Medical Teacher, Vol. 27, No. 3, 2005, pp. 207–213

Taylor & Francis
Taylor & Francis Group

MIRIAM FRIEDMAN BEN-DAVID MEMORIAL ISSUE

Ensuring global standards for medical graduates: a pilot study of international standard-setting

DAVID T. STERN[1,5,6], MIRIAM FRIEDMAN BEN-DAVID[2], ANDRE DE CHAMPLAIN[3], BRIAN HODGES[4], ANDRZEJ WOJTCZAK[5] & M. ROY SCHWARZ[6]

[1]Departments of Internal Medicine and Medical Education, University of Michigan Medical School and the VA Ann Arbor Healthcare System; [2]Tel Aviv University Sackler School of Medicine; [3]National Board of Medical Examiners; [4]Wilson Centre for Research in Education, Department of Psychiatry, University of Toronto; [5]Institute for International Medical Education; [6]China Medical Board of New York

Miriam Friedman Ben-David was involved very early in the process of analyzing data from the IIME assessment. The IIME realized that while student-level standards had been set at a local and even national level, no standards had before been set using international panels with varying specialties. Her ideas about how to construct the standard-setting panels and how to organize the sessions created the foundations of this work. More than that, her grace and intellect in running the standard-setting exercises over 3 days with a group of dedicated and experienced physicians demonstrated not only her wisdom and insight, but also her great abilities as a facilitator, and leader in medical education.

SUMMARY *Increasing physician and patient mobility has led to a move toward internationalization of standards for physician competence. The Institute for International Medical Education proposed a set of outcome-based standards for student performance, which were then measured using three assessment tools in eight leading schools in China: a 150-item multiple-choice examination, a 15-station OSCE and a 16-item faculty observation form. The purpose of this study was to empanel a group of experts to determine whether international student-level performance standards could be set. The IIME convened an international panel of experts in student education with specialty and geographic diversity. The group was split into two, with each sub-group establishing standards independently. After a discussion of the borderline student, the sub-groups established minimally acceptable cut-off scores for performance on the multiple-choice examination (Angoff and Hofstee methods), the OSCE station and global rating performance (modified Angoff method and holistic criterion reference), and faculty observation domains (holistic criterion reference). Panelists within each group set very similar standards for performance. In addition, the two independent parallel panels generated nearly identical performance standards. Cut-off scores changed little before and after being shown pilot data but standard deviations diminished. International experts agreed on a minimum set of competences for medical student performance. In addition, they were able to set consistent performance standards with multiple examination types. This provides an initial basis against which to compare physician performance internationally.*

Background

In 1999, the Institute for International Medical Education (IIME) was founded by the China Medical Board of New York to create, assess, validate and set standards for a set of competences of medical school graduates internationally. With the understanding that each locality would necessarily have its own training, and many outcomes that are not universal, the IIME convened a meeting of international medical education authorities from 12 nations (IIME Core Committee) to identify minimal essential requirements for medical education. After 18 months of meetings and consideration, along with reference to many of the existing national outcome documents, the IIME Global Minimum Essential Requirements (GMER) were completed, consisting of 60 competences in seven domains (Core Committee, Institute for International Medical Education, 2002). These competences were intended to reflect student abilities, but ultimately intended to evaluate the quality of medical schools, not medical students (Stern *et al.*, 2003).

Following the identification of the GMER, the Chinese ministries of education and health invited the IIME to work

Correspondence: David T. Stern, MD PhD, 300 North Ingalls, 7E02, Ann Arbor, MI 48109-0429, USA. Email: dstern@umich.edu

ISSN 0142–159X print/ISSN 1466–187X online/00/000207-213 © 2005 Taylor & Francis Group Ltd
DOI: 10.1080/01421590500129571

with eight leading medical schools in China, using the IIME blueprint to see if it would be possible to evaluate graduating medical students using the newly elaborated GMER schema. After convening an international panel on assessment, the IIME chose three tools for assessment: a multiple-choice examination (MCQ), a multi-station objective structured clinical examination (OSCE), and a longitudinal faculty observation of students in clinical settings. GMER items were then mapped onto these assessment tools, as described elsewhere (Stern *et al.*, 2003).

The exam specifications were then brought to China, where international experts in each of these examination types held training meetings with Chinese counterparts. An examination blueprint was developed, and MCQ items, OSCE stations, and a faculty observation form were developed to assess the various GMER competences. The examination forms were then reviewed by the IIME Core Committee (representing perspectives of 14 countries and six regions of the world), who approved the examination as designed.

In developing a written assessment it was noted that while all schools in China had used the MCQ format in the past, questions were usually of a simple recall type that privileged memorization rather than problem-solving, and so sample case-based questions were supplied for development of a new bank of questions. In developing in-training evaluations, it was noted that while faculty observations of student performance are routine in China, it is usually of the oral examination single observation type, and not longitudinal observation. Thus faculty from each site were trained to teach other faculty at their home institutions on how to use the newly created in-training evaluation forms. Finally, in developing performance-based assessment, it was noted that only two schools had OSCE experience prior to this examination, so workshops were held over a six-month period, in which faculty were trained in the organization, recruitment, training and evaluation procedures for OSCEs.

In October 2003, eight schools in China simultaneously administered the new MCQ and OSCE examinations. The faculty observations occurred over a three-month period from August to October 2003, with a minimum of three unique faculty ratings per student (one per month). The MCQ and OSCE examinations at each school were observed by international experts in medical education. All data were secured by these individuals, and brought back to the IIME for further analysis. Over 200 000 data points were collected on a total of 384 students at eight schools.

Having collected these data, the next step in analysis was to attempt to define an international standard for the large set of data collected from the three forms of students' assessment. There are many different techniques for setting standards described in the literature, and all rely on expert judgments of a group of individuals who can understand and interpret examination performance in light of the developmental, cultural and other characteristics of those being assessed. This is a particularly interesting challenge for an assessment that purports to measure competences that are 'international' and presumably applicable across different contexts. The purpose of this report is to describe the methods and results of our efforts to undertake such a process of setting international standards.

Methods

Participants

In any standard-setting process, a critical element is the choice of the panel that will set the standard (Norcini, 2003). Because the target examinee population in this project was graduating medical students, we selected individuals who had close contact and experience with medical students or early postgraduate students. In addition, diversity of specialty and geography were important. Because the IIME Core and Advisory Committees have both expertise in education and geographic diversity, they were solicited for suggestions on who might comprise this committee. Eleven individuals were chosen representing all regions of the world, and most specialties of medicine (see Table 1).

These individuals were split into two groups (A and B) with one representative from basic sciences, general practice, specialist practice and medical practice in China in each group. All individuals were sent materials in advance of the meeting, including papers on the IIME project, standard setting and sample examination materials from the three exam instruments.

In the opening session of the standard-setting meeting, there was a review of the IIME project, a review of standard-setting methods, and a discussion of what constitutes a 'borderline' student. Following this plenary session, the groups were divided, so that they would have no influence on each other's standard setting decisions. Only after all exercises were completed did the participants see the standards set by the other group.

MCQ standard-setting procedures

An Angoff (Cizek, 1996) standard-setting exercise was completed separately for each of the two panels of judges. In the initial round, panelists were asked to estimate, item by item, the percentage of minimally proficient examinees who would answer each item correctly. These percentages were averaged for each expert judge. For each panel, the mean

Table 1. Standard-setting participants

Name	Country	Specialty
Alejandro CRAVIOTO	Mexico	Pediatrics
Rhee FINCHER	US	Internal Medicine
Gary MIRES	UK	Obstetrics/Gynecology
Jadwiga MIRECKA	Poland	Histology
Nadia MIKHAEL	Canada/Egypt	Pathology/Surgery
Chao NI	China	Cardiology
Joan PRAT	Spain	Physiology
Alberto RESTREPO	Colombia	Orthopedic Surgery
Onike RODRIGUES	Ghana/ Sierra Leone	Pediatrics
Janet SEGGIE	South Africa	Nephrology/ Internal Medicine
Xuehong WAN	China	Gastroenterology/ Internal Medicine

value across judges and items was computed. Finally, the initial Angoff cut-off score on the examination was estimated by averaging estimates across all 11 judges and 143 items.

In a second round, panelists were provided with decile plots for each item. These plots illustrated the percentage of examinees who correctly answered each item of the China Pilot Examination as a function of the decile in which candidates were located. The deciles were formed by dividing the total test population into 10 strata based on their overall test score. As was the case in the initial round, judges were asked to review all items with these supplementary performance data and revise, if so desired, their initial estimate of the percentage of minimally proficient examinees that would correctly answer each item. It is important to stress that revising initial judgments was strictly left up to each panelist.

The final estimate for a given item corresponded to either the revised estimate or the initial value, in the instance where no change was instituted following the second review. As was the case in the initial round, each panelist's cut-off score was obtained by averaging item level estimates. The overall revised Angoff cut-off score was computed by averaging all individual judge ratings across all items.

It is also important to underscore that each panel initially completed a practice round which focused on assigning judgments for a set of 12 exemplars, selected to reflect each GMER category. Following initial Angoff judgments, panelists were invited to discuss their ratings, especially with regard to items that displayed a larger amount of variability. Panelists were then provided with the same 12 items in addition to accompanying decile plots and asked to provide final judgments. An additional discussion also ensued in the hopes that panelists would agree on the general characteristics of the minimally proficient or borderline examinee.

After completion of the Angoff judgments, panelists were asked to follow the Hofstee method (Norcini, 2003) to make judgments about the maximum, minimum and ideal percentage cut-off score and maximum, minimum and ideal percentage of students passing the examination. This process was completed for questions clustered into five major domains of the GMER, with data from student performance readily available for consideration.

OSCE standard-setting procedures

The modified Angoff procedure (Cizek, 1996) was employed for the 10 OSCE stations. All participants attended an introductory session focusing on the concept of a borderline examinee. This session was unique in the sense that panelists referred to borderline characteristics in their own country, thus a comprehensive international description of borderline examinees was generated as a product.

The standard-setting procedures followed the common Angoff procedures:

1. Panel reviewed case material for each station.
2. Panel provided an estimate answering the question: 'How many checklist items will a borderline candidate answer correctly?' Estimates were given separately for the History and Physical examination checklists.
3. Panel discussed estimates with special emphasis on cultural aspects of the station.
4. Panel provided a second estimate.
5. Station score distribution drawn from a sample of students was shown to the panel. The impact of the set standard on student failure rate was discussed.

It was decided to show the actual performance data only after the second and final panel's estimates to avoid influence of the Chinese performance data on the international standard. This standard-setting procedure was repeated for all 10 stations with the two separate groups of panelists.

OSCE global rating at the station level

In each OSCE station, students were rated by an observer on a five-point scale assessing their performance on three dimensions: Interviewing skills, Physical examination and Communication skills. The rating scales for each dimension were anchored with performance descriptors only for points 1, 3 and 5 of the scale. The same rating scale and the same descriptors were repeated for all OSCE stations.

A holistic criterion-referenced standard-setting procedure was employed (Hambleton, 1995), in which panelists were asked to judge on a 1 to 5 scale which point on the scale represents the performance of a minimally competent candidate. Thus, it became a cut-off-point score for each one of the three dimensions of all OSCE stations. Panelist provided a judgment for each of the three dimensions on the 1–5 scale using the first OSCE station as a contextual example. A discussion of the first rating was followed by a second rating, after which the actual performance data were shown to panelists. After the standards were set for all 10 OSCE stations for the history and physical examination, panelists again reviewed their estimates on the three dimensions and provided a third rating. The third rating allowed panelists to review the generic aspect of the global rating in light of the review process for the 10 OSCE stations. These standard-setting procedures were estimated separately by the two groups of panelists.

Faculty observation standard-setting procedures

A similar holistic criterion-referenced standard-setting approach was employed with the faculty observation scale. In this case there were three domains:

- Professionalism (contained seven major items);
- Communication skills (contained six major items);
- Scientific approach (contained three major items).

On each of the domain items, students were rated on a 1–5 scale. Each item domain score was anchored for behavioural descriptors solely for points 1, 3 and 5 of the scale. The panelists were asked to provide a holistic profile (Hambleton, 1995) rating of a minimally competent candidate considering the scoring descriptors for each item domain. The relative importance of the item to the whole domain concept was also taken into account. In some instances, due to the international aspect of the task, panelists had to ignore the item descriptors within an item domain, and conceptualize

descriptors that might fit their own culture. Panelists provided for each domain a sum of their judgment on all items. The sum was divided by the number of items in the domain and averaged across panelists. Consequently a standard was set on each domain. All three standard domains were considered as conjunctive.

A discussion of the first rating was followed by a second rating. Once the standard was set, the score distributions on each of the three domains for all students from the eight Chinese medical schools were shown to the panelists. The actual student performance data of the Chinese distribution were not shown to the panelists after the first rating to avoid unduly influencing the international standard. This process was repeated with two separate groups of panelists for the purpose of validating the obtained standards.

Results

MCQs

Figure 1 provides initial Angoff cut-off score values for both panels of judges across five domains of multiple-choice assessment. Panel A cut-off score values ranged from 58.5% for the Population and Health Systems GMER domain to 66.8% for the Clinical Skills domain. Panel B cut-off score values ranged from 62.7% for the Scientific Foundation of Medicine domain to 68.3% for the Professional Values, Attitudes, Behaviour and Ethics domain. Overall, initial Angoff cut-off score values varied from 62.6% for the Population Health and Health Systems domain to 67.6% for the Professional Values, Attitudes, Behaviour and Ethics domain.

Revised Angoff cut-off score values from Panel A ranged from 56.1% for the Scientific Foundation of Medicine to 59.8% for the Clinical Skills domain while Panel B cut-off score values varied from 54.6% for the Scientific Foundation of Medicine domain to 58.0% for the Clinical Skills domain. Overall revised Angoff cut-off score values ranged from 55.3% for the Scientific Foundation of Medicine domain to 58.8% for the Clinical Skills domain.

Differences in Angoff cut-off score values between both panels were less than 2% for six domains, whether

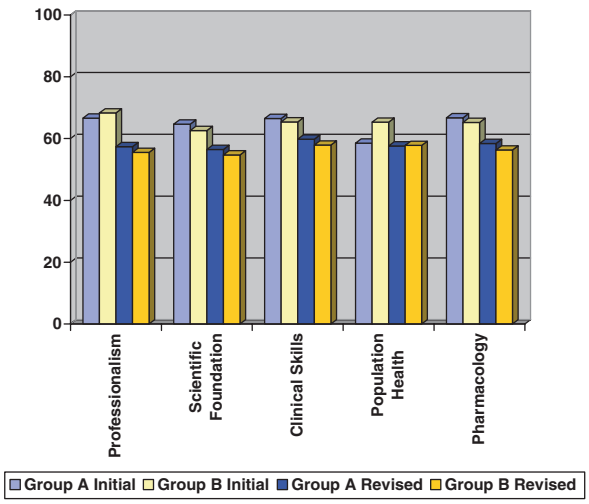

Figure 1. MCQ Angoff standards in five domains.

for initial or revised judgments. The largest difference between both panels was noted for the Population, Health and Health Systems domain where cut-off scores varied by 6.8%. It is important to indicate that this domain contained very few items (13) and as such, the difference in cut-off score between both panels amounts to less than one item.

Hofstee Judgements for both groups are shown in Table 2. As with the Angoff judgments, there was close similarity in the cut-off points identified by each group independently.

OSCE

As with the multiple-choice examination questions, all panelists provided very similar cut-off points for the OSCE examinations both by station and by global rating (Figures 2–5). Initial ratings and subsequent ratings were quite similar, without a trend to increase or decrease the cut-off score. There were, however, consistent trends to decreasing variation in the panelist scores comparing first and second ratings (Table 3).

Faculty observations

As with both MCQs and the OSCE ratings, the parallel panels provided similar standards. There was no consistent trend in revision of faculty observation standards, except that the initial standards were adjusted, on average, very little by either group (average of 0.1 points on the five-point scale). However, the standard deviations within group and among all raters fell on revision from 0.55 to 0.38 (see Figure 6 and Table 3).

Discussion

This paper describes the first effort to set an international standard for three assessment formats based on a set of pre-defined global physician competences. This project is, necessarily, the first step in validating the concept of international outcome standards for physician competence. With further administration of this assessment prototype in other countries, aggregate performance of students from different countries may provide better international performance data for future standard-setting procedures. Ultimately, validation of the international standard depends on collecting at least one more set of data from China, and also data from students in other countries.

The details of this manuscript outline the remarkably comparable standards set between two independent panels across a wide degree of geography and specialty. Panelists, in the process of standard-setting, reviewed the quality of the test materials, which resulted in only a few suggested minor changes—an indication of the high quality of the assessment materials employed in this study. The success of this project helps support the idea that it is possible to undertake a process of an international standard-setting for medical student assessments.

One concern about this process has been whether local variation in practice can adequately be captured in an international program. First, it should be re-emphasized that the Global Minimum Essential Requirements are not

Table 2. Hofstee Judgments

		Min % Corr.	Max % Corr.	Min % Fail	Max % Fail
Professionalism	Panel 1	44.00	64.00	6.20	21.00
	Panel 2	53.33	74.17	7.83	24.17
	Mean	49.09	69.54	7.09	22.73
Scientific foundations	Panel 1	49.00	67.00	7.40	25.00
	Panel 2	49.17	70.83	10.83	24.17
	Mean	49.09	69.10	9.27	24.55
Clinical skills	Panel 1	51.00	69.00	6.60	19.00
	Panel 2	55.83	75.00	8.67	18.33
	Mean	53.64	72.27	7.73	18.64
Population health	Panel 1	44.00	63.00	6.40	21.00
	Panel 2	48.33	65.83	7.00	21.67
	Mean	46.36	64.55	6.73	21.36
Pharmacology	Panel 1	47.00	65.00	6.40	20.00
	Panel 2	51.67	70.83	8.67	20.00
	Mean	49.55	68.18	7.64	20.00

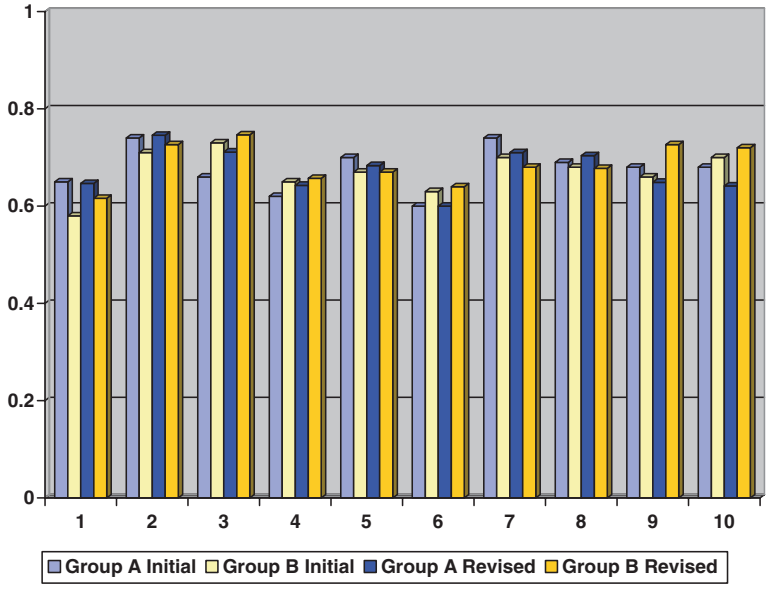

Figure 2. OSCE content standards for 10 OSCE stations.

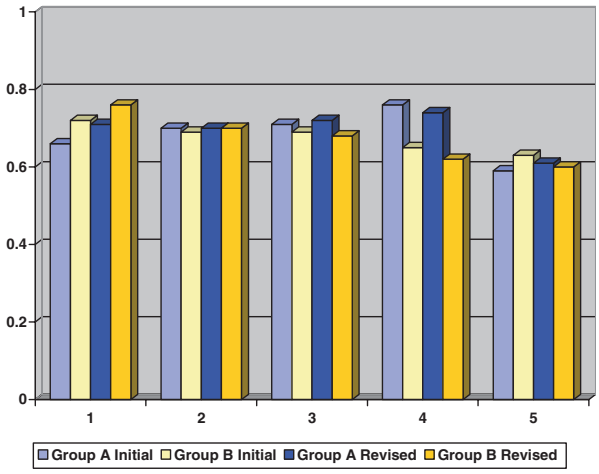

Figure 3. OSCE physical examination standards for five stations.

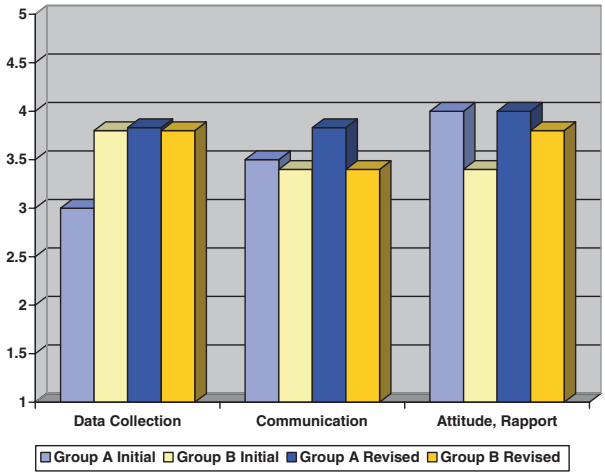

Figure 4. OSCE global content standards for three items (10 stations).

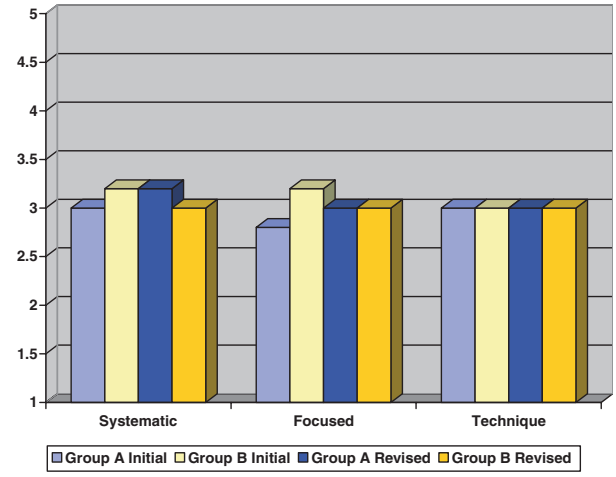

Figure 5. OSCE global physical examination standards for three items (four stations).

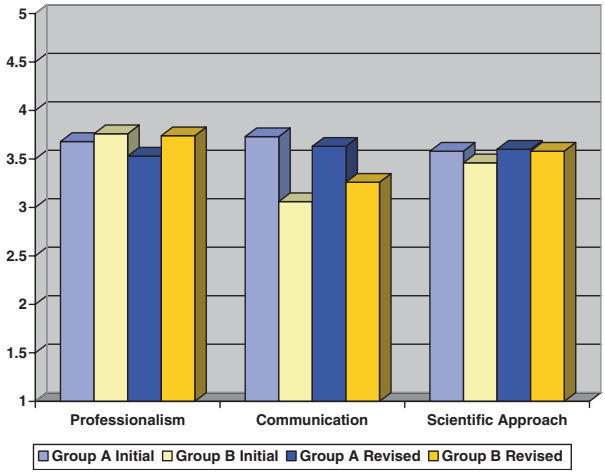

Figure 6. Faculty observation standards in three domains.

Table 3. Average standard deviations

	First rating	Second rating
MCQ		
A	0.12	0.07
B	0.14	0.08
OSCE		
History—A	7.67	3.27
History—B	9.30	5.52
Physical Exam—A	3.20	1.05
Physical Exam—B	2.55	1.92
Global OSCE rating		
Communication—A	0.17	0.34
Communication—B	0.51	0.41
Examination—A	0.25	0.14
Examination—B	0.18	0.00
Faculty observations		
Professionalism—A	0.4	0.19
Professionalism—B	0.73	0.62
Communication—A	0.48	0.28
Communication—B	0.47	0.39
Scientific thinking—A	0.56	0.27
Scientific thinking—B	0.39	0.38

Note: MCQ = Multiple-choice examination; OSCE = objective structured clinical examination; A, B = standard-setting Group A and Group B.

intended to describe the complete doctor (medical graduate) in any country. These competences represent perhaps 40–60% of what constitutes competence for practice in any one location because of reasonable differences in local, regional and national healthcare expectations and practices. Second, most of the dimensions are to be measured at a level that allows for cultural relativism. 'Effective patient communication' may vary from country to country, but when measured by patients or external observers with knowledge of the individual patients' expectations, a consistent measure of competence can be observed.

Ultimately, we hope that the assurance of high-quality medical graduates is an important component of ensuring high-quality care for patients. Naturally, the domain of 'internationalism' and 'globalization' raises concerns for many people. For example, it has been argued that a freer exchange of health professionals between countries might lead to a net loss of highly trained health professionals from developing countries to Western countries facing personnel shortages. This scenario creates a terrible cycle of expense and loss in countries where healthcare needs are great (Joint Learning Initiative, 2004). For those who look for evidence of such a claim, consider the reverse. Can a doctor with low levels of competence provide good care? To the concern that higher quality education will lead to 'brain-drain' from developing countries, or to a multi-tiered medical care system in developing countries, it should be remembered that the GMER ensures only minimum competence, and would not provide evidence of the ability of a graduate to practice in any location, including his/her country of origin. In addition, the GMER, in domains of professionalism and population health, evaluates the degree to which the graduate is committed to basic principles of the profession including service to those in most need, and a dedication to serving local populations. Finally, issues of competence are separable from those of human resources, which should be addressed by multifaceted political and economic initiatives, and not through ignoring standards of competence (Joint Learning Initiative, 2004).

The assessment of global competences of medical graduates is in its infancy. Standard-setting for these assessments is even less well developed. With evaluation of schools in many countries, the validation of the principles of assessment and the setting of international standards, we hope to further refine both the standards themselves and the procedures for ensuring them. Ultimately, we hope that such standards will lead to both better health, and a firmer commitment on the part of the profession to ensuring and maintaining the competence of physicians worldwide, recognizing that there are many historical, political and economic issues that require careful attention to the process.

Practice points

- Student-level standard setting is a common practice in the setting of high-stakes assessment. Standards have been set at local and national levels but not international levels.
- This study describes a process for international standard-setting of student performance, with evidence that a diverse group of individuals can achieve similar standards.
- Further validation of these standards will be necessary with additional examinations in other countries.

Notes on contributors

David T. Stern, MD PhD, is Associate Professor of Internal Medicine and Medical Education at the University of Michigan Medical School. He is the director of assessment for the IIME, and coordinated the assessment of global minimum essential requirements outlined in this manuscript.

Andre De Champlain is Senior Psychometrician at the National Board of Medical Examiners. He has published extensively in both the psychometric and medical education literatures where his areas of interest are performance assessment, standard setting and cross-cultural measurement. He has consulted on a number of international initiatives in Japan, China, Panama, France and the UK.

Brian Hodges is Associate Professor and Vice-Chair in the Department of Psychiatry at the University of Toronto where he is also Director of the Wilson Centre for Research in Education. He has undertaken research and consulted widely on performance-based assessment for organizations such as the International Medical Graduate Program, Medical Council of Canada, Royal College of Physicians and Surgeons of Canada, American Board of Psychiatry and Neurology and other international organizations in Japan, Pakistan, China, Poland, New Zealand, Switzerland and Israel.

Andrzej Wojtczak, MD, is Director of the Institute for International Medical Education in New York, Professor Emeritus in the School of Public Health and Social Medicine in Warsaw, and visiting Professor at Kwansei Gakuin University, Sanda, Hyogo, Japan. Previously, he was Director of the WHO Research Center for Health in Kobe, Japan, and held the position of AMEE President.

M. Roy Schwarz is currently President of the China Medical Board of New York, Inc. He is the former Funding Director of the WAMI Program, Dean, School of Medicine, University of Colorado and Senior Vice-President of American Medical Association. He has published on tele-education, decentralized medical education and the Global Minimum Essential Requirements of medical education.

References

Cizek, G.J. (1996) Standard setting guidelines, *Educational Measurement Issues and Practice*, 15, pp. 12–21.

Core Committee, Institute for International Medical Education (2002) Global minimum essential requirement in medical education, *Medical Teacher*, 24, pp. 130–135.

Hambleton, R.K. (1995) Setting standard on performance assessments: promising new methods and technical issues, paper presented at the meeting of the American Psychological Association, New York, August.

Joint Learning Initiative (2004) Human Resources for Health. President and Fellows of Harvard College.

Norcini, J.J. (2003). Setting standards on educational tests, *Medical Education*, 37, pp. 464–469.

Stern, D.T., Wojtczak, A. & Schwarz, M.R. (2003) The assessment of global minimum essential requirements in medical education, *Medical Teacher*, 25, pp. 589–595.