ELSEVIER

THE CEYLON MEDICAL JOURNAL

Established 1887
The Official Publication of the Sri Lanka Medical Association
Volume 51, No. 3, September 2006
Quarterly ISSN 0009-0875

All communications should be addressed to **The Editors, CMJ**

Continued Overleaf

Standard setting for medical exams

'Standard' is defined as "any measure by which one judges a thing as authentic, good, or *adequate*, or the degree to which it is authentic, good, or *adequate*. Standard applies to any authoritative rule, … or measure used to determine the … value, quality, level, or degree of a thing" [1].

In examinations we are 'measuring' the candidates. 'Standard setting' is the process of deciding the pass mark or a cut-point for an examination. It usually involves a group of experts. The judgement of the experts in the group is used to set the standard.

Why standard setting?

Why is standard setting necessary? The reasons relate to the level of difficulty of the exam. If an unduly difficult exam is set, some students who might otherwise have passed will fail, when the pass mark is always fixed at 50%. If an unduly easy exam is set, some students will pass who might otherwise have failed. Standard setting allows for variations in difficulty of an exam to be taken into consideration in deciding the pass mark.

Standard setting ensures that the student who passes an exam has mastered the core knowledge and competencies that are assessed by that exam. The aim of standard setting is to separate the 'competent' candidates from the 'non-competent'.

Absolute and relative standard setting

There are two types of standard setting: absolute or criterion-referenced standard setting and relative or norm-referenced standard setting.

1. Absolute or criterion-referenced standard setting

A candidate passes the assessment who achieves the level of competence set by the experts. If all the candidates achieve the desired competence level, all will pass the exam, and, if no candidates achieve the set competence level, no one will pass.

This method of standard setting should be used when the examination outcome is related to accreditation or promotion of the candidate to a higher level of learning or training.

There are two methods of applying a criterion-referenced standard (Table 1).

(a) Conjunctive standards: To pass the whole exam, the candidate needs to achieve the standard for each test component; e.g. each Objective Structured Clinical Examination (OSCE) station.

Zulfiqar Ahmed Bhutta FRCPCH, PhD
Karachi, Pakistan

Samiran Nundy FRCS, FRCP
New Delhi, India

N Medappa MD
New Delhi, India

Jane Smith BA, MSc
London, UK

Anita KM Zaidi MMBS, SM
Karachi, Pakistan

David Warrell MD, FRCP
Oxford, UK

Published by

Elsevier
A Division of Reed Elsevier India Pvt. Ltd.
Sri Pratap Udyog, 274, Captain Gaur Marg,
Sriniwaspuri,
New Delhi – 110065, India.
Tel: +91-11-2644 7160-64
Fax: +91-11-2644 7156
Website: www.elsevier.com

Printed by

Ananda Press,
Colombo 13, Sri Lanka
Tel: +94 11 2435975

This journal is indexed in BIOSIS, CAB International, EMBASE, and Index Medicus

(b) Compensatory standards: The candidates can score below the set standard in one assessment component and compensate for this by scoring highly in another component and pass the overall assessment.

2. Relative or norm-referenced standard setting

Norm referencing involves grading and ranking the candidates. A fixed percentage of candidates (e.g. the top 60%), as determined by the experts, will pass the assessment, irrespective of the level of competence they have shown at the assessment. The implication is that when relative standards are applied some incompetent candidates may pass or some competent candidates may fail.

Norm referencing is helpful if the aim of the exam is to select the best candidates (e.g. selection to a medical faculty or a postgraduate course) and in awarding prizes or medals.

From an educational standpoint, however, norm-referenced standard setting methods have the following drawbacks [2].

(i) Standards are not content related; i.e. mastery of curriculum content may not be required.
(ii) A fixed percentage of candidates may pass/fail each year, although some of the failing candidates at a given assessment may be more competent than those who passed at a previous administration of the same/similar assessment.
(iii) Level of candidates' ability may influence the standard; i.e. the candidate success is determined by the capability of the other candidates.
(iv) The standard is not known in advance, so the candidates may not be able to prepare adequately for the exam.
(v) Diagnostic feedback regarding the candidate competence/performance may be difficult to provide as the required standard in each item is not known.

It is not uncommon to have *combinations* of criterion and norm-referenced standard setting methods (i.e. compromise methods); e.g. Hofstee method [2–4]. Both criterion and norm-referenced standards may be employed *together*, when there is a need to rank candidates who have achieved the standard.

Different standard setting methods

A plethora of standard setting methods exist, indicating that there is not one universally applicable method. There are two categories of standard setting methods: one focuses on the test (i.e. test-centred methods); the other focuses on the candidates (i.e. examinee-centred methods).

(i) Test-centred methods

The examiners focus on the assessment items individually, to hypothetically decide how the candidates will fare at each assessment item.

Angoff method: Examiners determine the probability of a *borderline* candidate answering each item of the assessment correctly [5]. One system defines the borderline candidate as the 'minimally competent' (i.e. *just*-passing) candidate, whereas the other identifies the borderline candidate as one who is neither qualified nor unqualified to pass (i.e. on-the-fence candidate). For the purposes of this article we will consider the borderline candidate as the candidate who '*just* passes' the exam.

Ebel method: Examiners first categorise all the test items into several categories [6]. For example, in the modified Ebel method, the categories are termed as: 'essential', 'important' and 'indicated' test items [7]. The proportion or percentage of test items in each category that a borderline candidate could answer correctly is then estimated. This estimated proportion is multiplied by the number of items in that category to arrive at the pass mark

Table 1. **A comparison between conjunctive and compensatory standards**

Conjunctive standards	*Compensatory standards*
The candidate needs to achieve competence (i.e. pass) in all parts of the assessment.	The candidate needs only to pass the overall assessment; i.e. can compensate for a low mark in one part.
Should be adopted only if the individual test items (assessment parts) are high in reliability; conjunctive standards for unreliable test parts will result in unreasonable failures.	Can be adopted even if individual test parts are low in reliability; the assessment as a whole having high or moderate reliability is sufficient.
Can be used in an exam with several assessment parts, if the individual assessment parts are dissimilar to each other, assess unrelated curriculum content, or assess different competencies/constructs.	Can be adopted in an exam with several assessment parts, if the assessment components are similar to or positively correlate with each other; compensation will not result in loss of assessment information.
Provide clear diagnostic feedback to the candidate.	Diagnostic feedback to the candidate is less clear and less detailed.
May result in multiple failures.	Multiple failures are less likely.
Must consider ways of dealing with multiple failures (e.g. remediation and options for re-sitting the exam), before adopting conjunctive standards.	Re-sitting considerations are not a must. However, the consequences of compensating across test parts need to be considered.

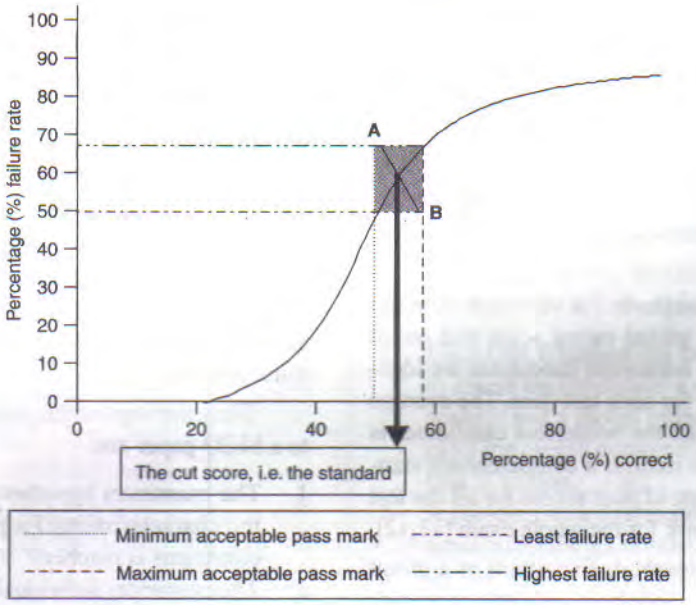

Figure 1. **Hofstee method of standard setting.**

for the category. The pass marks of all the categories are summed to obtain the pass mark for the whole assessment. This method considers the *quality or value* of individual test items (i.e. how important test items are), whereas the Angoff method considers only the item *difficulty*.

Hofstee method: Examiners estimate the lowest acceptable pass score, the highest acceptable pass score, the lowest acceptable failure rate and highest acceptable failure rate for an examination [8]. The examiner averages of these four estimates are then superimposed on a graph (Figure 1) containing actual candidate exam scores (with '% fail' on *y* axis and '% score' on *x* axis). The average examiner estimates create a rectangle (shown shaded in Figure 1), which the actual candidate exam score curve bisects. The pass mark is the point of intersection of the actual candidate score curve with the diagonal (AB) drawn from upper left to lower right corners of the rectangle (created by the examiner estimates).

Nedelsky method (for MCQs): First, the examiners identify and exclude the incorrect options that they think the minimally competent candidate will recognise; e.g. in a one-from-five Multiple Choice Question (MCQ), if the examiners think the minimally competent candidate will recognise two incorrect options, these are excluded. Next, the examiners count the number of options remaining; i.e. three in this example. Then, the pass mark for the MCQ is calculated as one over the number of options remaining. Therefore, one over three (33.3%) will be the pass mark for this single MCQ. Similar pass marks for all the other MCQs in the exam paper are totaled to determine the pass mark of the whole MCQ exam. This method is only suitable for selected response tests as it depends on the presence of distractors; i.e. incorrect options [9].

Jaeger method: A number of different, representative panels of stakeholders (not only a single panel of examiners) decide whether a just-passing candidate could

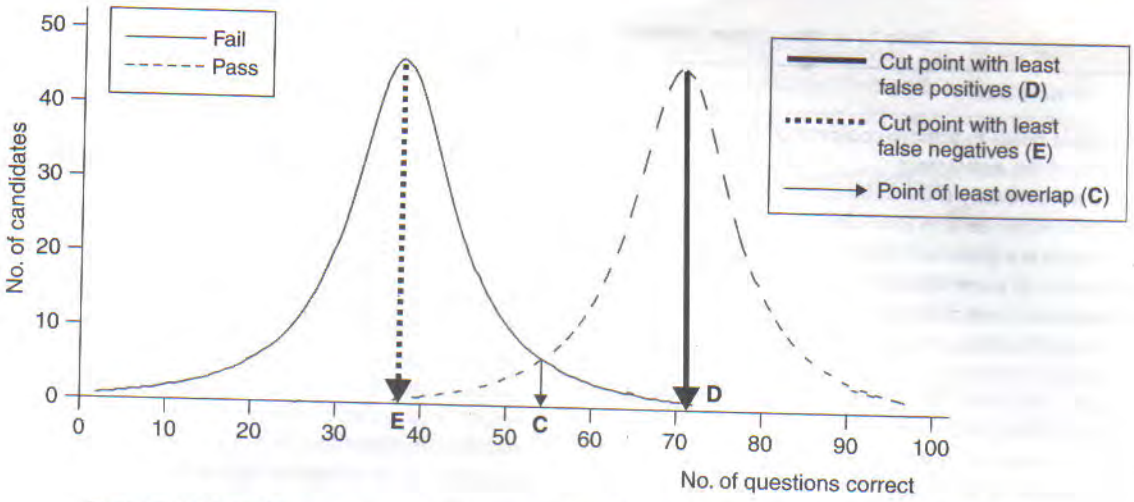

Figure 2. **Contrasting group standard setting method for a 100-item MCQ test carried out on a random sample of 50 students.**

answer each test item correctly. In an iterative process the judgment of each panel is fed into the judgments of the other panels to arrive at the pass mark [10].

(ii) Examinee-centred methods

The examiners concentrate on the actual (not hypothetical) borderline candidate score for a given assessment, as below.

Borderline group method: For each test item the candidate is scored on a global rating scale and on an itemized rating scale. The borderline candidates are identified by the global rating for each test item. The average of the itemised scores of all the borderline candidates is the pass score for that test item. If a compensatory standard is used, the summation of pass scores for all the test items provides the pass mark for the whole exam [11,12].

Contrasting group method: Examiners as a group select a random sample of candidates and categorize each candidate into 'pass' or 'fail' groups, based on their written exam answers for the whole examination. The test scores of the 'pass' and 'fail' groups are then plotted separately on the same graph, with 'questions correct' on x axis and 'number of examinees' on the y axis. A point on the graph is chosen as the pass mark, depending on the relationship between the two distributions (i.e. pass and fail), to suit different exam purposes; e.g. the point of least overlap ('C' in Figure 2), which provides the maximum discrimination between the two groups; to minimize false positives (the candidates who pass, but should have failed; 'D' in Figure 2) or false negatives (the candidates who fail, but should have passed; 'E' in Figure 2). In medicine it is not recommended to increase the false positives for reasons of patient safety.

Examples from practice

Two widely used methods of standard setting (one each from test-centred and examinee-centred standard setting methods) are described in detail below.

(a) Modified Angoff technique

Angoff [5] developed a technique to decide the pass mark for a multi-component (i.e. multi-item) assessment, such as a MCQ exam. This is the most frequently used method to set standards in the health professions' assessments.

In the modified Angoff technique, the examiners are additionally supplied with the actual test scores of previous assessments, to facilitate determining the probability of borderline candidates answering a given question correctly.

Thus, the steps of this process (Table 2), when applied to a MCQ paper are:

1. The examiners hypothetically visualize and discuss the characteristics of a just-passing candidate until consensus is reached.
2. The examiners individually consider each MCQ, one at a time.
3. Each examiner estimates the probability of a just-passing candidate answering each MCQ correctly. The probability lies within the range of 0 to 1, where 0 = no probability for the just-passing candidate getting the right answer, and 1 = 100% chance of the just-passing candidate answering correctly.
4. The probabilities per examiner for all the MCQs are summed.
5. The total per examiner is divided by the number of questions; i.e. the mean probability per examiner.
6. The mean probabilities of all the examiners are summed.
7. The sum of mean probabilities is divided by the number of examiners.
8. Examiners are provided with actual, previous test results to reconsider if necessary, their earlier probability estimates.
9. Examiners may revise their initial probabilities, in the light of the information provided on previous assessment.

Table 2. **Application of the modified Angoff technique, using five examiners, to a 10-question MCQ paper**

MCQ number	*Probability of a just passing candidate passing*										*Total probability per examiner*	*Mean per examiner*
	MCQ 1	*MCQ 2*	*MCQ 3*	*MCQ 4*	*MCQ 5*	*MCQ 6*	*MCQ 7*	*MCQ 8*	*MCQ 9*	*MCQ 10*		
Examiner 1	0.20	0.65	0.51	0.70	0.40	0.72	0.32	0.56	0.62	0.55	5.23	0.52
Examiner 2	0.15	0.58	0.45	0.75	0.38	0.75	0.35	0.54	0.65	0.53	5.13	0.51
Examiner 3	0.18	0.59	0.48	0.77	0.45	0.69	0.40	0.51	0.59	0.58	5.24	0.52
Examiner 4	0.25	0.63	0.55	0.80	0.48	0.78	0.39	0.58	0.68	0.49	5.63	0.56
Examiner 5	0.19	0.66	0.52	0.79	0.38	0.82	0.33	0.60	0.66	0.48	5.43	0.54
Total pass mark of all five examiners (sum total of mean probabilities)												2.65
Mean pass mark; i.e. the pass/fail standard												0.53
Pass/fail standard as a percentage cut score												53%

10. If one or more examiners change their initial probabilities, steps 4 to 7 are repeated to re-calculate the pass mark for the entire MCQ examination.

The modified Angoff process evaluates the difficulty of each test item and sets the pass mark accordingly. Hence, it is a criterion-based method of standard setting. It has stood the test of time, as it has been used widely for some time in the health professions' assessment. The difficulty of visualizing the hypothetical just-passing candidate, a criticism attributed to the classical Angoff method [11], has been reduced in the modified Angoff procedure by introduction of the past assessment results.

Visualizing the hypothetical just-passing candidate, however, still remains a problem even after providing the examiners with past exam scores of candidates. It is time consuming. The examiners have additional work in setting the standard, in contrast to the borderline group technique where there is no additional work for the examiners. Critics fear that the modified Angoff technique overlooks the overall candidate competence when deciding the pass mark [15,16], in contrast to the examinee-centred methods, which rely on the global ratings or the overall candidate achievement to calculate the pass mark [17].

(b) Borderline group technique

The steps of the borderline group technique, as applied to an OSCE, are explained below.

1. Each OSCE station has two assessment instruments to assess the candidate: a checklist or itemised rating scale(s) and a global rating scale. Global rating indicates the overall competence of a candidate at a given OSCE station.
2. A pre-determined reference point on the global rating scale indicates the borderline candidate competence.
3. At the end of the OSCE all the borderline candidates are identified using the global ratings.
4. The checklist scores, or all the itemised rating scores, of all the borderline candidates (as identified by step 3 above, using the global ratings) for that particular station are either added up or arranged in descending order.
5. The mean or the median checklist/itemized rating score of the borderline group of candidates forms the pass mark for each station [2,18].
6. If compensatory standards are applied, the pass marks of all the stations will be added to arrive at the pass mark for the entire OSCE.
7. One standard error of measurement is added to the pass mark to minimise the possibility of non-competent examinees passing, in examinations where patient safety is a consideration.

No additional examiner time or deliberation is needed to set the pass mark as the examiners identify the borderline candidates during the course of the exam through global ratings. A software package can be used to calculate the pass mark after the exam. Identification of the

'hypothetical' borderline candidate is not required as actual candidate scores are used to identify the borderline candidates. The overall ability of the candidate is taken in to account when setting standards. However, this technique is not suitable for an assessment with a small number of candidates, and it cannot be used in assessments where an alternative score/reference point (e.g. global rating) is not available.

Conclusions

There are four principles that any standard setting method should follow [7]:

1. Standard setting calls for 'expert judgement'. Such judgement can be moderated by different standard setting approaches to arrive at the most reliable and fair pass mark for an assessment. In other words, "setting standard will always be arbitrary, but need not be capricious" [7].
2. Unless there is a specific reason (e.g. to award a prize or to select the best candidates for a course with limited places) absolute standard setting is preferred to relative standard setting.
3. Multiple, experienced examiners should be employed to set the standard.
4. Where feasible, examiners should be provided with actual examinee data (either past or present) as setting standards without actual data may result in unrealistic pass scores and unreasonable pass or failure rates.

The method followed, however, is not so important as long as the method is fit-for-purpose, is based on informed judgement, demonstrates due diligence, is supported by research, and is easily explained and implemented [19].

References

1. Gove PB; ed. Webster's third new international dictionary. Massachusetts: G & C Merriam Company; 1976: 2223.
2. Friedman Ben-David M. AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher* 2000; **22**: 120–30.
3. De Gruijter DNM. Compromise models for establishing examination standards. *Journal of Educational Measurement* 1985; **22**: 263–9.
4. Norcini JJ. Setting standards on educational tests. *Medical Education* 2003; **37**: 464–9.
5. Angoff WH. Scales, norms and equivalent scores. In: Thorndike RL; ed. Educational Measurement. Washington DC: American Council on Education; 1971.
6. Ebel RL. Essentials of Educational Measurement. New Jersey: Englewood Cliffs, Prentice-Hall; 1972.
7. Case SM, Swanson DB. Constructing written test questions for basic and clinical sciences. 3rd ed. Philadelphia, USA: National Board of Medical Examiners (NBME); 2001. Available at: http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf (accessed on 27 December 2005).
8. Hofstee WKB. Een alternatief voor normhandhaving bij toetsen. *Nederlands Tijdschrift voor de Psychologie* 1973; **28**: 215–27.
9. Nedelsky L. Absolute grading standards for objective tests. *Educational and Psychological Measurement* 1954; **14**: 3–19.
10. Jaeger RM. An interactive structures judgment process for establishing standards on competency test: theory and application. *Educational Evaluation and Policy Analysis* 1982; **4**: 461–76.
11. Smee SM, Blackmore DE. Setting standards for an OSCE. *Medical Education* 2001; **35**: 1009–10.
12. Wilkinson TJ, Newble DI, Frampton C. Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. *Medical Education* 2001; **35**: 1043–9.
13. Zieky MJ. So much has changed. How the setting of cut-scores has evolved since the 1980s. In: Cizek GJ; ed. Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2001: 19–52.
14. Kaufman DM, Mann KV, Muijtjens AMM, van der Vleuten CPM. A comparison of standard setting procedures for an OSCE in undergraduate medical education. *Academic Medicine* 2000; **75**: 267–71.
15. Cusimano MD. Standard setting in medical education. *Academic Medicine* 1996; **71**: 112–20.
16. Hambleton RK, Jaeger RM, Plake BS, Mills C. Setting performance standards on complex educational assessments. *Applied Psychological Measurement* 2000; **24**: 335–66.
17. Dauphinee WD, Blackmore DE, Smee SM, Rothman AI, Reznick RK. Using judgments of physician examiners in setting standards for a national multi-centre high stakes OSCE. *Advances in Health Science Education: Theory in Practice* 1997; **2**: 201–11.
18. Cusimano MD, Rothman AI. Consistency of standards and stability of pass/fail decisions with examinee-based standard setting methods in small-scale objective structured clinical examination. *Academic Medicine* 2004; **79**: S25–S27.
19. Norcini J. Standard setting. In: Dent JA, Harden RM; eds. A practical guide for medical teachers. London: Elsevier Churchill Livingstone; 2005: 293–301.

GG Ponnamperuma, *Lecturer in Medical Education, Medical Education Development and Research Centre, Faculty of Medicine, University of Colombo (presently at Centre for Medical Education, University of Dundee, Scotland, UK), and* **MH Davis**, *Director and Professor of Medical Education, Centre for Medical Education, University of Dundee, Scotland, UK.* Correspondence: GGP, *e-mail: <g.ponnamperuma@dundee.ac.uk>* *(Competing interests: none declared).*