APPENDIX C

A GLOSSARY FOR STANDARD SETTING

Materials for this Glossary are drawn from JMcL's postgraduate teaching material. The first person singular (rather than plural – 'I' rather than 'We') is used in this Appendix. Terms shown in bold are given as Glossary entries.

Angoff Standard Setting Procedure.

Expert informed experienced Judges estimate the proportion of the group of minimally competent candidates who would respond correctly for each item, then record, repeat and cumulate for the test as a whole. See **Modified Angoff.**

Arbitrary

The word 'arbitrary' has several meanings, one of which certainly is 'capricious', but another accords with 'judgement', as in 'arbitration'. See **Standard**.

Beuk Compromise Standard Setting Method (Beuk, 1984)

A **Compromise** standard setting method, in the same class as **Hofstee** and **De Gruijter** approaches (q.v.), with the following principles:

1. Each member of the standard setting committee forms an opinion of (a) what the passing (cut) **score** should be, and (b) what the pass rate should be.

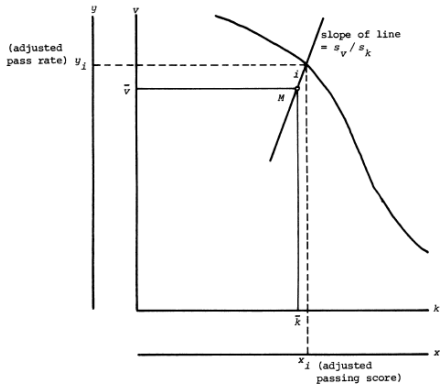
2. The relative emphasis given to the two types of judgments should be in proportion to the extent to which the members of the committee agree with each other.

The means and standard deviations of the pass rates and cut scores are then calculated, and a linear relationship between them of the following form is assumed, where 'SD' = standard deviation:

Estimated pass % = (estimated pass % SD/ estimated cut score SD) (estimated cut score – estimated cut score Mean) + estimated pass % Mean.

On a graph of scores versus pass rate, the estimated cut score and estimated pass rate are plotted (call this point M). Then the relationship between cut scores and pass rates for the candidates as a decreasing curvilinear function is plotted (see Figure C1). Finally, a line is drawn through M with the slope (estimated pass rate SD/ estimated cut score SD). Where it intercepts the plot of actual student performance is the compromise cut score and pass rate.

Figure C1 (from Beuk, 1984)



Bookmark standard setting methods (Karantonis & Sireci, 2006)

'Bookmark' methods are reasonably widely used in a variety of testing environments. They belong to the category of criterion referenced methods relating to test items. The first step is to construct an Ordered Item Booklet in which questions are ranked in order of their difficulty. This information is generally, and most appropriately, determined by Item Response Theory approaches, although this assumes that questions can be continuously graded on a monotonic scale. The expert panel then identify where in the Ordered Item Booklet different boundaries lie (such as 'Borderline', 'Satisfactory', and 'Excellent'), and these estimates can be averaged. An advantage of Bookmark is therefore that it readily lends itself to multiple cut points as part on one exercise. A further described advantage is that Selected and Constructed Response questions can be integrated into the OIB as a single continuum.

Evidence of the validity and reliability of Bookmark methods remains rather sparse. In addition, the argument that it allows the incorporation of Selected and Constructed Response questions seems tenuous – the same argument could be used for Angoff or Ebel Methods with the appropriate modifications. Most challenging is the assumption of monotonicity. Many professional assessments explore different areas of expertise, and there may be Case Specificity within these areas. This is most evident in assessment methods such as OSCEs, where it is hard to imagine Bookmark methods being applied.

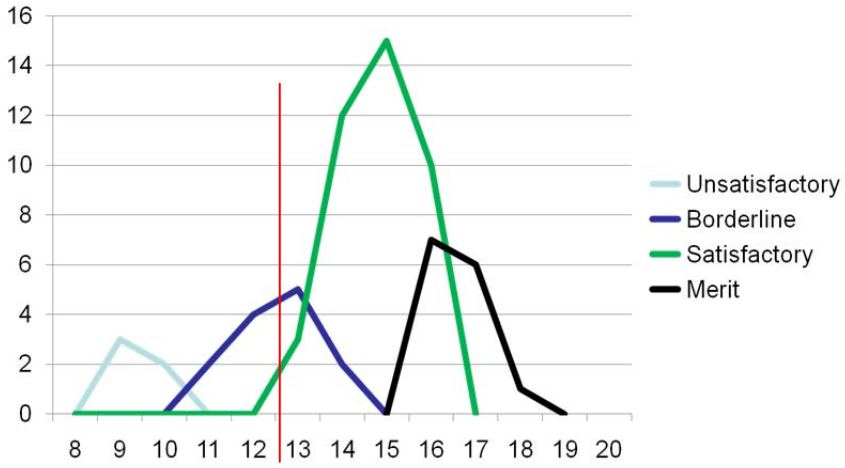
Borderline Group(s)

This method has the following steps (as envisaged in an OSCE setting)

- The assessor awards score points on the basis of a checklist.
- The assessor makes a "Global Judgement" (e.g. 'Borderline', 'Pass', 'Fail', 'Merit', etc.) independently of the score, as in Figure C2.
- Scores in each 'Global Judgement' category are plotted against frequency, with scores as the abscissa, and frequency as the ordinate

- The mean or median of the 'Borderline' Group is chosen as the cut score. In some variants, 'Borderline Pass' and 'Borderline Fail' are the grades awarded. Here the cut score is the average of the respective means. In the figure below, the red line indicates the cut score.

Figure C2 (University of Durham data)



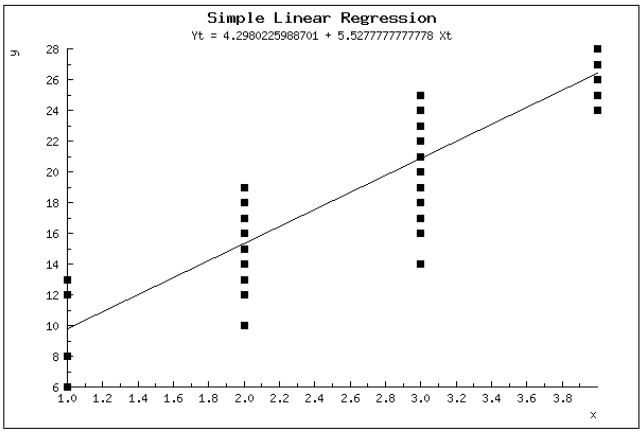
Borderline Regression

This method has the following steps (as envisaged in an OSCE setting)

- The assessor awards score points on the basis of a checklist
- The assessor makes a "Global Judgement" (e.g. 'Borderline',' Pass', 'Fail', 'Merit' etc.) independently of the score
- Scores in each 'Global Judgement' category are plotted as a scatter plot against frequency, with grades as the abscissa, and candidate scores as the ordinate as in Figure C2
- The regression line through those points is calculated
- Where the regression line cuts the 'Borderline' grade is chosen as the cut score
- In the example below, points 1,2, 3 and 4 on the abscissa represent the Grades 'Unsatisfactory', 'Borderline', 'Satisfactory' and 'Merit' respectively.

Figure C3 (University of Durham data)

Here, the values on the abscissa correspond to grades (e.g. Unsatisfactory, Borderline, Satisfactory, Merit), from left to right respectively



Cohen Standard Setting Method

This method was proposed by Cohen-Schotanus and Van der Vleuten (2010). Essentially, the recommended method requires taking 60% of the score of the 95th percentile candidate as the cut score. The rationale is stated as being that the top scoring students show less variability than the generality of students. The process is indicated in the original paper as reducing variability in cut scores compared to a normative process, and reducing variability in pass rates compared to a fixed pass rate (described in the paper as a criterion referenced approach). The 'Cohen' method is described by the authors as a compromise method.

There are a number of criticisms that can be advanced against the principles underlying this approach. Although it is described as a compromise method (here, a compromise between norm and criterion referenced methods), there is no criterion based element actually present. It is in fact entirely normative, and the argument for its use must rest entirely on its superiority to other norm referenced methods (such as setting a cut score at 1 SEM below the mean, as explored in the citation). However, since it is based on the performance of one particular top student, as opposed to the mean of the distribution, this superiority is at least debatable. It may indeed give lower variability, but this assumes that it is an accurate representation of the ability of the groups involved – that the groups do not vary in ability in a way which would require varying cut scores and pass rates. In other words, it privileges reliability above validity. Indeed, no evidence on validity is adduced in the original paper. It is certainly inexpensive, and is probably the least demanding in terms of examiner time. However, I doubt that it is sufficiently well evidenced in terms of validity at the moment to be recommendable for use in high stakes settings.

A modification of the Cohen method has been proposed which attempts to address this problem by incorporating criterion referenced weightings (Taylor, 2011) but such information will rarely be available – if it is then the criterion referenced value would be preferred.

Compensation

Full Compensation Model: every assessment category or domain is aggregated to give the final pass score

No Compensation Model: every assessment category must be passed separately e.g. Knowledge, Skills, and Behaviours. This may be described as a 'Conjunctive' approach.

Partial Compensation Model: there is compensation within or between categories or domains.

There are theoretical grounds (and some evidence) for believing that full compensation models may increase false positives.

Computer Adaptive Testing

Characterisation of the properties of assessment items allows the use of Computer Adaptive Testing (CAT). In this approach, candidates follow a path through the assessment which is modified by their performance. In other words, if candidates get the first question right, they get a harder one, if they get it wrong they get an easier one. This reaches reliable estimates of candidates' ability (not 'knowledge' since knowledge has a case-specific component) quite quickly. For example, the reliability of a 100 item CAT might be the same as a two hundred item conventional assessment.

Computer Assisted Testing

Computer Assisted testing means testing delivered by means of a computer, rather than on paper. Advantages include the possibility of using multi-media and 'unfolding' questions, where a scenario develops through a number of steps. The software can also test-equate and score questions, giving rapid feedback to the candidate, and rapid information to the assessors. Disadvantages relate to security of software and availability of hardware.

Compromise Standard Setting Methods: See Hofstee , Beuk, De Gruijter

In the original literature, these methods are described as being compromises between criterion and norm referenced approaches, and this might be the case if they are used to set one off standards. In my view, however, in cases where the judges have the opportunity to observe candidates, they are best repeated on each test occasion, rather than being used on a once-and-for-all basis. In such cases, they can be viewed as a compromise between criterion referenced judgements on test items and judgements on test takers. The instruction to judges might be "Given what you know of this test paper (e.g. easy or difficult) what do you think the cut score should be?", and conversely "Given what you know of this group of candidates (e.g. a good group or a poor group compared to normal expectations) what do you think the pass rate should be?" This of course relies on the judges having experience of the group, but in many local standard setting environments, this is indeed the case.

Connoisseurship (Bleakley et al, 2003)

The term 'gut feeling' sounds derogatory, but in the end is the basis for most standard setting decisions. The term 'connoisseurship' has been proposed as a more accurate, and more seemly, description of the process of **Expert** judgement.

Context referenced

It is a mistake to imagine that criteria are absolute and unchanging (see for instance the Flynn effect). It might be possible that some candidates suffer an unfair and systematic negative bias, for instance. It may be appropriate to consider their performance in the light of these biases, and make adjustments accordingly. I propose that this is called 'context referencing'.

Standards may change with contexts

- E.g. by place (private selective schools versus deprived state comprehensives)
- e.g. by time – there is a widely held view that assessments are getting easier (there is a higher proportion of first class and upper second class degrees in the UK , and increasing grades at A levels)

It may be possible to use an **Angoff** procedure with guidance on what the 'minimally competent candidate' is like from each setting.

Contrasting Groups

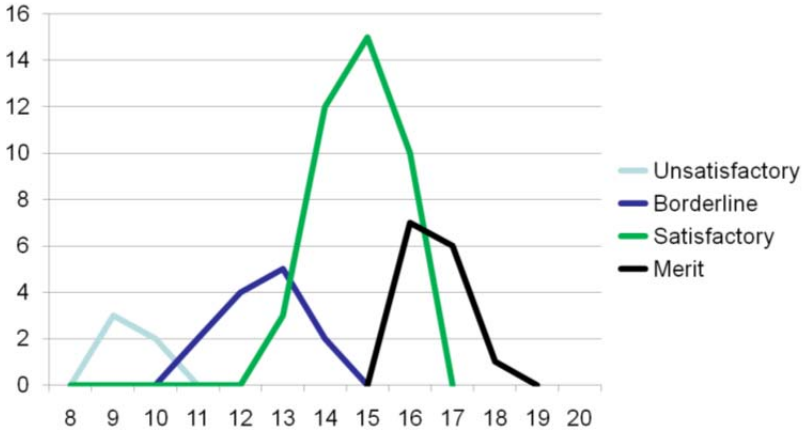
This is performed initially in the same way as Borderline Groups, but the cut score is determined by inspection of all the data. Cut score can be chosen to maximise sensitivity and specificity together (the intercept of two groups) or where the upper and lower boundaries intercept the abscissa.

For contrasting groups we have to assume that the variances of the groups are normal and equal in order to use parametric methods; otherwise we have to use nonparametric QDF methods.

This method has the following steps (as envisaged in an OSCE setting):

- The assessor awards score points on the basis of a checklist (as in OSCEs, Foundation white space or Mini-PAT)
- The assessor makes a "Global Judgement" (e.g. 'Borderline', Pass, Fail, Merit etc) independently of the score
- Scores in each 'Global Judgement' category are plotted against frequency, with scores as the abscissa, and frequency as the ordinate as in Figure C4
- The intercept between the 'Borderline' grade and the 'Satisfactory' grade is chosen as the cut score (alternatively, the higher or lower intercept of the Borderline grade with the abscissa could be chosen).

Figure C4 (University of Durham data)



Criterion Referenced

Based (in principle) on some absolute standard of knowledge or performance. See also **Norm Referenced, Context Referenced**.

The usual steps are:

- First, *define* a group of experts
    - (knowledge of subject, knowledge of context, knowledge of assessment, knowledge of students)
- Then establish the minimum required size of the expert group
    - Frequently taken to be about 8, some evidence that 10 are needed if there is no feedback on item or candidate performance, 6 if there is.
- The experts may make judgements on *test items (rational)* or *test takers* (*pragmatic*)
- In principle, judgements on test items are *prospective,* judgements on test takers are *retrospective*

Criterion referenced judgements on test items: See **Angoff, Ebel, Bookmark, Nedelsky.**

Criterion referenced judgements on test takers: See **Borderline Group, Contrasting Groups, Borderline Regression, "Up and Down" Method.**

'Critical Action' or 'Critical Approach' Standard Setting Method

In this approach, the assessors identify the actions which are viewed as critical to passing a station through a process of discussion (Ferrel, 1996; Payne et al, 2008). Candidates then must perform all of these actions in order to pass the station. One such study (Richter Lagha et al, 2012) identified history taking, physical examination, and patient education as the critical actions in a group of 6 OSCE stations. However, in this paper, very low reliability was obtained.

This approach has similarities to the idea of summing behaviours across stations as used in PACES (Elder et al, 2011) and is the exact inverse of deducing factors retrospectively by exploratory factor analysis (Chesser et al, 2004).

Defensible

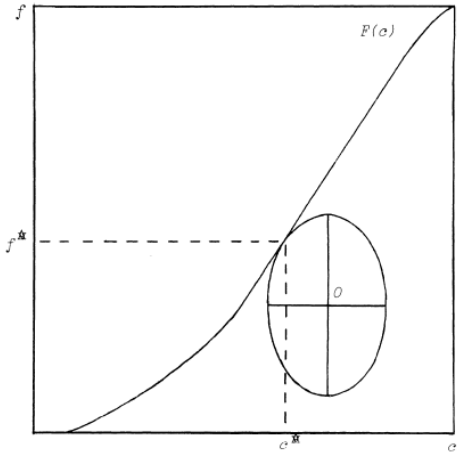
If assessment is always arbitrary, what are the characteristics of a "Defensible" standard? Norcini and Shea (1997) suggest the following exploratory questions:

- Are the judges credible?
- Is the method used supported by a body of research evidence and data?
- Is the method practicable (too complex a method can lead to errors)
- Can 'Due Diligence' be demonstrated? (i.e. exam security, lack of bias)
- What are the outcomes? ("If you have an outcome which violates common sense then there is something wrong with the standard")

De Gruijter Compromise Standard Setting Method (De Gruijter, 1985)

In the De Gruijter method (De Gruijter, 1985), the judges are asked to estimate the means and standard deviations of the pass rates and cut scores, and then give their estimates of the uncertainty of their estimates. As in Beuk's method, they then draw a graph of scores versus pass rate, plot the estimated cut score and estimated pass rate, and call this point M. Then they plot the relationship between cut scores and pass rates for the candidates as a decreasing curvilinear function (Figure C5). The uncertainty estimates are then used to draw the ellipse of all possible values around M, and where this ellipse touches the plot of student performance is the optimal compromise between cut score and pass rate.

Figure C5 (De Gruijter, 1985)



Differential Item Functioning (DIF) (Zumbo, 2007)

This addresses the question of whether tests are 'fair' 'equitable' between different groups. It was previously known as 'Item Bias' but this is a loaded term, assuming the outcome. In the current understanding, DIF includes Item Impact (real differences in latent trait) and Item Bias.

The main methods are Mantel-Haenszel (MH) and logistic regression (LogR) methods – conditional, probability models. It is also possible to use **Item Response Theory** by exploring differences in Item Characteristic Curves. The new generation of thinking focuses on 'why?' rather than just measurement. Since this is often not obvious from the Item, we need to also consider the 'testing situation'. See **Context Referencing.**

Domains of Assessment

Bloom's Taxonomy (Bloom, 1956) suggests that there are three major Assessment domains: Declarative Knowledge, Procedural Skills, and Behaviours. There are excellent assessments for knowledge, reasonable tools for skills, but few practicable or valid measures for behaviours.

Ebel Standard Setting Procedure

This commences as in Angoff Methods by determine the proportion of the minimally competent candidates who would respond correctly for each item, but then weights items for difficulty and importance, repeating and cumulating for the test as a whole.

Angoff methods assume that difficulty and importance are co-distributed. If this is not the case (and one can readily think of items which are easy and important, and others which are difficult but trivial) we need another way of dealing with it. The Ebel method assigns difficulty and importance on two orthogonal axes. The provision of a range of items of varying difficulty can therefore be negotiated against the provision of a range of items of differential importance. In principal, in terms of relevance, questions can be Essential, Important, Acceptable and Questionable – the last category

should be eliminated during test design! In terms of difficulty, they may be rated Easy, Medium or Hard.

Error estimates

Determining the cut score alone may not be enough – we may need a measure of uncertainty. The Standard Error of Measurement (SEM) (qv) is often used. It may be added to the cut score to reduce false positives, or may define a Borderline Group who receive a 'second look.

Experts, Expert Panel

The idea of the 'expert' involved in standard setting can be defined in different ways ((http://www.edmeasurement.net/5221/Angoff%20and%20Ebel%20SS%20%20-%20TDA.pdf)). However, I propose a simpler definition. The individual expert must be an expert in the domain under assessment, must have at least a basic understanding of assessment processes (including the particular assessment under consideration), and most crucially of all, be thoroughly familiar with the level at which candidates should be expected to operate. This requires familiarity with the normal capabilities of those working at the level of the candidates. Criterion referenced methods fail when there are unrealistic positive or negative expectations of the appropriate level of performance by the candidates.

"Failure to Fail" (Cleland et al, 2008)

There is a well known phenomenon whereby assessors belive in their heart that a candidate should fail, but none the less award a passing grade. There are a number of reasons why they might do this.

- 'I liked them so I passed them/ I didn't like them so I'm trying to be fair"
- "What will happen to them if I fail them?"
- "What will others think?" (compliance and complicity)
- "Whose fault is it? ("Mine as a poor teacher?")
- "Am I sure this is a fail?" (concerns about the assessment and standards)
- "What will happen to me if I fail them?"

False Negative

This term refers to candidates whose 'true score' would meet or exceed the required threshold, but whose actual score (the 'true score' plus the 'error score) on a particular occasion does not reach the threshold. The implication is that those candidates would be appropriate to go into practice, but do not have the opportunity.

False Positive

This term refers to candidates whose 'true score' would not meet or exceed the required threshold, but whose actual score (the 'true score' plus the 'error score) on a particular occasion does reach the threshold. The implication is that those candidates would not be appropriate to go into practice.

Flynn Effect

I.Q. around the world appears to be rising by about three points per decade. The Wechsler Intelligence Scale for Children is re-normed every generation or so, leading to swings in the number of children with 'special educational needs'.

Grade

A Grade represents a qualitative description of performance on an assessment (see **Score**). For instance, 'acceptable' and 'unacceptable' might be awarded or more complex outcomes such as 'unsatisfactory', 'borderline', 'satisfactory' and 'merit'. There is no fixed relationship between a score and a grade (so the pass mark is not always 50%!) The term 'mark' conflates the concepts of score and grade, and is avoided in this report. 'Cut score' is frequently used as defining the boundary between one grade and another, in preference to 'pass mark'.

High Stakes

When important consequences arise from an assessment, it is generally described as 'high stakes'. Summative assessments in medicine are almost by definition high stakes, and this is certainly true for PLAB.

A high stakes exam should be clearly defined as to purpose. It should be 'blue printed' i.e. matched against a body of knowledge which must itself be defined in advance. The development of assessment items requires assessors to be trained, benchmarked and audited. Assessment items should be field tested, and there should be a feedback loop which allows for performance (see below) to be evaluated. The size of the assessment must be suitable to the task. Appropriate standard setting methods must be employed, involving expert staff. Storage and delivery of the assessment items must be secure.

To deliver a national level high stakes exam, an organisation capable of obtaining, testing and administering the equivalence questions in a professional, competent and confidential way would need to be established. This would require selection, training, benchmarking and auditing of question setters. It would be necessary to create a question bank in which performance details of questions was recorded, and to select questions from the bank by means of a blueprint. Since questions would have to be sent to a variety of environments, secure means of communication would have to be established.

Hofstee Compromise Standard Setting Method

In this, judges are asked 4 questions:

1. What is the minimum acceptable cut score?
2. What is the maximum acceptable cut score?
3. What is the minimum acceptable fail rate?
4. What is the maximum acceptable fail rate?

These values are averaged. Then the cumulative sum of candidate scores is plotted, with the scores as the abscissa and number of candidates as the ordinate. The Hofstee limits of these are drawn, and in the rectangle thus generated, the cross diagonal from top left to bottom is drawn. Where it intercepts the plot of cumulative number of candidates is the cut score. See **Compromise Methods, Beuk, De Gruitjer.**

Hofstee can be viewed as a 'safety net 'method, particularly where an assessment is new and/or the assessors are inexperienced. It might be best to confine to the pass/fail boundary – it is trickier at upper and lower bounds where the number of students is low and exceptional performance is rare.

It is possible to retrospectively use Hofstee methods if an approach such as Angoff fails a reality check.

Item Performance

Assessment items can be more or less easy. This property is called *Facility*. If the question is easy, then most candidates can answer it correctly (high facility). Conversely, if a question is difficult, few students can answer it (low facility).

The *Discrimination* of a question shows the range of responses it receives. It might be helpful to think of discrimination as being like the standard deviation of the distribution of the answers, while facility is in some ways like the mean.

Finally, a question may be answered correctly by strong candidates and incorrectly by weak candidates. This can be thought of as a correlation (and for MCQs, is calculated as the *Point Biserial*). The situation of interest occurs when strong candidates tend to get an individual item wrong, suggesting that there is something wrong with the item.

A sophisticated way of looking at the performance of each individual assessment item is *Item Response Theory*. This approach is used by professional testing organisations, such as the Australian Counsel for Educational Research (ACER) and the National Board of Medical Examiners (NBME) in the USA.

Once the performance of individual items has been determined, these can be combined in various ways according to the purpose of the assessment. For instance, a competency assessment can be designed to be most sensitive in the pass-fail zone, while a discriminator assessment might combine items with a much wider range of facilities and strong discrimination properties.

Item Response Theory – see 'Reliability'

Low Stakes

A test which does not in itself lead to serious consequences. It is frequently considered that lower assessment standards may be required of a low stakes test. A number of low stakes assessments may be aggregated to give a 'high stakes' outcome. In such cases an approach such as Generalisability Theory must be used to confirm that a sufficient number of tests are employed to give valid and reliable outcomes.

Mapmark Standard Setting Method

This is a variant of the Bookmark method (Schulz and Mitzel, 2009). Items can be grouped by topic and represented on a visual display with a scale which corresponds to absolute difficulty. Topics can then be scored differently from each other, providing valuable feedback to setters, but also to candidates on their performance.

'Modified Angoff' Standard Setting Procedure:

There are many varieties of Angoff procedures. One can, for instance, have several (generally not more than three) rounds of ratings, with discussion between each round. The aim is to help assessors build a common understanding of the process, and begin to approach consensus (or at least reduce variability). Between rounds of ratings, retrospective information might be introduced – such as the facility of the question, and/or the proportion of candidates currently failing. This is time consuming and may lead to hawks or doves exerting undue influence, as well as not being the point of using the Angoff procedure in the first place.

Nedelsky Standard Setting Procedure:

Expert, informed, experienced Judges determine how many distractors the minimally competent student can eliminate, and the Item Score is the reciprocal of the remainder.

The Nedelsky Method differs from the Angoff approach in that it focuses on each individual alternative, rather than the item as a whole (and may therefore be more robust with regard to poor quality options). It has been criticised on the basis that it forces certain weightings on questions (for instance, a question cannot be weighted between 1 and 0.5. Moreover, there is some evidence that candidates do not primarily answer questions by eliminating, but rather may have a preferred option against which they test alternatives. Gross (1975) adapted the Nedelsky formula to take account of these factors, and Maguire et al (1992) explored the consequences of using the Gross modification in the Medical Council of Canada qualifying examination. They conclude that this is a credible and stable method of standard setting.

Norm referenced; normative

These terms refer to standard setting based on how an examinee performs against a reference population (e.g. those who took the test). See **Criterion Referenced; Context Referenced**

It has generally fallen out of favour in high stakes medical testing, as it is seen as being arbitrary and dependent on the cohort. However, it is still entirely appropriate where a set number of places have to be filled (as a ranking). It is more reliable than criterion referencing, especially with high performing students, and is more robust than criterion referencing (hawks and doves often agree on the relative ranking, but disagree on the absolute grading). It may be used serially as in a Progress Test. An interesting question is whether a minimum number of candidates are required for its employment, and there is no clear context-independent answer to this.

Purposes – Competency and Discrimination

Assessments can be intended either to assess competence ('do all candidates meet a minimum standard?') or to discriminate between candidates ('where do candidates fall with respect to each other on a particular scale?'). Each assessment should be designed for its purpose. For instance, a competence assessment should be most sensitive at the borderline between pass and fail. Discriminator assessments, by contrast, may be designed to be most sensitive in the middle of the range, where most candidates are found. And, naturally, the scoring and reporting scales are different for each kind of assessment. For competence assessments, only two scale points are required – pass/fail, competent/not competent, both for individual assessment items and for the assessment items as a whole. For discriminator assessments, many more points are necessary, and the fineness of the scale required relates to the number of candidates and the intended purposes of the discrimination.

Competency Assessments benefit from Criterion Referencing approaches, while Discriminator Assessments benefit from Norm Referencing.

Purposes – Formative and Summative

Similarly, the distinction between formative and summative purposes is well known – formative assessments offer feedback to candidates and summative assessments determine progression. A widely agreed assessment principle is that formative and summative tests should be kept separate. For instance, Stern (2006) says "*Evaluators must decide the purpose of evaluation prior to developing an evaluation system…Educators planning both formative and summative assessments should use separate and independent systems*". However, all summative assessments can have formative consequences.

Receiver Operating Characteristic Curves

This term is familiar from screening tests. Setting a cut score represents just such a diagnostic test, for which one can define Sensitivity and Specificity if a Gold Standard is present. If one plots all possible cut scores against their corresponding sensitivity and specificity, then one can explore their relationship. Sensitivity is graphed against (1 – Specificity) (so the plot goes from "completely sensitive but not at all specific" to "completely specific but not at all sensitive"). 'Optimum' cut score

is the point closest to the top left hand corner. This is a retrospective approach - one needs the Gold Standard first.

In Martin & Jolly (2002), the 'gold standard' is more than one subsequent failure.

Reliability

Reliability is the degree to which an assessment measures with consistency.

There are several different ways of approaching this.

In Classical Test Theory (also known as Classical Measurement Theory, 'True Score' Theory), it is assumed that any given Score consists of a True Score plus an Error. The error is treated as being of one kind, and it is assumed that the Error can be estimated. Typical tools for exploring this kind of error are Test-Retest estimates, Cronbach's Alpha and tests of inter-rater reliability such as Kappa.

In Generalisability Theory, errors are treated as arising from a number of sources, each of which can be explored and measured separately. More technically, it considers all sources of error (factors) and their interactions, e.g. candidate, marker, item, student-with-item, marker-with item, marker-with-student, and marker-with-student-with-item.

In Item Response Theory, the underlying construct is that there is a relationship between the probability of a candidate answering the question correctly, and the ability of the student. This is expressed as the Item Characteristic Curve. This sophisticated, powerful but complex interpretation is widely but probably exclusively used in national and large commercial testing organisations.

Score

A Score is the raw performance on an assessment (see **Grade**). There is no fixed relationship between a score and a grade. The term 'mark' conflates the concepts of score and grade, and is avoided in this report. 'Cut score' is frequently used as defining the boundary between one grade and another.

Standard

A standard is a statement about whether an examination performance is good enough for a particular purpose. It is based on expert judgement against a social or educational construct, and in that sense, as Case and Swanson (1996) state: "Standard setting is always arbitrary but should never be capricious". See 'Arbitrary' in this regard.

Traditional Method ('Reverse Angoff'):

It is possible to argue that even in the apparent absence of a formal standard setting method, an informal approach which I have suggested is called 'Reverse Angoff' is employed. In this, a draft Item is proposed and then the expert panel adjusts the wording of the Item until it is suitable to pass the correct proportion of candidates (for example, such that a minimally competent candidate would score 50% and a fail candidate 49%). Like Angoff methods, it also relies on good knowledge of the average student. This may underlie the historical use of fixed cut scores, which therefore may not be

as arbitrary as they seem, especially if the assessment is more consistent than the students (which is not always true in medicine).

Up and Down Standard Setting Method

This method is suitable for complex responses such as essays, which have already been scored according to a checklist, then placed in rank order. One selects a sample of candidates near the expected cut score, and assessors agree whether each candidate 'passes' or 'fails' on an expert judgement. If it is graded as a 'pass', the assessors move down the score ranking to next lowest score. If it is a 'fail' then the next highest ranking is selected. This is iterated until a 'zone of passing' is defined. The cut score can be taken, for instance, as the mid-point of this zone.

Utility

The Utility of an assessment was helpfully summarised by van der Vleuten (1996) as

Utility = V x R x E x A x C

where

V = Validity

R = Reliability

E = Educational Impact

A = Acceptability

C = Cost

However, this might better be described as a general relationship than an equation, and the construct of Defensibility (capable of withstanding professional or legal challenge) should be added. Hence, a better formulation might be:

Utility is a function of Validity, Reliability, Educational Impact, Acceptability and Cost and Defensibility.

Validity

Overall, Validity is the degree to which a test measures what it is intended to measure. It relates to Reliability in somewhat complex ways - a measure with low Reliability is sometimes described as being excluded from having high Validity - but Reliability and Validity cannot be traded off against one another in a simple way as is sometimes assumed.

There are a variety of sub-types of validity. Their meanings may sometimes be controversial, but the following operational definitions are used here.

Face Validity: Whether an item makes sense to a panel of experts. One can usefully ask this of one item or question.

Content Validity: Whether the items in an assessment accurately represent the domain being tested e.g. fair sampling. One can usefully ask this of one test or group of items.

Criterion Validity: Drawing inferences between scale scores and some other measure of the same construct. One can usefully ask this of one or more tests.

There are two sub-varieties of criterion validity:

> **Concurrent Validity** is when correlation of one measurement is observed against another measure of known or supposed validity at the same time.
>
> **Predictive Validity** is when correlation of one measurement is observed against another measure of known or supposed validity at a future time.

Construct Validity: A test of the underlying construct. One can usefully ask this of one or more tests. This is the hardest to understand, but an example of a construct is that in a test, higher scores will be progressively obtained by those with increasing levels of expertise. So a test of construct validity would be to give a medical performance test to 1st year students, 5th Year students, Foundation Year 2 doctors, registrars and consultants.

Convergent Construct Validity should be positive where tests are assumed to measure the same construct and Divergent Construct Validity should be negative where tests are assumed to measure different constructs.

References

Beuk CH. (1984) A method for reaching compromise between absolute and relative standards in examinations. *J Educ Measure*, 21:147-152.

Ben-David MF. (2000) AMEE guide no. 18: Standard setting in student assessment. *Medical Teacher,* 22: 120-130.

Boulet, JR, De Champlain AF, McKinley DW. (2003). Setting defensible performance standards on OSCEs and standardized patient examinations. *Medical Teacher,* 25: 245-249.

Bleakley A, Farrow R, Gould D, Marshall R. (2003) Medical Education Complex tasks with an aesthetic component: Making sense of clinical reasoning: judgement and the evidence of the senses *Medical* l 37:544–552

Bloom BS. (1956) "Taxonomy of educational objectives: The classification of educational goals." Handbook I, Cognitive Domain. New York: Longmans, Green, 1956.

Case SM, Swanson DB (1996) Constructing written test questions for the basic and clinical sciences national Board of Medical Examiners, Philadelphia.

Chesser et al (2004) Factor analysis can be a useful standard setting tool in a high stakes OSCE assessment *Medical Education,* 38: 825-31

Cizek GJ, Bunch MB. (Eds). (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.

Cleland J et al (2008) Is it me or is it them? Factors that influence the passing of underperforming students. *Medical Education*, 42: 800-809.

Cohen-Schotanus J, Van der Vleuten C. (2010). A standard setting method with the best performing students as point of reference: Practical and affordable. *Medical Teacher,* 32**,** 154-160.

De Gruijter DNM. (1985) Compromise models for establishing examination standards. *J Educ Measure* 22: 263-269.

Elder A, McManus IC, McAlpine L, Dacre J. (2011) What skills are tested in the new PACES examination? *Ann.Acad.Med.Singapore,* 40:119-125.

Ferrel BG. (1996) A critical elements approach to developing checklists for a clinical performance exam. *Medical Education, Online* 1:5.

Gross LJ. (1975) Setting cut off scores on credentialing exams. A refinement of the Nedelsky procedure. *Evaluation and the health profession,* 8: 469-493.

Hambleton RK, Plake BS. (1995) Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education,* 8, 41-55.

Karantonis A, Sireci SG. (2006) The Bookmark standard setting method: a literature review. *Educational Measurement: Issues and Practice, 4*-12.

Maguire T, Skakun E, Harley C. (1992). Setting standards for multiple-choice items in clinical reasoning. *Evaluation and the Health Professions,* 15: 434-452.

Martin IG, Jolly BC. (2002) Predictive validity and estimated cut score of an objective structured clinical examination (OSCE) used as an assessment of clinical skills at the end of the first clinical year, *Medical Education*, 36: 418-425

Norcini JJ. (2003). Setting standards on educational tests. *Medical Education, 37,* 464-469.

Norcini JJ. Shea JA. (1997) The credibility and comparability of standards. *Applied Measurement in Education,* 10*:* 39-59.

Payne NJ, Bradely EB, Heald EB et al (2008) Sharpening the eye of the OSCE with critical action analysis. *Academic Medicine,* 83: 900-905.

Richter Lagha et al, (2012) A Comparison of two standard-Setting approaches in high stakes clinical performance assessment using generalizability theory. *Academic Medicine,* 87: 8; 1-6.

Schulz EM, Mitzel,H. (2009). A Mapmark method of standard setting as implemented for the National Assessment Governing Board. In E. V.Smith, Jr., & G. E.Stone (Eds.), Applications of Rasch measurement in criterion-reference testing: Practice analysis to score reporting. Maple Grove, MN: JAM Press.

Stern DT, Friedman Ben-David M, Norcini J, Wojtczak A, Schwarz MR. (2006) Setting school-level outcome standards. *Medical Education*, 40: 166-172.

Taylor CA. (2011).Development of a modified Cohen method of standard setting. *Medical Teacher,* 33: e678-682.

van der Vleuten C: The assessment of professional competence: developments, research and practical implications. Advances in Health Science Education; 1996;1: 41 – 67**.**

Zumbo BD. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly,* 4: 223–233.