Innovations in Global Health Professions Education 2013

International assessment of medical students: Should it matter anymore where the school is located?

Author

Donald E. Melnick[1,*]

Abstract

With assessment systems that are adequate, robust, comprehensive, as well as responsive to local and regional needs, should the location of the medical education institution be irrelevant? Adequate assessment is determined by local needs, along with accepted minimum global standards of practice. If an assessment system is robust, it should be able to predict future behavior and performance to some degree. A comprehensive system would include assessment of all relevant competencies. In order to achieve comprehensiveness, new approaches are needed to demonstrate mastery of competencies that is now inferred from medical school and graduate medical education participation. These are likely to require a novel approach to assessment — gathering natural, real world data longitudinally rather than only through point-in-time tests. Increasingly the world of assessment may be able to provide tools and data that offer individualized assurances of competence.

Introduction

Medical education is becoming global, as is medical care. In our increasingly global village, should we care about the institution where a doctor was educated in making decisions about a doctor's fitness to practice or to pursue additional education?

One premise is that with adequate, robust, comprehensive assessment systems that are responsive to local and regional needs, the geographic location of the educational institution should be irrelevant. The institution is always a proxy — one large step removed — from the real point of interest: whether or not the individual is competent to practice in a specific context. Even the best institutions may produce incompetent doctors, and the worst may produce competent doctors. We know that accreditation and institutional reputation do not guarantee competence for an individual health professional.

If that premise were really true, would we care about accreditation of the medical school, its location, or even if an applicant for practice attended a formal educational program? In the United States, in an example from another field, 11 states allow admission

[1] President and Chief Executive Officer, National Board of Medical Examiners (NBME)

* Email: DMelnick@nbme.org

Submitted: 11 November 2013

Accepted: 14 November 2013

Cite this article as:
Melnick DE. International assessment of medical students: Should it matter anymore where the school is located? *Innovations in Global Medical and Health Education* 2013:5
http://dx.doi.org/10.5339/igmhe.2013.5

to the bar to practice law for those qualified through law office experience, correspondence courses, or online learning. Perhaps the most famous self-taught, apprentice lawyer was Abraham Lincoln. Why should we use a proxy like medical school quality metrics, rather than answering the specific question about competence, when deciding whether or not a person is capable of practicing medicine?

To answer this question, we should deconstruct the premise, examining each of its characteristics.

What does it mean to have an "adequate" assessment system? In answering that question, we look at locally determined needs. We must always balance the minimum standards of practice in relationship to workforce and care-access parameters. We also must give attention to competence as defined by local cultural and linguistic needs, and we must incorporate variation in standards of practice, such as custom, culture, and disease prevalence.

What does it mean for a system to be "robust"? How predictive of future behavior and performance are the assessments? Assessment should have reliability, validity, comprehensiveness, and non-compensatory minimum pass points for selected competency areas. Those are words with tremendous import in the world of measurement. Reliability is not just a psychometric impediment to the use of tests imagined by medical educators. It is necessary because making judgments based on "noise" rather than "signal" helps no one. For example, score variability may be caused by the examiner rather than the examinee; these are not reliable measures of examinee competence and therefore cannot be valid. We must be able to depend on the score produced by an assessment program and assure ourselves that the results are consistent and meaningful. The pass/fail decision must be dependable, and we must be able to identify those professionals who do not possess knowledge, skills, and attitudes that are adequate to practice safely and effectively, now or in the future.

We want to know that an assessment system is "comprehensive." Often, accredited medical education serves as a proxy for unassessed areas of interest. The ideal program of assessment would include assessment of all relevant competencies, which is likely to require a system that gathers information longitudinally, as well as at a point in time, particularly for skills and attitudes. We need to know about attitudes because they are displayed in the aggregate of daily experience and are not fully trustworthy in the confined space of an examination. Knowledge and skills are important in the context of practice, more than in the abstraction of a test. So, the full range of relevant competencies includes knowledge, skills, behaviors, and performance in practice environments.

If various competencies are considered to be uniquely important or to have different levels of importance, the assessment system should take this into account in establishing performance standards. Is it acceptable for high performance in one competency domain to offset low performance in another? If so, combining measures in a single decision is sensible; if not, minimum performance standards should be established independently for each competency area.

Current state of the art

Adequacy

Can we balance minimum standards with locally defined societal needs? Yes, if score scales have reasonable precision across their range, passing standards can be adjusted based on local needs, setting the minimum standard at a level that admits the best qualified to meet that region's workforce needs while assuring minimally competent, safe care for patients. Do the standards of the assessment match local workforce and care-access parameters? Yes, if the assessment has score scales with precision across the score range, as is true for many current large-scale exams, the passing standard can be shifted to present a higher bar, with higher quality for patients, or a lower bar, with higher access for patients.

Can we tailor assessment to local cultural and linguistic needs? Yes, as demonstrated by the United States Medical Licensing Examination Step 2 Clinical Skills (USMLE CS) exam, which tests the ability to establish rapport with 12 diverse patients representative of local United States patient care needs, and the ability to communicate effectively in English in a medical care context. Demonstrating

culturally sensitive patient care skills, including communication and interpersonal relationships, is now required for all those seeking licensure in the United States. In another example, specific variations in medical practice can be accommodated by adjusting the content of examinations. Test items relating to tuberculosis screening using purified protein derivative (PPD) skin testing may be relevant in the United States, but they will function poorly in Europe where most patients have received BCG vaccine against tuberculosis and are reactive to the skin test when uninfected. Simply tailoring the test blueprint to these variations assures that the assessment is relevant.

Can we identify and provide assessment that is sensitive to local standards of practice? Yes. NBME international assessment development efforts have repeatedly utilized United States-based blueprints and item banks as a starting point, and in a straightforward process identified the large overlap where United States-based content taxonomy and test items are fully appropriate, and those areas where additions to or modification, deletion, or revision of test content is necessary. NBME has worked with colleagues in France, England, Japan, South Korea, Taiwan, Italy, Portugal, Panama, Switzerland, Australia, New Zealand, and Singapore, among others, to determine local standards of practice. Typically 80 – 95 percent of knowledge content is identical for knowledge exams.

The differences are largely related to disease prevalence issues (for example, tropical disease), variations in health systems (specialty consultants do not provide front line acute care in England), or differences in medical practice (BCG immunization to prevent tuberculosis and PPD skin testing). NBME has demonstrated that the creation of a core knowledge exam augmented by modules that address unique local needs is feasible by repeatedly developing core assessment tools supplemented by locally relevant content.

Robustness

Reliability

Can we depend on the score produced by an assessment program? Is it a stable predictor of future behavior? How consistent are the results of the same individuals taking different forms of a test that assesses the same competency? Reliability is likely to vary across the range of the score scale. For making pass/fail decisions, the issue is less how consistent scores are across the range than how precise and dependable is the pass/fail decision. Formal assessment systems — knowledge tests using multiple choice questions (MCQs), essays, and other test formats — can achieve high reliability; clinical simulation assessment with various simulation formats, including standardized patients (SPs), achieve lower but acceptable levels of reliability.

This is less true of measures that are derived from real world observation. A number of factors may reduce the reliability of these measures, such as rater variability in workplace observation. However, currently used, less formal intramural medical school systems of behavioral observation generally lack reliability as well. Various forms of observational assessment, such as multisource feedback, can be implemented as large scale assessments, but work remains to be done to determine ways to optimize reliability of the scores and decisions arising from these tools.

Validity

Current large-scale assessments — mostly MCQ and CS exams — are often disparaged as having little relevance to subsequent quality of practice. While the evidence base is slim, in aggregate there is a fair amount of evidence that these exams are valid predictors, including extensive research by Tamblyn, Holmboe and Norcini, among others. There is less direct data on simulation/CS exams. There is little validity information on behavioral observation. Demonstrating validity is difficult. We need to do a better job of providing evidence to support the validity of these assessment tools, but I am not aware of any study that provides validity evidence that graduation from an accredited medical school assures effective medical practice.

Figure 1, based on work from Tamblyn et al.,[1] shows how high stakes licensure examination scores can predict future clinical performance.

Comprehensiveness

Many taxonomies of competencies necessary for effective medical practice have been developed around the world. While these have different structures, they are essentially identical in content. In a mapping exercise undertaken by the author and his colleagues, competency standards developed by several entities were compared not at the "header" level of the taxonomy, but at the lowest level of definition. When comparing the taxonomies developed by the Accreditation Council for Graduate Medical Education (ACGME) and the American Board of Medical Specialties (ABMS), the Royal College of Physicians and Surgeons of Canada (CanMeds), the United Kingdom's Good Medical Practice (GMP), the International Institute of Medical Education's (IIME) Global Minimum Essential Requirements, and Institute of Medicine's proposed taxonomy (IOM), 99% of the descriptive text from each document mapped easily to the headers in the other variable taxonomies.

Using the dominant United States taxonomy (core competencies developed by ACGME and ABMS), we do a terrific job in assessing **medical knowledge** in USMLE. We do a fair job — through MCQs, Computer-based Case Simulation (CCS), and CS exams — of assessing patient care, although a number of dimensions, such as procedural skills, are missing. We do a fair and improving job of assessing communication skills through Step 2 CS. We do little to assess professional behavior, systems-based practice, and practice-based learning and improvement. Of course, there is limited evidence that medical schools assess these with any more success than do national assessment systems and scanty evidence that students are held responsible for mastery of the knowledge and skills associated with these competencies.

In order to achieve comprehensiveness, new approaches are needed to demonstrate mastery of competencies that is now inferred from medical school and graduate medical education participation. In our experience, these are likely to require a novel approach to assessment – gathering natural, real world data longitudinally rather than through point-in-time tests. We must seek improved approaches to observational assessment, such as better rater training and recording in greater proximity to observation. Large-scale applications have been implemented. A good example is in the United Kingdom's Foundations assessment. Such systems have the potential to create reliable and valid measures of professional behavior, and the augmentation of assessment of communication

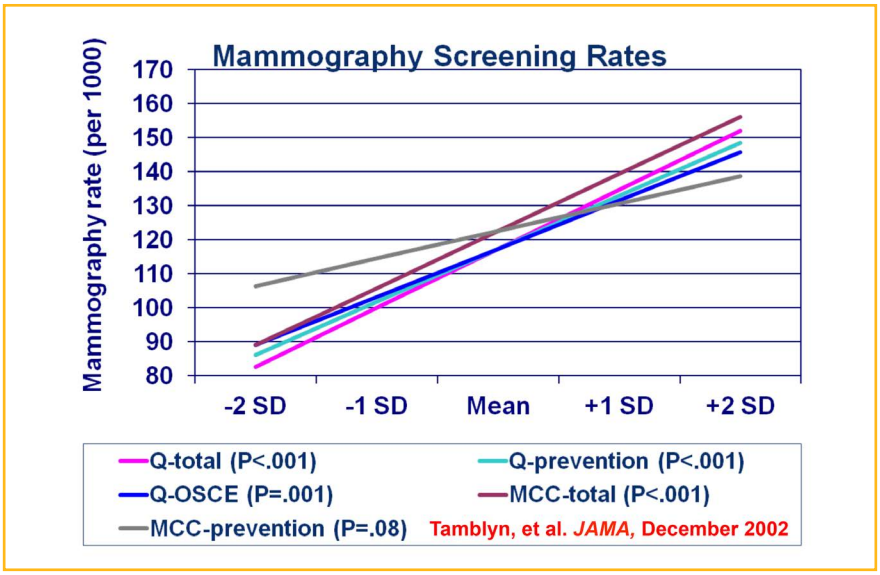

Figure 1. Mammography rate per 1000. Physicians achieving higher scores on examinations had higher rates of mammography screening, an indication of clinical proficiency.

skills. Systems that gather data about educational and practice experience in a standardized format might support inferences about practice-based learning and improvement and systems-based practice. While early stages of development are promising, these kinds of systems require more work.

A system of assessment that would obviate the need to rely on proxy quality measures of medical education institutions would require a sharply different approach. Each student would need to develop a portfolio of evidence that aggregates assessment information from real-world observation with information from test events. The portfolio would need to document achievement of learning milestones relevant to the geography and culture in which practice was envisioned. These assessments would also provide independent evidence of mastery of each relevant competency.

Traditional tests — using well-established means of assessing knowledge and its application, clinical reasoning, and some components of clinical performance [Figure 2] — would be supplemented by a rich, longitudinal record documenting educational and patient care experiences, self-reflection on the learning process, observations of behavior in the real world, and measures of the outcomes of individual performance. New statistical tools are needed to allow aggregations of these data from disparate sources into a cohesive profile of competency that can be shown to be valid and reliable and to establish standards of performance relevant to the planned locus of practice.

In this simplified graph of key results from Norcini et al.,[2] patient outcomes are examined regarding mortality from cardiac disease in Pennsylvania hospitals. It shows that the patients of non-United States citizen medical graduates (IMGs) had lower odds of mortality than patients of United States-trained (USMGs) or United States citizen IMGs. The international graduates came from many medical schools and many countries — 391 schools in 79 countries, so there is certainly a high variability in training. However, after successfully completing United States assessments and United States-based graduate medical education (GME), their patients fared slightly better than graduates of accredited United States medical schools.

Several hypotheses are possible. Students from schools not necessarily meeting United States accreditation standards, when succeeding at a broad assessment of competence, perform as well as or better than students from known, accredited medical

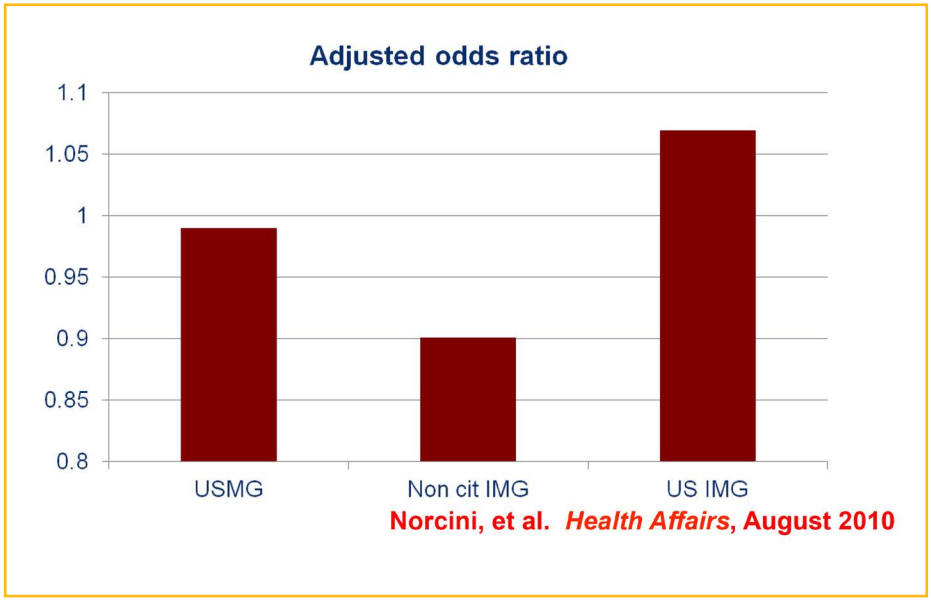

Figure 2. IMG Versus USMG Patient Mortality.

schools. Physicians who complete accredited GME eliminate any differential effect of the undergraduate educational quality.

Conclusion

Let's look back at the initial premise — with adequate, robust, comprehensive assessment systems that are responsive to local and regional needs, the geographic location (or credentials) of the educational institution should be irrelevant in decisions to license or employ health professionals — and determine what conclusion, if any, we can reach. I do not believe we are currently ready to evaluate a physician's fitness to practice in a specific environment in isolation from the proxy information about competence derived from quality measures of the doctor's medical education. But increasingly the world of assessment can provide tools and data that offer individualized assurances of competence.

A future is within reach in which an individual who believes he or she is capable of practicing medicine in a chosen jurisdiction anywhere in the world, without regard to the source of education, could document competency in the core, globally-common domains, augmented by assessment in the domains uniquely relevant to that jurisdiction in a manner that would satisfy the patient protection and quality assurance roles of licensure authorities. It remains a topic of debate as to whether this would be useful or productive for our profession, for our educational institutions, or more importantly, for our patients.

References

1. Tamblyn R, Abrahamowicz M, Dauphinee WD, Hanley JA, Norcini J, Girard G, Grand'Maison P, Brailovsky C. Association between licensure examination scores and practice in primary care. *JAMA*. 2002;228(23):3019 – 3026.

2. Norcini JJ, Boulet JR, Dauphinee WD, Oepalek A, Krantz ID, Anderson ST. Evaluating the quality of care provided by graduates of international medical schools. *Health Aff (Millwood)*. 2010;29(8):1461 – 1468.