



A TRADITION OF
INDEPENDENT
THINKING



Overview of Standard Setting

DR HELEN HYNES, MEDICAL EDUCATION UNIT, UCC

This presentation, produced by the Medical Education Unit in University College Cork, explains the rationale for using standard setting in assessments in medical education and gives an overview of standard setting methods that have been described and validated in the medical education literature.

We have produced 2 further presentations which describe standard setting methods used in Multiple Choice examinations and in OSCE examinations.



Why not just use 50%?

Using a pre-fixed pass mark was the traditional approach.

Many other departments still use this.

Why can't we?

Traditionally universities just used a set pass mark for examinations. In many faculties this has been 40%. In medical education, the traditional pass mark has been 50%. So why can't we just keep on doing what we did in the past?



Purpose of a Medical Education?



Medical education aims to produce graduates who are competent and ready to begin supervised medical practice. But how do we know our graduates are competent? If they get 50% in their exam does that mean they are competent?



How much knowledge is enough? Is 50% enough? How comfortable would we be with the idea that our doctor got 50% in their exams?

50% does not really tell us anything. Maybe it was a very hard exam - maybe it was a very easy exam.

How can we make sure that the doctors we graduate are actually competent?

To answer that question we need to review what are the standards elsewhere.



Who says we need to standard set?

World Federation for Medical Education ¹



Medical Council

Association for Medical Education in Europe ^{2,3,4}

We'll start with the World Federation of Medical Education or (WFME)¹.

The WFME is an internationally recognised organisation which publishes basic standards, which must be met by all universities offering undergraduate medical degrees. It also produces Quality Development Standards, which constitute best practice in medical education, and to which medical schools should aspire.

The Irish Medical Council uses the WFME standards in its accreditation of Medical Schools. The Medical Council is our regulator, and we must adhere to its recommendations.

The Association for Medical Education in Europe (AMEE) is a worldwide organisation with members in ninety countries on five continents. AMEE promotes international excellence in education in the healthcare professions by multiple methods, including promoting the use of evidence-informed education, and setting standards for excellence in healthcare professions education. AMEE publish guides giving practical advice and current thinking on important topics in healthcare education ^{2,3,4}.



World Federation of Medical Education: Basic Medical Education Standards ¹

3.1. ASSESSMENT METHODS :

Basic standards: The medical school must

- define, state and publish the principles, methods and practices used for assessment of its students, including the criteria for setting pass marks, grade boundaries and number of allowed retakes. (B 3.1.1)
- use a wide range of assessment methods and formats according to their “assessment utility”*. (B 3.1.3)

Quality development standards: The medical school should

- evaluate and document the reliability and validity of assessment methods. (Q 3.1.1)

* “Assessment utility” is a term combining validity, reliability, educational impact, acceptability and efficiency of the assessment methods and formats.

So looking at what the WFME says about assessment:

Among their basic standards for assessment they state that:

The medical school must


- define, state and publish the principles, methods and practices used for assessment of its students, including the criteria for setting pass marks, grade boundaries and number of allowed retakes. (B 3.1.1)
- use a wide range of assessment methods and formats according to their “assessment utility” * (B 3.1.3).

In their Quality development standards they state that:

The medical school should:

- evaluate and document the reliability and validity of assessment methods. (Q 3.1.1)

*“Assessment utility” is a term combining validity, reliability, educational impact, acceptability and efficiency of the assessment methods and formats.



Association for Medical Education in Europe (AMEE)

AMEE Guide 18 ²: Standard setting in student assessment

- "Historical precedent cut off scores (eg 50%) are no longer acceptable in high stakes examination"
- "Standard setting on a written examination is essential for good testing practice."

AMEE Guide 37 ³: [Setting and maintaining standards in multiple choice examinations](#)

AMEE Guide 85 ⁴: How to set standards on performance-based examinations

The Association for medical Education in Europe has published 3 guides relating to standard setting in medical Education.

In the first of these, AMEE Guide 18, they say that:

"Historical precedent cut off scores (eg 50%) are no longer acceptable in high stakes examination"

"Standard setting on a written examination is essential for good testing practice." ²

AMEE Guide 37 describes: Setting and maintaining standards in multiple choice examinations ³.

AMEE Guide 85 describes: How to set standards on performance-based examinations, for example in Clinical OSCEs ⁴.



Licensing Exams / Postgraduate Exams

PRES – Medical Council's licensing exam for International Medical Graduates (IMGs)

PLAB – UK's General Medical Council licensing exam for IMGs

USMLE - Federation of State Medical Boards (FSMB) and National Board of Medical Examiners (NBME) – licensing exam for all medical graduates

ICGP – MICGP exams

RCPI – MRCPI exams

Joint Committee of Intercollegiate examinations – FRCS exams

UK Academy of Medical Royal Colleges

This is why medical licensing organisations such as:

- the Medical Council;
 - the UK's General Medical Council;
 - the American USMLE;
- professional bodies such as:
- the Irish College of General Practitioners;
 - the Royal College of Physicians in Ireland;
 - the joint committee of intercollegiate examinations (who are responsible for surgical exams); and
 - many other international postgraduate training bodies and high standard medical school around the world

- have moved away from a fixed pass mark to find better ways to ensure that their graduates are competent.



Literature

Norcini JJ. Setting standards on educational tests. *Medical Education* 2003; 37(5):464–9 ⁵

Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher* 2010; 33:3 ⁶

“Predetermined fixed standards are increasingly difficult to justify and defend and indeed have been described as “the least defensible approach to standard setting” in the context of medical education” - Cohen Schotanus J and Van der Vleuten CP, *Medical Teacher* 2010 ⁷

There are a number of excellent resources in the literature describing the pedagogy behind standard setting, and also a growing body of evidence for various standard setting methods methods which will be described later in the presentation.

Norcini JJ. Setting standards on educational tests. *Medical Education* 2003; 37(5):464–9 ⁵

Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher* 2010; 33:3 ⁶

“Predetermined fixed standards are increasingly difficult to justify and defend and indeed have been described as “the least defensible approach to standard setting” in the context of medical education” - Cohen Schotanus J and Van der Vleuten CP, *Medical Teacher* 2010 ⁷



Absolute Standards Vs Relative Standards ⁵

2 types of standard setting:

- Relative standard – also known as norm referenced.
- Absolute standards – also known as criterion referenced

Decision between them is related to the purpose of the test.

In 2003 Norcini published *Standard Setting in Educational Tests* in the *Journal Medical Education*. ⁵

Norcini divides standard setting into 2 broad types – absolute standards and relative standards, and he says that which of them we choose to use, depends on the purpose of the test.



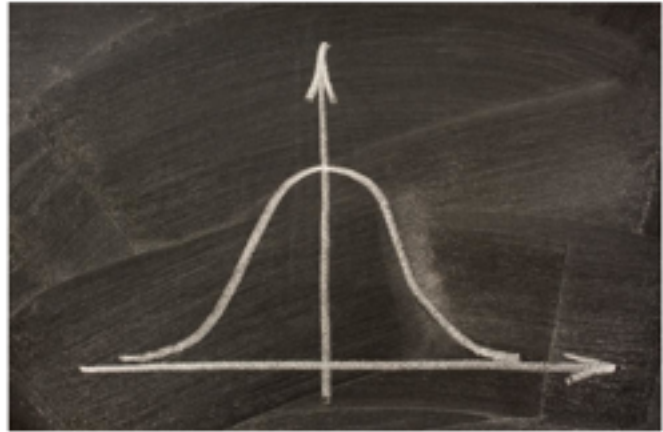
Relative Standards

Relative standards are expressed as a number or percentage of examinees.

The pass mark is set at some predefined limit based on all of the candidates' performances on the test.

For example we might decide to pass the top 50 performers or decide to fail everyone who is in the lowest 5th percentile.

A type of relative standard we are all familiar with is the Bell curve.



Relative standards are expressed as a number or percentage of examinees who will pass or fail the exam.

The pass mark is set at some predefined limit based on all of the candidates' performances on the test.

For example we might decide to pass the top 50 performers or decide to fail everyone who is in the lowest 5th percentile.

A type of relative standard we are all familiar with is the Bell curve.



Relative Standards

Relative standards are most appropriate for examinations where the purpose is to identify a certain number of examinees.

This includes tests that are used to select the highest or lowest scorers for admissions or placement, where a limited number of students can be accommodated.

Standard setting is based on the test results.

Weaknesses of relative/norm-referenced standards

- Does not take account of difficulty of the test or competence of the students
- Standards are not content related
- Examinees' ability may influence the standard.
- A fixed number of candidates may fail the test even if all the examinees are excellent
- A fixed number of candidates will pass and get honours grades in the test if all the candidates are weak.
- Standard is not known in advance.

Relative standards are most appropriate for examinations where the purpose is to identify a certain number of examinees.

This includes tests that are used to select the highest or lowest scorers for admission or placement, where a limited number of students can be accommodated.

Using these methods, standard setting is based on the test results.

Some of the problems using relative/norm-referenced standards are:

- They do not take account of difficulty of the test or competence of the students
- Standards are not content related
- Examinees' ability may influence the standard.
- A fixed number of candidates may fail the test even if all the examinees are excellent
- A fixed number of candidates will pass and get honours grades in the test even if all the candidates are weak.
- The standard is not known in advance.



Relative – The Leaving Cert ⁸

			# of students	H1	H2	H3	H4	H5	H6	H7	H8
ENGLISH	Higher	2019	40217	2.9	10	20.4	27.6	23.8	12	2.6	0.6
ENGLISH	Higher	2018	38283	2.9	10	20	27.9	24.2	12.2	2.3	0.5
ENGLISH	Higher	2017	38769	2.9	10.7	20.6	26.8	23.3	12.7	2.4	0.6

Leaving Cert Higher English Results, 2017-2019

			# of students	H1	H2	H3	H4	H5	H6	H7	H8
GEOGRAPHY	Higher	2019	19982	3.7	13.1	20.7	23.1	20.2	13.7	4.6	1
GEOGRAPHY	Higher	2018	19506	3.5	12.5	19.7	23.3	20.8	14.1	4.9	1.1
GEOGRAPHY	Higher	2017	19293	3.4	12.7	21.8	24	20.1	12.9	4.2	0.9

Leaving Cert Higher Geography Results, 2017-2019

Relative standards are defensible when there is a very large cohort and the performance is not expected to vary greatly from year to year.

We see it used in the Leaving Certificate exam. ⁸

Here we have the percentage of students who achieved the various grades from H1 to H8 in both English and Geography in the Leaving Certificate in 2017, 2018 and 2019.

We can see that within each subject the percentages getting each grade is very similar from year to year, although the percentages differ in different subjects.

And it works fine given the large number of candidates, but if you used this method with a small number of candidates, it would be less defensible because it does not take into account the ability of the candidates.



Absolute Standards / Criterion Referenced

The standard is set based on certain pre-defined criteria. Eg:

- Predefined criteria where the candidate demonstrates competency or mastery
- Often relies on the concept of the Borderline (Minimally Competent) Candidate

Absolute standards are most appropriate for tests of competence, where the purpose is to establish that the examinees know enough for a particular purpose.

These include final or exit examinations and tests for certification and licensure.

In absolute / criterion referenced standards the standard is set based on certain pre-defined criteria such as predefined criteria where the candidate demonstrates competency or mastery. ⁵

Absolute methods often rely on the concept of the Borderline (Minimally Competent) Candidate.

Absolute standards are most appropriate for tests of competence, where the purpose is to establish that the examinees know enough for a particular purpose.

These include final or exit examinations and tests for certification and licensure.



Which methods should I use?

Absolute / Criterion Referenced

- Modified Angoff – MCQ ^{2,3,4,5,9}
- Ebel – MCQ ^{5,10}
- Borderline Regression analysis – OSCE ⁴
- Borderline Group method – OSCE ⁴

Relative

- Cohen – “Reality Check” ⁷
- Bell Curve – not recommended

Compromise

- Hofstee ^{2,3,5,11}

Multiple different methods have been described in the medical literature.

They include:

Absolute / Criterion Referenced

- Modified Angoff – useful with MCQ examinations ^{2,3,4,5,9}
- Ebel – also suitable for MCQ examinations ^{5,10}
- Borderline Regression analysis – useful with OSCE ⁴
- Borderline Group method – also can be used with OSCE ⁴

Relative

- Cohen – useful for lower stakes examinations when the cohort is sufficiently large – can also be used with other methods as a “Reality Check” ⁷
- Bell Curve – not recommended

Compromise

- Hofstee ^{2,3,5,11}



Modified Angoff ⁹

A group of **expert judges** makes estimates of how a **borderline candidate** would perform on each item in the examination.

The experts must:

- be knowledgeable about the standard of the candidate population, the subject matter.
- understand what has been taught about the subject matter to this cohort of students.

The module coordinator / exam coordinator must train the experts on the concept of the borderline candidate, and must explain to the expert judges what the students have been taught.

The borderline candidate is described as that candidate who is minimally competent.

This method is evidence based and is very suited to MCQs and extended matching questions.

When using the Modified Angoff method ⁹, a group of **expert judges** makes estimates of how a **borderline candidate** would perform on each item in the examination.

The experts must:

- be knowledgeable about the standard of the candidate population, the subject matter.
- understand what has been taught about the subject matter to this cohort of students.

The module coordinator / exam coordinator must train the experts on the concept of the borderline candidate, and must explain to the expert judges what the students have been taught.

The borderline candidate is described as that candidate who is minimally competent. This method is evidence based and is very suited to MCQs and extended matching questions.



Ebel ¹⁰

A team of expert judges reviews the test. They rate each item on 2 dimensions –

- Difficulty : easy / medium / hard
- Importance: essential / important / acceptable (nice to know)

	Easy	Medium	Hard
Essential			
Important			
Acceptable			

Examiners agree on the definition of the minimally competent / borderline candidate.

Examiners estimate the percentage of questions in each of the 9 categories that a minimally competent candidate would answer correctly.

The pass mark is calculated based on the number of questions in each category and the likelihood of a borderline candidate correctly answering questions from each category.

There is a reasonable body of evidence behind this method and its use is suited to MCQ type examinations.

When using the Ebel Method ¹⁰, a team of expert judges reviews the test. They rate each item on 2 dimensions –

- Difficulty : easy / medium / hard
- Importance: essential / important / acceptable (nice to know)

Examiners agree on the definition of the minimally competent / borderline candidate.

Examiners estimate the percentage of questions in each of the 9 categories that a minimally competent candidate would answer correctly.

The pass mark is calculated based on the number of questions in each category and the likelihood of a borderline candidate correctly answering questions from each category.

There is a reasonable body of evidence behind this method and its use is suited to MCQ type examinations.



Cohen ⁷

This is a relative standard – the pass mark is based on the performance of the highest scoring students in the class.

The pass mark is set at 60% of highest achiever's score (or 60% of the mean of the top 3 highest achievers' scores, or 60% of the 90th / 95th centile.)

Need to have at least 100 students to use this method with confidence

Fast, easy to calculate

Useful as a "reality check" if other methods are used.

Acceptable to use with lower stakes exams.

This is a relative standard – the pass mark is based on the performance of the highest scoring students in the class.

The pass mark is set at 60% of highest achiever's score (or 60% of the mean of the top 3 highest achievers' scores, or 60% of the 90th / 95th centile.)

One would need to have at least 100 students to use this method with confidence.

It is fast and easy to calculate.

It is useful as a "reality check" if used in combination with other methods. It is acceptable to use with lower stakes exams.



Hofstee ^{2,3,5,11}

A compromise between relative and absolute standards.

The examiners estimate 4 values:

- The minimum acceptable failure rate
- The maximum acceptable failure rate
- The minimum pass mark (cutscore), even if all examinees failed
- The maximum passmark (cutscore), even if all examinees passed

Their responses serve as the focus for discussion, with all being free to change their estimates.

These minimum and maximum failure rates and percent correct scores are averaged across panelists and projected onto the actual score distribution to derive a passing score.

The Hofstee Method ^{2,3,11,15} is described as a compromise between relative and absolute standards.

The examiners estimate 4 values:

- The minimum acceptable failure rate
- The maximum acceptable failure rate
- The minimum pass mark (cutscore), even if all examinees failed
- The maximum passmark (cutscore), even if all examinees passed

Their responses serve as the focus for discussion, with all being free to change their estimates.

These minimum and maximum failure rates and percent correct scores are averaged across panelists and projected onto the actual score distribution to derive a passing score.



More Information

More information on:

- the Modified Angoff;
- Ebel;
- Cohen; and
- Hofstee Methods

of standard setting are included in our presentation on Standard Setting for MCQ examinations.



More information on:

- the Modified Angoff;
- Ebel;
- Cohen; and
- Hofstee Methods

of standard setting are included in our presentation on Standard Setting for MCQ examinations.



Borderline Regression Analysis ⁴

Used to set the standard in OSCE examinations.

A checklist is used on which the examiner records whether predefined actions were carried out or not. The examiner separately rates the student's overall performance as Fail, Borderline, Pass, Good Pass, Outstanding.

All candidates' results are plotted on Excel (equivalent software). A scatter plot is produced which allows for a regression line to be drawn.

The pass mark is calculated using the slope of the regression line.

More information on this method is included in our presentation on Setting the standards in OSCE examinations.

Borderline Regression Analysis ⁴ is an evidence based method used for standard setting OSCE examinations.

A checklist is used on which the examiner records whether predefined actions were carried out or not. The examiner separately rates the student's overall performance as Fail, Borderline, Pass, Good Pass, Outstanding.

All candidates' results are plotted on Excel (or equivalent software). A scatter plot is produced which allows for a regression line to be drawn.

The pass mark is calculated using the slope of the regression line.

More information on this method is included in our presentation on Setting the standards in OSCE examinations.



Guidelines for choosing a method ⁵

The method must:

- Produce standards consistent with the purpose of the test
- Rely on informed expert judgement
- Demonstrate due diligence
- Be easy to explain and implement
- Be supported by a body of research

Whichever method you choose, the method must:

- Produce standards consistent with the purpose of the test
- Rely on informed expert judgement
- Demonstrate due diligence
- Be easy to explain and implement
- Be supported by a body of research



Other Presentations

Standard Setting for Multiple Choice Examinations

Standard Setting for OSCE Examinations

The Medical Education Unit has also prepared 2 further presentations on standard setting. One of these relates to standard setting for MCQs, the other relates to standard setting for OSCEs. Further details are included in these presentations of how to carry out standard setting using the methods described above.

References

1. [WPM: Basic Standards for Medical Education, 2015.](#)
2. Ben-David MF. AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher* 2000; 22(2):120-30
3. Bandaranayake RC. AMEE Guide No. 37: Setting and maintaining standards in multiple choice examinations. *Med Teacher* 2008;30(3-10):836-45
4. McKinley DW, Norcini JJ. AMEE Guide 85: How to set standards on performance-based examinations. *Medical Teacher* 2004; 26(2):97-110
5. Norcini JJ. Setting standards on educational tests. *Medical Education* 2003; 37(5):464-9
6. Norcini J, Anderson B, Bellala V, Burch V, Costa M, Duvviler R, Galbraith R, Hays R, Kent A, Ferrott V, Roberts T. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher* 2010; 33:3
7. Cohen-Schotanus J and Van der Vlieten CP. A standard setting method with the best performing students as point of reference: practical and affordable. *Medical Teacher* 2000; 32(2):354-60
8. [Learning Certificate Examination Statistics.](#)
9. Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington DC: American Council on Education
10. Ebel RL. (1972) *Essential of educational measurement*. Englewood Cliffs (NJ): Prentice Hall; p. 622.
11. Hofstee WBB. The case for compromise in educational selection and grading. In: Anderson SB, Helminck JS, editors. *On Educational Testing*. San Francisco: Jossey-Bass; 1983. pp. 309-27.