an outside observer or content expert. Finally, with the growing use of OSCEs for summative purposes (e.g. certification, licensure), special attention must be paid to fairness issues. Since the same test form cannot be used day after day, examinee scores must be 'equated', taking into account the psychometric properties of scores from individual cases and individual SPs. To date, the CSA has been one of the highest-volume, high-stakes, standardized patient assessments to be developed and successfully administered. In 2003 alone, over 11 500 IMGs were tested. The early conceptual framework for this assessment was synthesized from the research endeavours of several notable individuals, including,*

Correspondence: Gerald P. Whelan, MD, Educational Commission for Foreign Medical Graduates, Clinical Skills Assessment Program, 3624 Market Street, 3rd Floor, Philadelphia, PA 19104-2685, USA. Email: gwhelan@ecfmg.org

ISSN 0142–159X print/ISSN 1466–187X online/00/000200-206 © 2005 Taylor & Francis Group Ltd
DOI: 10.1080/01421590500126296

amongst many others, Harden et al. 1975, Swanson & Stillman, 1990, Newble & Swanson, 1988, Vu et al. 1992 and Colliver, 1995. The early prototype administrations of the CSA, including many operational research studies, were supported and guided by Dr Friedman Ben-David, Friedman et al. 1991, 1993, Stillman et al. 1992, and Sutnick et al. 1993, 1995.

Introduction

In the late 1980s and early 1990s, medical educators in the US and around the world began to express concerns that even though there had been incredible technological advances in medicine, some of the most fundamental doctoring skills were being lost among graduating medical students and residents. This uneasiness coincided with a growing negative public sentiment and was backed by an increase in the number of patients who were dissatisfied with the quality of their interactions with their physicians. As a result, the Liaison Committee for Medical Education (LCME), the entity responsible for accrediting programs of medical education in the US and Canada, introduced, and continues to update, requirements for teaching and assessing basic clinical skills in undergraduate medical curricula (Liaison Committee on Medical Education, 2003).

The Educational Commission for Foreign Medical Graduates (ECFMG) is responsible for certifying the readiness of graduates of international medical schools to enter accredited graduate medical education programs in the US. The certification process includes verification of credentials and a series of examinations (Educational Commission for Foreign Medical Graduates, 2004). In 1992, when steps 1 and 2 of the United States Medical Licensing Examination (USMLE™) were introduced (Melnick *et al.*, 2002), ECFMG adopted them as the examinations it would use for certification, thus making the basic science and clinical science examination requirements for IMGs identical to those for US graduates. Unfortunately, with regard to the concern regarding clinical skills, the actions of the LCME would be unlikely to have any impact on medical schools outside the US and Canada; nor was there any single international agency that could require that clinical skills be incorporated into curricula. Absent of any assurance that clinical skills were in fact being taught in international medical schools or, if instruction was provided, how well the students were performing, the obvious alternative was to resort to an outcome assessment. Hence the concept of what would become the Clinical Skills Assessment (CSA®) emerged.

In the years prior to the conceptualization of CSA, a great deal of work had already been done to validate Objective Structured Clinical Examinations (OSCEs) and standardized patient assessments as effective tools for teaching and evaluating the basic clinical skills of medical students and other health practitioners (Robb & Rothman, 1985; Newble & Swanson, 1988; Stillman *et al.*, 1992; Vu & Barrows, 1994; Vu *et al.*, 1994). However, many of these studies were conducted at medical schools with relatively small numbers of participants, under conditions involving questionable motivation of test takers, and, often, incorporating non-standard administration designs. As a result, the generalizabilty of the findings to other examinations, especially those with high volumes and whose purpose was certification or licensure, could be questioned. In addition, there was no clear evidence that any single scoring framework was psychometrically superior or, based on logistical constraints, more efficient for deriving examinee ability estimates (Norcini *et al.*, 1993; Finkbiner *et al.*, 1994). As a result, since 1998, there have been numerous additional studies conducted to more clearly establish the psychometric adequacy of scores from OSCEs and SP evaluations (Boulet *et al.*, 2003a; Norcini & Boulet, 2003), and to firmly establish their role in training medical students (Barzansky & Etzel, 2004).

For IMGs, ECFMG certification is a *sine qua non* for entry into the American graduate medical education (GME) system as well as for licensure in any state in the US. Hence, as a compulsory component of ECFMG certification, CSA became a 'high-stakes' examination; failing candidates could not be awarded an ECFMG certificate, could therefore not enter GME and, without this graduate training, could not obtain an unrestricted license to practice medicine. In addition, since thousands of IMGs engage in the certification process each year, the CSA, starting in 1998, became the highest volume, continuously administered, SP examination in history. Before the first candidate arrived at the test center, the challenge facing ECFMG was to design and implement an approach to clinical skills assessment that would be logistically feasible and psychometrically defensible and still yield fair, consistent and meaningful scores.

Purpose

The purpose of this paper is to outline the processes used to produce scores for candidates taking the CSA. Since many of the decisions were based on the currently available literature, we will also provide a synopsis of some of the major issues concerning the generation of scores for a high-stakes performance-based assessment: (1) test length (number of encounters); (2) checklists; (3) rating scales; (4) post-encounter exercises; (5) case construction and equating. This is preceded by a short introduction to the structure of clinical skills assessments, including the development of CSA.

Structure of clinical skills assessments/OSCEs

Since their initial development over 30 years ago (Harden *et al.*, 1975; Harden & Gleeson, 1979), OSCEs and clinical skills assessments have taken many forms. Generally, individuals who participated in these assessments and evaluations completed a series of clinical tasks, usually over fixed time intervals. Initially, these types of assessments were used primarily for formative purposes; more recently there has been a marked growth in the development and administration of clinical skills examinations for certification and licensure purposes (ECFMG, 1999; Tombleson *et al.*, 2000; Medical Council of Canada, 2002; Federation of State Medical Boards & National Board of Medical Examiners, 2003). From 1998 to 2004, the CSA was administered as part of the certification process for international medical graduates (IMGs) who wished to enter accredited residency programs

in the US. Over the six-year period, there were 43 642 CSA administrations, including 37 930 first-time takers. In 2004, the USMLE™ Step 2 Clinical Skills (CS) examination was instituted for all US medical students. International medical graduates are now required to take and pass this examination, similar in content and structure to the CSA, as a requirement for entry into an accredited residency program in the US.

Development of CSA

The initial developmental work for the CSA began in the late 1980s and carried forward through to 1998. Various pilot studied were completed, both within the US and internationally (Stillman *et al.*, 1992; Sutnick *et al.* 1993; Sutnick & Wilson, 1994; Sutnick *et al.*, 1995; Ben David *et al.*, 1997; Boulet *et al.*, 1998a; Ziv *et al.*, 1998). These studies were conducted to test the logistics of CSA administration, gather evidence to support the validity of the assessment, and develop appropriate and meaningful scoring algorithms.

Although early pilot administrations of the CSA prototype took several forms, including various test lengths and differing measurement tools, the assessment, at least in general, had several key components related to scoring. These key components will be specifically discussed at length later in the paper. First, the assessment was composed of several clinical encounters with standardized patients, where every examinee (participant, candidate) rotated through every encounter. Multiple encounters were used to ensure that reasonably accurate estimates of an examinee's clinical skills could be obtained. Second, for each encounter there was a content-specific checklist that the standardized patient was required to complete. Here, the literature suggested that an objective evaluation could be obtained by documenting what an examinee asked and, given the presenting patient complaint(s), which physical examination maneuvers were chosen to be performed, and whether these were done correctly (Cunnington *et al.*, 1997; Pangaro *et al.*, 1997). The checklists were constructed by expert committees, and comprised relevant history-taking questions and physical examination maneuvers. Third, in addition to completing the case-specific checklist, SPs were also required to evaluate examinees on their interpersonal skills and spoken English proficiency. The communication scores were based on rating scales, with score categories ranging from poor to excellent. Fourth, following each clinical encounter, the examinee was required to provide a written summary of the clinical encounter in the form of a patient note. Initially, all examinee notes were handwritten; in 2003, examinees were offered the option of typing some or all of their notes. This written (or typed) exercise was included to ensure that examinees could summarize and synthesize the information gathered in the clinical encounter. Finally, to provide overall examinee component scores, the encounter scores were averaged over the number of stations completed. As part of this process, the examinee scores were 'equated'. It was recognized that certain encounters were more difficult than others, and select SPs were more stringent than others. Without adjustment, depending on particular cases on the test form, an examinee could potentially be advantaged, or disadvantaged, by testing on one day as opposed to the next.

Key scoring issues

Test length (how many scored encounters do you need?)

One of the primary scoring decisions relates to the specific length of the assessment. To obtain a reasonably reliable estimate of an examinee's performance, multiple clinical encounters are required. Although this issue has been studied extensively in the literature (van der Vleuten *et al.*, 1991; Gimpel *et al.*, 2003; Govaerts *et al.*, 2002), there is only limited guidance on how, given the varied purposes of different assessments, to determine the appropriate test length. Furthermore, if the reliability standards required for typical high-stakes multiple-choice examinations are applied to these types of performance assessments, there are few, if any, that will make the grade. However, it is still possible to develop a reasonably reliable assessment of clinical skills as long as there are sufficient numbers of encounters, the encounters are not overly content-specific, the scoring systems are adequate, and potential non-random sources of measurement error (e.g. overly lenient or strict scorers) are adequately controlled. It should also be noted that, while there may be concerns regarding the reliability of scores from OSCEs and clinical skills assessments, especially where they are used for high-stakes decisions, the varied clinical content, often associated with individual performance fluctuations, may actually result in a more valid assessment (Norman *et al.*, 1991; Norcini & Boulet, 2003).

Based on our early research and the prevailing literature, we decided that the test length for CSA would be 10 scored encounters. Depending on the component of interest, this resulted in generalizability coefficients ($\rho_{xx'}$) ranging from about 0.70 to 0.90 (Boulet *et al.*, 1998a; Boulet *et al.*, 2003a; Boulet *et al.*, 2004). However, for some components (i.e. physical examination) the reliability of the measure was not adequate to warrant deriving a separable, independent, component score. Therefore, all checklist items for a given encounter, including both physical examination and history taking, were added together, and subsequently averaged over encounters. The combination of history-taking and physical examination was called data gathering (DG). Other measured components (e.g. interpersonal skills (IPS), spoken English proficiency [ENG]), although based on 'subjective' SP ratings (described later), were sufficiently reliable ($\rho_{xx'} > 0.80$) over a 10-encounter assessment to produce independent component scores.

Checklists (how can we document the examinee's performance?)

As mentioned previously, each case (encounter) had a content-specific checklist. These checklists ranged in length from 14 to 31 items. Approximately one-third of all items were related to the physical examination. The choice of using checklists for documenting examinees was based on a number of factors, including psychometrics, logistics and cost. In terms of psychometrics, there have been numerous studies supporting the use of case-specific checklists (Cunnington *et al.*, 1997; Gorter *et al.*, 2000; Boulet *et al.*, 2002). Furthermore, although some organizations use physician examiners (Medical Council of Canada, 2002), thereby enhancing the acceptability of the assessment, they are difficult to recruit and costly to train. Moreover, from a relative scoring perspective, the holistic judgments of

physician examiners have been shown to be similar to the aggregate checklist scores and communication ratings provided from trained SPs (Boulet *et al.*, 2002). From a logistical and cost perspective, the expected examinee volume for the CSA (approximately 5000–10 000/year) made it impossible to entertain using physician examiners. Nevertheless, while the use of SP-scored checklists could potentially undermine the acceptability of the assessment, at least from the examinee's perspective, physicians were intimately involved in constructing the scenarios, determining the checklist content, and rating the quality of the post-encounter exercise (discussed later).

Often, when constructing the checklist, there was considerable debate concerning what questions the physician should ask and which, if any, focused physical examination maneuvers should be performed. However, through a Delphi process, a draft checklist was secured for pre-testing. Initially, some of the test development committee members lobbied for weighting certain checklist items. While this strategy may have lessened the debate regarding what should be on the checklist, it is unclear whether more valid examinee scores could be obtained (Lyons & Payne, 1975; Streiner *et al.*, 1993; Allen & Locker, 1997). Furthermore, since physician questions and examination maneuvers are interdependent, and with the very large number of checklist items, it is unlikely that weighting certain items, unless extreme weights are used, will lead to a different rank-ordering of examinee scores. More important, if one considers the checklist to be a sampling of the key questions and physical examination maneuvers, and this sampling is reasonably broad, checklist item weights should have little overall effect when making competence decisions, especially when scores are based on aggregating performances over 10 encounters.

Rating scales (how do you best assess doctor–patient communication skills?)

Over the past 10 years, there have been increased efforts aimed at improving and assessing the abilities of physicians to communicate with patients (Albanese, 2000; Cegala & Lenzmeier, 2002). Not only has the public demanded these skills, but the prevailing research also suggests that important healthcare outcomes (e.g. patient compliance with preventive treatments) can be achieved by improving physician skills with respect to rapport, empathy and listening (Ong *et al.*, 1995; Beck *et al.*, 2002). As a result, the ECFMG decided to incorporate doctor–patient communication (COM) skills as a separable, conjunctive element of CSA; examinees who did not demonstrate adequate COM skills would fail CSA, regardless of their performance on other parts of the assessment.

Since examinees could fail CSA based only on their COM skills, it was important to have a defensible, valid assessment of COM skills. At the time that the ECFMG CSA was developed, numerous rating scales were available for assessing the broad domain of doctor–patient communication. In general, these could be classified by their scoring method, either holistic or analytic. For the analytic scales, the assessor (SP, observer, physician, etc.) simply indicated whether or not a particular element (e.g. introduced self, eye contact, closed the patient encounter) was accomplished. The sum, or weighted aggregate, of these behaviours/elements yielded an assessment score. Unfortunately, it is not clear whether the sum of some set of discrete elements adequately captures the complex interplay between a physician and a patient (Mazor *et al.*, 2005). As a result, a plethora of behaviourally anchored rating scales have been developed and discussed in the literature, one of the first being the Arizona Clinical Interview Rating Scale (Stillman *et al.*, 1977). This influx of assessment scales, while satisfying the need for valid and reliable measurement tools, presented other challenges: Should an existing scale be adopted? Who should do the rating?

Given the high-stakes nature of the assessment, its novelty in terms of examinee volume and the specific nature of the test population, we decided to develop and validate our own doctor–patient communication rating instrument, composed of measures of both interpersonal skills and spoken English-language proficiency. The initial instrument was based on a synthesis of the prevailing literature, and required separate evaluations of the relevant constructs (Boulet *et al.*, 1998a; Boulet *et al.*, 2001). For interpersonal skills, four domains were identified as being important for doctor–patient communication: interviewing and collecting information, counseling and delivering information, personal manner, and rapport. For spoken proficiency, the scale required a single rating and that was based, in general, on comprehensibility. For both interpersonal skills and spoken English proficiency, behaviourally anchored rating scales were developed. The second important decision about the assessment of communications skills concerned who should provide the ratings. Although the assessment of doctor–patient communication skills can be accomplished by a physician, or other observer, and could be done 'live' or via videotape review, the task of training and monitoring these individuals, especially in the context of a high-stakes certification exam, is arduous and costly. More important, it is unclear whether someone watching the interplay between a doctor and a patient can adequately measure the complex, multidimensional, asynchronous nature of the communication. Many aspects of this communication, especially those that are non-verbal, are best assessed by the patient or, in this case, the person trained to be the patient.

Post-encounter exercise (how do you know that the physician can correctly interpret the data that he/she collected in the interview?)

It is quite common for standardized patient assessments and OSCEs to have some type of post-encounter exercise. This exercise can take many forms, including a short quiz, a review of test results (e.g. X-ray reading), or a general synthesis of the information gathered in the encounter. Given the purpose of the CSA, we decided to require examinees to document and interpret the clinical findings from the simulated encounter in the form of a patient note. Although alternative types of exercises were viable, they tend to measure skills that could be best assessed with other examination formats, and without the logistical complexity of administering a simulated patient encounter. Furthermore, to ensure that IMGs are ready to enter graduate medical education programs in the US, they must be able to communicate their findings, in writing, to other members of the healthcare team.

In early field trials, the patient note exercise was scored analytically. During the case development process a list of 'keywords' related to the simulated patient's medical

complaint were generated. The scoring of the patient note was accomplished by identifying all the 'keywords', or suitable facsimiles, that were present in the written text. This method was found to yield reliable scores, even if non-physicians were used to identify whether key elements were, or were not, present on the note (Ben David *et al.*, 1997; Boulet *et al.*, 1998b). Unfortunately, there was some question as to the validity of scores generated in this fashion; an examinee could list the relevant patient findings (i.e. keywords), generate a differential diagnosis and suggest a diagnostic management plan without providing any suitable linkage between these parts. Also, test-wise candidates could inflate their scores by simply identifying lots of keywords in hopes that at least some would be correct. As a result, one could get credit for numerous key elements yet fail to reveal a clear and appropriate decision-making process. In addition, given the plethora of egregious actions, it was impossible under the analytic framework to incorporate any sort of negative scoring for actions, or transcribed inaccuracies, that may result in harm to the patient. Therefore, we decided to use a holistic approach to scoring the patient note: validity, rather than reliability, considerations were going to be the driving force in the development of the CSA (Boulet *et al.*, 2000; Slater & Boulet, 2001). While scoring via this method is often criticized for being subjective and therefore less reliable, with proper rater training, moderately reproducible, unbiased scores can be obtained (Boulet *et al.*, 2004; Boulet & McKinley, 2005). Further, as markers of clinical competence, these holistic scores may be more valid. Equally important, holistic, or global scoring, of decision-making processes, especially when the content of the exercise includes medical jargon and abbreviations, demands expert physician raters. Since scores for other parts of the examination were based on SP documentation of data-gathering skills and SP rating of communication skills, we made the decision, albeit an expensive one, to have physicians score all the patient notes. For a high-stakes examinations such as the CSA, having an examinee's pass/fail status determined, in part, by his/her peers is certainly very important because the validity, credibility and acceptance of the exam is enhanced.

Case construction and equating (how do you ensure that all examinees are treated fairly?)

Most standardized patient assessments and OSCEs are used for formative purposes. Medical students, or residents, complete a series of cases and are provided with specific information regarding their strengths and weaknesses. In addition, since constructing scenarios can be costly, most students see the same cases and standardized patients. While this administration strategy may lead to exposure of case materials and, depending on information-sharing strategies, the improved performance of students who are assessed later, the formative nature of these assessments diminishes the need for exhaustive security measures, including the assignment of students to different sets of cases. For high-stakes assessments such as CSA, where testing took place almost every day, it was extremely important that examinees were neither advantaged, nor disadvantaged, by prior knowledge of test content. While SP-based research suggests only marginal score gains for exposed material (Boulet *et al.*, 2003a), the perception that some examinees may have an advantage over others can ultimately detract from the validity of the assessment.

For the CSA, ECFMG took a number of steps to ensure that all examinees had an equal opportunity to show their 'true' ability on the assessment. First, over 200 clinical scenarios were developed, allowing test administrators to shuffle content on a continual basis. Second, the scoring keys (checklists) were closely guarded, and regular quality assurance analyses performed to detect potential exposure (Boulet *et al.*, 2003b). Third, and most important, statistical strategies were adopted to adjust candidate scores for minor variations in CSA difficulty from one day to the next due to changing cases and SPs (Swanson *et al.*, 1999). By collecting performance data on case difficulty, rater/SP stringency, etc., we could establish whether the test form encountered by one examinee was more or less difficult than the one encountered by another examinee. For individual cases, this information was collected from a live exam administration, using an unscored or 'calibration' encounter. By adding the case difficulties and rater stringencies over cases, one can determine how much an individual's score needs to be adjusted. Simply put, examinees who encountered relatively difficult forms had a few points added; those who had relatively easy forms had a few points subtracted.

Notes on contributors

GERALD WHELAN, MD, served as Vice-President for Clinical Skills Assessment throughout the time that CSA was being developed and was administered by ECFMG. He came to the clinical skills arena after extensive experience with the American Board of Emergency Medicine's certification examinations, including an oral simulated patient encounter model.

JOHN (JACK) BOULET, PhD, was recruited by Dr Friedman Ben-David to be the Psychometrician for the CSA. He later became the Director of Test Development and Research and was responsible for the ongoing development and validation of test materials.

DANETTE MCKINLEY was Associate Psychometrician for Clinical Skills Assessment from its implementation in 1998 through to ECFMG administration. She joined ECFMG after more than 10 years' experience in psychometrics at other testing organizations, with specialization in licensure and certification examinations

JOHN J. NORCINI, PhD, has been President and CEO of the Foundation for Advancement of International Medical Education and Research (FAIMER®) since May 2002. Dr Norcini's principal academic interest is in the area of the assessment of physician performance. He is on the editorial boards of six peer-reviewed journals in educational measurement and medical education and has published extensively.

MARTA VAN ZANTEN originally served as assistant to the CSA Co-Director at ECFMG, from a background in higher education, teaching English as a second language, and public health.

RONALD K. HAMBLETON, PhD, was a psychometric consultant for the CSA project. He is a Distinguished University Professor at the University of Massachusetts, Amherst.

WILLIAM P. BURDICK, MD MSEd, has taught clinical skills to medical students for over 20 years, and was involved in the development of the ECFMG Clinical Skills Assessment as a consultant. For the past six years, he has been Assistant Vice-President for Assessment Services.

STEVEN J. PEITZMAN, MD, was consultant then part-time staff during development and implementation by ECFMG of the CSA. A nephrologist, historian and educator, he helped establish the standardized patient program at the Medical College of Pennsylvania, and has lectured and published on the teaching and assessment of competence in physical examination.

References

ALBANESE, M. (2000) The decline and fall of humanism in medical education, *Medical Education*, 34, pp. 596–597.

ALLEN, P.F. & LOCKER, D. (1997) Do item weights matter? An assessment using the oral health impact profile, *Community Dental Health*, 14, pp. 133–138.

BARZANSKY, B. & ETZEL, S.I. (2004) Educational programs in US medical schools, 2003–2004, *Journal of the American Medical Association*, 292, pp. 1025–1031.

BECK, R.S., DAUGHTRIDGE, R. & SLOANE, P.D. (2002) Physician–patient communication in the primary care office: a systematic review, *Journal of the American Board of Family Practice*, 15, pp. 25–38.

BEN DAVID, M.F., BOULET, J.R., BURDICK, W.P., ZIV, A., HAMBLETON, R.K. & GARY, N.E. (1997) Issues of validity and reliability concerning who scores the post-encounter patient-progress note, *Academic Medicine*, 72(10, Suppl. 1), S79–S81.

BOULET, J.R., BEN DAVID, M.F., ZIV, A., BURDICK, W.P., CURTIS, M., PEITZMAN, S. & GARY, N.E. (1998a) Using standardized patients to assess the interpersonal skills of physicians, *Academic Medicine*, 73(10 Suppl.), pp. S94–S96.

BOULET, J.R., MCKINLEY, D.W., NORCINI, J.J. & WHELAN, G.P. (2002) Assessing the comparability of standardized patient and physician evaluations of clinical skills, *Advances in Health Sciences Education, Theory & Practice*, 7, pp. 85–97.

BOULET, J.R., MCKINLEY, D.W., WHELAN, G.P. & HAMBLETON, R.K. (2003a) Quality assurance methods for performance-based assessments, *Advances in Health Sciences Education, Theory & Practice*, 8, pp. 27–47.

BOULET, J.R., MCKINLEY, D.W., WHELAN, G.P. & HAMBLETON, R.K. (2003b) The effect of task exposure on repeat candidate scores in a high-stakes standardized patient assessment, *Teaching and Learning in Medicine*, 15, pp. 227–232.

BOULET, J.R., REBBECCHI, T.A., DENTON, E.C., MCKINLEY, D.W. & WHELAN, G.P. (2004) Assessing the written communication skills of medical school graduates, *Advances in Health Sciences Education, Theory & Practice*, 9, pp. 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, pp. 767–773.

BOULET, J.R., FRIEDMAN BEN-DAVID, M., HAMBLETON, R.K., BURDICK, W.P., ZIV, A. & GARY, N.E. (1998b) An investigation of the sources of measurement error in the post-encounter written scores from standardized patient examinations, *Advances in Health Sciences Education, Theory & Practice*, 3, pp. 89–100.

BOULET, J.R., FRIEDMAN BEN-DAVID, M., ZIV, A., BURDICK, W.P. & GARY, N.E. (2000) The use of holistic scoring for post-encounter written exercises, in: Melnick. D (Ed.) *Proceedings of the Eighth Ottawa Conference of Medical Education and Assessment*, pp. 254–260 (Philadelphia, National Board of Medical Examiners).

BOULET, J.R. & MCKINLEY, D.W. (2005) Investigating gender-related construct-irrelevant components of scores on a written assessment exercise of a high-stakes certification assessment, *Advances in Health Sciences Education, Theory & Practice*, in press.

BOULET, J., DE CHAMPLAIN, A. & MCKINLEY, D. (2003) Setting defensible performance standards on OSCEs and standardized patient examinations, *Medical Teacher*, 25, pp. 245–249.

CEGALA, D.J. & LENZMEIER, B.S. (2002) Physician communication skills training: a review of theoretical backgrounds, objectives and skills, *Medical Education*, 36, pp. 1004–1016.

COLLIVER, J.A. (1995) Validation of standardized-patient assessment: a meaning for clinical competence, *Academic Medicine*, 70, pp. 1062–1064.

CUNNINGTON, J.P.W., NEVILLE, A.J. & NORMAN, G.R. (1997) The risks of thoroughness: reliability and validity of global ratings and checklists in an OSCE, *Advances in Health Sciences Education*, 1, pp. 227–233.

ECFMG (1999) *Clinical Skills Assessment (CSA) Candidate Orientation Manual*, pp. 1–32 (Philadelphia, PA, ECFMG).

EDUCATIONAL COMMISSION FOR FOREIGN MEDICAL GRADUATES, (2004) *ECFMG Certification Fact Sheet* (Philadelphia, PA, ECFMG).

FEDERATION OF STATE MEDICAL BOARDS, I. & NATIONAL BOARD OF MEDICAL EXAMINERS (2003) *2004 USMLE Step 2 CS Content Description and General Information Booklet* (Philadelphia, FSMB and NBME).

FINKBINER, R.G., FLETCHER, E.A., ORR, N. & KLASS, D.J. (1994) Question format and scoring methods for standardized patient interstation exercises, in: A.I. Rothman & R. Cohen (Eds) *Proceedings of the Sixth Ottawa Conference on Medical Education*, pp. 343–345 (Toronto, University of Toronto Bookstore).

FRIEDMAN, M., SUTNICK, A.I., STILLMAN, P.L., NORCINI, J.J., ANDERSON, S.M., WILLIAMS, R.G., HENNING, G. & REEVES, M.J. (1991) The use of standardized patients to evaluate the spoken-English proficiency of foreign medical graduates, *Academic Medicine*, 66(9, Suppl.), pp. S61–S63.

FRIEDMAN, M., SUTNICK, A.I., STILLMAN, P.L., REGAN, M.B. & NORCINI, J.J. (1993) The relationship of spoken-English proficiencies of foreign medical school graduates to their clinical competence, *Academic Medicine*, 68(10, Suppl.), pp. S1–S3.

GIMPEL, J.R., BOULET, D.O. & ERRICHETTI, A.M. (2003) Evaluating the clinical skills of osteopathic medical students, *Journal of the American Osteopathic Association*, 103, pp. 267–279.

GORTER, S., RETHANS, J.J., SCHERPBIER, A., VAN DER HELUDE, D., HOUBEN, H., VAN DER VLEUTEN, C. & VAN DER LINDEN, L.S. (2000) Developing case-specific checklists for standardized-patient-based assessments in internal medicine: a review of the literature, *Academic Medicine*, 75, pp. 1130–1137.

GOVAERTS, M.J., VAN DER VLEUTEN, C.P. & SCHUWIRTH, L.W. (2002) Optimising the reproducibility of a performance–based assessment test in midwifery education, *Advances in Health Sciences Education, Theory & Practice*, 7, pp. 133–145.

HARDEN, R.M. & GLEESON, F.A. (1979) Assessment of clinical competence using an objective structured clinical examination (OSCE), *Medical Education*, 13, pp. 41–54.

HARDEN, R.M., STEVENSON, M., DOWNIE, W.W. & WILSON, G.U. (1975) Assessment of clinical competence using objective structured examination, *British Medical Journal*, 1, pp. 447–451.

LIAISON COMMITTEE ON MEDICAL EDUCATION (2003) *Functions and Structure of a Medical School: standards for Accreditation of Medical Education Programs leading to the MD Degree* (Washington, DC, Association of American Medical Colleges).

LYONS, T.F. & PAYNE, B.C. (1975) The use of item importance weights in assessing physician performance with predetermined criteria indices, *Medical Care*, 13, pp. 432–439.

MAZOR, K.M., OCKENE, J.K., ROGERS, J., CARLIN, M.M. & QUIRK, M.E. (2005) The relationship between checklist scores on a communication OSCE and analogue patients' perceptions of communication, *Advances in Health Sciences Education, Theory & Practice*, in press.

MEDICAL COUNCIL of CANADA (2002) *Qualifying Examination Part II, Information Pamphlet* (Ottawa, Ontario, Canada, Medical Council of Canada).

MELNICK, D.E., DILLON, G.F. & SWANSON, D.B. (2002) Medical licensing examinations in the United States, *Journal of Dental Education,* 66, 595–599.

NEWBLE, D.I. & SWANSON, D.B. (1988) Psychometric characteristics of the objective structured clinical examination, *Medical Education*, 22, pp. 325–334.

NORCINI, J. & BOULET, J. (2003) Methodological issues in the use of standardized patients for assessment, *Teaching and Learning in Medicine*, 15, pp. 293–297.

NORCINI, J.J., STILLMAN, P.L., SUTNICK, A.I., REGAN, M.B., HALEY, H.-L.A., WILLIAMS, R.G. & FRIEDMAN, M. (1993) Scoring and standard setting with standardized patients, *Evaluation and the Health Professions*, 16, pp. 322–332.

NORMAN, G.R., VAN DER VLEUTEN, C.P. & DE GRAAFF, E. (1991) Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability, *Medical Education*, 25, pp. 119–126.

ONG, L.M., DE HAES, J.C., HOOS, A.M. & LAMMES, F.B. (1995) Doctor–patient communication: a review of the literature, *Social Science Medicine*, 40, pp. 903–918.

PANGARO, L.N., WORTH-DICKSTEIN, H., MACMILLAN, M.K., KLASS, D.J. & SHATZER, J.H. (1997) Performance of standardized examinees in a

standardized-patient examination of clinical skills, *Academic Medicine*, 72, pp. 1008–1011.

Robb, K.V. & Rothman, A.I. (1985) The assessment of clinical skills in general medical residents—comparison of the objective structured clinical examination to a conventional oral examination, *Annals RCPSC*, 18, pp. 235–238.

Slater, S.C. & Boulet, J.R. (2001) Predicting holistic ratings of written performance assessments from analytic scoring, *Advances in Health Sciences Education, Theory & Practice*, 6, pp. 103–119.

Stillman, P.L., Regan, M.B., Haley, H.L., Norcini, J.J., Friedman, M. & Sutnick, A.I. (1992) The use of a patient note to evaluate clinical skills of first-year residents who are graduates of foreign medical schools, *Academic Medicine*, 67(10 Suppl.), pp. S57–S59.

Stillman, P.L., Brown, D.R., Redfield, D.L. & Sabers, D.L. (1977) Construct validation of the Arizona Clinical Interview Rating Scale, *Educational and Psychological Measurement*, 37, pp. 1031–1039.

Streiner, D.L., Goldberg, J.O. & Miller, H.R. (1993) MCMI-II item weights: their lack of effectiveness, *Journal of Personality Assessment*, 60, pp. 471–476.

Sutnick, A.I., Friedman, M., Wilson, M.P., Razin, S., Sperling, O., Gutman, D. & Glick, S. (1995) Use of ECFMG certifying examinations for national comparisons: performance of graduates of Israeli medical schools in basic and clinical science examinations, *Israel Journal of Medical Science*, 31, pp. 250–254.

Sutnick, A.I., Stillman, P.L., Norcini, J.J., Friedman, M., Regan, M.B., Williams, R.G., Kachur, E.K., Haggerty, M.A. & Wilson, M.P. (1993) ECFMG assessment of clinical competence of graduates of foreign medical schools. Educational Commission for Foreign Medical Graduates, *Journal of the American Medical Association*, 270, pp. 1041–1045.

Sutnick, A.I. & Wilson M.P. (1994) ECFMG approach to global assessment of clinical competence. *AMSE Newsletter*, 8–10.

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

Swanson, D.B. & Stillman, P.L. (1990) Use of standardized patients for teaching and assessing clinical skills, *Evaluation and the Health Professions*, 13, pp. 79–103.

Tombleson, P., Fox, R.A. & Dacre, J.A. (2000) Defining the content for the objective structured clinical examination component of the professional and linguistic assessments board examination: development of a blueprint, *Medical Education*, 34, pp. 566–572.

van der Vleuten, C.P., Norman, G.R. & De Graaff, E. (1991) Pitfalls in the pursuit of objectivity: issues of reliability, *Medical Education*, 25, pp. 110–118.

Vu, N.V. & Barrows, H.S. (1994) Use of standardized patients in clinical assessments: Recent developments and measurement findings, *Educational Researcher*, 23, pp. 23–30.

Vu, N.V., Barrows, H.S., Marcy, M.L., Verhulst, S.J., Colliver, J.A. & Travis, T. (1992) Six years of comprehensive, clinical, performance-based assessment using standardized patients at the Southern Illinois University School of Medicine, *Academic Medicine*, 67, pp. 42–50.

Vu, N.V., Marcy, M.L., Barnhart, A.J., Colliver, J.A., Henkle, J.Q., Hodgson, K., Schrage, J.P. & Travis, T.A. (1994) Further evidence of construct validity of standardized patient–based performance examinations, *Teaching & Learning in Medicine*, 6, pp. 255–259.

Ziv, A., Ben David, M.F., Sutnick, A.I. & Gary, N.E. (1998) Lessons learned from six years of international administrations of the ECFMG's SP-based clinical skills assessment, *Academic Medicine*, 73, pp. 84–91.