

THE INTRODUCTION OF CLINICAL SKILLS ASSESSMENT INTO THE UNITED STATES MEDICAL LICENSING EXAMINATION (USMLE): A DESCRIPTION OF USMLE STEP 2 CLINICAL SKILLS (CS)

*Richard E. Hawkins, M.D., Deputy Vice President, Assessment Programs, National Board of Medical Examiners
Clinical Skills Evaluation Collaboration (CSEC)*

ABSTRACT

In June 2004, Step 2 Clinical Skills (CS) was introduced into the United States Medical Licensing Examination (USMLE). The purpose of USMLE Step 2 CS is to ensure successful candidates for licensure in the United States possess the clinical skills that are essential for safe and effective patient care. Ensuring high quality in such a large-scale, performance-based test requires meticulous attention to detail at multiple levels in preparing for implementation. These levels include: case and test development, standardized patient training, quality assurance, scoring and standard setting. The authors describe the efforts undertaken to ensure the examination provides for a fair assessment of individual examinee performance with regard to those fundamental patient-centered skills.

In June 2004, an examination component designed to assess clinical skills was introduced into the United States Medical Licensing Examination (USMLE). USMLE Step 2 Clinical Skills (CS) consists of a multiple-station, standardized patient-based examination developed and administered through the collaborative efforts of the National Board of Medical Examiners (NBME) and the Educational Commission for Foreign Medical Graduates (ECFMG); and with the cooperation of the Federation of State Medical Boards, co-sponsor with the NBME of the USMLE. The purpose of Step 2 CS is to ensure successful candidates for licensure in the United States possess the fundamental clinical skills essential for safe and effective patient care. These clinical skills include taking a medical history, performing a physical examination, communicating effectively with patients, clearly and accurately documenting the findings and diagnostic impressions from the clinical encounter, and identifying appropriate initial diagnostic studies.

The examination typically consists of 12 encounters with

standardized patients (SPs) who portray common and important medical problems. SPs are individuals trained to portray a patient problem or condition in representing a realistic clinical situation. More than four decades of research supports the reliability of scores using SP-based assessments of clinical skills. Similarly, extensive research shows that scores from SP-based examinations yield valid interpretations of examinee performance. For USMLE Step 2 CS, examinee data gathering performance is documented by SPs using checklists constructed to represent the appropriate history and physical examination components for a specific patient presentation. Communication, interpersonal skills and spoken English proficiency are evaluated by SPs using rating scales that are based on the communication, interpersonal and spoken language skills that should be exercised in all patient encounters. The patient note, recorded by an examinee after the encounter, is rated by a physician. In formulating a score for each note, patient note raters are extensively trained to consider the extent to which important information is recorded for the specific case, as well as the quality of the written documentation. Examinees are scored on three examination components: the Integrated Clinical Encounter (ICE), which reflects the examinee's skill in data gathering and in completing the patient note; Communication and Interpersonal Skills (CIS); and Spoken English Proficiency (SEP). Examinees must pass all three components on the same attempt to pass Step 2 CS.

USMLE Step 2 CS is administered at five regional centers located across the United States. These centers operate year round, administering exams up to seven days per week, with as many as three testing sessions per day, depending upon examinee volume. It is estimated that approximately 30,000 examinees, comprising students and graduates of both U.S. and international medical schools, will take Step 2 CS annually. This delivery model was chosen because it repre-

sented an achievable balance of quality control, cost-efficiency and examinee convenience, while still permitting attainment of the high psychometric standard needed for a licensing examination.

As expected, significant operational challenges were encountered in developing the infrastructure for this delivery model, including test center construction, recruitment, hiring and training of personnel, and configuration of the audiovisual and information technology backbone necessary to support simultaneous, multiple-site administrations. Interwoven with these logistic challenges, yet extending well beyond them in terms of their complexity, were the potential impediments to attaining the psychometric standard required for high-stakes examinations.

Ensuring fairness and consistency in examination delivery and scoring across multiple sites and administrations requires meticulous attention to detail and appropriate planning at several levels, including identification of the ideal scoring approach, case and test development, blueprint design and construction of equivalent test forms, standardized patient (and standardized patient trainer) training, quality assurance, equating and standard setting.

In developing the scoring instruments and planning the rating process for each of the examination components, the intent was to assure the reliability of scores and the validity of the interpretations that would be based upon those scores. However, feasibility and credibility were also important qualities in terms of acceptability to the medical profession and other stakeholders. Through the application of checklists and global ratings provided by SPs and ratings of the patient note by physician raters, these criteria are met. For Step 2 CS, dichotomous checklists completed by SPs are a reliable and efficient means for determining whether examinees obtain the essential history and physical examination data for a given clinical encounter. Checklist scores (a process measure) are complemented by patient note ratings (a clinical outcome measure) in appraising data-gathering proficiency. Physician ratings of the patient note provide credibility; examinees are evaluated by medical "experts". In addition, the note rating process is cost-effective since rater training and scoring can be centralized. Global rating scales provide an ideal measure of interpersonal and communication skills and spoken English proficiency. These generic rating scales are constructed to represent a patient's perspective within the context of a clinical encounter, thus identifying the patient (SP) as the expert rater of examinee performance within these domains. Here,

the USMLE engages members of the public (SPs) as participants in scoring the examination, providing them with the training and instruments to assess skills critical to patient satisfaction and provision of high quality care.

The examination blueprint for Step 2 CS is designed to reflect the broad distribution of patients new residents are likely to encounter in different settings and contexts for graduate training programs in the United States. These include common and important clinical problems of varying acuity. In the aggregate, the exam also includes a broadly representative sample of patients of different ages, genders and ethnicities. To ensure delivery of content equivalent examinations at the five sites, rules for creating individual test forms have been defined, and software programs have been developed to select SP-case combinations at each site, and for each exam administration, that yield content-representative test forms.

Committees of content experts, representing various aspects of the medical profession (clinicians, academic faculty, state medical board members) participate in decision-making regarding test design and in the development of individual cases. This provides a level of assurance that examination content and difficulty are appropriate for the intended purpose of the examination. The case development process is iterative in nature, with individual test development committee members leading discussion on specific cases. Practice encounters with SPs and other committee members unfamiliar with the particular case inform case refinement and checklist development. Once the initial case development is complete, SPs are identified to portray the case in un-scored stations during live Step 2 CS examinations at one or more test sites. After completing a number of encounters in un-scored stations, interdisciplinary review committees consider both examinee and SP performance data in making decisions regarding continued use of the case and the need for specific refinements in SP training and/or checklist structure. Such a process will also occur periodically during ongoing live administrations of each case.

During the case development process, committees identify specific patient characteristics that are used to recruit SPs for each case. Blueprint category requirements, as well as case-relevant clinical criteria, guide selection of the age distribution, gender and race/ethnicity for SP-case combination. Additionally, for specific cases, patient Body Mass Index and other physical features, such as the presence of various physical findings or scars, may represent

characteristics that the committees believe are inappropriate for the proper presentation of the case; these characteristics are used for excluding some potential SPs. Once all relevant characteristics are defined, it is important that SP recruitment and selection be consistent within and across testing sites.

Despite the application of specific rules for creating test forms, developing cases and selecting SPs, an important threat to examination equivalence is the potential variability in performance and scoring of SPs both within and between testing sites and over time. To ensure consistency strict procedures and standards have been adopted for both SP and SP trainer performance. Additionally, staff undertook two projects critically important to maintaining consistency in SP training procedures, case portrayal and scoring across multiple sites:

1. The use of identical, multimedia-based training materials and methods administered electronically (via a process called "E-case") at all sites guarantees consistency in delivery of SP training materials and instructions. Periodic modifications and updates of test and case materials are transmitted simultaneously to all sites, automatically replacing existing versions. This electronic system allows for SPs and SP trainers to access current training and case materials between and during administrations of the examination as necessary. This system also supports the periodic delivery of quizzes to gauge SP readiness for entry into un-scored stations and then into live examinations, and to monitor accuracy on an ongoing basis. The use of electronic methods also provides an additional level of security by minimizing the availability of paper materials and controlling exposure of case materials and scoring instruments.
2. The SP Trainer Training Academy is conducted at NBME headquarters in Philadelphia. Faculty members include external consultants (expert SP trainers) and test development and SP training staff. At these sessions, all trainers receive identical, intensive instruction in SP training methods, including the use of the E-case for on-site SP training. Additionally, central "Academy" staff periodically travel to each site to co-facilitate various training sessions as an additional measure to ensure consistency in training and performance.

While the above innovations enhance consistency in SP performance within and across sites by requiring strict

adherence to training procedures and successful completion of periodic assessments, stringent quality assurance approaches are necessary to maintain optimal performance. Quality assurance begins with a sign-off process designed to ensure SP readiness to participate in live examinations. SPs are required to complete a minimum number of portrayals in un-scored stations and must demonstrate consistency and accuracy in recording and scoring examinee actions. Once an adequate level of portrayal and scoring is established, the SP can be used as part of the live examination. However, quality assurance continues with strict monitoring of compliance with guidelines for portrayal consistency and scoring accuracy. This is accomplished via random review of live and recorded SP performances, both by local site trainers and central test development staff. Lastly, a series of ongoing score-based analyses provide continuous statistical monitoring and feedback on SP performances over time for each score component.

Despite rigorous training methods and stringent approaches to quality assurance, it is unavoidable that residual systematic differences in case difficulty and SP stringency will remain within and across testing sites. For this reason, equating procedures have been developed to adjust for differences in the difficulty of test forms between and within sites, and over time. Separate video-based ratings of checklist and CIS/SEP performance by SPs from across multiple sites provide a basis for placing scores across centers on a common scale. These ratings are also used to monitor consistency in portrayal and accuracy in scoring across sites and SPs. For patient note ratings specifically, since these are performed at a single location, ratings only need to be adjusted for differences in stringency among individual raters.

The standard setting process for Step 2 CS is similar to that used for the other USMLE Steps, involving the medical education and practice communities. Individuals from these groups define expectations for examinee performance through an extensive review of data from live administrations of the examination. Surveys completed by medical school administrators, undergraduate and graduate medical educators, clinicians, state medical board members and examinees themselves provide an important perspective on the prevalence of substandard clinical skills. In addition, panels of faculty and clinicians review samples of examinee/SP encounters and provide their opinion on a minimal standard for performance in each of the areas tested. The USMLE Step 2 Committee, comprising leaders in medical education and licensing, review these data and other

related information in identifying minimally acceptable performance levels for all of the Step 2 CS components.

In conclusion, the addition of clinical skills assessment into the USMLE provides the public with another level of assurance regarding the clinical skills of individuals granted a license to practice medicine in the United States. The inclusion of such a large scale, performance-based examination has posed unique logistical and psychometric challenges. We have described the efforts undertaken to ensure that the examination provides for fair assessment of individual examinee performance with respect to those fundamental patient-centered skills necessary for safe and effective health care.

CONTRIBUTING AUTHORS

The following individuals from the National Board of Medical Examiners and the Educational Commission for Foreign Medical Graduates were responsible for developing and implementing the programs and processes described in this article. All participated in the writing or critical review of this article.

National Board of Medical Examiners

Richard E. Hawkins, M.D., Deputy Vice President, Assessment Programs

David B. Swanson, Ph.D., Deputy Vice President, Professional Services

Gerard F. Dillon, Ph.D., Associate Vice President, USMLE

Brian E. Clauser, Ed.D., Associate Vice President, Measurement Consulting Services

Ann M. King, Manager, SP Test Material Development

Peter V. Scoles, M.D., Senior Vice President, Assessment Programs

Educational Commission for Foreign Medical Graduates

Gerald P. Whelan, M.D., Vice President, Assessment Services

William P. Burdick, M.D., M.S. Ed., Assistant Vice President, Assessment Services

John R. Boulet, Ph.D., Assistant Vice President, Research and Evaluation

Ann G. Homan, Director, Center Operations, Assessment Services