Setting school-level outcome standards

David T Stern,[1,2] Miriam Friedman Ben-David,[3] John Norcini,[4] Andrzej Wojtczak[5] & M Roy Schwarz[6]

BACKGROUND To establish international standards for medical schools, an appropriate panel of experts must decide on performance standards. A pilot test of such standards was set in the context of a multidimensional (multiple-choice question examination, objective structured clinical examination, faculty observation) examination at 8 leading schools in China.

METHODS A group of 16 medical education leaders from a broad array of countries met over a 3-day period. These individuals considered competency domains, examination items, and the percentage of students who could fall below a cut-off score if the school was still to be considered as meeting competencies. This 2-step process started with a discussion of the borderline school and the relative difficulty of a borderline school in achieving acceptable standards in a given competency domain. Committee members then estimated the percentage of students falling below the standard that is tolerable at a borderline school and were allowed to revise their ratings after viewing pilot data.

RESULTS Tolerable failure rates ranged from 10% to 26% across competency domains and examination types. As with other standard-setting exercises, standard deviations from initial to final estimates of the tolerable failure rates fell, but the cut-off scores did not change significantly. Final, but not initial cut-off scores were correlated with student failure rates ($r = 0.59$, $P = 0.03$).

DISCUSSION This paper describes a method to set school-level outcome standards at an international level based on prior established standard-setting methods. Further refinement of this process and validation using other examinations in other countries will be needed to achieve accurate international standards.

KEYWORDS schools, medical/*standards; education, medical, undergraduate/*standards; educational measurement/*standards; China; clinical competence/*standards; reference standards; international co-operation; feasibility studies.

*Medical Education 2006; **40**: 166–172*
doi:10.1111/j.1365-2929.2005.02374.x

INTRODUCTION

National and international rhetoric on outcome-based education has been running far ahead of both practical examples of implementation and the educational methods necessary to ensure the quality of such standards. Yet individual medical schools,[1] countries[2] and international organisations[3,4] are working to implement the necessary means through which educational outcomes can be assessed.

At the international level, the Institute for International Medical Education (IIME) outcome-based assessment project has identified a set of internationally defined outcomes for undergraduate medical education[3] and the means with which to evaluate these outcomes.[5] The IIME implemented the outcomes assessment in a pilot project in China during the autumn of 2003. The underlying principle of the IIME project is that students are the outcome of medical education; therefore the aggregate performance of graduating students defines the areas of strength and weakness in the medical education experience provided by the school.

[1]Departments of Internal Medicine and Medical Education, University of Michigan Medical School, Ann Arbor, Michigan, USA
[2]Veterans Administration Ann Arbor Healthcare System, Ann Arbor, Michigan, USA
[3]Tel Aviv University Sackler School of Medicine, Tel Aviv, Israel
[4]Foundation for the Advancement of International Medical Education, Philadelphia, PA, USA
[5]Institute for International Medical Education, White Plains, NY, USA
[6]China Medical Board of New York, New York, New York, USA

Correspondence: David T Stern MD, PhD, 300 North Ingalls, Room 7E02, Box 0429, Ann Arbor, Michigan 48109-0429, USA.
Tel: 00 1 734 615 8341; Fax: 00 1 734 936 8944; E-mail: dstern@umich.edu

Overview

What is already known on this subject

National and local authorities are calling for evidence of educational outcomes. Aggregating student examination scores to the level of schools leaves open the question of how to set school-level performance standards.

What this study adds

This study describes a method for setting international school-level outcome standards and how those standards were applied in the context of an 8-school examination in China.

Suggestions for further research

Further validation of this method and the standards set by this method are necessary. Additional research on school-level outcomes will be needed as calls for school accountability continue.

Based on existing work from many countries (General Medical Council, UK; CANMeds, Canada; Scottish Doctor, Scotland, UK; Association of American Medical Colleges, USA) an international panel of medical education experts worked over 18 months to develop a set of 60 minimum and essential outcome-based competencies for graduating medical students.[3] These competencies were written such that they could all be measured, and were categorised into 7 domains:

1 professional behaviour and ethics;
2 scientific foundations;
3 communication skills;
4 clinical skills;
5 population health;
6 information management, and
7 scientific thinking.

A group of international assessment experts subsequently met to identify the best methods for evaluating these competencies and created a blueprint for the examination using 3 assessment methods: the multiple-choice question (MCQ) examination, the objective structured clinical examination (OSCE), and longitudinal faculty assessment.[5] This blueprint was then used to create a set of examinations in China. This project examined all Year 7 (graduating) students at 8 leading medical schools in China, using a 150-item MCQ examination, a 15-station OSCE, and a 16-item faculty observation form used at least once per month for 3 months on each student. Each assessment type measured multiple domains of competence (Table 1). When this set of assessments had been completed, the IIME team had over 200 000 data points on graduating students at 8 schools, but no standard against which to measure these outcomes. Using established educational methods, student-level standards were set for these examinations.[6] Although this was the first set of international examination standards set in medical education of which we are aware, the process went at least as smoothly as our experiences with similar standard-setting processes at our individual institutions.

As much as student-level standards are essential to determine if graduates are to attain the expected competences for the next phase of their training, it is also essential to evaluate the quality of the school and its capacity to deliver an educational programme that facilitates the attainment of outcomes. Consequently, outcome-based standards seek student-level standards as well as school-level standards. While student-level standard setting has been successfully performed in many contexts, we are not aware of any efforts to provide international school-level standards. The purpose of student-level standard setting is to identify the cut-off point above which a *student* can be considered competent. The purpose of school-level standards is to identify the cut-off point above which the *school* can be considered competent. The assessment on which the student cut-off score is applied is a particular examination or component of an examination. The assessment on which the school cut-off score is applied is an aggregate of student performances on a

Table 1 Methods used to examine IIME competency domains

	MCQ	OSCE	Observation
Professionalism	X	X	X
Scientific foundations	X		
Communication		X	X
Clinical skills	X	X	
Population health	X		
Information management		X	
Scientific thinking			X

MCQ = single best answer multiple-choice questions
OSCE = objective structured clinical examination
Observation = longitudinal faculty observations in clinical settings

particular examination or component of an examination.

For example, imagine an OSCE on which there is a communication skills composite score (averaged over multiple stations) ranging from 1 (low performance) to 5 (high performance). A student-level standard-setting panel reviews the OSCE stations and scale, and arrives at a cut-off score of 3.9. Any individual student who scores above this would be considered to meet the standard; below it, the student would be considered to have an educational weakness that needs attention. At the school level, an aggregate of student performances might indicate that 1%, 10% or 50% of students fall below the student-level standard. At what point would you consider the aggregate student performance to indicate that the school does not meet standards? Certainly, if 1 or 2 students fall below standards, the school should not be held accountable. But what if 10%, 20% or 50% of students fall below standards? At some point along this continuum, the school-level standard should be set.

The purpose of this paper is to describe a process through which this school-level standard can be set using pilot data and information from the IIME pilot examination in China.

METHODS

Participants

In any standard-setting process, a critical element is the choice of those who will set the standard.[7] In this project, the target was the competency of the school, so we selected individuals who had close contact with and substantial experience in evaluating medical schools. Having served as deans of medical schools, health ministry advisors, or on external review committees, most panellists had had the opportunity to observe and evaluate a wide range in quality of medical schools internationally prior to this project. In addition, panellists were chosen to create geographic diversity. The 16 individuals had been employed as educational leaders in 13 different countries – some were doctors, some were in the basic sciences, and all were experts in medical education. This group had worked together since 2000, on both writing the IIME Global Minimum Essential Requirements[3] and guiding their assessment.

All individuals were sent materials in advance of the meeting, including papers on the IIME project and standard setting, and sample examination materials from the 3 examination instruments.

Standard-setting procedures

The opening session of the meeting included a review of the IIME project, a review of standard-setting methods, and a description of the student-level standards set by a different panel the previous month.[6] In this presentation, the details of participants, standard-setting processes and outcomes were provided. Decisions about which cut-off scores to use (Angoff, Hofstee or combined[7]) were reviewed, and the student-level standards were approved by the Core Committee.

The key task for setting school-level standards was to get a committee of international experts to simultaneously consider the competency domains and examination items, and the percentage of students who could fall below a cut-off score while still allowing the school to be considered as meeting competencies.

The standard-setting process at the school level was comprised of 2 main steps.

1 Committee members must develop an understanding of:
- what constitutes a borderline school;
- the relative difficulty for a borderline school to achieve acceptable standards in a given competency domain;
- the assessment materials used to assess the outcome domain, and
- student-level standards set on assessment instruments.

2 Committee members must estimate the percentages of students falling below the standard in a given domain that are tolerable in a borderline school.

Step 1

What constitutes a borderline school?

Author MFB-D facilitated a discussion of the 'borderline school' in a manner similar to that used during Angoff method discussions of the 'borderline student'.[7] Most committee members had served on accreditation committees, and the borderline discussion quickly generated a profile of the borderline school agreed by all international participants. For example, borderline schools might exist because they accept students with lower initial abilities, because their programmes of education are incomplete, or because the school invests inadequate resources in education. Keeping the performance of students from a borderline school in mind was a critical element of the standard-setting process.

The relative difficulty for a borderline school to achieve acceptable standards for a given domain

In standard-setting procedures, panellists are provided with information that helps them decide how realistic their estimates might be. The committee members were asked to consider the question: In a borderline school, what will be the likelihood of achieving minimal competency in each of the domains? The exercise was an attempt to help committee members consider the constraints of borderline schools, recognise the existing realities and recognise that not all domains might be fully attainable. In doing so, committee members made decisions about issues such as what is 'attainable' and what is 'tolerable' for rating each item domain.

Assessment materials

Committee members reviewed all test items assigned to each domain by assessment type. If there were multiple methods for assessing a single domain, this process was repeated for each measurement type (MCQ, OSCE, faculty observation).

Student-level standards set on the assessment materials for each domain

Committee members reviewed the student-level standards (cut-off score), for each of the assessment instruments.

Step 2

1 Committee members provided initial estimates of the percentage of students falling below the standard that is tolerable in a borderline school for each of the assessment instruments per domain.
2 Initial ratings were projected on a screen and discussion of high and low ratings took place.
3 Anonymous school-level data from 8 leading schools in China were shown, indicating the consequences of the school-level cut-off scores on the percentages of students falling below the domain standard for each school (Fig. 1, Table 2).
4 Final (revised) ratings were discussed and defined.
5 Ratings were applied to school-level data, with fewer failures than the cut-off score constituting 'strength' and more failures constituting a 'need for improvement'.

The standard-setting process took 3 days to complete. Data analysis was performed using SPSS. Traditional measures of reliability were not possible to calculate with only 1 observation per domain per rater. Comparisons between school-level standards,

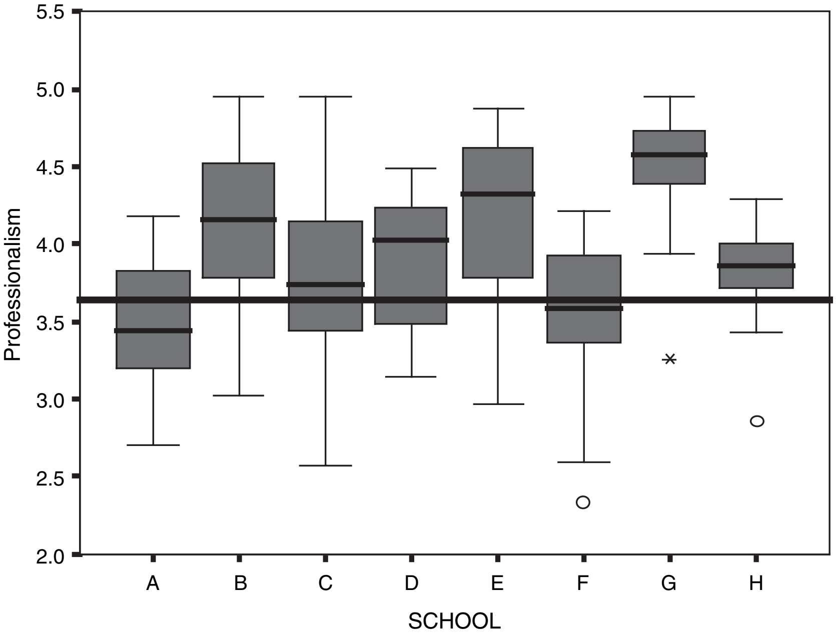

Figure 1 Sample consequential data shown after first round of rating. The heavy line on the graph represents the student-level standard set by a prior panel.

student-level standards, and student failure rates were performed using Spearman correlation coefficients.

Table 2 These data were shown simultaneously to the school-level standard-setting panel to aid in consideration of the consequences of 'tolerable' failure rates

School	Percentage of students below standard
A	69
B	19
C	40
D	32
E	15
F	50
G	1
H	22

RESULTS

Across all modes of examination, initial tolerable failure rates ranged from 10% to 26%, reflecting the raters' estimation of how well the domain was sampled, the accuracy of the assessment method, and the allowable percentage of failing students in a competent school. (Table 3) These initial cut-off scores were set after review of domains and examination materials, but before reviewing student or school failure rates. With a single exception, there was little difference between initial and subsequent ratings after reviewing school failure rates.

In this standard-setting exercise, standard deviations generally decreased from the initial to the final ratings, demonstrating group consensus development. Raters were expected to consider variation in the student-level cut-off scores as they set school-level standards. To measure this effect, we found strong correlations between the initial and revised school-level cut-off scores and student-level cut-off scores (initial cut-off scores $r = -0.48$, $P = 0.09$; revised cut-off scores $r = -0.69$, $P = 0.01$) Raters were also expected to review the examination material and adjust school-level cut-off scores in some relation to expected actual student failure rates. The Spearman correlation coefficient between initial school-level cut-off scores and student failure rates by domain was not significant (Spearman $r = 0.374$, $P = 0.21$, $n = 13$ ratings); however, the relationship between revised school-level cut-off scores and student failure rates was substantial (Spearman $r = 0.59$, $P = 0.03$). Average rater stringency ranged from 12.5% to 20% tolerable failure rates, indicating a fair degree of agreement on the overall range of acceptable failure in a competent institution.

DISCUSSION

This paper describes a method that can be used to set school-level outcome standards at an international

Table 3 Initial and revised school-level ratings

Examination type	Domain	Initial Mean (SD)	Revised Mean (SD)
MCQ			
	Professionalism	25.7 (6.7)	24.2 (6.1)
	Scientific foundation	22.7 (7.2)	21.4 (5.5)
	Therapeutics	19.1 (4.4)	18.6 (3.9)
	Clinical skills	11.36 (3.2)	12.27 (4.1)
	Population health	19.1 (3.0)	19.1 (3.0)
OSCE			
	Communication skills		
	Data collection	11.4 (3.2)	11.8 (3.4)
	Patient communication	15.0 (5.8)	13.2 (5.6)
	Professionalism		
	Attitude and rapport	16.3 (5.3)	9.2 (4.2)
	Clinical skills		
	Physical examination	10.5 (3.5)	10.0 (3.2)
	Information management		
	Literature searching	9.6 (4.2)	9.6 (4.2)
Faculty observations			
	Professionalism	11.8 (3.4)	14.1 (4.9)
	Communication	17.3 (4.9)	17.7 (4.1)
	Critical thinking	15.5 (4.1)	15.5 (3.4)

level. While the examination methods and procedures will need further refinement and improvement, along with validation using other examinations in other countries, the procedures set out in this paper provide a blueprint for how the profession can achieve internationally agreed standards for performance.

Although these procedures are not designed to produce results adequate for high-stakes, individual student reliability, medical students can use data from an examination like this to determine whether their performance approximates international standards. Medical schools can use data from this evaluation to determine both the baseline strengths and weaknesses of their programmes, as well as the impact of educational interventions with follow-up evaluation. Ministries and medical school organisations can use aggregate results across schools to determine funding priorities across institutions without concern for local biases.

The strengths of this process include its similarity to other standard-setting procedures, with the extension to a larger unit of analysis. The procedure itself is familiar to medical educators, but the task is a bit more complex in that it requires panel members to simultaneously consider 3 dimensions: the domain assessed; student performance, and the reflection of aggregate performance on school quality. Regardless, this panel was able to quickly understand and engage in this task, providing reasonable and consistent ratings of acceptable examination performance with which to rate schools.

Setting standards is not an exact science. It relies upon expert judgement and the combined opinions of multiple individuals. Therefore, the most important component in the process is the selection of the experts who set these standards, and we took great care to invite individuals who had the necessary expertise. That said, the standards reflect the opinions of these individuals, and future standard-setting with other panels is required for full validation. In addition, the standards were set on assessment instruments that have varying degrees of precision. While the panellists took this examination-type variability into account in setting standards, further work on developing reliable instruments will improve the overall process. Ultimately, this procedure and the IIME project itself require further validation with additional testing material from other countries. It is likely that this initial set of school-level cut-off scores will be adjusted over time with the addition of new data. However, given the small adjustments in scores after viewing consequential data, it is unlikely that a panel similar to that described in this paper will come to significantly more stringent or lenient conclusions about school-level performance.

External independent assessment of competencies provides valuable information for schools and countries that are being held increasingly accountable for the future performance of their graduates.[8] These outcomes are important because of the degree to which they meet an international standard. The standard-setting procedures outlined in the paper provide educators with the tools with which to assess school-level outcomes. Ensuring international standards in medical schools is a critical step towards developing a competent global health care workforce and to promoting quality medical care for patients worldwide.

Contributors: DS and MFBD contributed to the study design, data collection, data analysis and the drafting of the manuscript. JN contributed to the study design, data analysis, and the drafting of the manuscript. AW and MRS contributed to the study design, data collection, and drafting of the manuscript. All authors participated in the writing of the paper.
Acknowledgements: the standard-setting panel used for this work was the Core Committee of the Institute for International Medical Education. The participants were: Elizabeth Armstrong (USA); Raja Bandaranayake (Bahrain/Australia); Albert Oriol Bosch (Spain); Alejandro Cravioto (Mexico); Joseph Gonnella (USA); John Hamilton (UK); Ronald Harden (Scotland); David Hawkins (Canada); Jose Patino (Columbia); Pablo Pulido (Venezuela); M Roy Schwarz (USA); David Stern (USA); J P Van Niekerk (South Africa); Andrzej Wojtczak (Poland); Tai Yao (China), and Tongfu Zhou (China).
Funding: the China Medical Board of New York.
Conflicts of interest: none.
Ethical approval: not applicable for research using anonymous student data.

REFERENCES

1 Harden R, Crosby JR, Davis MH, Friedman Ben-David M. Outcome-based education from competency to meta competency. *Med Teacher* 1999;**21**:546–52.
2 Royal College of Physicians and Surgeons of Canada. CanMEDS Framework. http://rcpsc.medical.org/canmeds/index.php. [Accessed 20 December 2004.]
3 Core Committee Institute for International Medical Education. Global minimum essential requirement in medical education. *Med Teacher* 2002;**24**:130–5.

4 World Federation for Medical Education. WFME Global Standards for Quality Improvement in English. http://www.wfme.org/. [Accessed 3 June 2004.]

5 Stern DT, Wojtczak A, Schwarz MR. The assessment of global minimum essential requirements in medical education. *Med Teacher* 2003;**25**:589–95.

6 Stern DT, Friedman Ben-David M, Hodges B, De Champlain A, Wojtczak A, Schwarz MR. Ensuring global standards for medical graduates: a pilot study of international standard-setting. *Med Teacher* 2005;**27** (3):207–13.

7 Norcini JJ. Setting standards on educational tests. *Med Educ* 2003;**37**:464–9.

8 Schwarz MR, Wojtczak A. Global minimum essential requirements: a road towards competence-oriented medical education. *Med Teacher* 2002;**24**:125–9.

Received 3 February 2005; editorial comments to authors 7 April 2005; accepted for publication 27 May 2005