An Independent Review of the Prescribing Safety Assessment

Professor John Charles McLachlan
UCLAN SCHOOL OF MEDICINE
MARCH 2019

Index

Executive Summary

This Independent Review of the Prescribing Safety Assessment (PSA) was commissioned by MSC Assessment and the British Pharmacological Society.

Consideration was required of the method of Standard Setting employed, the Validity and Reliability of the PSA, the length of the assessment and the setting of the pass mark, item development and the administrative processes underpinning the delivery of the PSA.

The methodology included an ethnographic approach to anonymously interviewing stakeholders, observation of standard setting and item development sessions, review of PSA documentation, and a Rapid Review of the international literature. Interim conclusions were discussed confidentially with a group of international experts. Anonymised raw data from an administration of the test was reviewed and re-analysed.

The outcomes consist of this, the main body of the Review, two Appendices providing respectively a Glossary of terms and a classification and guide to standard setting methodologies, and an attachment containing a checklist of the full list of issues raised in the original brief.

My summary conclusions are that the processes underlying the item development, standard setting and delivery of the PSA are of a high standard, and are comparable with other national level tests. The PSA can demonstrate Face and Content Validity, though further work will be necessary, now that the PSA has been running for a sufficiently long time, to demonstrate Predictive Validity and Construct Validity. Reliability meets widely accepted standards, although it is challenging to calculate accurately given the multi-component nature of the test itself. The standard setting method used (Modified Angoff) conforms to international practice, and remains probably the current best option. The PSA fulfils the claim made for it on the official website that it allows candidates to demonstrate their competencies in relation to the safe and effective use of medicines. The interviews indicated that the PSA has a positive Educational Impact, and increases the attention paid by students and faculty to accurate prescribing.

Overall, I conclude that while no single test could be a *determinant* of patient safety, the PSA is almost certainly a significant contributor to patient safety.

The following detailed Recommendations were made, although the full Review may need to be consulted to understand their context.

Recommendation 1: That all items pass the 'cover test'.

Recommendation 2: That all items with Discrimination below 0.2 are given consideration by the Assessment Board to explore why this might be the case.

Recommendation 3: That a Predictive Validity study is retrospectively conducted through UKMED

Recommendation 4: That a Construct Validity study is organised through the PSA Assessment Partnership.

Recommendation 5: That Generalisability Theory approaches to Reliability be explored, as part of the Annual Psychometric Report

Recommendation 6: That Item Response Theory be explored as a way of generating a Test Information Curve.

Recommendation 7: Angoff Standard Setting should continue in use, while receiving regular review by the PSA psychometrician

Recommendation 8: That the Standard Setting Board explores the possibility of using shared electronic scoring rather than verbal scoring.

Recommendation 9: That Item Facility and Discrimination be routinely reviewed for all items where such data are available.

Recommendation 10: That the Standard Setting Panel considers a formal process of 'flags' so that issues raised during standard setting are addressed by the Assessment Board.

Recommendation 11: That the 'Yes/No' Angoff methodology be included in the initial discussion of the Just Passing candidate.

Recommendation 12: That consideration be given to standard setting panel members scoring items in advance of the meetings

Recommendation 13: That the PSA Executive Board considers reducing 'new item generation' to 300 per year

Recommendation 14: That the PSA remains separate from the MLA

Recommendation 15: That the PSA Executive works with the GMC MLA content mapping team to resolve areas of overlap

Recommendation 16: That the PSA should be extended to IMGs who are exempt from PLAB by reason of 'internship' or other experience

1. Introduction

This independent Review on the Prescribing Safety Assessment (PSA) was commissioned by the MSC Assessment and the British Pharmacological Society.

In the brief for this Review, I was given an extensive list of queries (see Attachment 1). However, these do not correspond to the most natural way of addressing the issues involved, so I have re-structured these for the Review itself. I believe that all the salient issues are addressed, however, an Attachment 1 is included to address the few minor points not covered.

The PSA is a computer delivered 60-item test, delivered in the first instance to final year undergraduate medical students. It is a 'moderate stakes' test, in that failure at the first attempt does not prevent a candidate from proceeding to Foundation Year 1, but it must be passed successfully during Foundation Year 1 in order to progress further. I do not intend to document the PSA in detail here, since it is well described on the PSA website.

Methodology

In view of the time and resource constraints, I adopted an ethnographic approach to the task, involving observation of key activities, and an informal interview strategy with stakeholders, based on opportunistic and snowball sampling. Interviews and discussions were anonymous, confidential and participant-led, rather than structured by the interviewer. I carried out a review of PSA documentation, including minutes and reports, and a rapid review of the international literature on standard setting. Anonymised raw data from an administration of the test was reviewed and re-analysed. Interim conclusions were discussed confidentially with a group of international experts. The task also gave me the opportunity to reflect deeply on psychometric principles, in ways which may be of value in other contexts.

In order to keep the main body of the Review relatively succinct, two Appendices are provided, including a Glossary of Terms and a Guide to Standard Setting Methods. There is also an attachment which is a checklist of the issues raised in the original brief. The Appendices draw on my previous publications and teaching.

2. Is the PSA *Valid*?

Validity is a complex and disputed concept (See Glossary for more detail), but I will focus on four kinds of validity in this Review. These are *Face*, *Content*, *Predictive* and *Construct Validity.*
Face Validity relates to whether an item is credible to a panel of experts. One can usefully ask this of one item or question.
Content Validity relates to whether the items in an assessment accurately and proportionately represent the domain being tested (see Glossary for *Domain*) and is therefore asked of the test as a whole.
Predictive Validity in this context relates to whether the test is a positive predictor of later clinical practice.
Construct Validity is concerned with the larger inclusive question of whether a test measures what it intends to measure. In some definitions, it includes reliability as a sub-category, although I prefer to keep these separate. I will interpret it here it its narrow sense.

Does the PSA show Face Validity?

I attended a session of the PSA Assessment Board, at which items and the structure of the tests are discussed. In summary, the Board was composed of an appropriate number of subject matter experts (see Glossary), with a variety of backgrounds. The meeting was conducted to a high standard, with extensive and appropriate consideration given to each item.

Overall, I conclude that the process meets the requirement for demonstrating item Face Validity in accordance with best practice nationally and internationally. I make two specific recommendations for further improvement in practice.

Recommendation 1: That all items pass the 'cover test'. (See Glossary)

Recommendation 2: That all items with Discrimination below 0.2 are given consideration by the Assessment Board to explore why this might be the case.

Does the PSA show Content Validity?

The structure of the test as a whole is a part of the original design. I observed nothing to suggest that the subject matter experts could challenge the design, by, for instance, suggesting major areas which were not covered. *Overall, I conclude that the process currently meets the requirement for demonstrating Content Validity in accordance with best practice nationally and internationally.*

An issue may arise with the development of the GMC's Medical Licensing Assessment, particularly the Applied Knowledge Test. This may overlap with some areas of the current PSA – see Section 10.

Does the PSA show Predictive Validity?

This has not yet been established, which is understandable in view of the date at which the PSA was introduced. However, sufficient time has passed since the introduction of the PSA, that candidates are in practice, and performance data will continue to accumulate as long as no radical changes are made to the structure of the PSA. In general, there is internationally consistent evidence that performance on tests such as USMLE and the Medical Council of Canada Qualifying Examination is indeed predictive of later clinical practice, and the development of the UK Medical Education Database under the aegis of the General Medical Council (GMC) and the Medical Schools Council empowers such studies.

Recommendation 3: That a Predictive Validity study is retrospectively conducted through UKMED

Such a study could for instance compare PSA scores with outcome measures such as normal progression at ARCP, Royal College assessment, and Fitness to Practise issues. It would be most valuable to be able to include referrals to the former NCAS (now Practitioner Performance Advice, with NCAS falling under the aegis of NHS Resolution). Since UKMED may charge a data access fee for such studies, it would be helpful if the research call offered to at least cover the costs of accessing UKMED, and perhaps also the costs of open access publication.

Does the PSA show Construct Validity?

Even in the narrowest sense of the term 'Construct Validity' this cannot currently be answered. Along with a study of predictive validity, a construct validity study would be of great value. One way in which

this might be conducted is to invite current Foundation Year 1 doctors, Foundation Year 2 doctors, Registrars and Consultants, to undertake the PSA concurrently with a normal cohort of medical student, to compare their relative performance.

Recommendation 4: that a Construct Validity study is organised through the PSA Assessment Partnership.

Such research studies would benefit from being promoted in a targeted way to researchers in the field, perhaps through the Medical Education Societies. Since I am advising that such research be carried out, I recuse myself from carrying out the research, so that there is no question of a conflict of interest.

Validity and Claims

A key aspect of modern thinking about Validity relates to claims made about what inferences can be drawn from a test (See Glossary).

The PSA makes a relatively modest claim with regard to these inferences: as described on the PSA website, it indicates that "the Prescribing Safety Assessment allows candidates to demonstrate their competencies in relation to the safe and effective use of medicines". I believe that this claim is justifiable. By contrast, the GMC make the claims for the planned Medical Licensing Assessment that "those who obtain registration with a licence to practise medicine in the UK [have met] a common threshold for safe practice" and that the CPSA will "demonstrate that an individual is capable of functioning safely as they enter clinical practice in the UK". These claims are much less sustainable.

3. Is the PSA Reliable?

Currently, reliability of the PSA is largely approached through Classical Test Theory approaches (see Glossary). The two metrics used are Cronbach's α and the Standard Error of Measurement (SEM).

However, Cronbach's α is a weak measure for this type of assessment. It is largely a measure of internal consistency, and very high values are often a measure of redundancy rather than reliability. For example, if a test contained the same question repeated 100 times, it would have a very high value of Cronbach's α! It is also influenced by the magnitude of the variance. High variance will lead to a higher value of Cronbach's α. This leads to the counterintuitive consequence that we could increase the value of Cronbach's α for a given test, by adding low performing candidates to the test cohort, even though the test remains unchanged!

This is partly corrected for by using the Standard Error of Measurement, since this includes a measure of dispersion as a multiplier in the formula. But the SEM is still a function of Cronbach's α in Classical Test Theory approaches, and therefore the problem persists to some extent.

This problem is exacerbated by the nature of the PSA, which by design has 8 different sections, each exploring a different aspect of prescribing. These vary from arithmetical calculations to explanations to patients. It seems unlikely that a common latent trait underlies all of these components.

Given these contributions to variability, it seems positive that the PSA records values of Cronbach's as high as it actually does (averaging 0.722 across all four testlets in the 2018 Psychometric Report, but with one testlet at 0.692).

Are there better ways of measuring Reliability? The two major alternatives are Generalisability Theory and Item Response Theory (IRT: see Glossary). However, Item Response Theory relies on a crucial assumption: that there is a latent trait which predicts in continuous fashion the likelihood that a candidate will answer correctly. I believe that the multi-component nature of the PSA is at odds with this assumption. There is extensive evidence in the medical education literature for the phenomenon of case or content specificity. This implies that how a candidate performs in one scenario does not necessarily predict how they will perform in another scenario.

Despite this, IRT approaches are frequently viewed as representing 'best practice' in assessment. I believe that this is due to the fact that a small input of data can seem to generate a larger volume of output data! The reason for this is that the underlying assumption of IRT allows a single value (such as Facility) to predict an Item Characteristic Curve, with its associated values. Further, since IRT is usually described in mathematical form, it is sufficiently opaque to the casual reader that its assumptions are not obvious.

From this it would seem to follow that Generalisability Theory might be the best way of assessing the Reliability of the PSA. Use of Generalisability Theory would also permit a Decision or D Study to be carried out, perhaps pointing to particular sources of variability that could be addressed.

However, uses of Item Response Theory should not be completely ruled out, since despite its assumptions it would be useful in generating a Test Information Curve, showing where the PSA is currently most sensitive.

Recommendation 5: That Generalisability Theory approaches to Reliability be explored, as part of the Annual Psychometric Report

Recommendation 6: That Item Response Theory be explored as a way of generating a Test Information Curve.

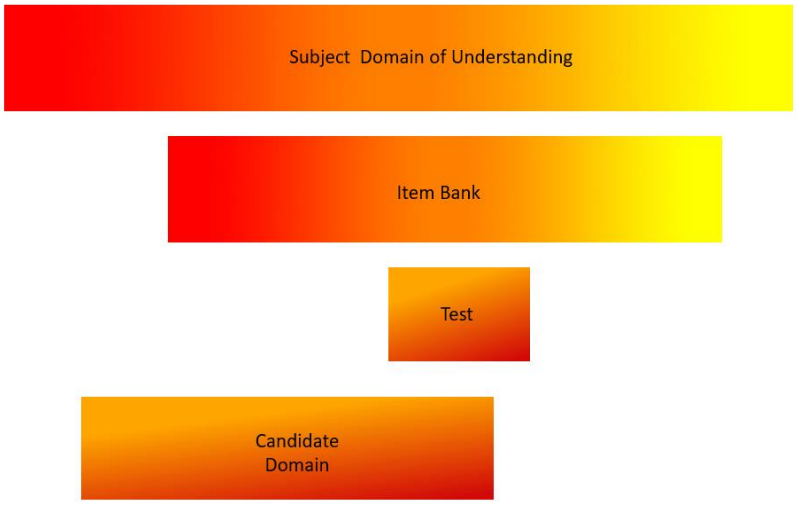
There are other sources of variability which are not addressed by standard approaches to variability, and are important here and in Section 6 on the 'pass mark'.

In any assessment, there is a body of information on which candidates might be assessed. I will call this the 'Subject Domain of Understanding', or Subject Domain for short. I prefer 'Understanding to 'Knowledge', particularly with regard to the Subject Domain for the PSA, which I believe includes elements of skill as well as declarative knowledge. In this case, the Subject Domain is intended to be 'Anything that an F1 might know'. This is hard to imagine, but can perhaps be visualised: the PSA Assessment Board included a variety of experienced individuals, some medically qualified, some pharmacists. The collective expertise and understanding of these individuals as to what an F1 doctor ought to know probably represents the Subject Domain very reasonably. However, the Subject Domain cannot readily be codified for the benefit of candidates.

It may help to represent the Subject Domain visually, as is done below.

However, the Item Bank is smaller than the Subject Domain, by a certain amount (discussed again in Section 6). The Test itself is a relatively small sampling from the Item Bank.

Each candidate possesses a personal Domain of understanding, which is generally smaller than the Subject Domain, but crucially, may not be identical with the Item Bank. These relationships are presented in the Figure below.



The Item bank is drawn entirely from the Subject Domain, and the Test is drawn from the Item Bank. Unsurprisingly, the Candidate Domain and the Test do not fully overlap: the Test will contain items that the candidate cannot answer. But the candidate domain need not overlap with the Item bank either. The candidate may have learned things which are not in the Bank, and therefore cannot be sampled. The candidate's Test Score cannot therefore be viewed as an estimate of how much of the Subject Domain they are possessed of.

The relative proportions of the Item Bank, the Test and the Candidate Domain to the Subject Domain determine the magnitude of the resulting error. The probability that an item in the test matches the candidate's understanding is the product of these respective probabilities.

The relationship between the number of items in the Bank, the number of items in the test, the number of items a candidate could answer correctly, and the number of items the candidate gets correct can be represented by a hypergeometric distribution.

$$p_X(k) = \Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

N is the number of items in the bank
K is the number of items the candidate can answer correctly
n is the number of items in the test
k is the number of items the candidate gets correct on the test

This would allow calculation of the variance around the actual candidate score on the test compared to the bank. But there would be a separate calculation based on the sizes of the relationships between the Bank, the Domain and the Test.

My general conclusion would be that standard estimates of 'reliability' based on internal consistency are generally underestimates. This will also impact on the meaning of the pass mark.

4. Is Modified Angoff the 'best' standard setting method to use?

There is no ideal standard setting measure: moreover, different standard setting methods tend to give varying outcomes. Appendix 2 provides an extended list of possibilities.

The current methodology, generally described as 'Modified Angoff', is widely used in comparable professional settings such as the GMC's Professional and Linguistic Ability tests, and therefore has credibility. It is familiar to the standard setting staff, and has been used since the inception of the PSA, contributing to the body of custom and practice which helps ensure the smooth operation of the process.

A major challenge to the Angoff method is the difficulty of conceptualising the Minimally Competent Candidate. The literature even at times uses the term 'Borderline Candidate' as a synonym, even though this is quite a different construct. This was addressed in the standard setting meeting by a preliminary discussion among the panel Members lasting for one hour, and the use of a defined form of words to describe what is named here as the 'Just Passing Candidate':

Just passing candidate:

- Makes ***no lethal or very serious errors***, but may make a few significant errors
- Is familiar with most common treatments and ***has no major gaps in knowledge*** that could compromise patient safety
- When prescribing, may make SOME suboptimal treatment and therapeutic decisions, but ***not life-threatening*** ones
- Has just about acceptable ***ability to differentiate*** and communicate most appropriate lines of treatments
- Is competent in using the British National Formulary
- Is aware, but not ALWAYS consistently, of patient groups at risk e.g. elderly, pregnant
- Is competent at ***basic one- and two-step calculations***, BUT not consistently competent at others.

This represents good practice in addressing the challenge of construct conceptualisation. To be clear, the definition should not itself be re-negotiated each year. It is very good of its kind, and should remain the benchmark. However, it is valuable to explore its meaning again each year, particularly for new panel members, but also to ensure that all panel members have a common understanding of what the definition means.

In practical terms, a major drawback is cost. Three days were set aside for an extensive Panel of reviewers to meet face to face, and items had been read in advance by the reviewers.

The operation of Angoff methodology after these steps is relatively simple to understand: however, the problems arising from it are also easy to understand! The same could not be said for some other

methods of standard setting, whose technical complexity may militate against a quick appreciation of their flaws.

Recommendation 7: Angoff Standard Setting should continue in use, while receiving regular review by the PSA psychometrician

Alternatives to Angoff include the Ebel methodology (see Glossary) and those relying on Item Response Theory. However, in addition to the assumptions described above, IRT itself is not a standard setting method. The two most widely used standard setting methods using IRT are Bookmark and Mapmark (see Glossary), and these merely use IRT as a way of ranking items, not of actually setting the standard, which is still done by expert reviewers, as in Angoff methods.

The Ebel methodology was considered as part of this Review, since it may be less time consuming to administer. However, it poses challenges of its own. It still requires visualisation of the concept of the Minimally Competent Candidate, but additionally requires the classification of items in terms of difficulty and importance. In addition, the current Angoff standard setting approach can usefully be informed by performance data from previously used item (See Recommendation 2). This benefit would be lost on changing to a new standard setting method. No compelling argument for a change to the Ebel method presented itself to me.

5. How robust are the item development, standard setting, and general administration processes?

The Assessment and Standard Setting Boards are conducted in an appropriate and professional manner, and to a high standard, and are suitable to their tasks. Recommendations 1 and 2, already described, represent proposed modifications to the item development processes. The following recommendations relate to the Standard Setting process.

Recommendation 8: That the Standard Setting Board explores the possibility of using shared electronic scoring rather than verbal scoring.

Recommendation 9: That Item Facility and Discrimination be routinely reviewed for all items where such data are available.

Recommendation 10: That the Standard Setting Panel considers a formal process of 'flags' so that issues raised during standard setting are addressed by the Assessment Board.

Recommendation 11: That the 'Yes/No' Angoff methodology be included in the initial discussion of the Just Passing candidate.

Recommendation 12: That consideration be given to standard setting panel members scoring items in advance of the meetings

I explored other processes such as exam security, the appeals process, and the process of reasonable adjustments, through discussion with the relevant members of the PSA partnership. The reasonable adjustment process corresponds with those used in other high stakes setting. It is also the subject of a separate report, in the discussion of which I was able to share. The Appeals process seems to have been

little used, which is a positive marker for candidate content. Exam security corresponded to the level normally applied in other high stakes tests.

The PSA relies upon an 'in house' item banking system, which has both benefits and challenges. An in house system allows the bank to be customised to the needs of the organisation as compared to a commercial 'off the shelf' system. By contrast, a commercial system is regularly updated and maintained, while an in house system in general may be vulnerable to changes in personnel over time. These general reflections do not represent a recommendation for a change.

6. Is the Pass Mark too low? (Or, "Shouldn't the pass mark be 100%?")

As indicated above, there is a variety of sources of variance in a test outcome.

First, no one individual question (Item) has perfect Validity or Reliability. Even if unusually, an item were perfect when it was written, it may not be perfect by the time it is delivered, since medicine is in a state of constant change!

Second, no test is perfect, and indeed, there is variance between the different papers undertaken on different occasions in the PSA.

Third, candidates perform differently on different occasions, so even if they had sat the 'same' test on different dates (if that were possible) then they would perform differently.

Fourth, and most crucially, is the relationship between the Subject Domain and the Candidate Domain described above in Section 3. In medical students, we are dealing with learners who have not yet entered practice. If it were possible for beginners to know everything, then we would have no need of career development. Although the PSA is understandably generally described as a high-stakes test, it might be more accurately described as a medium-stakes entry test, since 'failure' does not initially prevent progress to the next stage of a student's career path. The remediation and resit process takes place during the process of actually doing the job. In this phase F1s are under supervision by a variety of more expert colleagues such as pharmacists and other doctors. The relationship between the Subject Domain and the Candidate domain cannot therefore be 100% or anything like it. Indeed, to my mind the surprising thing is how well students perform on the PSA given all these sources of variance– the median score is around 75%. With this, and the sources of variance in mind, the pass marks set for the PSA (frequently around 63%) seem appropriate.

This argument has technical components, and if I were to attempt to summarise it in lay terms, it would be as follows.

'Medicine' is a huge subject, and nobody knows it all. And the best kind of learning happens through experience. But beginners by definition don't have much experience.
We train medical students hard for 5 years, and they know a surprising amount, but they can't know it all, and the questions in the exam are chosen at random from 'everything they *might* know'.
Even excellent experienced doctors wouldn't always score 100%! The fact that the 'average student' scores about 75% is pretty astonishing. Once they have qualified, 'Beginner doctors' are supervised very closely in their first and early years, so that when they make mistakes, there is a process for spotting and correcting them. Such supervision remains necessary so no assessment (or learning experience) could guarantee that a learner would be as safe to practice as an experienced expert. Clinical judgement is

profoundly informed by experience, as is well established in the literature. It cannot be claimed for the PSA that those who pass it are always 'safe': on the contrary, high performers will still make mistakes in practice with a lower probability than low performers. Novices need time and practice to be experts.

7. How do stakeholders perceive the PSA?

The following groups of stakeholders were consulted: current medical students, teachers in medical schools with responsibility for preparing students for the PSA, junior doctors, educational supervisors and more senior individuals such as Directors of Medical Education, and pharmacists. BMA students were also approached for responses. The methodology was ethnographic in nature, featuring opportunistic and snowball sampling, commensurate with the timetable and resources available. Anonymity and confidentiality were assured to all participants. This was not a full qualitative research study.

Students in general were positive about the PSA, although the time pressure was felt to be 'unrealistic'. (There is in fact no evidence that the PSA shows speededness). The level of some items was thought to be above that of an FI, and students reported some local delivery issues (computer access and network stability).

In general, students described a positive Educational Impact of the PSA, in that it led to increased study, learning and practice on prescribing issues.

Teaching staff were also generally positive about the PSA, both about the assessment itself and the impact on student engagement with prescribing. There was variation in the amount of philosophy of teaching, with generic 'teaching about prescribing' and one end of the spectrum and 'teaching specifically to the PSA' at the other.

Post-qualification medical staff such as educational supervisors were also generally positive, especially with regard to the possibility of remediation during the F1 year for those who had failed. A number of special programmes for F1 doctors who were still having difficulty were described.
Pharmacists in general were slightly more sceptical, noting that they still observed prescribing errors on the part of F1 doctors, and one noted that she felt there had been no clear difference before and after the introduction of the PSA.

I also contacted a number of internationally eminent psychometric experts to discuss standard setting in high stakes exams in general. Their views have informed the discussion in this review, but the views expressed are my own.

8. Is the Item Bank big enough?

This is hard to confirm, but I will advance two lines of argument. The first is that the existing bank is comparable with other banks which attempt to address much larger Domains of Understanding. A useful comparison is to the bank size of the Medical Schools Council Assessment Alliance, where the relevant domain is everything the medical student should know in terms of declarative knowledge over the last three years of the programme (early years are less well developed in the MSCAA bank). The current PSA bank is also comparable to the PLAB Part 1 Bank.

The second argument inverts the logic laid out in Section 3 above as to the relative sizes of the Subject Domain and the Bank. If the median candidate can score 75%, the Bank *must* be relatively large compared to the Subject Domain.
I therefore believe that the bank has reached an appropriate size, and now largely needs maintenance (a) for currency and new items, and (b) to guard against item exposure. I am conscious that Item Development at the current rate is a significant component of the cost of delivering the PSA, and while it has been necessary in the early stages, the time has come to reconsider the requirement for new items.

Recommendation 13: That the PSA Executive Board consider reducing 'new item generation' to 300 per year

As an informal comment short of a recommendation, I suspect that many of the PSA items would lend themselves to Automatic Item Generation (see Glossary) and this may be worth exploring in the future. And there may be opportunities to share item development with other bodies in the field such as the General Pharmaceutical Council's Registration Assessment.

9. Does the PSA ensure safe practice of those passing?

A single test of this kind cannot achieve this goal. Even if candidates could be dichotomised into 'safe' and 'unsafe', measurement error would mean that these could not be unequivocally distinguished
In practice, even this dichotomy is problematic: instead, a valid test might establish the *probability* that a candidate will be safe in practice, other things being equal. As indicated above, the claim made for the PSA, that the "test allows candidates to demonstrate their competencies in relation to the safe and effective use of medicines" is realistic and supportable. There is a tacit claim of many such tests that they can determine who should proceed to the next stage of their career. It is a strength of the PSA that it does not immediately make this claim, and unsuccessful candidates may progress and gather experience in real practice before re-attempting the test.

What can be said is that other national level tests such as the USMLE and the Canadian equivalent have predictive evidence supporting their ability to predict clinical practice. This reinforces the rationale of Recommendation 3.

We may appropriately ask, by what rationale is an assessment standard established as 'defensible' in legal terms (See Glossary for citation). This can be summarised as follows:

1. The standard setters are credible, i.e. they know the subject, know the correct level for the candidates, and are familiar with assessment methodology.
2. The standard setting method used is supported by a body of research evidence and data
3. The method used is practicable (since too complex a method can lead to errors)
4. 'Due Diligence' in exam delivery can be demonstrated, in for instance item bank and exam security
5. The outcomes are reasonable in nature
6. The process is equitable between candidate categories

The last of these points is difficult to answer. It is possible to calculate Differential Item Functioning (see Glossary) for each item in a test, but across virtually all large scale tests of attainment there is evidence of differential performance of different population groups. This is almost certainly due to long standing social imbalances of various kinds, and is not readily addressable within a single test. Long term societal

change will be necessary to eliminate these imbalances. However, 'equitable' in this context may be taken to mean that all candidates have an equal opportunity as far as the test itself, and its corresponding reasonable adjustments, are concerned. With this proviso, the PSA meets the requirements for defensibility.

Given the impact on student behaviour in terms of study and learning, and the evidence gathered in this review, the PSA is almost certainly a contributor to patient safety. However, no single test could be a *determinant* of patient safety.

Similarly, we could ask if the operation of the PSA should re-assure the public.

It is necessary to expand on who 'the public' are in this context. There are five possibilities.
It could be the general UK public, as surveyed by opinion polls. I know of no surveys which express concern about the level of training of UK-trained doctors. There are expressed concerns about the performance of non-UK trained doctors.

The popular media may be considered as part of the public in this context. However, news reporting of health issues is a contentious issue, and positions may be adopted which are not based on sustainable evidence. To this extent, satisfying the popular media is a chimaera.

Elected representatives, such as MPs and Government, represent another public interest group. There have been expressed concerns about medical errors, including prescribing errors, and patient safety. However, the PSA may be taken to represent an attempt to alleviate these concerns.

The law as interpreted by the courts is a further marker of public interests. Again, I know of no cases raised against the PSA, and such cases are unlikely to focus on the content or outcomes of the PSA itself, since it is not the remit of the courts to overrule considered academic judgements. Rather, courts will intervene where there are failures of due process; and currently the PSA has observed due process in its operations.

Finally, there are the stakeholder groups that work with the PSA partnership, and I have seen evidence of full but co-operative discussions with these stakeholders, leading to mutual understanding.

In the end, however, the decision as to whether or not the PSA performs to the public's satisfaction can be described as political, or if the term is preferred, societal.

10. What should be the relationship between the PSA and the MLA?

Currently, in addition to carrying out this independent review of the PSA, I am also serving on the GMC's Medical Licensing Assessment (MLA) Applied Knowledge Test Expert Reference Group. The relationship between these two is therefore of particular interest to me. Although this was not part of my formal remit, the imminent arrival of the MLA seems to indicate that it would be remiss not to consider the relationship between the two.

A key question is whether the construct or latent trait tested by the current PSA is the same as that which will be tested by the MLA. If the latent traits are believed to be the same, then there is an argument for subsuming the PSA into the MLA AKT.

In my view, however, they are different. The trait or traits tested by the PSA include a considerable skills element, in, for instance, writing a prescription, carrying out calculations, and reviewing prescriptions. Skills of calculation can be carried out independently of declarative knowledge – it would be quite possible to set a calculation which featured an imaginary drug. The Equip study confirmed that graduates who had necessarily performed adequately in declarative knowledge still had difficulty in writing correct prescriptions. The differing formats employed in the PSA from the format proposed for the MLA AKT also suggest different characteristics are under test. Informally, it is quite common for individuals to expand 'PSA' to 'Prescribing Skills Assessment' in conversation!

Certainly, in terms of public perception, I believe that the public, however defined, would identify prescribing skills as separate from general medical knowledge. I also believe it would be assigned a high value as a necessary set of skills for a newly qualifying doctor.

In any case, if the PSA were incorporated into the MLA, it would necessarily be a small component, and therefore of markedly lower reliability. Moreover, a *full compensation* model would operate, in which lack of knowledge on prescribing could be compensated for by knowledge of other aspects of the curriculum.

However, there may be some overlap, since there are elements of the PSA which arguably are declarative knowledge. I would recommend that the PSA works with the GMC, on a long term developmental basis, to optimise the relationship between the PSA and the GMC: the relative weighting of some of the sections might be increased, for instance, in line with evidence that prescription *writing* is particularly challenging for students at this level. The required amount of prescription writing in the PSA might usefully be increased, at the expense of some more generic knowledge based material, if discussion suggested this was useful. Overall, however, I would strongly recommend that the PSA remains a separate assessment.

Recommendation 14: That the PSA remains separate from the MLA

Recommendation 15: That the PSA Executive works with the GMC MLA content mapping team to resolve areas of overlap

Recommendation 16: That the PSA should be extended to IMGs who are exempt from PLAB by reason of 'internship' or other experience

I have previously made these arguments both through the MLA Expert Reference Group, and in correspondence with PSA members, and note that the GMC website[1] currently states:

"Will the MLA include the current situational judgement test (SJT) and prescribing safety assessment (PSA)?"

"No. These will remain separate".

[1] https://www.gmc-uk.org/education/standards-guidance-and-curricula/projects/medical-licensing-assessment/the-mla-and-uk-students
Accessed 07/03/19

11. Summary Conclusions

My summary conclusions are that the processes underlying the item development, standard setting and delivery of the PSA are of a high standard, and are comparable with other national level tests. The PSA can demonstrate Face and Content Validity, though further work will be necessary, now that the PSA has been running for a sufficiently long time, to demonstrate Predictive Validity and Construct Validity. Reliability meets widely accepted standards, although it is challenging to calculate accurately given the multi-component nature of the test itself. The standard setting method used (Modified Angoff) conforms to international practice, and remains probably the current best option. The PSA fulfils the claim made for it on the official website that it allows candidates to demonstrate their competencies in relation to the safe and effective use of medicines. The interviews indicated that the PSA has a positive Educational Impact, and increases the attention paid by students and faculty to accurate prescribing.

Overall, I conclude that while no single test could be a *determinant* of patient safety, the PSA is almost certainly a significant contributor to patient safety.

Acknowledgements

Grateful thanks are due to Lee Page, Veronica Davids and Dr Lynne Bollington for their prompt and helpful replies to the various queries and requests made to them during the course of this Review.

About the author

I am currently Professor of Medical Education and formerly Deputy Head of School at UCLan Medical School. I am a GMC Associate, and have carried out a number of projects commissioned by the GMC, HEE, the Department of Health, and others. Currently I serve on the GMC Applied Knowledge Test for the proposed Medical Licensing Assessment, and on the Recruitment Development Group of the UK Foundation Programme Office. Previously I was a Board Member of the UK Clinical Aptitude Test, and Editor-in-Chief of the journal *Medical Education*. I have no conflicts of interest which impact on this project.

My involvement in particular with the setting and administration of the UKFPO Situational Judgement Test, UK Clinical Aptitude Test, and the GMC's PLAB mean I am able to directly compare the PSA with other national tests aimed at future doctors.

In addition to having published widely in assessment research, perhaps of value to this project is that I have been involved in teaching and assessing medical students over many years.

Since I recommend modest funding for several research projects, I recuse myself from benefiting in terms of future research funding from any of the recommendations in this Review.

Appendix 1. Glossary of Terms

Arbitrary

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

Automatic Item Generation

A number of approaches have been made to generating multiple versions of MCQs from a 'stem bank' of keys and distractors (see Lai, H et all, 2017, for example). Here, work is required to generate the appropriate template, but once this has been done, a significant number of versions from the stem bank can be generated. This might lend itself particularly well to calculation items of the kind used in the PSA.

Case Specificity

This is the well established phenomenon that "Regardless of the assessment method used, performance on one case does not predict performance on other cases" (Swanson and Roberts, 2016). Also known as Content Specificity.

Compensation (see also Conjunction, Conjunctive)

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.

Competency see Purposes – Competency and Discrimination

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.

Conjunction, Conjunctive

In order to ensure that there is not full compensation between different components of an assessment, it is a common practice to put in place conjunctive rules (also known as **profile criteria**). These add an additional requirement for 'passing' the test to that of exceeding the cut score. Typically, they may be of the form that candidates must also pass a minimum proportion of items. For instance, the GMC PLAB Part 2 OSCE requires that candidates pass a minimum of 11 out of 18 stations.

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**, in which scores on IQ tests increase year on year, requiring regular re-calibration). 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.

Classical Test Theory

In Classical Test Theory (also known as Classical Measurement Theory and 'True Score' Theory), it is assumed that any observed 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.

Cover test

The cover test is an aspect of good MCQ item design, in which during the design process the item is reviewed by subject matter experts, without access to the alternatives. If the experts can successfully deduce the correct response from the stem and instructions alone, the item can be said to have passed the cover test. This ensures that items are consistent in the nature of the alternatives (e.g. all dealing with diagnoses, or all dealing with management) and that the desired correct response is indeed the best available.

Decision Study (D Study) see Generalisability Theory (G Theory)

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)
- Are the outcomes reasonable? ("If you have an outcome which violates common sense then there is something wrong with the standard")

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.

It is 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 Referenced.**

Discrimination see Item Performance or Purposes – Competency and Discrimination

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. A revision of the taxonomy (Anderson and Krathwohl, 2001) enjoys wide-spread acceptance. In the Declarative Knowledge Domain, 'creating' becomes the top level of the hierarchy. The taxonomy is useful both for setting assessments at different levels, and for scoring constructed response items.

I have also developed the term 'Subject Domain of Understanding' to encompass all that the candidate might be examined on. In some circumstances this might correspond to the syllabus of the course, but in health care in particular, it may be very much wider than a formal syllabus, and incorporate almost anything that the learner may encounter in clinical practice.

Error estimates

Determining the cut score alone may not be enough – we may need a measure of uncertainty. The **Standard Error of Measurement** (SEM) 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.

Expert, 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 are 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.

Facility see Item Performance

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

There is a well known phenomenon whereby assessors believe 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?" ("Will my decision be challenged, perhaps with accusations of bias?")

False Negative

This term refers to candidates whose 'true score' would meet or exceed the required threshold, but whose observed score (the 'true score' plus the 'error score in **Classical Test Theory**) 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. This has implications for the **Sensitivity** and **Specificity** of a test.

False Positive

This term refers to candidates whose 'true score' would not meet or exceed the required threshold, but whose observed score (the 'true score' plus the 'error score in **Classical Test Theory**) on a particular occasion does reach the threshold. The implication is that those candidates would not be appropriate to go into practice. This has implications for the **Sensitivity** and **Specificity** of a test.

In this 'True and False Positive terminology', there is a tacit acceptance of the idea that candidates can be dichotomised into positive and negative categories. In reality, the situation is much more complicated, as described in the main body of the PSA Review.

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'.

Formative see Purposes – Formative and Summative

Generalisability Theory (G Theory)

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. It can then be used to identify where the major sources of error occur, and attempts to remediate these can be made. For instance, if marker variance is the biggest source of error, then marker training might be stepped up.

G Theory calculations can be extended to perform a Decision or D Study. This models the effect of altering the parameters, for instance increasing the number of items in the test, or increasing the number of markers for any given item.

Grade (see Marks and Grades)

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 (see Low 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.

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 or Item-Total correlation*. If the item under consideration is removed from the total, this may be described as the *Item-Rest correlation*. A 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.**

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 (IRT)

In Item Response Theory (Bond and Fox, 2007), the underlying construct is that there is a relationship between the probability of a candidate answering the question correctly, and the ability of the candidate. In Rasch's words (Rasch, 1960),

"A person having a greater ability than another person should have a greater probability of solving any item of the type in question".

This relationship can be expressed for any item is expressed as an Item Characteristic Curve, showing this relationship. This sophisticated, powerful but complex interpretation is widely 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. It purports to be able to offer an absolute difficulty of any item, irrespective of the cohort which undertook the assessment, and is therefore viewed as very powerful. It empowers the calculation of a Test Information Curve, which indicates where a test is most sensitive.

However, to my mind, its inherent weakness is expressed in the definition "*any* item of the type in question" (my emphasis). This is not in accord with the substantial body of evidence for Case Specificity in medicine. I wonder if IRT appeals most to mathematicians, because in this field 'items of the type in question' have powerful similarities, and perhaps the assumption holds good. In the practice of medicine, 'items' vary enormously, and the best way to comprehend them is through experience.

Item-Total, Item-Rest see **Item Performance**

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.

Point BiSerial see **Item Performance**

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. This is at odds with the 'everyday usage' meaning of reliability as indicating 'that which can be relied on', and the confusion arising from these two meanings is widespread. The meaning of 'that which can be relied on' is actually closer to Validity in these technical usages. For instance, the reliability of a test in which the same item were asked 100 times would be perfect, but the validity would be very low. Conversely, a reliability coefficient for a test is no more an absolute marker of quality than a candidate's score on a test is an absolute marker of their ability. It would be perfectly reasonable to apply an error range to a reliability estimate itself, as to use it to as a way of calculating error ranges!

It is a common trope that 'reliability' estimates should exceed 0.7 or even 0.8 for a test (Roberts et al, 2006). These ranges should be used as clues, not rules: they depend on context.

There are several different ways of approaching reliability.

In **Classical Test Theory**, it is assumed that any observed 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.

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 candidate.

Score (see Marks and Grades)

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

Sensitivity

It is often instructive to consider assessments as screening tests, as these are used in medicine itself. The following definitions can be made.

True Positive (TP): Candidates who pass and deserve to pass

False Positive (FP): Candidates who pass and deserve to fail

True Negative (TN): Candidates who fail and deserve to fail

False Negative (FN): Candidates who fail and deserve to pass

From these we can calculate:

Sensitivity = TP/(TP+FN)

Specificity = TN/(TN+FP)

Accuracy = (TP+TN)/Total

Candidates generally wish assessments to be highly Sensitive (minimising False Negatives). But if False Positives are more expensive than False Negatives, then society may want an assessment to be highly Specific (minimising False Positives).

In calculating sensitivity and specificity, it is necessary to have a 'Gold Standard' against which to assess True and False Positives and Negatives. See **Receiver Operating Characteristic Curves**

Specificity See **Sensitivity**

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". Although 'arbitrary' and 'capricious' are often treated as synonyms (in English Law, for instance) an alternative reading of arbitrary goes back to its Latin root, and can be taken to mean 'decided by

arbitration. In other words, there are no absolute standards: these are always socially constructed. See 'Arbitrary' in this regard.

Subject Matter Experts see **Expert, Expert Panel**

Summative see **Purposes – Formative and Summative**

True Positive See **Sensitivity**

True Negative See **Sensitivity**

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

A simple working definition of Validity is that it 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 as is sometimes assumed. I would phrase the relationship as "Validity cannot be *determined* for an item of low reliability".

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 is credible to a panel of experts. One can therefore usefully ask this of one item (or 'question' – the term 'item' is generally preferred in this Glossary for a single component of an assessment). Credibility relates to whether the item is unambiguous, plausible, and, in the context of medical education, relates to actual clinical practice.

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. In the context of medical education, predictive validity relates to a positive correlation between tests during training and late clinical performance, however defined.

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.

Modern thinking on Validity is rather complex (Newton and Baird, 2016), with important contributions made, for instance, by Messick (1995) and Kane (2013). Messick (1989) indicates that "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment". Kane would suggest a two stage approach, requiring an interpretation argument indicating the inferences and assumptions leading from candidate performance to candidate outcomes, and a validity argument exploring these assumptions in detail. Shaw and Crisp (2015) helpfully suggest five 'validation questions' that are pertinent to this process. These are:

Do the tasks elicit performances that reflect the intended constructs?

Are the scores/grades dependable measures of the intended constructs?

Do the tasks adequately sample the constructs that are set out as important within the syllabus?

Do the constructs sampled give an indication of broader competence within and beyond the subject?

Do scores/grades give an indication of success in further study or employment such that they can be used to make appropriate decisions?

It will be seen that the further research recommended in the Report, along with the Report itself, addresses many of these questions.

References

Anderson, L. W. and Krathwohl, D. R., et al (Eds.) (2001) *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* Allyn & Bacon. Boston, MA (Pearson Education Group)

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.
Bond TG, Fox CM (2007) Applying the Rasch Model. 2nd Ed. Lawrence Elbaum Associates, New Jersey, London.
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* I 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.
Kane, M.T. (2013) 'Validating the interpretations and uses of test scores'. Journal of Educational Measurement, 50 (1), 1–73.

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

Krathwohl DR. (2002) A revision of Bloom's taxonomy: an overview. *Theory into Practice*, **41**:212-218.

Lai, H., Gierl, M.J., Touchie, C., Pugh, D., Boulais, A.P. and De Champlain, A., 2016. Using automatic item generation to improve the quality of MCQ distractors. *Teaching and learning in medicine*, *28*(2), pp.166-173.

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

McLachlan, JC & Whiten, S.C. (2000) Marks, scores and grades: scaling and aggregating student assessment outcomes. *Medical Education*. **34**: 788-797.

Messick, S. (1989) 'Validity'. In Linn, R.L. (ed.) Educational Measurement. 3rd ed. New York: American Council on Education/Macmillan, 13–104.

Messick, S. (1995) 'Standards of validity and the validity of standards in performance assessment'. Educational Measurement: Issues and Practice, 14 (4), 5–8. M

Newton, P.E. and Baird, J.A., 2016. The great validity debate. Assessment in Education: policy, principles and practice 23: 173-177.

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.
Papageorgiou, S., 2010. Investigating the decision-making process of standard setting participants. *Language Testing*, *27*(2), pp.261-282.
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.

Roberts C et al (2006) Assuring the quality of high-stakes undergraduate assessment of clinical competence. *Medical Teacher* **28**: 535-543.

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.

Shaw, S. and Crisp, V. (2015) 'Reflections on a framework for validation: Five years on'. Research Matters, 19, 31–7.

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

Swanson, D.B. and Roberts, T.E., 2016. assessing achievement. *Medical Education*, *50*, pp.101-114.

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.

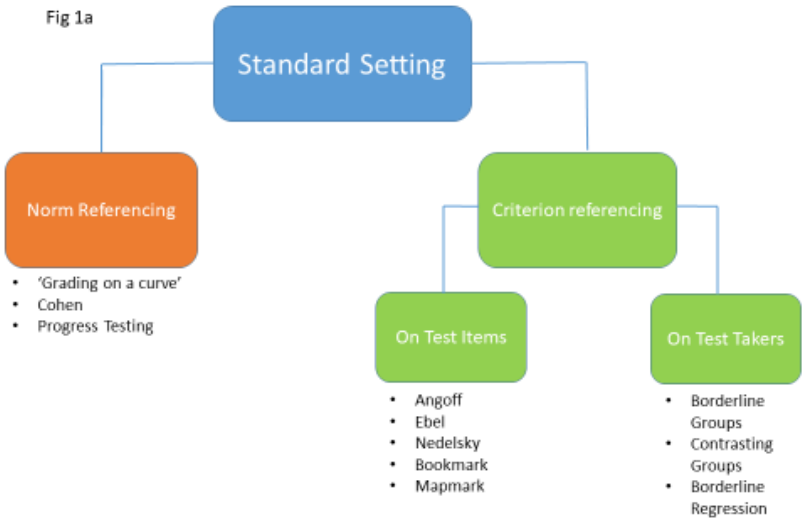
Zumbo BD. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly,* 4: 223–233.

Appendix 2: Standard Setting Methods, A Classification and Guide

1. Introduction

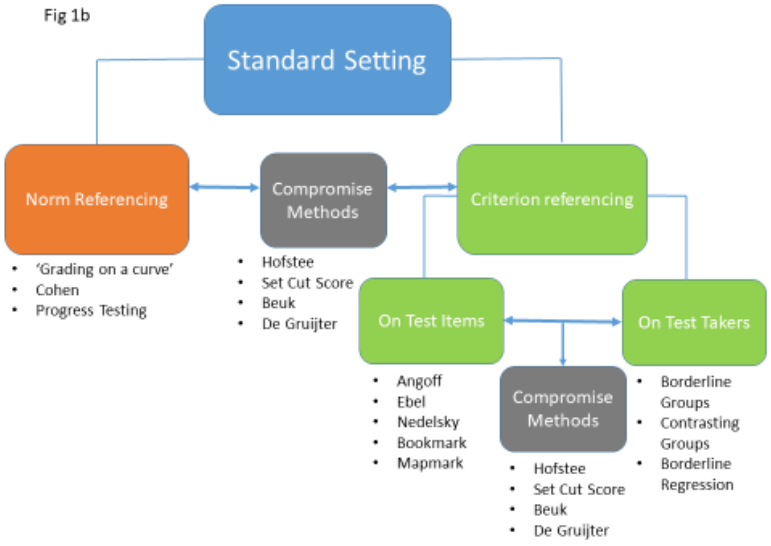
There are two super-categories of standard setting methods. These are norm-referenced and criterion-referenced approaches. Criterion-referenced methods also fall into two major categories: those based on test items, and those based on test takers. See Fig 1a, in which I have named some common examples of each.

Fig 1a



Then there are 'compromise' methods. I indicate that these can be compromises between test items and test takers, or compromises between norm- and criterion-based methods, depending on the context in which these are used. These, with named examples, are shown in Fig. 1b.

Fig 1b



Perhaps controversially, I place the 'fixed cut score' (e.g. 50%) approach familiar in UK Universities, in the compromise category, for reasons explained below.

Test equating is not a standard setting method in itself – it relies on a previous standard setting method of some kind to establish its base line, and is therefore not included in Figs 1a and 1b.

These categories, and individual methods, are described in more detail below.

2. Norm-referenced or 'normative' approaches

These terms refer to standard setting based on how an examinee performs against a reference population (generally those who took the test).

It is not widely used in high stakes medical testing, as it is seen as being dependent on the cohort. However, it is still entirely appropriate where a set number of places have to be filled. Major examples are selection for healthcare programmes such as medicine (e.g. the UK Clinical Aptitude Test), and choice-distribution tasks such as assigning UK medical school graduates to their first post (e.g. the UK Foundation Programme Office recruitment process).

Norm referencing is more *reliable* than criterion referencing, especially with high performing students, since it only requires assessors to *rank* candidates, which can be done from raw scores, without also making a subjective *rating* decision (what level of performance a score represents). Stringent assessors ('hawks') and lenient assessors ('doves') often agree on the relative ranking of candidates, but disagree on the absolute rating.

Norm referencing may be used serially in a multiple Progress Test setting, where evidence suggests that candidates who consistently perform in the lowest levels of a series of norm referenced tests, would also perform at fail levels in criterion referenced tests.

Other advantages include the ability to compare student performance across a number of different tests or different assessors.

An interesting question is whether a minimum number of candidates are required in employing norm referencing approaches, and there is no clear context-independent answer to this. It is known that when apparently equated test forms are used on different occasions, outcomes may also vary (e.g. different PLAB sittings).

2.1 'Grading on a curve'

This terminology is often employed in the US for norm-referencing purposes. Typically, raw scores are represented as a normal distribution, effectively with standard deviation boundaries determining the grades awarded. The grades may be represented by letters, and even further divided by modifiers.

2.2 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, although I would disagree with this – see below). The 'Cohen' method is described by the authors as a compromise method. However, 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). However, since it is based on the performance of just 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 method available in terms of examiner time, hence its attraction. However, I do not believe that it is sufficiently well evidenced in terms of validity, or indeed of sufficient credibility, to be recommended 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.

2.3 Norm referencing in Progress Testing

Progress testing is a form of longitudinal examination which samples at regular intervals from the complete domain of knowledge considered a requirement for medical students on completion of the undergraduate programme.

Because there is a large degree of random sampling from a large domain of knowledge, performance is highly case-specific, especially in the early stages. Criterion referencing is therefore very difficult. It has been argued instead that candidates who perform poorly on one test due to this chance effect, but are possessed of a sufficient sample of knowledge for their stage, are unlikely to repeatedly perform poorly. Conversely, a candidate who repeatedly performs poorly on sampling, is likely to possess an inadequate sample of knowledge. Thus repeated poor performance, as adjudged by norm referencing (say, 2 standard deviations below the mean, over 4 consecutive occasions) is likely to represent poor performance by criterion referenced standards.

3. Criterion Referenced methodologies

Criterion based standard setting is (in principle) based on some absolute standard of knowledge or performance. As described previously, of course, there are no absolute standards, and therefore this really represents an expert consensus on the standard reached. This is generally done in set ways.

The usual steps are:

- First, *define* a group of experts
  - (these require knowledge of subject, knowledge of context, knowledge of assessment, and knowledge of the appropriate student level)
- 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* or *test takers*
- In principle, judgements on test items are *prospective,* judgements on test takers are *retrospective*

Judgements referenced on *test items* include Angoff, Ebel, Nedelsky, Bookmark and Mapmark, approaches, and those referenced on *test takers,* include Borderline Groups, Contrasting Groups, Borderline Regression, and the "Up and Down" Method.

3.1 Judgements referenced on Test Items

3.1.1 Angoff Standard Setting Procedure

As described above, '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. An imagined example is shown in the Table below.

	Judge 1	Judge 2	Judge 3	Judge 4	Mean
Item 1	55%	47%	45%	52%	50%
Item 2	41%	36%	32%	38%	37%
Item 3	78%	72%	68%	69%	72%
Sum of item means					159%
Mean	58%	52%	48%	53%	Cut Score = 53%

If therefore there were 100 items with these difficulties, the cut score would be 53%. Note I have chosen values here such that Judge 1 is the most stringent in their expectations, while Judge 3 has lower expectations. Such 'hawks' and 'doves' are not uncommonly observed in real situations. In general, even the most 'doveish' judge should not go below 20% in a 'one best of five' MCQ, since this is the average value which would be obtained by guessing alone.

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.

The original standard setting described by Angoff is often referred to as 'Yes/No Angoff' and is particularly suitable for constructed response items, while the version known as 'Modified Version' is actually a footnote in the original article, and is most suitable for selected response items. 'Yes/No' Angoff asks what score a minimally competent candidate would be expected to achieve on a continuously graded scale.

3.1.2 Ebel Standard Setting Procedure

Angoff methods tacitly 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) then this may be important to take into account in standard setting.

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.

A grid is therefore generated of the following form.

	Easy	Medium	Hard
Essential	3	3	2
Important	3	2	4
Acceptable	2	2	3
Questionable			

And, after 'questionable' items have been deleted, the proportion of minimally competent candidates who are expected to be able to answer each category is determined by the expert reference group.

	Easy	Medium	Hard
Essential	95%	85%	80%
Important	75%	70%	65%
Acceptable	60%	55%	50%

And the product of each box is summed to give the final cut score.

	Proportion right (%)	No. Questions	Product	Mean
Essential				
Easy	95	3	285	
Medium	85	3	255	
Hard	80	2	160	
Important				
Easy	75	3	225	
Medium	70	2	140	
Hard	65	4	260	
Acceptable				
Easy	60	2	120	
Medium	55	2	110	
Hard	50	3	150	
Sum		24	1705	1705/24 = 71

3.1.3 Nedelsky Standard Setting Procedure

In this process, the expert, informed, experienced judges determine how many distractors the minimally competent student can eliminate, and the Item Score is the reciprocal of the remainder.

Item number	Options in the Item	Eliminate	Reciprocal
1	5	3	0.5
2	5	1	0.25
3	5	4	1
4	5	2	0.33
5	5	3	0.5
	Total		2.58

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 when there are poor quality distractors). It has been criticised on the basis that it forces particular weightings on questions (for instance, an item 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.

3.1.4 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 items are ranked in order of their difficulty. This can be done in advance by

inspection of the items, although a common approach is to use Item Response Theory based on previous administrations of the items to establish item difficulty. At first glance, this might suggest that Bookmark is a retrospective method based on test takers. However, Item Response Theory claims to identify absolute difficulty of items across all possible tests, so I have placed in the category of prospective tests based on test items. If Facility were used after a test to rank the items, then this would indeed be a retrospective method based on test takers. The usual caveats about Item Response Theory, that it assumes a single latent trait and continuous distribution of ability and performance, apply.

The expert panel then identifies 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 of one exercise. A further described advantage is that Selected and Constructed Response questions can be integrated into the Ordered Item Booklet 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.

3.1.5 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.

3.1.6 '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).

3.2 Judgements referenced on Test takers

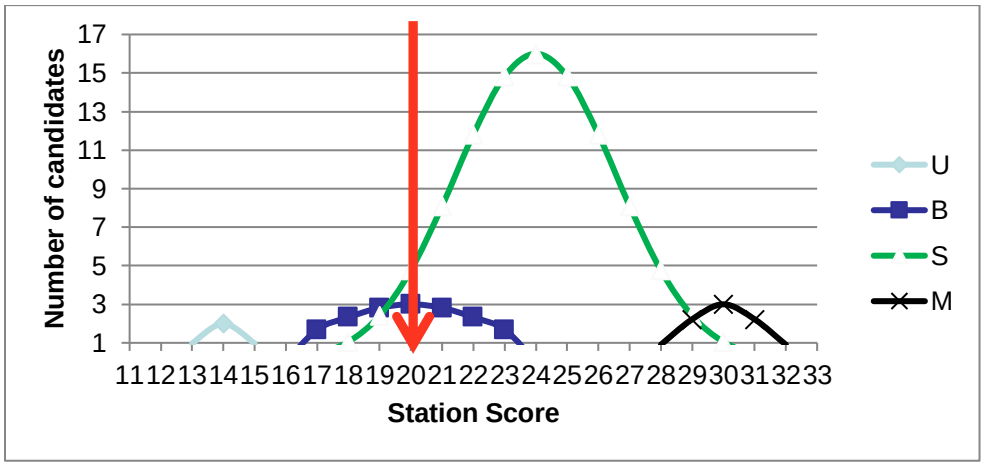
While judgements on test items are often applied to MCQs, judgements on test takers are often applied to observational tests such as OSCEs. The reason for this is that the judges in an OSCE actually observe the candidates performing complex tasks, and their judgements made at the time are informed by greater richness than would be the case with an MCQ. There is evidence that 'expert judgments' made on a gestalt basis are actually better than check list scores, and the methods described below combine expert judgement and check list scores in an attempt to optimise the value of both.
Three methods, Borderline Groups, Contrasting Groups, and Borderline Regression, share a common approach to scoring which is described below in the context of an OSCE.

- 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 are then plotted against judgements

Evidently, the outcomes are strongly dependent on the specific terms used for the Global Judgements. Here 'Borderline' is chosen as the example, since it avoids the assessor having to make a pass/fail judgement when they are in any doubt about it being a clear fail.

3.2.1 Borderline Group(s)

- Scores in each 'Global Judgement' category are plotted against candidate 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.



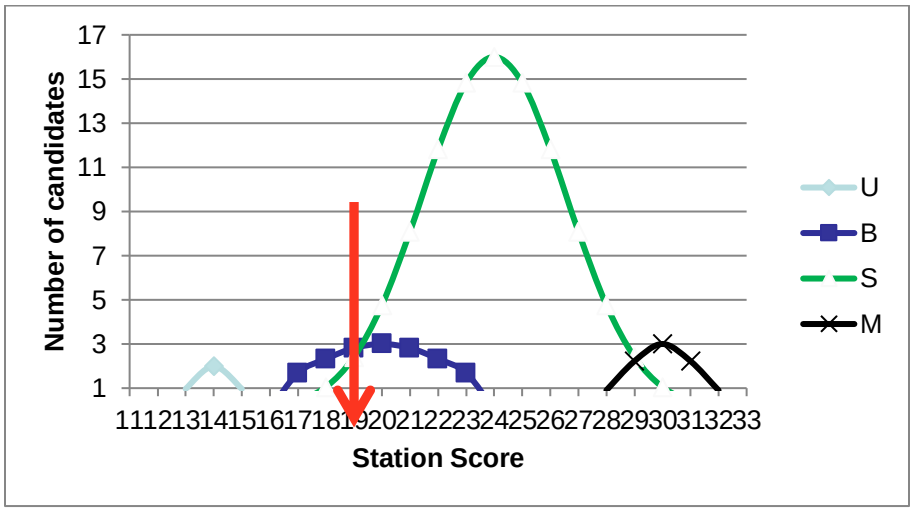
The values on the abscissa correspond to grades (e.g. Unsatisfactory, Borderline, Satisfactory, Merit), from left to right respectively

3.2.2 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. A 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.

- Scores in each 'Global Judgement' category are plotted against frequency, with scores as the abscissa, and frequency as the ordinate as with Borderline Groups
- 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)

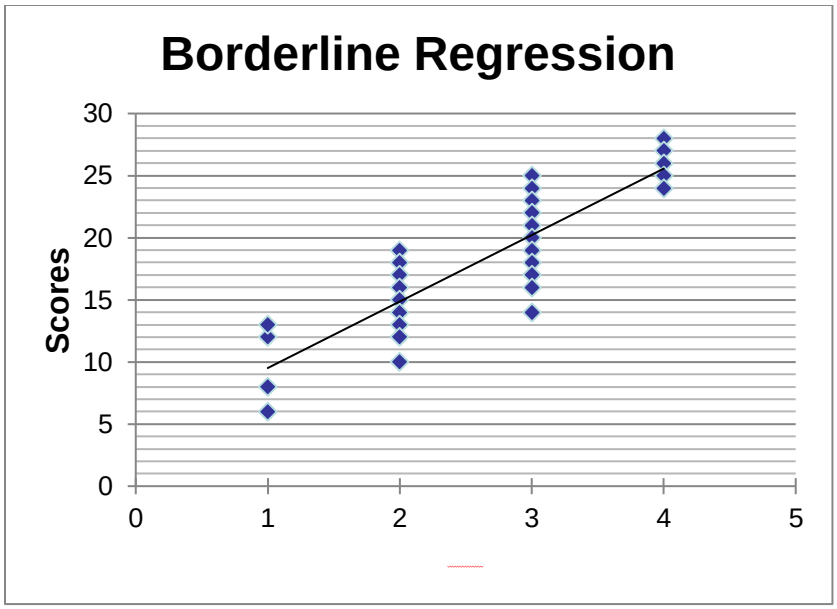


3.2.3 Borderline Regression

Following the same two initial steps as the Borderline and Contrasting Groups methods,

- 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.
- 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.

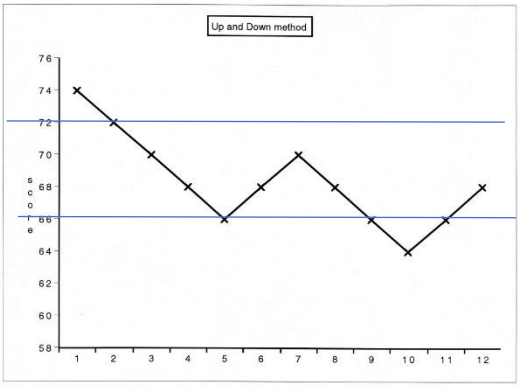
Here, the values on the abscissa correspond to grades (e.g. Unsatisfactory, Borderline, Satisfactory, Merit), from left to right respectively



3.2.4 'Up and Down' Standard Setting Method

This method is suitable for constructed response items such as essays or Short Answer Questions, which have already been scored according to a checklist, then placed in rank order. A sample of candidates near the expected cut score is selected, and assessors agree whether each candidate 'passes' or 'fails' through their expert judgement. If it is graded as a 'pass', the assessors move down the score ranking to the 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.

	Score	Grade
Candidate 1	74	Pass
Candidate 2	72	Pass
Candidate 3	70	Pass
Candidate 4	68	Pass
Candidate 5	66	Fail
Candidate 6	68	Fail
Candidate 7	70	Pass
Candidate 8	68	Pass
Candidate 9	66	Pass
Candidate 10	64	Fail
Candidate 11	66	Fail
Candidate 12	68	?



4. Compromise Standard Setting Methods

It is possible to consider both the assessment and the candidates at the same time, and these are generally referred to as compromise methods.

The instruction to assessors might be “Given what you perceive of this test paper (e.g. is it 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. are they a good group or a poor group compared to normal expectations) what do you think the pass rate should be?”

In such cases where assessors have the opportunity to observe a particular group of candidates, or are given information on the performance of the candidates, on a particular occasion, then this represents a compromise between test items and test takers, under the general heading of criterion referenced methodologies.

If, however, the candidate information is generic to all occasions (‘on average, we expect between 5 and 10% of candidates to fail’) rather than specific to a particular occasion, then this would represent a compromise between norm referenced and criterion referenced methods.

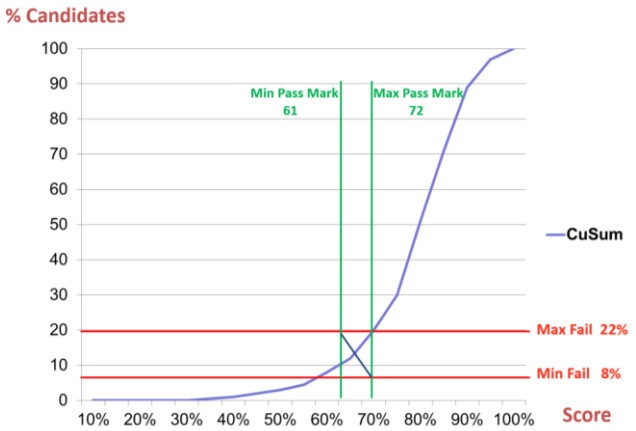
It should be noted that the lines can become blurred between approaches. For instance, if, during an Angoff approach, information about the consequences of the decisions is introduced, and used to influence the decisions of the assessors on an iterative basis, then Angoff is being used as a compromise method, rather than as one based on test items alone, as it is usually described.

4.1 Hofstee Compromise Standard Setting Method

In the Hofstee approach, assessors 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 across the assessors, and set upper and lower bounds for the cut score and fail rate. Then the cumulative sum of candidate scores is plotted, with the scores as the abscissa and number of candidates as the ordinate. The Hofstee boundaries of these are drawn on the graph, and in the rectangle thus generated, the cross diagonal from top left to bottom is drawn. Where it intercepts the plot of the cumulative number of candidates is the cut score and final pass rate.



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 Hofstee decisions to the pass/fail boundary – it is trickier at upper 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. In my experience, it takes staff some three full iterations (i.e. three years, in many settings), to become sufficiently familiar with Angoff approaches to be able to use them effectively.

4.2 Fixed Cut Scores and the 'Reverse Angoff' approach

At first glance, the traditional UK University approach of using fixed cut scores (typically 50%, as the pass mark, and 70% as the merit score) as the cut score for passing, does not seem like a compromise approach. Rather it appears as an arbitrary decision, perhaps predicated on the idea that the 'passing' candidate should be possessed of about half of the required domain on knowledge. Yet in general, over many decades, there has been a convincing impression that this apparently arbitrary process has indeed recognised merits and demerits.

However, 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 the candidates under consideration (for example, such that a minimally competent candidate would score 50% and a fail candidate would score below this level). In other words, instead of adjusting the estimate of the proportion of minimally competent candidates who would succeed in a fixed item, the item is adjusted so that 50% of minimally competent candidates would get it correct.

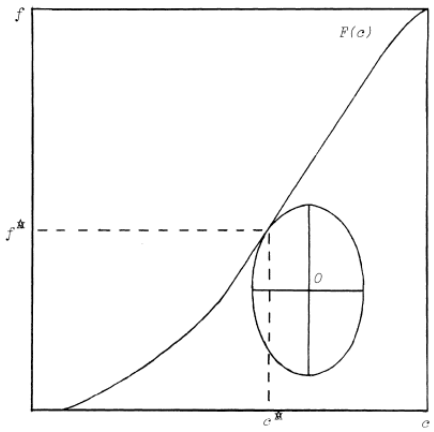
This matches the common experience of anyone who has sat through traditional 'exam review meetings', where questions are described as 'too easy' or 'too challenging', and modified accordingly.

I have classed this as a compromise method, since it is most frequently applied retrospectively in the context of constructed response items such as Short Answer Questions or essays, where the assessor first determines a 'grade' in their own mind, and then awards the corresponding 'mark' (e.g. 50%).

4.3 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 from De Gruijter, 1985



4.4 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.

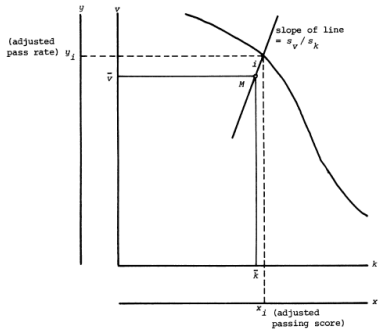
2. The relative emphasis given to the two types of judgements should be in proportion to the extent to which the members of the committee agree with each other.

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 from Beuk, 1984



References

Angoff, W.H. (1971). Scales, norms and equivalent scores. Educational Measurements. Washington, DC: American Council on Education

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* I 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 (1996) The assessment of professional competence: developments, research and practical implications. *Advances in Health Science Education* 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.

Attachment 1: The Detailed Brief.

1. Standard Setting Process

1.1 What improvements might be made to the PSA's standard setting processes?

See Recommendations 1 and 2

1.2 Is Modified Angoff the most appropriate method of standard setting?

Currently this is probably the most appropriate methodology

1.3 What alternative approaches to Angoff might be considered?

Ebel methodology and IRT dependent methods such as Bookmark: Review Section 4

1.4 What processes should be used to define the description of the 'just passing' candidate?

A robust process is in place: Review Section 4

1.5 **Is the process for test equating appropriate, (e.g. should test equating be performed as part** of the post-assessment review process)?

Yes. It is important to establish that a common standard exists across all delivery dates. Hence test equating is a requirement.

1.6 **Does the PSA standard setting process meet best international practice in medical** assessments of this type?

Essentially, yes, although a number of detailed Recommendations have been made as to improvements. It is important to note that other methods would set different standards and have different outcomes: there is no absolute 'Gold Standard' in standard setting.

1.7 What benefits might arise from aligning all assessment events on a single day?

If this relates to the practice of setting all four testlets on one day, I know of no literature which relates to this. However, it does offer a way of comparing the testlets across a single cohort, and therefore probably improves the process of test equating.

There are different approaches to single versus multiple testing days or occasions. It is possible to have each candidate sit the assessment at a different time, as is done for UCAT. This requires that each item in a large bank has been standard set in advance, and that candidates can draw at random from the bank in each testlet. Multiple secure proctored testing sites must be provided for candidates, who then book themselves in for a date of their own choosing. However, the resource requirements for this are unlikely ever to be available for the PSA.

Currently, the PSA offers four test dates per year. Offering fewer test dates reduces the number of test forms required, which in turn reduces the need for item generation. However, it is always necessary to

offer a second test date, for those candidates who for good reason are unable to attend the initial test date, as is done for the UK Foundation Programme Office Situational Judgement Test.

Reducing the number of test days however, is something which has a potential impact on stakeholders, and on resource provision, and would need to be extensively canvassed with students, medical schools and others before implementation.

2. Reliability, assessment length and pass mark

2.1 What improvements might be made to the PSA to ensure that the reliability and external validity of the assessment are at a level that is acceptable for a high stakes examination?

See Recommendation 5 with regard to reliability and Recommendations 3 and with regard to Validity.

2.2 What compromises might have to be made in an effort to increase the internal consistency of the PSA?

'Consistency' is not the issue: it is the latent trait under study that is of key importance. Otherwise, this is just a way of describing Reliability. There may be more than one latent trait under study, especially given the diverse nature of the items. Reliability is therefore difficult to measure by Classical Test Theory methods.

2.3 Is there room for compromise on the reliability statistic?

The measures of reliability may not be optimal, and other methods may be required.

2.4 How might the apparent tension between selecting items that discriminate well (in order to achieve a high reliability coefficient, Cronbach's alpha), and the impact that this approach has on lowering the pass mark, in what is intended to be a safety assessment, be addressed?

A test information curve would be helpful in this regard. My view is that it is valuable to include items which are important, even if they are not perceived as difficult. Students should be tested on the most clinically significant aspects of prescribing, even if they have been extensively taught on this area. To do otherwise risks turning the test into one of the less important aspects of prescribing.

2.5 Might a strategy of repeated testing overcome the limitations of having a shorter assessment with a lower Alpha value?

Repeat (short term second look) testing has benefits (increased reliability) and costs (multiple test forms, requiring equating). In my view, the costs outweigh the benefits for the PSA.

3. Item and assessment development

3.1 Is the item bank of an appropriate size to service the future needs of the assessment?

In my view, yes.

3.2 Are the content authors trained and supported?

Yes

3.3 Is the item bank secure with access rights appropriately controlled?

Discussion with the assessment team indicates that security conforms to other similar national tests.

3.4 Are the quality assurance and peer review methods sufficient to maintain quality in a rapidly developing area of practice?

Yes

3.5 Could the input of important national bodies (e.g. NICE, BNF) improve the quality of the assessment?

I have suggested engagement with GPhC. I am not clear how NICE and BNF would engage in practical terms.

4. Administration and delivery

4.1 What improvements might be made to the administration and delivery of the PSA? Do the candidates have long enough to prepare?

Yes, as revealed in discussion with students and Prescribing Safety Assessment leads.

4.2 Are the candidates given sufficient information regarding the PSA in advance of sitting?

Yes. Discussion with students and Prescribing Safety Assessment leads

4.3 Is the user journey through the interface appropriate?

Yes. Discussion with students and Prescribing Safety Assessment leads

4.4 **Do local PSA teams have sufficient information and support to run the assessment? Should there be more practice content?**

Yes. Discussion with Prescribing Safety Assessment leads. More practice content – always!

4.5 Is access to the BNF sufficient?

Yes, though familiarity with the use of the *electronic* BNF seems to be key, from the student and teacher perspectives, so it can be used in a timely manner.

4.6 Should other sources of information be permitted?

Discussion with students and Prescribing Safety Assessment leads, but it depends on 'what happens in practice'; I would say provisionally no, since this can slow down weaker candidates, who need to look more up, when they are under stress.

4.7 Are disadvantaged groups appropriately supported?

Depends on (a) appropriateness of Special Circumstances policy (some cultural biases are inbuilt into all higher assessment processes in the UK (See DIF in the Glossary)). There were some comments to the effect that the test was intended for a UK setting, but this could be viewed as an appropriate intention. The common structure for reasonable adjustments is helpful. There was no evidence that candidates

with adjustments are penalized. It would be valuable to continue to promote diversity among item writers and standard setters.

5. Governance and management

5.1 What improvements might be made to the PSA's governance and management policies and processes?

No specific recommendations

5.2 Are the data handling processes defensible?

Relying on 'in house' written bank poses some risks, as well as benefits. Developers may move away, and individualistic solutions to problems may not be transparent to replacements. But generally, yes.

5.3 Is there an appropriate and transparent appeals process in place?

Yes, but note that candidates properly cannot appeal against academic decisions, only process steps.

5.4 Are all stakeholders able to make opinions known?

Yes.

5.5 Are the financial and accounting processes transparent?

Yes

6. Given the above, is the PSA a valid assessment of prescribing competency?

In the end, only Predictive Validity can answer this question, so further research is required.

6.1 **To what extent has the PSA fulfilled its original objectives as a reliable pass/fail assessment of minimal competence to prescribe safely at the boundary between medical school and work in the NHS as a Foundation Year 1 doctor?**

This is not the stated claim on the PSA website, and it would represent a claim too far for any single test.