Critical Issues in Dental Education

Medical Licensing Examinations in the United States

Donald E. Melnick, M.D.; Gerard F. Dillon, Ph.D.; David B. Swanson, Ph.D.

Dr. Melnick is President of the National Board of Medical Examiners; Dr. Dillon is Associate Vice President, USMLE, National Board of Medical Examiners; and Dr. Swanson is Associate Vice President, Test Development Services, National Board of Medical Examiners. Direct correspondence to Dr. Melnick at the National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104; 215-590-9546 phone; 215-590-9755 fax; dmelnick@nbme.org.

Submitted for publication 3/28/02; accepted 4/8/02

Editor’s Note: For decades, academicians in medicine have grappled with finding a balance between a fair, thorough format for licensing exams and meeting the needs and preserving the ethical considerations of the patients involved in these exams. A debate on the involvement of patients in dental licensing exams is currently occurring as well at the national, regional, and state levels. Here we present a set of four articles contributing to this debate: the first, by Dr. Melnick et al., traces the history of medical licensing exams in the United States; it is followed by Dr. Pattalochi’s arguments for retaining human subjects in dental licensing examinations, Dr. Formicola’s counterpoint to Dr. Pattalochi’s position, and a commentary on all three articles by Dr. Kalkwarf. I would like to strongly encourage readers to express their opinions on this important issue in dental education in the form of Letters to the Editor.

In this article we discuss examination requirements for medical licensure in the United States, focusing on the exam components related to assessment of hands-on clinical skills with patients and assessment of medical decision-making skills. The first section provides a brief history of medical licensing examinations, highlighting some of the methods that have been used in attempts to measure clinical and decision-making skills. The next two sections provide an overview of the United States Medical Licensing Examination® (USMLE™), which has been in place for the past decade. The last section takes a look at the use of standardized patients for assessment of clinical skills; this new component of USMLE is scheduled for introduction in 2004 or 2005.

A Brief History of Medical Licensing Examinations in the United States

America’s forebears in Europe had developed systems of examination for medical licensure well before the Colonial era. However, these systems generally were not adopted in the American colonies. The first emulation of European systems of examination for licensure appeared in the 1760s, with colonies and then new states authorizing the examination of practitioners, usually through one of the learned societies of professionals. These first efforts to ensure a minimum quality of medical practitioners eroded steadily in the first half of the nineteenth century through the period of “Jacksonian democracy,” which eschewed governmental regulation. During that time, the medical school provided the authority to practice to its graduates, and the proliferation of proprietary, poor quality medical schools was reflected in the credentials of physicians. Not until after the Civil War, when military and political leaders discovered the huge variation among medical practitioners, did states once again seek to regulate entry into the practice of medicine. Texas established the first modern medical licensing authority in 1873, and nearly all states had established licensing boards and examinations by the turn of the century.1 The failure of medical schools to provide a reasonable assurance of minimum quality explains our current system of checks and balances, with examination for licensure residing outside the medical schools. States relied on various bodies to conduct examinations: public health agencies of the state, the state boards, and state medical societies were the most common groups charged with this task.

As the twentieth century began, every state required an examination to obtain a license to practice
These examinations varied widely in scope and form—from an interview to fairly rigorous oral, written, and practical examinations and observation of clinical encounters. Because of the widely variable examinations and examination standards, reciprocal recognition of one state’s license by another was a rarity. The first proposals to create a national examination system appeared in 1902; the National Board of Medical Examiners (NBME®) was founded in 1915, administering its first examinations in 1916. This voluntary examination program grew in parallel with state examinations until the 1960s.

The first NBME examinations were weeklong extravaganzas, incorporating essay, laboratory, oral, practical, and bedside components. In 1922, the NBME’s examination program was restructured. The first component in the new structure tested basic biomedical sciences through essay questions. The second component tested fundamentals of clinical medicine through essay questions. The final component included observed patient encounters and an oral examination. This structure persisted until the early 1950s. As the NBME embraced the emerging science of psychometrics, it replaced essay questions with newly developed “selected-response” questions, using many formats of multiple-choice questions. The reproducibility of scores with this test method was much higher than for the old essay examinations, and the bias introduced by different essay scorers was also eliminated. The first NBME examinations were weeklong extravaganzas, incorporating essay, laboratory, oral, practical, and bedside components. In 1922, the NBME’s examination program was restructured. The first component in the new structure tested basic biomedical sciences through essay questions. The second component tested fundamentals of clinical medicine through essay questions. The final component included observed patient encounters and an oral examination. This structure persisted until the early 1950s. As the NBME embraced the emerging science of psychometrics, it replaced essay questions with newly developed “selected-response” questions, using many formats of multiple-choice questions. The reproducibility of scores with this test method was much higher than for the old essay examinations, and the bias introduced by different essay scorers was also eliminated.

Studies of the final clinical component—a bedside oral examination—in the late 1950s clearly documented its psychometric inadequacies. The scores were found to provide more information about the examiner than they did about the examinee. Agreement between examiners observing examinees with different patients was at near-chance levels. The test relied on a small number of cases, and this limited sampling contributed to unreliability. Because the clinical examination did not meet the NBME’s standards for reliability and validity, the observed clinical examination was eliminated in 1964. The high cost and logistical difficulties of this examination for the rapidly growing postwar cohort of medical students also contributed to its demise. The new final clinical examination experimented with a number of test formats over the ensuing decades, all intended to recover the assessment of clinical skills that was lost with the elimination of the observed clinical examination. Initially, motion pictures of clinical encounters were projected to examinees, who answered multiple-choice questions based on the portrayed encounter. Later, latent-image patient management problems were introduced. These multi-step problems allowed the examinee to uncover hidden information about a patient at several points in the case, usually diagnosis, management, and follow-up. Hiding information permitted the examinees to use different approaches to patient care.

Unfortunately, each of these innovative assessment methods failed to achieve acceptable levels of reliability and validity, and each was eventually dropped from the examination. The key problem of the 1960s motion picture examination was the difficulty of achieving standardized testing conditions for projection of the films. Latent image patient management problems were affected by the small sample of clinical cases in any examination. Additionally, the format itself introduced error. Requiring the examinee to select an unspecified number of choices from relatively long lists seemed to measure as much about the examinee’s propensity to act or not act in an uncertain situation as it did about the clinical decision-making skills of the examinee.

By the late 1980s, all components of the NBME examinations were in multiple-choice format. As problems with the intended proxies for an observed patient encounter increased, the NBME began a long period of research and development on computer-based patient management simulations. Three decades of research effort eventually culminated in the inclusion of this uncued, interactive patient management simulation—Primum®—in the national medical licensing examination in 1999. The NBME began to explore standardized patients in the late 1970s. Research continues on this promising method for assessing patient-centered skills. These assessment methods are more fully described below.

In the late 1960s, the Federation of State Medical Boards (FSMB) petitioned the NBME to develop a single examination that could be used by state licensing boards to meet licensure examination needs that were not fully satisfied by the NBME certifying examinations. The NBME developed the Federation Licensing Examination (FLEX) from test materials prepared for NBME certifying examinations. In the mid-1980s, the link between the NBME certifying examinations and FLEX was severed, with the “new” FLEX built from items developed specifically for it. From the late 1960s through 1994, the NBME certifying examinations and the FLEX existed as parallel licensing examination systems, and the state-developed examinations disappeared.
The Current U.S. Medical Licensing Examination

In the early 1990s, efforts to develop a single examination pathway for the initial medical license culminated in the introduction of the USMLE. All individuals seeking a license to practice (allopathic) medicine in the United States must take USMLE, and all U.S. medical licensing jurisdictions accept USMLE scores and recommended pass/fail decisions as satisfying examination requirements for licensure. This program, developed jointly by the NBME and the FSMB, consists of three components: Steps 1, 2, and 3.

Step 1 assesses whether the examinee understands and can apply important concepts of the sciences basic to the practice of medicine, with special emphasis on principles and mechanisms underlying health, disease, and modes of therapy. This examination is generally taken upon completion of formal training in the basic medical sciences, which, for most students at U.S. medical schools, is near the end of their second year of training. Next, Step 2 assesses whether the examinee can apply medical knowledge and understanding of clinical science essential for the provision of patient care under supervision—that is, whether the examinee is ready to take on the responsibilities of a new resident. Step 2 is generally taken near the end of training in the clinical sciences that, for most U.S. students, is in the senior year of medical school. Finally, Step 3 assesses whether the examinee can apply medical knowledge and understanding of clinical science essential for the unsupervised practice of medicine. Step 3 provides a final assessment of readiness for independent responsibility in delivery of general medical care. It is typically taken after the examinee has acquired some postgraduate training or experience. Most graduates of U.S. schools sit for Step 3 within two years of graduation.5

Until 1999, all three steps were two-day, paper-and-pencil tests offered twice per year. In 1999, computer-based test administration was introduced, and paper-and-pencil tests were phased out. Step 1 is now a one-day, 350-item examination taken at Thomson Prometric™ centers around the world. Step 2 is also a one-day exam taken at Prometric centers; it consists of 400 items. Step 3 remains a two-day examination, consisting of a day and a half of multiple-choice questions (500 items) and a half-day (nine cases) of computer-based case simulations, which are described in more detail in the next section.

Assessment of Medical Decision-Making Skills in USMLE

As illustrated by the brief history of the medical licensure process, assessment of medical decision-making skills has been a high priority for a long time. In the current USMLE, two primary methods are used: patient-based multiple-choice questions (MCQs) and computer-based case simulations (CCS).

Patient-based MCQs have appeared in all three steps since the introduction of USMLE in 1992. In Step 1, these take the form of brief descriptions of patient care situations followed by questions challenging examinees to use their understanding of basic biomedical science to explain or predict patient findings.6 Roughly 60 percent of the items on Step 1 currently take this form. In Step 2, virtually all items begin with a description of a clinical situation; these are longer and less classic than the patient descriptions on Step 1. Examinees must differentiate important from incidental findings and indicate a clinical decision, generally a diagnosis or the next step in patient care.7 Step 3 test items also provide a robust description of a physician-patient encounter. Items are often presented in a format termed “case clusters,” in which a series of MCQs address different facets of an unfolding clinical situation. For examinees, the emphasis is on patient management, most often in an ambulatory setting.8 Across all three steps, the MCQs are best viewed as a form of low-fidelity patient simulations aimed at assessment of decision-making skills. In the future, these are likely to become higher fidelity, as all three steps take advantage of computer-based test administration to incorporate multimedia into item “stems,” enriching patient presentations.

The introduction of Primum CCS in USMLE Step 3 in 1999 was the first major change in the format of medical licensing examinations in more than a decade. In CCS, the examinee is presented with a brief description of a patient, including a chief complaint and a brief history.8,9 From that point forward, the case unfolds as the examinee works up and manages the computer-simulated patient, obtaining diagnostic information, ordering therapeutic interven-
tions, and monitoring patient progress. Any of several thousand diagnostic and therapeutic maneuvers can be specified in free text on an “order sheet.” As simulated time passes, the patient’s condition changes based on the underlying medical problem and the examinee’s interventions; results of tests are reported and the impact of interventions must be monitored. Examinees are scored on CCS using an algorithm that essentially compares their patient management strategies with policies obtained from experts. Examinees must balance thoroughness, efficiency, timeliness, and avoidance of risk in responding to clinical situations, with dangerous and unnecessary actions lowering scores. Step 3 pass/fail decisions are based upon a composite score formed by combining performance on MCQs and CCS, with CCS weighted somewhat less proportionally than the testing time allotted for it. Though CCS cases have proven expensive to develop, administer, and score, psychometric analyses have indicated that CCS measures something different than MCQs with a reasonable degree of precision.

**Future Plans for Assessment of Clinical Skills in USMLE**

There has been a strong desire to include an assessment of clinical skills with patients in the medical licensure process since the demise of bedside oral exams in the early 1960s; research on a replacement has been under way almost continually since that time. In 1999, the parents of USMLE—the NBME and FSMB—approved the addition of a clinical skills examination (CSE) to USMLE as soon as the reliability, validity, and feasibility of such an exam have been demonstrated.

Current plans call for final field testing of a USMLE CSE in 2002-03, with implementation scheduled for 2004 or 2005. The projected CSE design is very similar to the Clinical Skills Assessment (CSA®) developed by the Educational Commission for Foreign Medical Graduates (ECFMG) for testing international medical graduates. The proposed USMLE CSE will consist of a series of ten to twelve twenty-five-minute “stations,” each of which involves a “standardized patient” (SP). An SP is a lay person who has been trained to portray a patient role in a standardized, consistent fashion for purposes of assessment. SPs can be asymptomatic or can simulate physical findings (e.g., hyperactive reflexes, diminished breath sounds). During the first fifteen minutes of each station, the examinee interacts with an SP as though he or she were a real patient, taking a history, performing a focused physical examination, and providing patient education and counseling. During the last ten minutes of each station, the examinee prepares a written record of pertinent history and physical findings, a differential diagnosis, and a diagnostic workup plan. The SP concurrently documents the history and physical information obtained by the examinee and rates the examinee’s interpersonal skills. Examinees rotate from station to station, interviewing and examining each SP and preparing a written record of their findings for each case. USMLE CSE administrations are planned to take place year-round at five to seven regional sites.

Assessing the clinical skills of 25,000 examinees annually in a network of regional centers poses some difficult challenges. Like most practical exams, performance on one SP station is not strongly related to performance on other stations (this is often termed “content specificity”). As a consequence, tests must be fairly long or total scores will not provide an accurate, generalizable indicator of examinee proficiency. This, in turn, increases costs and logistical complexity. NBME and ECFMG research indicates that ten to twelve cases will yield adequately generalizable scores. Administration at multiple sites results in a danger that SP portrayal and case/test difficulty will vary from site to site, disadvantaging examinees who take the test at the more stringent sites. Placing scores from different forms of the test on the same scale can be difficult. Accounting for the challenge of multiple sites’ effects on score equivalency contributes to the high cost of these examinations. Test security is also a concern. Examinees may share information about the cases and SPs that they encountered. Some score components seem particularly vulnerable to this risk, while others are more robust. Finally, cost is, of course, a major issue. Assessments of clinical skills typically cost much more than their cognitive counterparts.

Despite these challenges, the experience of the ECFMG and the Medical Council of Canada, which uses a similar standardized patient-based exam in assessing Canadian physicians for licensure, indicates that the clinical skills exams identify potential physicians who lack minimum competency in this area. This remains true for many examinees who pass cognitive examinations. This finding, more than any other, warrants the inclusion of the clinical skills test in U.S. medical licensing examinations.
REFERENCES

11. Clauser BE, Margolis MJ, Swanson DB. An examination of the contribution of Primium case simulations to the USMLE Step 3 examination. Unpublished manuscript.
20. Wood T, Blackmore D, Dauphinee D. Results from the Medical Council of Canada’s qualifying examination: examining the validity of the licentiate procedure. Unpublished manuscript.