Durham
University
School of Medicine and Health

Developing an evidence base for the Professional and Linguistic Assessments Board (PLAB) Test

(Literature Review)

Professor John McLachlan

Professor Jan Illing

Ms Charlotte Rothwell

Dr Jane K Margetts

The Centre for Medical Education Research, Durham University

and

Dr Julian Archer

Dr Duncan Shrewsbury

Peninsula College of Medicine and Dentistry

Grateful Acknowledgments to Dave Swanson, John Norcini, Cees van der Vleuten, Lambert Schuwirth, Kathy Boursicot, Chris McManus and Fiona Patterson

Gill Morrow, Ali Malik, Paul Crampton, Dee Corbett, Tracy Straker, Heather Robb, Emma Crawford, Suzanne Chamberlain, John Foulkes and Adrian Freeman.

Contents

8. Appendices

- **Appendix A**

Table A.1. UK and International Professional Examination and Assessment Systems

- **Appendix B**

Sample letter to professional bodies

- **Appendix C**

Glossary of key terms

- **Appendix D**

Matrix of papers on standard setting

- **Appendix E**

Outcome of modelling of borderline regression

Executive Summary

Aims

- A literature review for the General Medical Council of best practice in examination and assessment methodology within the context of professional entrance examinations
- An independent review of the PLAB test with focus on robustness and fairness for assessment of international medical graduates (IMGs) applying for registration with the GMC

Objectives

- **Theme 1**
  The number of attempts candidates are allowed at examinations
- **Theme 2**
  The periods of validity of passes for candidates
- **Theme 3**
  Best practice in examination and assessment methodology

Methods

A variety of rapid review methods were employed, through formal literature searches, grey literature searches, snowballing of references from key papers and hand searches of the research team's personal data records. Searches were limited to English language papers but not limited by methodology. A number of international experts were contacted personally, as were both national and international professional bodies, including a number of non-medical bodies.

Findings

Themes 1 and 2

1. **The evidence base relied upon by other professional organisations is in general weak or absent**

   Fifty six relevant professional examinations and assessments were reviewed to identify what evidence, if any, was referred to in order to inform decisions about the number of re-sits allowed and periods of exam validity. This review identified only three organisations (GP NRO, RCGP, and UK Foundation Programme: see text for details) that clearly identified the evidence that informed their decisions. Other organisations referred to external guidance or their own internal consultation exercise. The evidence basis for the number of re-sits allowed was at best weak and generally absent.

2. **The most frequently permitted number of attempts in medical organisations is four in the UK and three internationally**
   The majority of UK medical organisations do not limit the number of re-sits, but when they do, four exam attempts (first exam plus three re-sits) are the most frequently permitted. This contrasts with non-medical and international medical organisations who tend to set a limit of three exam attempts with a *maximum* of four attempts.

3. **Moderately strong evidence indicates there is no significant benefit after four exam attempts**
   The evidence from the literature on re-sitting exams highlights that there are benefits to re-sitting. Re-sits particularly benefit candidates who only just fail. There is moderately strong evidence to indicate that there is no further benefit after four exam attempts (first exam plus three re-sits). The GMC's own data on PLAB 1 is consistent with this effect.

4. **Evidence on periods of exam validity highlights no clear pattern**
   Most often periods of validity are tied into individual exam structures and designs, such as requirements to pass one part only of an assessment within a particular period of time, making it very difficult to generalise to 'good practice'. There is no substantive evidence base to recommend a defined period of validity, even though this may be desirable.

5. **Moderate to strong evidence associates poor examination performance with later poor performance in practice**
   There is almost no evidence relating the number of re-sits directly to later performance in practice. But there is moderate to strong evidence to support a correlation between examination performance and later performance in practice.

6. **In relation to protected characteristics, evidence suggests (1) a decline in current knowledge with age, (2) females outperform males, (3) white candidates outperform ethnic minority UK graduates**
   However, there are inconsistencies in the data, which depends on context. Factors such as ethnicity and gender may also be interactive with each other.

7. **Evidence is lacking in relation to exam performance and other protected characteristics**
   There is a lack of research on the impact of protected characteristics other than age, gender and ethnicity.

8. **As re-sits increase so complexities in relation to question fatigue becomes more relevant**
   The evidence on size of a question bank and the robustness of the examination to avoid question fatigue is complex. Question fatigue will become more relevant as the number of re-sits increase, allowing candidates more chance of having seen the Items previously.

Theme 3

9. **Current Methods are defensible**

The methods currently used by the GMC to set cut scores for Part 1 and Part 2 have acceptable reliability and validity, and are defensible. However, we recommend that the GMC consider a number of alternative approaches, and collect further data on the existing methods.

Recommendations

1. Housekeeping recommendations

A. Pending strategic review, the GMC should retain Angoff and Borderline Groups standard setting methods, since both are well recognised and supported by evidence, and the addition of 1 Standard Error of Measurement (SEM) to the cut score for Part 2, as this helps reduce false positives in the crucial Skills component (See 5.2.5).

B. We recommend the GMC conduct analyses of existing PLAB results in the following areas:

1) Conduct an Item Response Theory analysis of PLAB results to allow (a) Test Equating or (b) a Computer Adaptive Testing approach if desired or (c) test optimisation near the cut score (see 5.2.5).

2) Conduct a Generalisability Theory Analysis of PLAB Part 2 on a routine basis, and subsequently a Decision Study, to calculate the number of OSCE stations appropriate for PLAB Part 2 (see 5.3.2).

3) Calculate Differential Item Functioning to explore the ways in which individual test items perform poorly (see 4.5).

C. We recommend the GMC collate and analyse further data relating to PLAB candidates in the following areas:

1) Data on the number of previous attempts at PLAB, and the time interval between candidates' attempts, in order to identify attrition or accumulation of knowledge in candidates, and to determine the appropriate period of validity of the tests (see 4.6.2).

2) Demographic data (including a voluntary section for candidates on all protected characteristics) to enable analysis of the influence these have on test outcomes (see 4.5).

3) Data on current level of grade or post when taking PLAB to enable better assessment of whether candidates are in fact well-matched to the Foundation Year 1 level of the exam (see 4.6.2).

D. We recommend the GMC statistically analyse the relationship between score performance in PLAB Part 1 and 2 and IELTS score performance (see 6.1).

E. We recommend the GMC publish on its website the data gathered in the recommendations above to demonstrate a culture of transparency to patients and to PLAB examinees, and to improve examinees' understanding of, and expectations of PLAB (see 5.8).

F. We recommend giving a more detailed breakdown of performance to PLAB OSCE examinees to enable them to improve (see 5.8).

2. Strategic recommendations

A. The GMC should limit the number of attempts at PLAB Part 1 and 2 to four, followed by a period adequate to allow further personal development should elapse before further attempts are permitted (see 4.2.8). The length of this period can be determined from the data collected in House Keeping Recommendation 3 (1) above.

B. The GMC should consider the following strategic approaches to examinations. These could be piloted alongside its existing approach and/or developed in the future.

1) Exploring the use of Situational Judgements Tests as part of the PLAB process (see 5.6.1)

2) The GMC should keep under review the possibility of the use of Computer Assisted Testing, Computer Adaptive Testing and Test Equating for PLAB Part 1, particularly in the light of the Australian Medical Council's approach (see 5.6.2).

C. In addition, the GMC should consider the following strategic approach to assessment: 'Interim' Registration followed by Workplace Based Assessment (which could include patient feedback) for a defined period as part of PLAB assessment. There could be a process of linking this with PLAB Part 1 and 2 results in a portfolio for overall assessment before grant of Full Registration (see 6.1).

3. Good practice from other professional bodies

The following are exemplar examination approaches for the GMC to consider, as described further and hyperlinked in Table A1, Appendix A.

A. We recommend considering a model for targetting bespoke further training as in the College of Emergency Medicine (CEM) Membership Examination (MCE) (see 6.1).

B. We recommend consideration of an approach whereby marginal 'passing' performance in PLAB leads to a 'second look' re-test (as opposed to a re-sit), following the Australian Medical Council Ltd Clinical Examination model (see 4.2.8).

C. We recommend giving consideration to an 'exceptional circumstances clause' to give flexibility for unusual candidate circumstances with regard to re-sitting rules. The model is the RCGP Applied Knowledge Test (AKT) and CSA exam, allowing one further attempt in exceptional circumstances (see 4.3.5).

4. Recommendations for key areas for further research

A. The GMC should conduct research into the relationship between re-sitter performance and subsequent performance in clinical practice (see 4.3.2).

B. We recommend the GMC examines through further research why some international medical graduates perform poorly, as seen in the National Clinical Assessment Service (NCAS) and GMC data, despite having passed PLAB (see 6.1).

C. We recommend that the GMC conduct qualitative research to improve its understanding of why some candidates re-sit PLAB so many times, how they interpret their failures, and what their subsequent strategies are (see 4.6.2).

Limitations

The information to answer some of the theme questions directly was often not available, so inferences were drawn, on the best available evidence, on a number of the questions posed, such as the relationship between re-sits, exam performance, and later clinical practice.

The time and resources available for this large and complex project were limited.

Web and grey literature searching is by its nature less systematic than formal literature searching and we may have missed pertinent information.

Since not all professional bodies replied by the deadline, some information remains ‘unverified’ in this particular sense.

1 Introduction

The GMC must ensure that overseas qualified doctors have adequate cognitive medical knowledge and practical skills before being registered to practise as a doctor in the UK, as well as capability in English. In order to ensure stakeholders have confidence in the registration process, and in the interests of patient safety, high quality assessment processes must be in place for these high stakes decisions. Following a number of literature reviews of the topics required, and extensive searching of the grey literature, this project describes, analyses, and critiques existing approaches nationally (including within other professions) and internationally, summarises the evidence available for various practices, reviews the literature on re-sits, standard setting and assessment more generally, and synthesises all this information into recommendations for consideration.

The project has been delivered by a team with considerable experience of the research methods proposed, in the field of medical education. International authorities have been consulted and have provided input at various stages in the process.

Detailed review questions, sources, inclusion and exclusion criteria, and the analysis structure are described below. A key focus was on defensibility: ensuring that the options recommended have a clear strategy by which they can be defended against internal and external challenge.

Since this Report does not represent a meta-analysis, a qualitative description of effect sizes is employed, using the conventional Cohen (1988)[i] delimiters: effect sizes below 0.3 are described as 'low' or 'small', around 0.5 are described as 'medium' or 'moderate', and around 0.8 are described as 'high', 'good' or 'strong'. Quality of studies is a different issue, particularly since Double Blinded Randomised Controlled Trials are rare in medical education and training, and many studies rely in whole or in part on qualitative data. We have categorised quality as 'High' (supported by sophisticated studies or analyses of data for quantitative data, and methodologically sound, triangulated multi-site studies for qualitative data); 'Moderate' (supported by sound if less sophisticated qualitative studies, or single site studies for qualitative data); and 'Low' (expressions of opinion or anecdotal accounts unsupported by data). In general, if we have cited a reference in the text, we have regarded it as 'Moderate' or 'High' in quality. Some references in the appendices may be included even if their quality is Low, for completeness sake.

If PLAB is fit for purpose, it could be asked why international medical graduates (IMGs) who have passed PLAB are over-represented in GMC disciplinary procedures, in NCAS data, and in failing GP National Recruitment Office (NRO) selection. Of course, the fault does not necessarily lie with PLAB or the IMGs: it could be that IMGs do not receive the induction and support they are justly entitled to expect. But the underlying causes of these phenomena merit exploration. And of course, some IMGs enter practice in the UK through routes other than PLAB – by having gained recognised higher level qualifications, for instance.

2 Ethical Status

Ethical approval was not required for this project. Various documents were shared with us in confidence, and we have abstracted information in general terms from these where appropriate, but have not provided information which could lead to the supplying organisations or individuals being identifiable. In several cases, we were kindly granted permission to be more specific. A variety of copyright Figures and Tables have been reproduced from other publications as academic fair usage, since this is an internal report. However, if this report is to be published more widely, written permission from the copyright owners will be required.

3 Literature Methodologies

Literature Review Strategies

This has been an unusually complex task, with a variety of overlapping themes. We have explored both interdependent and conjoint searches around each theme and bullet points. It was decided that a combined search strategy for Themes 1 and 2 (excluding protected characteristics where a separate search was deemed necessary to capture specific literature, see table 2). Advice was taken from a senior medical librarian at Durham University on combining the search terms used. These are described below.

A separate search strategy was undertaken for Theme 3. There was a considerable amount of web searching, hand searching from existing documents, and follow up of personal communication, associated with this project.

3.1 Database search for Theme 1 and 2

A literature search was conducted across the most relevant databases (Medline, Embase, ERIC and Web of Science). An initial search limiting results by inclusion/exclusion dates (Jan 2008 to March 2012) yielded only a small number of papers. Therefore searches have not been limited by specific dates. Searches did not limit results by methodology, but were limited to English language.

The following search terms have been used:

- Exam*
- OSCE*or Objective Structured Clinical Exams*
- Performance
- Professional competence
- Educational measurement
- Pass rate
- Attempt$2
- Retake $1
- Re-take$1

- Resit$1
- Re-sit$1
- Accreditation
- Licensure
- Certification
- ‘permitted attempt’
- Prof*

Table 1 shows the search strategy used for Theme 1 and 2 with duplications not removed.

Table 1: Theme 1 and 2 literature searches (excluding protected characteristics)

Search terms	**Database**	**Found**	**Relevant by title**
Medical education or continuing education or continuing or medical graduate & prof*or accreditation or licensure or certification or exam* & ‘permitted attempt’ or retake($1) or re-take ($1) or resit ($1), re-sit($1) or attempt($1) or repeat ($2)	Embase Ovid	244	12
	Medline Ovid (without medical education terms)	100	13
	ERIC	120	10
Medical Education or education or medical graduate or internship or residency AND prof* OR accreditation OR licensure OR certification OR exam AND “permitted attempts” OR retake OR resit OR repeat OR “multiple attempts”	Web of Science	506	2
exp education, medical, continuing/ or exp education, medical, graduate/ or "internship and residency"/IMGs & retake($1), re-take ($1), resit ($1), re-sit($1),	Medline Ovid	315	9

attempt($1), repeat ($2) & exam* or objective structured clinical exam* or osce$1)			
Snowballed references			3
Hand searched references			9

Papers were chosen initially on their title. If these papers were deemed relevant then they were judged on the relevance of topic when the abstract was read. The following inclusion/exclusion criteria were used to determine relevance of the articles.

INCLUSION CRITERIA FOR PAPERS

1. Is it about examination or assessment?
2. Is it in a clinical context? or (in high stakes environments)
3. Is the study about either:

   i) number of re-sits allowed? OR

   ii) evidence on number of re-sits? OR

   iii) evidence on relationship between number of re-sits and later professional practice (positive or negative)? OR

   iv) do people with 'protected characteristics' show particular trends with regard to number of re-sits and later practice?

4. Is the study about the length of time the exam pass is valid? AND

   Is the study about attrition of professional knowledge? OR

   The testing method? OR

   Consideration for people with 'protected characteristics'

EXCLUSION CRITERIA

1. Papers that are not on examination or assessments
2. Paper that are not in a clinical or high stakes context
3. Papers that are opinion based rather than evidence based
4. Papers that describe an examination/assessment without any evaluation.

In total 1285 papers were found using a combination of the above search terms in the main databases. Papers were then sifted out by relevance of title which left 46 papers. Duplications and any papers which were not relevant by the abstract were then removed which left 16 papers to read in more depth. Three additional papers were identified via snowballing and nine through hand searches.

Protected characteristics

Table 2 outlines the search strategy (with duplicates removed) used for the protected characteristics literature search. The papers were limited to those of English language and full-text availability.

Table 2: Protected Characteristics

Search terms for protected characteristics other than ethnicity	Databases: (via NHS Athens): AMED, BNI, EMBASE, HMIC, MEDLINE, PsycINFO, CINAHL and Health Business Elite Google Scholar was also searched.	
	Found	**Relevant**
(performance + assess*) AND (international medical graduates)	67	9
(performance + assess*) AND (medical student*)	4263	24
(performance + assess*) AND ((medical + trainee) or (junior doctor*))	385	3

In total a number of 36 papers were found for review using the above search strategies.

3.2 Database Search for Theme 3

Theme 3 literature searches were supported through considerable expertise in this area, through both original research, and through delivery of high level teaching, training and CPD courses on assessment theory and practice. Theme 3 was explored first through consultation with the Principal Investigator's extensive network of experts in the field of assessment and via contacts in regulatory bodies both nationally and internationally.

A systematic search of literature was also carried out in addition to the hand searches of the PI's pre-existing papers. This focused particularly on recent literature as the areas in theme 3 are where good practice is developmental and incremental and the most recent studies generally update and expand on previous historical practice. The search was conducted across the most relevant databases (Medline, Embase and Web of Science). The Searches in the three databases were limited to between 1990 and 2012. Searches were not limited by methodology but were limited to English language. Table 3 shows the search strategy (duplicates not removed).

Table 3: Theme 3 literature search

Search terms	**Database**	**Found**	**Relevant by title**
Standard setting & clinical	Embase	291	69
Standard setting & clinical	Web of Science	341	29
Standard setting and medical, education	Medline Ovid	81	67
'Standard setting' and medical or education	Google Scholar	63 (most about undergraduate, higher education or schools	4 (duplicates)

713 articles were found from searching the databases. The number of relevant articles by title from the three main databases was 165. The number of duplications in databases was 50. The number of articles relevant (with duplications removed) by title from the databases searched was 115. Abstracts from these papers were then read and the inclusion/exclusion criteria applied for relevance of articles. In addition to the systematic literature search the PI's personal literature on the topic of standard setting and assessment (consisting of 91 papers) was utilised and a hand search took place using the same inclusion/exclusion criteria. Eighteen of these papers were duplicates of the articles from the database searches. This yielded a total of 66 relevant articles for review.

The following inclusion/exclusion criteria were used to determine relevance of the articles.

Inclusion Criteria	Exclusion Criteria
Relevant to identified topic	Irrelevant to the identified topic
Relevant to any aspect of topic	No relevant aspect of topic addressed e.g. relies solely on U.S. Law
Adequate methodology	Inadequate methodology identifiable from the abstract
Informed opinion which adds to the picture of the topic.	Language other than English

3.3 Grey literature search

The grey literature searching methodology was as follows:-

- Commenced with a search of Medical Education England and Academy of Medical Royal Colleges to identify websites for UK medical professional bodies, and hand searched these.
- This was supported by one team member's personal experience in medical training, and knowledge of general practice.
- Continued with location of information for barristers, solicitors, architects, pilots, teachers, pharmacists, dentists, nurses and midwives, chosen on basis of professional status, risk, complexity, length of training, and health context.
- Completed with identification of international professional bodies for the countries reported in the RAND report[ii], and France, Netherlands, Eire, Canada, U.S.A., Australia and New Zealand.

Websites of Royal Colleges and other organisations (UK medical, UK non-medical and International) were searched directly to obtain the required information, in accordance with the table below.

Table 4

Number of professional examination and assessment systems identified through websites	57 relevant professional examination and assessment systems identified

Number of organisations with whom correspondence was opened to uncover any existing evidence base and policy	36 organisations

Correspondence was only opened with 36 organisations, out of 56, since some organisations operated the same exam (e.g. surgical colleges), and some organisations operated two examination and assessment systems (e.g. non medical organisations who are responsible for assessment of both (i) home and EEA students, and (ii) non-EEA students from non-preferred international countries e.g. Solicitors Regulation Authority.

Table 5

Number of additional papers relevant to Themes 1,2, and 3, identified from • the quoted evidence-base on websites, • evidence arising in correspondence, and • citation-following, and after de-duplication	27 papers (including some reports of conference proceedings and technical reports)

3.4 Direct contacts

In the early stages of the project, e-mails were sent to specific individuals with leading roles in medical education nationally and internationally, in order to identify materials or inputs lying within their personal experience. Replies were received from Professors Dave Swanson, John Norcini, Cees van der Vleuten, Lambert Schuwirth and Fiona Patterson, and these were extremely helpful in shaping the early stages of the study.

4 Theme 1 and 2 Findings

In each of the ‘Theme Findings’ Sections, the original GMC questions are repeated for clarity at the head of each Section, printed in blue. However, some of the ‘Theme 2’ questions (such as evidence relating to protected characteristics) are best addressed alongside Theme 1 issues: to do otherwise would require unnecessary duplication and lack of clarity and integration.

Theme 1 – An examination of available evidence on the number of attempts that candidates are allowed to sit examinations and assessments (for the purposes of gaining access to a profession in the UK, Europe and elsewhere in the world).

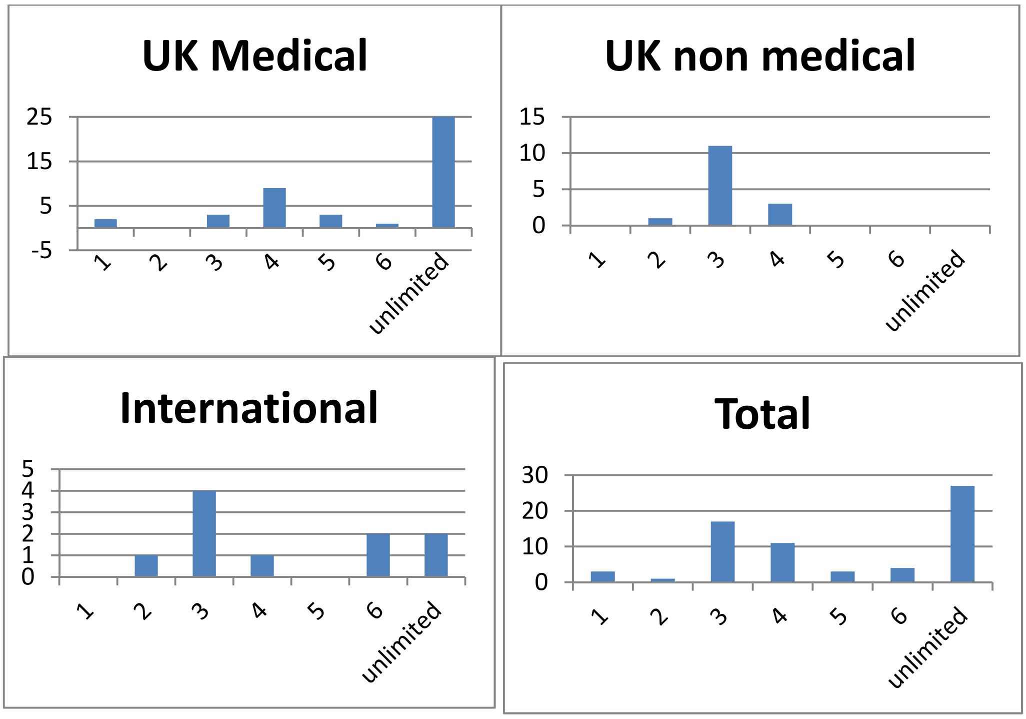
The review should explore the following themes, with particular reference to comparable professional and/or training examinations and assessments undertaken by other UK, European and international professional regulatory/examining bodies:

a. Is there any evidence on the number of attempts that other UK, European and international professional regulatory and examining bodies allow candidates to sit professional examinations and/or assessments? This should, where possible, include a commentary on the evidence base/rationale for the approaches taken.

4.1 Practice of Other Professional Bodies

The information on the 56 professional examination and assessment systems has been collated to provide information about the number of attempts and periods of validity used by other relevant organisations. This information has been collated into tabular form (Appendix A). The summary numerical data is presented in the form of histograms (Figure 1). References given in bold and in brackets in this section are to the corresponding sections of table A1, where they are described in more detail, and often with hyperlinks, not to the main reference list which is given in Harvard style with endnote links.

Figure 1. Data from Appendix A



The majority of UK medical organisations do not limit the number of re-sits, but when they do, four exam attempts (first exam plus three re-sits) are the most frequently permitted. This contrasts with non-medical and international medical organisations, the majority of whom set a limit of three attempts but some set a limit of four.

After considering the quantitative aspects of the data there are a number of significant issues to be considered. Firstly, only 5% of the organisations responsible for the 57 professional exam and assessment systems quoted an evidence-base for their exam design, although we heard from only 53% of the organisations we identified.

Table 6

Number of professional examination and assessment systems identified through websites	57
Number of non-medical and international examination and assessment systems investigated through websites	31
Number of organisations with whom correspondence was opened to uncover any existing evidence base and policy	36
Number of organisations who published evidence base on websites	3 of 57 (5%)
Number of replies received by deadline	19 of 36 (53%)
Number of international organisations identified	17
Number of replies from international organisations	4 of 17 (24%)

The three professional bodies who published their evidence-bases on their websites are the NRO (National Recruitment Office) for GP Training (**4**), the Royal College of General Practitioners (**7**), and The UK Foundation Programme Office (**2**). The majority of evidence found in the grey literature search came from these three bodies. However, these three bodies all have some relevance to the GMC PLAB exam, in that they are concerned with assessing knowledge and skills relevant to practicing doctors, albeit at a slightly higher level for the first two bodies.

For those bodies who replied to our correspondence, and from consideration of all the websites, we were able to gain some information about how decisions had been reached. However for the other bodies who did not reply to the letter, although no evidence-base was quoted, it cannot be assumed that none exists.

The histogram highlights that the majority of UK medical organisations do not limit the number of re-sits, but when they do, four attempts are the most frequently permitted, with six attempts as the maximum. This is in contrast to other non-medical and international organisations that tend to set a limit of three attempts with a maximum of four, with only two international organisations permitting an unlimited number of attempts.

4.1.1 Evidence provided

Table A1 in Appendix A contains the results of our investigations. Key suggestions and exemplar exam systems have been highlighted in this report's Conclusion, and in the Executive Summary and Recommendations, for the GMC to consider further. Hyperlinks to the relevant exam regulations are embedded within Table A1.

We were also able to identify common themes, particularly from The Royal Colleges. Various bodies reported being guided by policy, for example The Royal College of Anaesthetists (**5**) refers to The Tooke Report[iii] and to GMC Guidelines 2012[i]. Some bodies were guided by consensus, referring to common educational and professional practice, or to policy that had evolved over time. Others relied on consultation, for example The General Dental Council (**25**) held a consultative exercise in 2006 which led to the setting up of their Overseas Registration Exam, or by undertaking their own review and analysis of data; an example being the Intercollegiate Committee for Basic Surgical Examinations 2011 (**24**) after a review by Sheffield University. One international medical council (the Medical Council of New Zealand) (**57**) was influenced by a legal challenge in another jurisdiction (Australia)(**56**).

Most of the 19 professional bodies who provided information in correspondence were careful to take the opportunity to explain their exam regulations in more detail to us. Where possible, data in the Tables A.1 and A.2 has been marked 'verified' where confirmed by correspondence (received prior to the deadline). Condensing the data and reducing it to numbers results in a loss of detail for each exam design. It is worth making reference here to van der Vleuten's (2005)[iv] comment that 'a characteristic of an assessment method is not inherent in the method but depends on how and in what context assessment takes place'; thus, it is observed that the number of attempts and periods of validity for each and every exam are a part of examination design that has evolved over time, and these factors interact with particular training structures and the requirements of individual specialties, different professions, and different national cultures.

Taking the view of most professional bodies that their examinations or assessments function as a dynamic system it is apparent that if numbers of attempts and periods of validity are set rigidly rather than being fine-tuned to the type of exam, its objectives and its professional culture, the effectiveness of the examination is likely to be compromised. It flows from this that rigid adherence to the same set number of attempts or period of validity for all medical organisations is likely to be unsatisfactory.

Another point worth raising in connection with this data is that all professional bodies use similar language but it does not always mean exactly the same thing**.** 'Attempts', 'periods of validity', 'papers', and 'stages' can have quite different meanings. Exams are also often interdependent (e.g. candidates cannot take part 2 of an exam until part 1 has been taken) hence 'attempts' and 'validity' may depend on understanding the particular exam structure concerned.

Furthermore it should be noted that most medical assessment systems are in transition. The information in Table A1 often relates to new arrangements where there may be little experience of outcomes. Information may have changed even by the time of publication of this report; some exams are planned to change in the near future. For example, the Royal College of Paediatrics and Child Health (**11**) is introducing a new paediatrics Foundations of Practice Exam. For some professional bodies the length of training requires careful arrangements for transition from one set of exam regulations to another, for example the Regulations for the Intercollegiate Membership Examination of the Surgical Royal Colleges of Great Britain (revised March 2012) (**24**).

4.1.2 Evidence relating to UK postgraduate exams

Of those bodies with different aims from the GMC and PLAB, most are the Royal Colleges. McManus (2012) [v] notes failure rates in postgraduate exams are high, and bearing this in mind, and the unique specialty training routes the Royal Colleges individually set, these exams are difficult to compare to PLAB, but at the same time they inform us about UK medical education. The following are points of interest in considering the data in Table A1 at Appendix A for UK medical bodies (**2-24**)

4.1.2.1 Workplace-based assessment is seen as key to robust assessment

Some Royal Colleges have fully integrated workplace based assessment with examinations e.g. RCGP (**7**) and relied on a published evidence-base in so doing.

4.1.2.2 Remedial training is seen as an integral part of managing exam failure

The College of Emergency Medicine (**6**) has created a system whereby re-sit of its membership exam MCEM B is conditional on remediation and on the fourth attempt the approval of the candidate's trainer. The college reserves the right to halt re-sit attempts, possible because of local feedback from the candidate's supervisors.

4.1.2.3 Other approaches also act to disbar failing candidates

Royal College exams are run in parallel with specialty training hence the Annual Review of Competency Progression (ARCP) and the requirements of specialty training Person Specifications are additional ways of limiting the opportunities of weaker candidates by bringing to an end attempts to re-sit or closing a period of validity e.g. the Royal College of Physicians Membership Exam (MRCP(UK)) (**17**), and The Faculty of Occupational Medicine Part 1 Membership Exam (MFOM) (**9**).

Some exams are linked to a single annual recruitment round e.g. new Situational Judgement Tests (SJT) for UK Foundation Programme (**2**) and the National Recruitment Office for GP Training (**4**) assessments. This automatically removes the requirement for complex exam regulations.

The Royal College of Anaesthetists (**5**) has commented that the timing of exams within training and the number of exams available per year are also 'governing factors'[vi].

4.1.2.4 Royal College exams are unique to their training programmes

Some exams can be taken early after a medical degree, for example The Royal College of Obstetricians and Gynaecologists Part 1 exam (**8**), others not for, say, nine years later, at the end of training, for example the Royal College of Ophthalmologists Fellowship Assessment (**10**).

4.1.2.4 Some Royal Colleges have taken innovative approaches

The Royal College of Psychiatrists (**18**) states its validity period of 1643 days is linked to revalidation (Clisset G, personal communication) in being 4.5 years to which 6 months post-foundation experience in psychiatry is added. It is the only Royal College to state it has considered revalidation within its exam design.

The Royal College of General Practitioners has quoted evidence about involving lay people in selection to general practice training[vii]. While lay assessors would not be appropriate in the context of measuring clinical skills themselves, the GMC may wish to consider involving lay members as observers, to enable them to appreciate the rigour of the process, and therefore be able to provide independent testimony to enhance public confidence.

4.1.3 Evidence particularly relevant to PLAB

PLAB is of course different from many exams, in particular many Royal College Exams, in not being associated with a period of study or training. Twenty four UK Medical bodies (**1-24** in Appendix A) along with 2 UK Dental organisations (**25, 26**). Fourteen UK non-medical professional bodies were investigated (**27-40**). Seventeen international medical bodies were also considered (**41-57**). For the fourteen non-medical bodies and the General Dental Council, where possible or appropriate, two examination routes were considered: (i) home and EEA students, and (ii) non-EEA students from non-preferred international countries. From this group six organisations replied to our correspondence by the deadline.

Some organisations were in countries from which doctors tend to emigrate rather than being recruited from overseas, so their regulations were not aimed at addressing the assessment of such doctors.

Where information has been obtained it does provide useful evidence of current practice in this field, and the different approaches raise examples for the GMC to consider of both best practice, and practice that might not be sufficient to safeguard patient safety according to UK standards. This leads to an important question about the admission of international medical graduates through other EEA countries. This issue is raised by Sonderen et al[viii] in their 2009 paper about their exam system for international medical graduates (IMGs) in The Netherlands (**43**),

> "after registration in one country of the European Economic Area (EEA), IMG's have access as an independent health care provider to all other countries of the EEA without further national assessment of their skills"

4.1.3.1 Workplace Based Assessment is being adopted in a variety of areas

The Australian Medical Council Ltd (**56**) is piloting a move to this, replacing its Clinical Examination in four states, alongside the existing exam.

The Medical Council of New Zealand's New Zealand Registration Exam (NZREX) Exam (**57**) is part of a wider clinical process of 12 months' supervised clinical practice.

Some bodies tie an examination into a compulsory training period such as The General Pharmaceutical Council's Overseas Pharmacists' Assessment Programme (OSPAP) (**35**), and The Nursing and Midwifery Council's Overseas Nursing Programme and their Adaptation to Midwifery programme (**33**). This is discussed further in Section 6.1.

4.1.3.2 Examinations design can lead to individualised training recommendations for candidates

The Netherlands' procedure, the Dutch Assessment of Medical Competence of Foreign Medical Graduates (DAMCFG), examines a candidate's portfolio and skills assessment (**43**) before, where necessary, recommending additional training and how long such training should take.

4.1.3.3 Assessment of international medical graduates can be integrated with selection for all

The Medical Council of Canada (**55**) uses its Medical Council of Canada Evaluating Examination (MCCEE) for international medical graduates to also rank doctors in their Canadian Residency Matching Service) (CaRMS).

4.1.3.4. Safeguarding patients can be achieved in a number of additional ways

A critical incident rule is a useful safeguard. The New Zealand Registration Exam (NZREX) (**57**) has this rule, and, in the UK, The Bar Standards Board has an equivalent red light rule (**30**) to allow immediate disqualification of a candidate who demonstrates knowledge, behaviour or skills that are potentially dangerous.

The GMC has a 'Cause for Concern' procedure for PLAB 2, activated if "…. any examiner considers a candidate's behaviour gives rise to cause for concern, about the candidate's fitness to practise which cannot be addressed by the station's marking schedule"[ix], and this is to be commended.

A higher IELTS score is regarded as necessary by the New Zealand Medical Council (IELTS 7.5) (**57**) and by the Solicitors Regulation Authority (**40**) and Bar Standards Board (**30**).

The Australian Medical Council Ltd has a system whereby a marginal performance leads to further testing – a separate 'second look' rather than a re-sit. (**56**). The value of this approach is that eliminating those candidates who clearly passed and those who clearly failed means the 'second look' can be designed specifically for borderline candidates. This can give high reliability with some economies of scale. For instance, in principal, an 8 station OSCE followed by a second (different) 8 station OSCE for borderline candidate could achieve the reliability of a 16 station OSCE without the cost of administering a 16 station OSCE to all candidates[x]. Having a borderline category with re-examination reduces the possibility of false positives. However, this should be viewed as an alternative to the current [practice of adding 1 SEM to the PLAB Part 2 cut score, which also reduces false positives. Our Recommendation on this approach is that the GMC review the evidence from Australia when it is fully available.

Preference is given to first attempt candidates in The Australian Medical Council Ltd exams (**56**).

4.1.3.5. Organisations have made significant progress with evidence-based exam design

The USA (**54**) and Canada (**55**) have developed robust and validated procedures for assessment'.

The Netherlands (**43**) have based their design on a firm evidence base. Their exam now is likely to maintain its place at the forefront of this type of assessment since it is devolved to three universities where medical education is a leading academic subject of study.

This approach is described by Sonderen et al (2009)[xi], who review the introduction of a test combining language skills, cognitive knowledge and practical skills for IMGs wishing to work in the Netherlands. While at the point of publication, little evidence on the practical consequences of the test was available, and the numbers having undertaken it were too small for conclusions to be drawn, it does describe the implications and development of such a test in a clearly laid out manner. Standard setting in these early stages was by comparison with a reference group of Dutch graduates.

4.1.3.6 Special exemptions

We note that GMC guidance refers to exempt categories of persons, which includes spouses of EEA citizens. This means that IMGs may potentially be registered without any tests of professional competence, which we do not believe is conducive to patient safety. While this is described in the 'EC Rights Factsheet', we are concerned that a general right to employment may be viewed as a specific right to employment as a doctor. We suggest, short of making a formal recommendation, that the GMC re-examine the EC Rights Factsheet in the light of this observation.

4.1.4.Summary

Fifty six professional examinations and assessment systems were examined to identify what evidence, if any, was referenced to inform decisions about the number of attempts allowed and periods of exam validity. This review found only three organisations that clearly identified the evidence that informed their decisions. Other organisations referred to guidance or held internal consultation exercises. Evidence on the number of attempts allowed was at best weak and at worst absent. Evidence on periods of exam validity highlights no clear pattern. The majority of UK medical organisations do not limit the number of attempts, but when they do, four attempts are the most frequently permitted. This contrast with non-medical and international medical organisations who tend to set a limit of three attempts with a maximum of four.

4.2 Evidence on Re-sitting Assessments

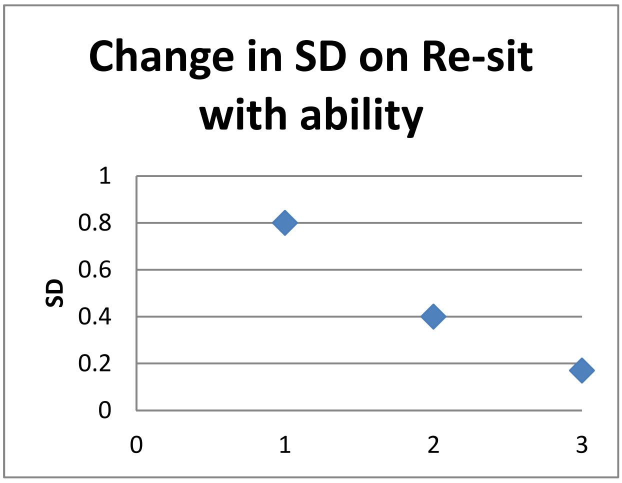
The following findings arise from the psychometric literature in general, sometimes not related to medicine, and sometimes based on studies of undergraduates and younger students, and based on a variety of test formats. Research work relating specifically to medicine is limited. In a discussion review, Ricketts (2010)[xii] commented that "there is no 'theory of re-sits' " and "there is much common practice but no evidence base for the interpretation of re-sit results.

4.2.1. Scores increase on re-sit

There are generally improvements on test scores on re-sitting (Matton et al, 2009)[xiii] . A meta-analysis (Kulik et al, 1984)[xiv] indicated that test scores increased by 0.42 Standard Deviations (SD) at the second administration of an identical test, and by 0.23 SD at the second administration of a parallel test.

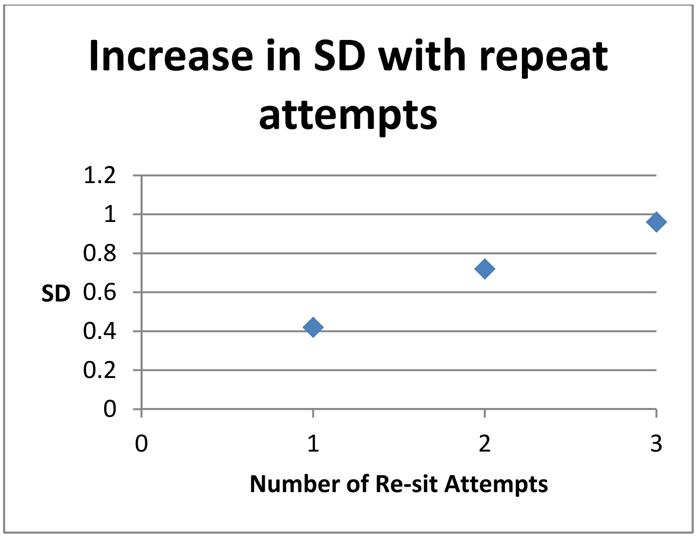
Interestingly, they also found a significant positive relationship between ability and the score increase observed on re-take. High ability re-sitters (i.e. those just below the cut score) had a practice benefit of 0.80 SD, middle ability re-sitters of 0.40 SD and low ability re-sitters of 0.17 SD.

Figure 2 a (drawn from data in Kulik et al, 1984). 'High' ability = 1, middle ability = 2, and low ability = 3 on the abscissa of this Figure.



They also observed a 'dose response' effect of multiple re-sittings, with SD gains of 0.42 from 1st to 2nd re-sit, 0.70 from 1st to 3rd re-sit, and 0.96 from 1st to 4th re-sit. These data, however, relate to **identical** test forms, and may not correspond to related test forms.

Figure 2 b (drawn from data in Kulik et al, 1984).



Geving et al (2005)[xv] reviewed various items of literature on re-sit performance, summarising one key finding as being that repeated exposure to items promotes score increases beyond those of latent trait change. However, their own study was based on candidates sitting a real estate qualification in the U.S. They concluded that test scores increased but “the number of retakes and score gains were inversely related, indicating that, after the second testing opportunity, score gains were not as great”. Unusually, and out of line with the literature, previous exposure to items did not seem to have a significant effect in this study, but length of time between test attempts was positively related to score, suggesting learning had taken place.

This was also the case in a study by Raymond et al (2009) [xvi] who looked at same form retest effects on credentialing exams for radiographers. 541 examinees who had previously failed a national certification exam on their first attempt were randomly assigned to receive either the same paper again or a different (but parallel) paper during their second attempt. The study found that the group who had received the same form as they had in their first attempt had a shorter response time. However there was no mean score difference found between the two cohorts.

4.2.2. Studies focused on International Medical Graduates

This is similar to findings by Boulet et al (2003)[xvii] where they investigated the performance of repeat candidates in a high stakes standardized patient assessment using both new and previously seen exam materials. Data for this study were taken from first time and repeat candidates who sat the Educational Commission for Foreign Medical Graduates (ECFMG) Clinical Skills Assessment (CSA®). Skills are examined in high fidelity simulated environments. There were significant (*p*<.01) increases in candidates scores between first attempt and second attempt (retaken within six months of initial CSA) candidates for all of the CSA components. It was found that repeat test takers do not achieve any advantage or disadvantage over first time takers if assessment content overlaps.

An American study by Swygert et al (2010)[xviii] investigated gains for repeat examinees, where they had experienced repeat information, for the Step 2 Clinical Skills exam. This is a performance scenario based assessment where candidates interact as doctors with standardized patients. A large data set (n=3045) of candidates who had failed their initial exam were retested. They found that there was a significant score increase in their second attempt in all four areas of the Step 2 CSA. However they observed no significant difference in candidates who had previous exposure to the exam information. In a paper by Boulet et al (2003)[xix] they reported that non-US IMGs do slightly worse when they are exposed to repeat information.

4.2.3 Improvements are often the result of practice

In a review, Hausknecht et al (2007)[xx] found an overall increase of about 0.25 SD, based on 107 studies of cognitive oriented tests; while Raymond et al (2007)[xxi] reported re-test effects of .79 and .48 on two knowledge tests administered to radiographers. Schleicher et al (2010)[xxii] reported a re-test effect of .15 SD on a job-knowledge test given to federal agency job applicants.

These are often considered as practice effects, and therefore construct irrelevant[xxiii]. A similar conclusion was reached by Matton et al (2009)[xxiv] with regard to aircraft pilot training, in that the changes were all due to situational effects (anxiety, familiarity, etc). However, a study of law enforcement applications (Hausknecht et al, 2002)[xxv] indicated that re-sitters increased their scores from first to second assessment and from second to third assessment. Interestingly, these authors state that in work-related follow up, “the number of tests necessary to gain entry into the organisation was positively associated with training performance and negatively associated with turnover probability”.

Candidates who had undertaken the US Department of Labor General Aptitude Test (which tested cognitive ability and physical dexterity) were re-tested after two weeks, with one sub-group sitting an identical test, and the other a parallel test (United States Department of Labor, 1970)[xxvi]. The effect size for the parallel test form (0.15 – 0.55) was significantly smaller than that for the identical test (0.32 – 0.74).

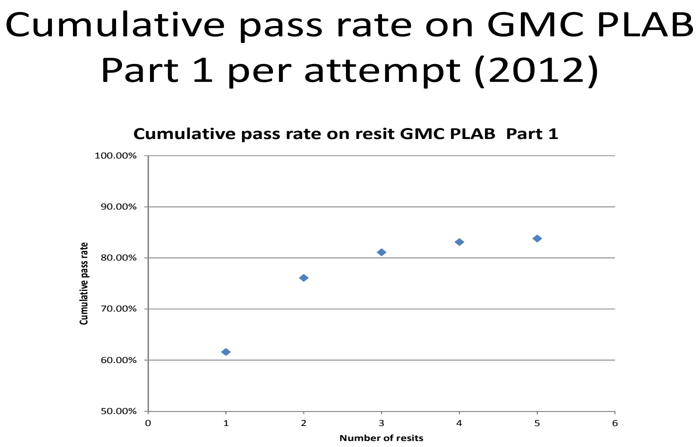
A study of candidates for medical school in Belgium (Lievens et al, 2005)[xii] showed that candidates performed significantly better on re-sitting knowledge, situational judgement and cognitive ability tests, with the improvement being most marked in the last of these[xxvii]. Subsequent exploration of performance showed that higher levels of performance were associated with a given score on the first attempt at the test for those who passed, but for re-sitters, the second attempt score for knowledge tests had higher validity than the first attempt score, suggesting there had been construct relevant improvement. However, cognitive ability tests showed the opposite effect, suggesting that in this case, score improvement was construct irrelevant. The study by Lievens et al (2005) [xii] includes fascinating insights into the consequences of retesting in a relevant setting, since they compared predictor scores taken in a medical school application with outcome variables for their performance as students. For cognitive knowledge tests they found that for first time test sitters, there was higher validity of test scores as compared with re-sitters, but for re-sitters, there was higher validity for their re-sit score than for their original score. (Such effects did not apply to knowledge tests and SJTs). This corresponds well to allowing at least one and perhaps two re-sits.

4.2.4. Evidence consistent with four attempts as an optimum

Hausknecht et al (2002)[xxviii] found the score increase on re-sit reaches a plateau between tests 3 and 4, and ascribed this to construct irrelevant features such as anxiety and test familiarity, as well as construct relevant features such as improved knowledge or skills.

Analysis from the GMC PLAB Part 1 examination is presented in Figure 3 below. This highlights that after the fourth re-sit, i.e. the 5th attempt, the curve of the pass rate flattens showing few or no additional passes.

Figure 3 (drawn from GMC data provided by Dr John Foulkes, psychometrician and statistician for the Part 1 exam, contracted to provide services to the GMC).



In 1992, McManus indicated that successful re-sit candidates in the MRCGP examination do not solely pass on the basis of constant ability mediated against chance, but that they do improve in performance on their second and third re-sit. After that, their performance declines, so limiting re-sit has some rationale on that basis[xxix]. This matches the data gathered by Hausknecht et al (2002)[xx], who also noted that performance increased significantly on the second and third attempts, but showed no further gain on the fourth attempt.

Bandaranayake and Buzzard (1994)[xxx] noted that, with regard to the Royal Australian College of Surgeons Part 1 exam, the probability of re-sit candidates passing remained fairly constant up to the fourth attempt and fell thereafter. The lower the candidates' original mark had been, the less likely they were to pass on subsequent attempts. Despite this, these authors do not recommend a limitation on the number of attempts a candidate may make, for reasons which are unclear.

Freeman and Wakeford (unpublished observations) reviewed re-sit performance of MRCGP candidates and found that the initial increases on scores on re-sits were not sustained, and recommend limiting the number of attempts to four. They also suggest that an increased passing score should be considered for re-sitters, although do not provide a rationale for selecting this approach rather than others.

4.2.5 Potential problems with unlimited re-sits

Cohen-Schotanus (1999) also commented on the paucity of literature on this subject, and comments "The more opportunities students have to repeat an exam, the less seriously they will prepare themselves for this exam"[xxxi]. However, this comment relates to undergraduate medical students, and is made as an assertion of opinion rather than ascribed to evidence, and his advice that it is best to give "no more than one opportunity per year to repeat the exam" must therefore be viewed in this light.

Probably the most in depth analysis of re-sit performance published so far on the consequences of multiple re-sits is that of McManus and Ludka (2012)[xxxii]. These authors analyse attempts from 2002/3 to 2010 at the Royal College of Physicians (UK) Membership tests, which are in three parts. Parts 1 and 2 are written (Best of Five MCQs): Part 3 (Practical Assessment of Clinical Examination Skills) is in the form of a practical multi station OSCE-style test, but using real patients. Since number of attempts has been unlimited, data is available on the consequences on candidate scores of large numbers of re-sits (with two candidates sitting Part 1 26 times, and one candidate requiring 35 attempts to pass all three Parts). Limitations include that the data is left and right truncated in that some candidates had sat the various parts before the study start date, and some candidates who had failed at the study end date would undertake further re-sits, but these are not likely to be significant factors. Using a variety of models, the authors suggest that scores increased up to the tenth attempt at Part 1, the fourth attempt at Part 2, and the sixth attempt at PACES. As a general conclusion, the authors suggest that there is no clear rational basis for limiting the number of re-sits. They suggest the possibility of increasing the cut score for candidates at each re-sitting (along the lines suggested by Millman (1989)[xxxiii] .

While these authors observed increases in scores on re-sit as described above, not due to chance, this does not necessarily reflect an improvement in physician competence. As described previously, there may be construct-irrelevant increases in scores (as described extensively in the literature – see above, particularly Matton et al, 2009)[xxxiv]. While the psychometric value of increasing the cut score on each re-sit occasion is clear, it is possible that there may be low acceptability of this approach, which may be seen to raise issues of equity: candidates attempting an exam for the *n*th time will have a higher cut score than candidates sitting it for the 1st time, and those 1st time sitters will therefore be deemed clinically competent on the basis of a lower cut score.

In a commentary on Pell at el (2012)[xxxv], Hays (2012)[xxxvi] refers to difficulties in remediation, and suggests "The alternative may be more practical and more controversial: be tougher with students who clearly fail".

4.2.6 Degree of failure predicts later success

Hays et al (2008)[xxxvii] explored re-sit performance by degree of severity of failure, banded by factors of Standard Error of Measurement (SEM) and found, as might be expected, that candidates who missed the pass mark by 1 SEM had a reasonable probability (83%) of passing on the second attempt: those who failed by 3 SEM had 100% probability of failing the re-sit or withdrawing. Interestingly, this might suggest that there is no single optimal strategy for the number of re-sits that should be allowed: that depends on the degree of 'failing'. On the other hand, it might also suggest a possible calculation to determine the number of re-sits that would allow all 'True Pass' candidates to proceed. This calculation could be compared with Bandaranayake and Buzzard's(1994)[xxxviii] observations, and also with theoretical calculations on passing probabilities (see Clauser et al, 2006[xxxix] and 4.2.7 below).

Raymond et al (2011)[xl] reviewed performance of re-takers on USMLE Step 2. This showed that first time failures had a markedly different factor structure than first time passers, but on their second attempt became more like first time passers. Comparison with subsequent clinically related performance showed that the re-sit score had more validity than the initial score, in findings similar to those of Lievens et al (2005)[xli] for knowledge tests.

Pell et al (2012) [xlii] show evidence that the performance of re-sit OSCE candidates declined across repeated attempts, despite conscious efforts at remediation. This implies limited value in retesting.

Tighe et al (2010)[xliii] demonstrate through Monte Carlo Simulation an example in which they make imaginary candidates sit a high reliability test, and then sit it again. While some 16% pass on the first attempt, only 11% pass on both occasions. However, by setting the cut score at 60%, while the mean score is 50%, this simulation increases the number of 'false negatives' (see Glossary, Appendix E for definition of this term). Not many professional exams consistently have a fail rate of 85% as in this model, and this would probably be viewed as a cause for concern in reality.

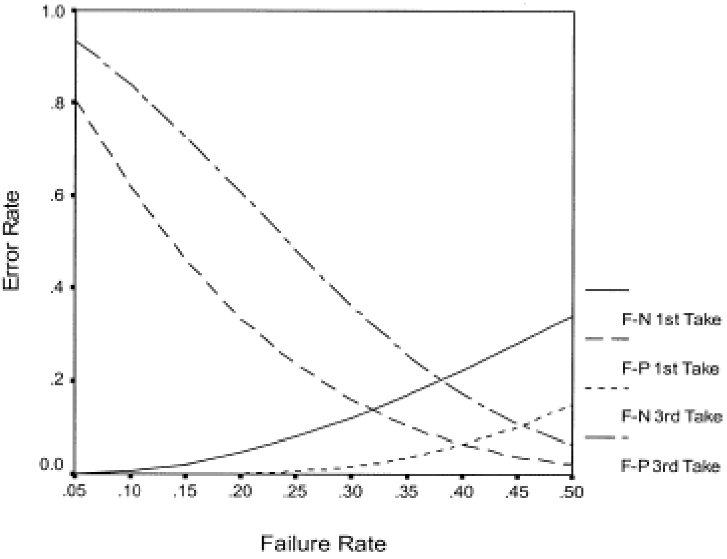
Clauser and Nungster (2001)[xliv] demonstrate that a test with a reliability of 0.92, applied to candidates, 90% of whom are proficient, will have a false positive rate of 20%, doubling after two re-sits, and at that point corresponding to the false positive rate of a test with a reliability of 0.69. As these authors indicate, false positive errors "may put the public at risk by allowing unqualified candidates to become licensed or certified". The false positive rate increases as reliability decreases, and decreases as the cut score increases, and consequently the fail rate increases. "One important strategy is to limit the number of retakes". Another is to increase the initial cut score, especially where the cost of a false positive is higher than that of a false negative. Finally, the paper explores the consequence of raising the cut score for re-sits, as suggested by Millman (1989)[xlv]. The effect of increasing the cut score by 0.25 SD per administration is considered, along with possible resistance to this approach. Although the value and strategy of increasing the cut score by this amount, and the reliability, cut scores, and abilities of the theoretical exams and candidates are arbitrarily chosen, the effect of increasing the cut score by 0.25 SD for each of 3 administrations is to reduce the false positive rate from 53% to 29%.

4.2.7 Theoretical models and risks associated with passing poor performers

Theoretical models (Clauser et al, 2006)[xlvi] indicate clearly that if either the cut score is reduced **or** the number of re-sits increased in an effort to reduce false negatives, then the false positives increase disproportionately (Figure 4).

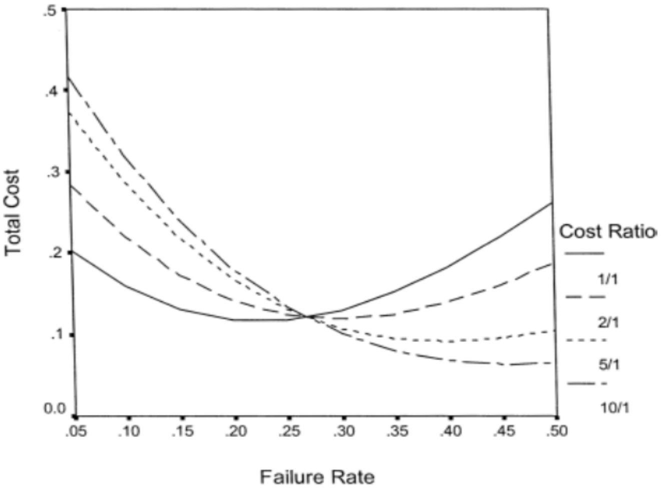
Figure 4 (from Clauser et al, 2006). The conditional false positive and false negative rates for a single administration of a test and for three administrations (i.e. 2 re-sits) of the same test

This diagram allows comparison of the consequences of three attempts as opposed to one attempt on false positive and false negative rates at different failure rates as represented on the x-axis. Note that the exam is getting harder as you move from left to right. If you compare the false positive rates on one attempt (F-P 1st Take, the 'dashed' line) with the false positives (F-P 3rd take, the 'dash-dot' line) you can see that there are many more false positive at all values. The fail rate in the GMC part 1 PLAB is about 25%: at this value the false positive rate has increased from about 0.2 on the first attempt to about 0.5 on the 3rd attempt.



This is because the situation is asymmetric. If a candidate fails, they are allowed to sit again: if they pass, they are not required to sit again, even though there will be false positives amongst those passing. The costs of this are also of interest, and depend on (a) the cost to the individual of a false negative, (b) the benefit to the individual of a false negative (you have to sit again, and may learn more), (c) the benefit to society of a false positive (if doctors are in massive under supply then any doctor may be better than none) and (d) the cost to society of a false positive (risk of poor medical treatment) (Figure 5).

Figure 5 (from Clauser et al, 2006). Costs from false positives increase as failure rates decrease, and the ratio of false positive costs to false negative costs rises.



Millman (1989)[xlvii] presents evidence of the impact of repeat testing on candidates of varying proficiency (and setting aside all factors other than exam reliability – in other words, neglecting factors such as candidate health and wellbeing, or the details of the testing environment). This indicates that candidates whose True Score is 5% above the cut score representing proficiency have a 90% chance of passing first time (hence with a 10% false negative rate). Borderline candidates whose true score equals the cut score have a better than 90% chance of passing after three re-sits. But candidates whose True Score is 5% below the proficiency cut score have a 15% chance of passing on the first attempt, and an 80% chance of passing after 10 attempts.

As Millman (1989)[xlviii] suggests, psychometrically, the best way to avoid false positives and negatives is to re-test 'Borderline' candidates repeatedly until it becomes clear where their True Score is likely to lie. However, this may not be practicable or acceptable to subject candidates who believe they have 'passed' (i.e. exceeded the cut score) to repeated and expensive tests which they might fail. Alternatively, the cut score may be increased on each re-sit attempt (in Millman's example, by about 1% on each administration), or the average score for each re-sit might be recorded, with the average being required to exceed the average cut score. Millman offers as a criticism of this last approach, that it may discourage a candidate with a very low initial score from re-sitting. We do not view this as a significant hazard; as we have said already, candidates who are significantly below the cut score have little chance of passing in any case, and ought to be discouraged from wasting their time and resources. More significant, perhaps, is the psychometrically complex meaning of averaging the cut scores across a number of tests. There is also the possibility that a small group of low-performing candidates would go on re-sitting indefinitely, in a vain attempt to raise their average. Millman indicates that averaging the two most recent attempts only, still brings about a significant decrease in the number of false positives, and this is probably more acceptable.

Where scores have significance beyond passing or failing (for instance, in University exams where 1st, 2nd and 3rd Class awards may be made) it is usual to cap the maximum score obtainable on re-sit, for exactly these reasons. However, this is not relevant for PLAB and other competency (rather than discriminator) exams.

Juul and Loewy, 1988 (cited in Millman, 1989)[xlix] report that in a medical education setting, lower scoring candidates select markedly more potentially dangerous choices than candidates scoring even slightly higher. Even borderline passers show worrying levels of dangerous choices. This reinforces the argument for some kind of detriment being applied to repeat re-sitters or at least, for a professional body, the need for a policy re-think in relation to repeat re-sitters (such as a requirement for targeted further training before being eligible for further attempts). A possible strategy may therefore be one which depends on the nature of the fail (and would lend itself to Bayesian approaches to the data). A candidate who misses the cut by >3 SEM might be allowed no re-sits, and a candidate who misses the cut score by 1 SEM might be allowed 3 re-sits, but perhaps not more, and a candidate who *passes* only by 1 SEM might also be required to re-sit. The strategy of requiring candidates who have nominally passed to re-test might initially be thought to be socially unacceptable, but is already in operation in Australia. For example in the Australian Medical Council Ltd Clinical Examination (**56**) for international medical graduates, where candidates whose performance is graded 'marginal' in a 16-station OSCE have to take a pass/fail retest consisting of 8 stations.

Another psychometrically valid approach to reducing false positives might be to increase the cut score on each re-sit attempt. This might also prove challenging in terms of acceptability to PLAB candidates, but perhaps not to those concerned primarily with patient safety.

Many organisations are 'risk averse', but for the wrong risk (Millman, 1989)[l]. They are sensitive to the immediate hazard of complaints of unfairness on the part of candidates, but insensitive to the less immediate hazards of poor practice by false positive candidates, or perceive it less likely that responsibility for poor practice will be ascribed to them as testing bodies.

4.2.8 Summary

The evidence from the literature on re-sitting exams highlights that there are benefits to re-sitting. Re-sits particularly benefit candidates who have higher ability (i.e. who fail narrowly). There is moderately strong evidence to indicate that there are no *significant* further benefits after four attempt (even if there is a small increase). The GMC's own data on PLAB 1 shows this effect (with only 0.7% of candidates passing on the 5th attempt). Equally, there are rising costs of false positives as the number of permitted re-sits increases, and these false positive costs are costs to the public and patients. Given that in medical post-graduate settings some individuals will continue to re-sit many times (see 4.2.5 above), there are also costs to candidates, we believe that an appropriate compromise is to limit the number of attempts permitted. Four attempts (i.e. 3 re-sits) is in line both with the evidence summarised above and the practice of a number of professional bodies. Selecting any set number of permitted attempts has a cost-benefit ratio, and we believe that recommending a limit of 4 is a reasonable compromise in the light of the literature on the subject. A further period in which personal development can take place should be required before further attempts are permitted.

Recommendation: The GMC should limit the number of attempts at PLAB Part 1 and 2 to four, followed by which a period adequate to allow further personal development should elapse before further attempts are permitted. The length of this period can be determined from the data collected in House Keeping Recommendation 3 (1).

Consideration, at least, should be given to re-testing borderline passing candidates to ensure patient safety.

Recommendation: We recommend consideration of an approach whereby marginal performance in PLAB leads to a 'second look' re-test (as opposed to a re-sit), following the Australian Medical Council Ltd Clinical Examination model.

4.3 Re-sit Attempts and Subsequent Performance: Relationship of Tests to Later Clinical Practice

> a. Is there any evidence of a correlation between the number of attempts to pass an examination/practical assessment and subsequent performance (with regard to knowledge/skills/conduct)?
>
> b. Is there any evidence to suggest that the number of attempts taken to pass an examination and/or assessment is a suitable predictor of competence? In other words, is there any evidence to suggest that candidates who require multiple attempts are more or less likely to encounter problems in professional practice than a candidate who passed on the first or second attempt? Again, this should, where possible, include a commentary on the evidence base/rationale for the approaches taken.

4.3.1 General Evidence

While specific information relating specifically to re-sits and future performance is lacking, there is considerable evidence that performance in written or practical assessments of medicine is to some degree predictive of performance in the practice of medicine. In an excellent meta-analysis of evidence in this area, Hamdy et al (2006)[li] concluded:

> "The studies included in the review and meta-analysis provided statistically significant mild to moderate correlations between medical school assessment measurements and performance in internship and residency. Basic science grades and clinical grades can predict residency performance".

The type of assessment corresponded to the skills later observed. NBME II scores correlated with NBME III scores, medical school clerkship grades correlated well with supervisor rating of residents; and OSCE scores correlated with supervisor rating of residents.

Table 7 (from Hamdy et al, 2006)

Predictor Variable	**Outcome Variable**	**Correlation**	**Confidence Interval**	**Descriptors**
NBME I	supervisor rating during residency	Pearson r = 0.22	0.13-0.30	positive significant low
NBME II	supervisor rating during residency	summary correlation coefficient r = 0.27	CI 0.16-0.38	positive significant low
Clerkship Grade Point Average	supervisor rating during residency	Pearson r = 0.28	CI 0.22-0.35	positive significant low
OSCE	supervisor rating during residency	Pearson r = 0.37	CI 0.22-0.50	positive significant low
Clerkship Grade Point Average	supervisor rating during residency	Pearson r = 0.28	0.22-0.35	positive significant low
NBME I	American Board of Medical Speciality Examination	Pearson r = 0.58	0.54 – 0.62	positive significant moderate
NBME II	American Board of Medical Speciality Examination	Pearson r = 0.61	CI 0.51-0.70	positive significant moderate

This data indicates on the basis of strong studies that there are statistically significant relationships between earlier performance in medical training and later performance. The effect sizes are generally small, but it is worth bearing in mind that education is a very complex process, with a great deal of individual variability: to see any phenomenon which is marked enough to give statistical significance is in itself rare. These values are comparable with other effect sizes which strongly influence medical and other training (e.g. selection for posts). Even so, some of the effect sizes, such as that between performance in the undergraduate NBME Part 1 and Part 2 assessments (national tests administered to all US medical undergraduates) and performance in subsequent Speciality Board Exams, reach the technical level of 'moderate', which means they are likely to be meaningful.

4.3.2 Evidence from individual studies

Ramsey et al (1989)[lii] demonstrated that American Board of Internal Medicine (ABIM) certified doctors scored more highly than non board certified internists on a knowledge test, and were also significantly more highly rated by peers with regard to clinical skills. Scores on this test correlated with ABIM scores and peer ratings, and subsequent written examination scores. The certified clinicians performed modestly better in terms of preventative care measures and some measures of patient outcome. Modest associations with patient scores or parameters of patient care were found.

Tamblyn et al (2002)[liii] compared the performance of 912 family physicians in Canadian licensing examinations with subsequent performance measured by a number of indices, such as appropriate prescribing, delivering continuity of care, and screening patients for serious illness. For instance, they noted that higher scores on drug knowledge were associated with lower rates of contra-indicated prescribing (relative risk 0.88). They concluded "Scores achieved on certification examinations and licensure examinations taken at the end of medical school show a sustained relationship, over 4 to 7 years, with indices of preventive care and acute and chronic disease management in primary care practice". This study confirmed and extended an earlier study which was confined to the first 18 months of practice (Tamblyn et al, 1998). Tamblyn et al (2007) compared performance on the Canadian Clinical Skills Examination (CSE), which was similar to USMLE Step 2 CS. Candidates who lay two standard deviations below the mean for communication skills in the CSE were significantly more likely to be the subject of non-trivial complaint in later practice.

Beard (2005)[liv] showed that use of a score checklist and global judgements allowed surgeons, surgical trainees and OR nurses to discriminate between expert and novice performance.

Holmboe et al (2008)[lv] explored the relationship between physicians' scores on the American Board of Internal Medicine's Maintenance of Certification examination and a variety of indices such as delivery of diabetes care, mammography and cardiovascular care. Their conclusions, like those of Tamblyn et al (2002), were stated unequivocally: "Our findings suggest that physician cognitive skills, as measured by a 'maintenance of certification examination', are associated with higher rates of processes of care for Medicare patients".

Wenghofer et al (2009)[lvi] showed that, using the Medical Council of Canada qualifying exams as the predictor variable, and peer assessment using a structured chart review and interview as the outcome variable, there was a relationship between provision of unacceptable quality of care and low grades in the written and practical parts of the test.

Mitchell et al (2011)[lvii]demonstrate that, for UK Foundation doctors, there is a significant association between low CBD and mini-CEX scores and being in difficulty, although the predictive value was relatively weak, and these authors suggest this makes it difficult to use as a summative test.

Hess et al (2011)[lviii] demonstrated that carefully calibrated scores on an on-line 'practice improvement module' set by the American Board of Internal Medicine Maintenance of Certification programme, associated with an Angoff standard setting procedure, identified a small outlier group of physicians who had significantly lower clinical competence and professional behaviour ratings when residents, lower examination scores, and were more likely to work in solo practice. No sensitivity or specificity data is available from this study.

Wakeford (2012)[lix] demonstrated a strong similarity between scores of international medical graduates in the MRCGP exams and in the GP National Recruitment Office validated tests. He concludes there are good grounds for believing that the GP NRO tests predict future clinical performance. The inference is then that the MRCGP tests predict future performance. Note, however, that this is only true for the *population* of international medical graduates – no data on individual performance is presented.

Perhaps the most pertinent reference (hence taken out of chronological order) is that of Southgate et al (2001)[lx]. This study compared the performance of 'reference' doctors (in good standing) with doctors referred to the GMC's procedures over concerns. Each group undertook a written knowledge test (EMQs), a simulated surgery test and an OSCE (drawn in part from PLAB). Standard setting was by Angoff, Contrasting Groups, and the highest score of any failing candidate respectively*[1]. Extremely significant differences were obtained between the groups on all three tests, with high specificity and acceptable sensitivity. The correlation between the knowledge test and the OSCE was 0.69 ($p < 0.01$) and between knowledge and simulated surgery 0.72 ($p < 0.01$). This demonstrates a strong relationship between performance on the test, and performance as a doctor, dichotomised between 'predicted competent' and 'predicted incompetent' states. The particular relevance here is that this study was conducted on UK doctors, with some of the materials actually being drawn from PLAB OSCE. It is therefore consistent with these data to conclude that performance on PLAB does indeed bear a relationship with performance as a doctor in practice, and therefore that re-sitters who have performed less well on PLAB by definition are likely to perform less well as doctors.

If therefore we accept that there is a relationship between being required to re-sit, and generally low scores on tests, and that there is a relationship between low scores on tests and subsequent poor performance in clinical practice, then we might plausibly infer that there is a relationship between re-sitting and subsequent poor performance. However, this inference requires independent study since we have not so far identified any studies of this kind.

[1] *This represents a 'Borderline Groups' style of approach, in which the 'Group' is that of failing candidates rather than Borderline candidates, and it is not the mean but the upper boundary which is selected as the cut score.

It is of considerable interest that Hausknecht et al (2002)[lxi] and Van Iddeking et al (2011)[xcvii] report that for those who re-sat tests, there was higher job related validity for their re-sit scores than their original test score: in other words, those who failed and did better on the re-sit were giving a more accurate view of their performance on the re-sit. These results may be highly context specific. This effect is particularly strong for certain groups, e.g. women and younger men (i.e. under 40), who are well represented among PLAB candidates. The effect was less strong for Black and Hispanic candidates, but this sub group analysis may have suffered from low group size and hence statistical power. We will consider protected characteristics further below.

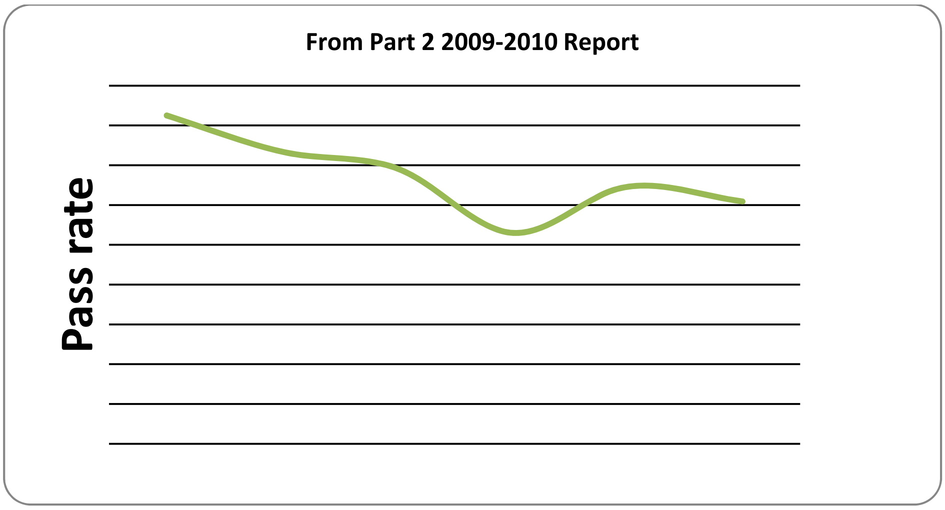
Recommendation: The GMC should conduct research into the relationship between re-sitter performance and subsequent performance in clinical practice.

4.3.3 Effect of Age on Performance

Another interesting association is that between scores on tests, which frequently decline with age, and performance as a clinician (which shows a similar decline, even with the factor of increasing experience). A review by Choudry et al (2005)[lxii] shows some variability but there is a general trend of decreasing performance with age over a wide range of clinical settings. This may relate to currency of knowledge as well as knowledge attrition. For instance, Norcini et al (2000)[lxiii] reviewed mortality from acute MI with age of cardiologists, internists and family practitioners. There was a consistent increase in mortality of 0.5% (SE 0.27%) for each year since medical school graduation across these specialities. Frequently the reported studies were stratified or dichotomised, so trends are less clear, but 40 years of age, or 20 years since graduation from medical school, were frequently chosen audit points at which statistical differences were detected.

GMC PLAB OSCE data from 2009-2011, re-plotted from figures provided by John Foulkes, shows a similar decline with age.

Figure 6 (Data from GMC Part 2, provided by Dr Suzanne Chamberlain, psychometrician and statistician for the Part 2 exam, contracted to provide services to the GMC.)

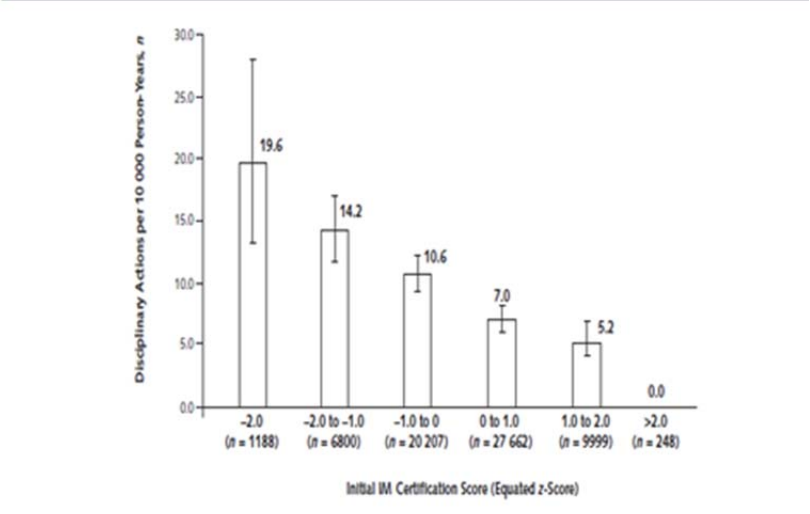


4.3.4 Test Scores also predict future professional behaviour

A most unexpected finding is that scores in the US Internal Medicine Certification test correlate negatively with the likelihood of disciplinary action in later careers, in a manner which is not merely statistically significant, but also shows a clear dose response curve (Figure 7, from Papadakis et al, 2008)[lxiv]. While it is not unexpected that there might be a relationship between knowledge and skills in early careers, as measured by tests, and later performance, it is much less expected that this extends to behaviours. One of us has argued elsewhere (McLachlan 2010) [lxv] that this may relate to *conscientiousness*: conscientious students are likely to perform better on tests and also to make conscientious doctors in later practice.

Figure 7 (from Papadakis et al, 2008)

Figure 2. Incidence of disciplinary actions, by certification on intial examination attempt.



4.3.5 Summary

The evidence to support a correlation between assessment performance and later performance in practice overall is moderate to strong. There is no evidence directly relating re-sit number to later clinical performance, but it is a reasonable inference that the two are negatively related.

There is also some evidence on a link between age and knowledge decay, and age and currency of knowledge, both declining with age.

We have already recommended that the GMC limit the number of re-sits permitted, in the interests of both patients and candidates, until a reasonable period has elapsed to allow for the possibility of further personal development (4.2.8).

Inevitably, there will be candidates with particular special circumstances (such as ill health) for whom allowance must be made.

Recommendation: We recommend giving consideration to an 'exceptional circumstances clause' to give flexibility for unusual candidate circumstances. The model is the RCGP Applied Knowledge Test (AKT) and CSA exam, allowing one further attempt in exceptional circumstances.

4.4 Trends with Regard to Candidates with Protected Characteristics

c. Does the evidence show any particular trends in outcomes for candidates who have 'protected characteristics' in terms of the number of attempts required to pass an assessment/examination, and the extent to which number is a predictor of subsequent problems in professional practice?

4.4.1 Introduction

Considering the nine protected characteristics in The Equality Act, references to three protected characteristics were found in our literature search: gender, race and disability.

It is perhaps worth noting that, at the present time, culture is not a protected characteristic. Hence there is a legitimate public and professional expectation that doctors seeking training or employment opportunities in the UK should have an appropriate degree of cultural awareness and be able to demonstrate this in an assessment. This is also reflected in the assessments of non-preferred doctors in most of the other seventeen countries in Table A.1 and A.2 (**41-57**), or, where it was not manifestly the case, this was usually either because a country had little expectation of attracting international doctors or had not fully developed the systems for managing the licensing of these doctors.

The evidence on these issues is confounded by inter-relationships identified between (for instance) race/ethnicity and gender in some research.

4.4.2 Gender

4.4.2.1 Females out performing males

Veloski et al (2000)[lxvi] showed that verbal ability had an increasing effect on performance as undergraduates moved through the training programme, while science knowledge had a diminishing effect. Women tended to outperform males, but the only consistent effect throughout was under performance from Asian American students.

Ferguson et al (2002) reported that a common finding in the literature is that females outperform males during medical training and in clinical assessments[lxvii] .

Wiskin et al (2004)[lxviii] found that female medical students performed better than males in communication skills assessments. However, the difference in performance was only statistically significant in stations simulating scenarios focused on vaginal discharge, female cancer, epigastric pain, haematology, addiction and chronic disease management.

A study by Boulet and McKinley (2005)[lxix] found that females performed better than males in written elements of clinical skills assessments at undergraduate and postgraduate levels.

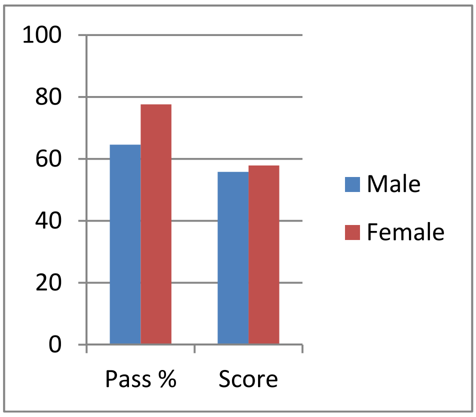
Dewhurst et al. (2007)[lxx] reported that female UK graduates performed best overall on performance on the MRCP(UK) compared to males, with white candidates performing better overall (discussed below). White males performed better than non-white women.

4.4.2.2 No gender difference

White and Welch (2012)[lxxi] found that male medical students initially performed better than female students on a Fundamentals of Laparoscopic Surgery task. However, after a brief training intervention, no between-gender differences were detected.

GMC PLAB data (part 2)(2009-2011)[lxxii] indicates a small gender disparity in terms of scores but a more significant disparity in terms of pass rates (see Figure 8).

Figure 8 (From GMC PLAB data (part 2) provided by Suzanne Chamberlain)



4.4.2.3 Males outperforming women

Dawson et al (1994)[lxxiii] compared the performance of four groups, classed as white, Asian/Pacific Islander, Hispanic and black, analysing also by gender, in the NBME Part 1 exams. Scores obtained by these groups were ranked in the order listed, and women performed less well than men in all categories.

A study by McManus et al (2008)[lxxiv] also reported that females were underperforming on Parts 1 and 2 of the Membership of the Royal College of Physicians examinations, again these examinations are MCQs. It is possible that the effect is related to examination type as the effect was reversed for the clinical assessment examination (PACES).

A study by Bowhay and Watmough (2009)[lxxv] reported that, while females outnumbered males in taking the MCQ section of the primary part of the Fellowship of the Royal College of Anaesthetists, they were not performing as well as males.

4.4.2.4 Conclusions

Female medical students outperform males in assessments of communication skills in key clinical scenarios. Male medical students from minority backgrounds are at risk of underperforming compared to their colleagues. Effective training in surgical techniques eliminates any difference in performance between males and females in surgical skills assessments.

4.4.3 Race

4.4.3.1 Non-white students tend to underperform in assessments

There is evidence both from the UK and US that ethnic minorities are more likely to underperform at medical school compared to white students (Ferguson et al 2002)[lxxvi]. However, the effect is confounded. In a study of final examination performance, UK ethnic minority students performed less well than UK White students, although non-UK ethnic minority students performed better than UK White students (McManus et al, 1996)[lxxvii]. This effect extended across all types of examinations including: MCQ, essay questions, clinical examinations and oral examinations. The observation that non-UK ethnic minority candidates performed better than UK white students suggests that the effect might not be due to simple prejudice.

Moore and Rhodenbaugh (2002)[lxxviii] conducted a survey of directors of surgery residency programmes, and found that domestic medical graduates were favoured in the applications process, and it was believed that international graduates were discriminated against when applying for training in general surgery.

Wass et al (2003)[lxxix] undertook a qualitative and quantitative exploration into the effects of ethnicity on the performance of undergraduate students. They found no evidence of discrimination. However, they found that male medical students from ethnic minority backgrounds performed less well than white students. This was particularly pertinent to communication skills, and their qualitative exploration demonstrates subtle differences in communication style that contribute towards this relative underperformance.

Van Zanten et al (2003)[lxxx] analysed performance of 30,000 international medical graduates undertaking the Clinical Skills Assessment of the ECFMG test. Performance was positively associated with female gender, recency since graduation, USMLE Part 1 and Part 2 scores, and 'English as a foreign language' scores (with native English speakers scoring most highly).

Dewhurst et al (2007)[lxxxi] reported that white UK medical graduate candidates achieved the highest pass rates in all parts of the MRCP(UK). Females performed particularly better on the Practical Assessment of Clinical Skills (PACES), with non-white males performing worst. Non-white candidates were found to perform poorly on both examination skills and communication and particularly poorly on the communication and ethics stations. This last point is worth highlighting as a finding in non-white UK graduates, highlighting potential cultural differences in interpretation as well as performance which one would expect to be even stronger in non-UK graduates.

A systematic review by Woolf et al (2011)[lxxxii] found evidence to suggest that 'non-white' medical students and trainee doctors performed less well in assessments than white candidates. They performed a meta-analysis of 23 research papers, demonstrating that the pattern is consistent between undergraduate, post-graduate and post-qualification training.

4.4.3.2 Male ethnic minority groups underperforming

As indicated above, Wass et al (2003)[lxxxiii] undertook a qualitative and quantitative exploration into the effects of ethnicity on the performance of undergraduate students. They found that male medical students from ethnic minority backgrounds performed less well than white students.

Yates et al (2006)[lxxxiv] commented that "not being white was a significant predictor of struggling" at medical school, while being male was also a risk factor.

4.4.3.3 No differences in performance

Hofmester et al (2009)[lxxxv] found that there was no difference in performance by ethnicity, gender or primary language among international medical graduates. Additionally, their performance in recruitment exercises for family medicine residency programmes was comparable to domestic applicants.

GMC PLAB Part 2 data (2009-2010)[lxxxvi] indicates no significant differences in pass rates whether analysed by nationality or ethnicity. We do not have this data for PLAB Part 1.

4.4.3.4 Conclusions

The investigations into ethnicity appear somewhat inconclusive. The largest investigations, as systematic review and meta-analysis, suggest that trainee doctors at all levels of their training from particular ethnic minorities underperform compared to their white peers.

4.4.4 Disability

Due to scarcity of evidence, the following discussion centres on dyslexia.

4.4.4.1 Dyslexia

We know that dyslexia is the most common Specific Learning Difficulty (SpLD) affecting between 3 and 10% of the general population, and up to 1.9% of medical students (Miles, 2004[lxxxvii]; BDA, 2012[lxxxviii]; Shrewsbury, 2011[lxxxix]). Data from the Universities and Colleges Admissions Service (UCAS) and the Higher Education Statistics Agency suggest that the number of students at UK medical schools who have declared a diagnosis of an SpLD has steadily increased since 2004 (Shrewsbury, 2011[xc] Gibson & Leinster, 2011[xci]).

Ricketts, Brice and Coombes (2010)[xcii] looked at the comparative performance in exams between dyslexic and non-dyslexic medical students to answer a different sort of research question: does exam format discriminate against candidates with an SpLD? Re-assuringly, perhaps, they found no evidence that this was the case, but noted an absence of research literature on this issue.

Gibson and Leinster (2011)[xciii], retrospectively analysing the exam results of medical students on a new undergraduate degree programme, found that dyslexic students performed less well than non-dyslexic peers in written assessments in their first year. They also report that dyslexic students who were not afforded extra time in exams performed less well in short answer question papers, but this difference was not found in other forms of assessment. Students with dyslexia performed less well than non-dyslexic students in OSCEs in their first year. Differences were noted in performance in OSCE stations, where dyslexic students performed less well in stations assessing practical skills and data interpretation. However, they found no difference in overall OSCE scores.

McKendree and Snowling (2011)[xciv], analysing a small set of data, concluded that dyslexic medical students in receipt of extra time in examinations did not perform significantly differently compared to their non-dyslexic colleagues.

4.4.4.2 Royal College Assessment Reports

The following UK Royal Colleges had accessible reports on their membership assessments:

RCA : no information of performance, or protected characteristics.

RCGP: data on performance and protected characteristics from 2007.

A small number of applications made by trainees with SpLD to the Royal College of General Practitioners (RCGP) for ‘reasonable adjustments’, such as extra time, in assessments. In 2007, only 3 (0.1%) applications were made, but by 2011 this figure had increased almost tenfold to 29 (0.9%). Evaluation of the performance of these candidates is not possible based on the data provided in the annual reports (RCGP, 2011).

No further information was available from the RCOphth, RCPath, or RCS.

A review undertaken by Prof. Dame Lesley Southgate, of the MFPH examination is available from the FPH (RCP). However, this does not give detail of performance in the faculty’s assessments.

The RCPsych declined to provide data, but suggested that all Royal Colleges have to provide this data in an annual report to the GMC and that, subsequently, this data may be accessible through the GMC. Other colleges declined to provide data due to the concern that the inherently small number of candidates reported with learning difficulties in the report could be, somehow, identified.

4.5 Summary on protected characteristics

No clear overall conclusions relating to protected characteristics can currently be drawn on the available evidence, in part because of its paucity, in part because of its inconsistency, and in part because of its context specificity. Whereas Ricketts et al (2010)[xcii] and Mckendree and Snowling (2011)[xciv] found no statistically significant differences in the performance between dyslexic and non-dyslexic students, Gibson and Leinster (2011)[xciii] did. Of significance, it appears that dyslexic students who do not receive additional time in assessments, underperform compared to those who do. A higher rate, or greater consistency, of provision of reasonable adjustments in the institutions included in Ricketts et al (2010)[xcii] and Mckendree and Snowling (2011)[xciv] could account for the lack of difference in performance between dyslexic and non-dyslexic students. Any differences that have been found are small and unstable, in that they appear to change over the progress of training. Studies such as these tend to be limited by the small sample sizes involved, which is a consequence of investigating a marginalised and under-represented minority.

The issue of SpLD in membership exams seems rather emotive to college assessments offices, and there is a great reluctance to provide data and suspicion that such data could lead to identification of individuals or used in a way to reflect poorly on the College. A small, but increasing, number of trainees are applying for, and being granted, extra time in the Applied Knowledge Test of the nMRCGP.

In an extensive review, Hough et al (2001)[xcv] reported ethnicity, age and gender differences, but also identified a number of factors such as measurement method, culture, test coaching, applicant perceptions such as test anxiety and perceptions of the validity of the test, and stereotype perceptions which can significantly reduce these differences.

In respect of gender, the literature highlights that females tend to outperform males. There are some studies that do not support this trend, indicating males may perform better in certain types of exams e.g. MCQ. The GMC PLAB Part 2 data indicates that females are more likely to pass an exam compared to males, however the difference in scores between males and females is small. We do not have this data for Part 1.

In respect of race, white candidates tend to outperform ethnic minority UK graduates (this trend seems to be reversed for non-UK undergraduates). The GMC PLAB Part 2 data for 2009-10 does not indicate any difference in pass rate when analysed by ethnicity. We do not have this data for Part 1.

Protected characteristics are an indicator for the targeted efforts to eliminate discrimination. In terms of employment, this may impact on recruitment strategies. However, in terms of professional qualification and regulation, it is clear that a protected characteristic cannot compromise patient safety.

Whilst the cultural mores to which an individual is enculturated may not translate to those that candidates sitting assessments in communication skills will be expected to demonstrate, there is an expectation that individuals seeking training or employment opportunities in the host country will make efforts to familiarise themselves with the customs therein.

Disabilities, such as enduring mental health conditions and Specific Learning Difficulties, however, do have the potential to interact with performance in assessments. Moreover, it is discriminatory and negligent, by law, to make no accommodation for such disabilities. It is common practice to allow extra time in written assessments for individuals with dyslexia. However, Gibson and Leinster (2011)[xciii] demonstrated that there are significant differences in performance in practical assessments too.

As regards disability, protected characteristics require targeted efforts to eliminate discrimination. However with this in mind finding a balance between ensuring patient safety and fairness to PLAB candidates can be seen to present an on-going challenge. In accordance with The Equality Act 2010 it is common, for example, to allow extra time in written assessments for individuals with dyslexia. However, Gibson and Leinster (2011[xcvi]) demonstrated that there are significant differences in performance in practical assessments too.

Certain classes of candidate therefore seem to benefit differentially from re-sits . In general, females, younger candidates and candidates reporting themselves as white tend to improve their performance on re-sit: males, older candidates, and candidates reporting themselves from an ethnic minority, do less well on re-sits (Van Iddeking et al 2005[xcvii]; Schleicher et al 2010[xcviii]). There is a lack of research on the impact of other protected characteristics.

One way to explore the potential bias within any assessment item is through Differential Item Functioning analysis (See Glossary). We therefore recommend that the GMC explore the impact of protected characteristics by collecting and analysing further data derived from an enhanced demographic data collection system obtained from PLAB assessments, as recommended. Of course, this would only identify Differential Functioning within the pool of IMGs, and in the ideal world, a reference group of white UK trained doctors should be used for comparison. However, there is probably sufficient diversity within the IMGs to give valuable results, particularly with regard to age and gender, but also to different ethnic groups.

Recommendation: We recommend the GMC collect demographic data during PLAB (including a voluntary section for candidates on all protected characteristics) to enable analysis of the influence these have on test outcomes.

Recommendation: We recommend that the GMC calculate Differential Item Functioning to explore the ways in which individual test items perform poorly.

4.6 Theme 2: Periods of validity and attrition of knowledge

An examination of available evidence on the periods of validity of passes in examinations and assessments (for the purpose of gaining access to a profession in the UK, Europe and elsewhere in the world).

1. The review should explore the following themes:

   a. Is there any evidence on the attrition rate of professional knowledge and skills, and the extent to which this varies between doctors, other health professionals, and non-health professionals undertaking comparable examinations/assessments?

b. How do periods of validity for examination/assessment 'passes' vary by profession (e.g. medical, non-medical, other). For example, is this related to:

i. The attrition rate of the relevant knowledge and / or skills examined / assessed?

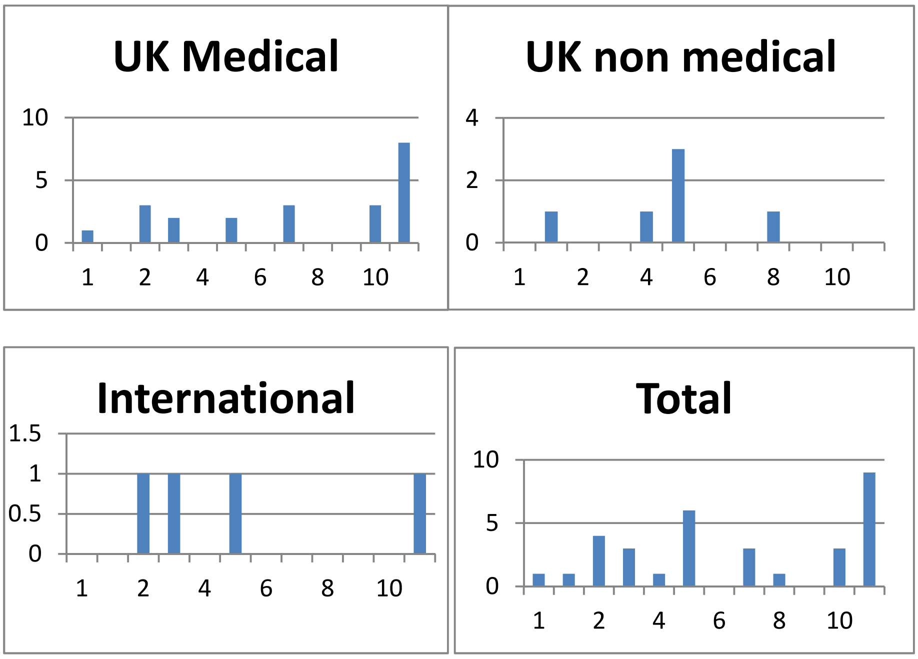
ii. The testing method (e.g. written assessment, practical assessment)?

iii. Consideration of people with 'protected characteristics'?

4.6.1 Periods of validity

The findings from study of the grey literature and from direct contact with national and international professional bodies is presented in table A1. The summary data from this table is presented in histogram form below (Figure 9).

Figure 9



These histograms show periods of validity in years on the x axis, and number of institutions setting such a limit for one of their assessments on the y axis, up to a maximum of 10 years. Unlabelled values beyond this represent no limit. So for 'UK medical' organisations the modal value is to have no limit. While presentation of this data in graphic form may lend an air of certainty to it, it should be borne in mind that periods of validity are used in very complex ways. Most often, periods of validity are tied into individual exam structures and designs, such as requirements to pass one part only of an assessment within a particular period of time, making it very difficult to generalise to 'good practice'.

4.6.2 Attrition of knowledge

There is an extensive literature on skills degradation (see for example Wisher et al, 1999[xcix]; Smith et al, 2008[c]; Kim et al, 2007[ci]). Skills decay depends on a variety of factors, including the nature of the skills (more complex tasks decay faster than simpler tasks; theoretical knowledge may decay more slowly than practical knowledge), the confidence and experience of the individual; the nature of the initial training; feedback provided; re-testing/re-training intervals, and so on. However, this is not entirely the question being asked here. Many of the individuals being tested in PLAB, particularly for Part 1, will still be in practice, and may therefore be improving their performance between re-sits; for all parts, candidates are likely to be engaged in further learning, either as individuals or as groups. And in addition, medical knowledge is not simply an issue of forgetting or retaining information: there is a key question of currency of knowledge, which requires acquisition of new knowledge, and the relinquishment of previously learned material. IMGs who have failed PLAB are therefore a unique group, with all three processes proceeding simultaneously.

We are unable to make a recommendation as to the appropriate period of validity of PLAB (or indeed other professional medical examination), due to a lack of significant evidence. It seems plausible to assume that there are issues of currency, attrition and learning proceeding in various individuals and even in the same individual at the same time. See also Section 5.1 for discussion of currency of the question bank.

As a result, for the purposes of PLAB, there is a need for a particular study of IMGs, with regard to the following questions, before this question can be answered.

- A quantitative study of the impact of the interval between re-sits has on PLAB scores in Parts 1 and 2
- A qualitative study of why candidates re-sit or fail to re-sit, and the strategies used by failing IMGs to prepare for their next attempt.

Recommendation: The GMC should analyse data on the number of previous attempts at PLAB, and the time interval between candidates' attempts, in order to identify attrition or accumulation of knowledge in candidates, and to determine the appropriate period of validity of the tests.

Recommendation: The GMC should collect data on current level of grade or post when taking PLAB to enable better assessment of whether candidates are in fact well-matched to the Foundation Year 1 level of the exam.

Recommendation: We recommend that the GMC conduct qualitative research to improve its understanding of why some candidates re-sit PLAB so many times, how they interpret their failures, and what their subsequent strategies are.

5. Theme 3 Findings

The review should explore the following:

a. Any evidence on best practice with regard to:

i. Question banks in written examinations and practical assessments undertaken by regulatory and examining bodies. Is there evidence of a correlation between the size of question banks and the robustness of an examination/assessment of knowledge and skills? This should include steps that can be taken to reduce question fatigue.

5.1 Question Banks

This is not a simple question[cii].

Firstly, it is not straightforward to define 'size'. This is more than merely the number of items in the test. It is also necessary to consider the number of blueprint categories, which may have to be considered as separate banks. There may also be overlap between items, where slightly different versions of the same question are present – for instance, the origin and insertion of a particular muscle might be numbered separately, but are essentially the same question.

It is not straightforward to define 'question fatigue'. Are faculty concerns about question re-use the relevant factor, or anecdotes from candidates, including organised subversion of the process by circulation of purported exam materials (see below)? Is it the increase in scores observed on re-sit? But as described, this contains error, construct irrelevant variation and construct relevant variation; see for instance Section 4.2.4). Calculations about likelihood of re-use may be carried out for any given bank, when the test size, bank size, and bank refreshment rate are known.

Finally, it is not even easy to define 'use'. With Computer Adaptive testing, mid-range questions with good discrimination are re-used more often than those at either end of the continuum.

Evidently, however, the size of the question bank will impact on the construct irrelevant aspect of re-sit frequency. If, for instance, 50% of questions are randomly re-used in each exam, then a candidate who sits an exam three times will already have seen perhaps 87% of the questions in previous tests.

In the Progress Tests administered by Maastricht Medical School (Schuwirth, L, personal communication) there is a bank of over 12,000 questions with approximately 10% refreshment each year, and with new questions being run on a pilot basis before full test incorporation, to allow their psychometric properties to be assessed. Since the typical Progress Test contains no more than 200 questions, this allows candidates to take away the test with them when they leave. This contributes to the formative aspects of the test, without threatening the summative power, and eliminates any attempts to subvert the process by candidates memorising items and sharing them with or selling them to other candidates to testing support companies.

There are a variety of resources available to candidates that purport to assist them in passing PLAB, for example a BMJ learning site [ciii] which lists at least 2390 EMQs and 680 SBAs. This site lists books, and other websites are easily found[civ]. These sites charge typically £25 for one month access or £30 for two month access. A free (though outdated) site[cv] does not charge, but explicitly indicates that questions are from previous tests, and invites contributions to be submitted to them. These must be viewed as quite conscious and overt attempts, not to subvert the process, but to provide an advantage to potential candidates in a way which might well be viewed as construct irrelevant (for example, how questions are formulated, length of test, etc).

There is also an issue of currency of the bank. Medical knowledge and practice grow organically, and change with evidence: For the GMC PLAB assessments, a question selection meeting is held for each Part 1 exam, attended by Part 1 Panel members across a range of specialties, and PLAB staff. The panel checks every item is accurate and up to date. The panel also reviews items after the exam, guided by a psychometrician, where items have performed unexpectedly, have low discrimination or extremes of facility. Part 2 stations are reviewed after each exam by the Part 2 Panel in light of the performance statistics. In addition, examiners are invited to provide feedback if they identify a problem with a station, including whether it is up to date.

The refreshment rate of the bank may shed some light on the appropriate period of validity of parts 1 and 2, and possibly calculations can usefully be made from this information.

In summary, the evidence of size of a question bank and the robustness of the examination to avoid question fatigue are complex. Issues to consider are unique questions compared with rewording of a question that demands the same or similar knowledge. Question fatigue will become more relevant as the number of re-sits increase, allowing candidates more chance of having seen the question previously.

5.2 Standard Setting

ii. Standard-setting. The pass mark for each Part 1 examination of PLAB is set using the Angoff method. The pass mark for each station and the examination in Part 2 (of PLAB) is set using the borderline group scoring method. Is there any evidence on best practice when determining the threshold for success in a written examination, or practical assessment, of knowledge and skills?

5.2.1 General Considerations

It is generally proposed that there is no absolute standard by which standards can be set. All standard setting is a social construct. As Messick (1994)[cvi] states:

> “Because standard-setting inevitably involves human judgment, a central issue is **who** is to make these judgments, that is, whose values are to be embodied in the standards. The cut-off points on the latent continuum do not possess any objective reality outside and independently of our minds. They are mental constructs, which can differ within different persons.
>
> Whether the levels themselves are set at the proper points is a most contentious issue and depends on the defensibility of the procedures used for determining them”.

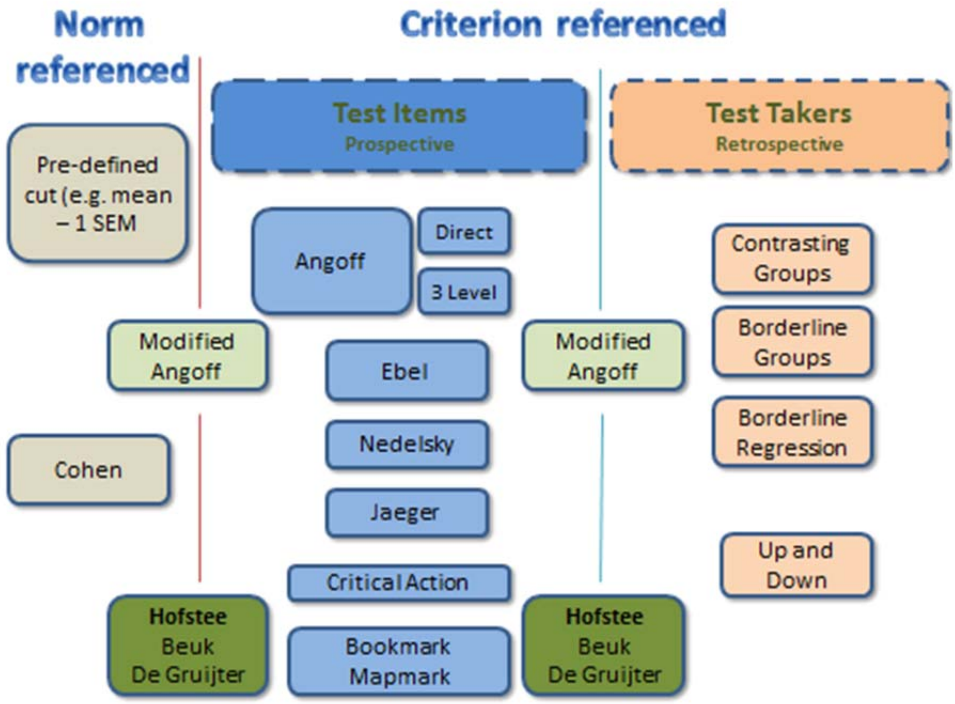
Perhaps as a consequence, it is generally observed that no one method of standard setting is superior to another[cvii] (Berk, 1986). Comparative studies identify considerable variation in the outcomes of different standard setting methods. As a consequence, defensibility of a particular standard setting method is often viewed as being an issue of process rather than outcome.

> “When a validation rests in part of the opinion or decisions of expert judges, observers or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The description of procedures should include any training and instruction provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth” (National Council for Measurement in Education, 1999) [cviii].

However, while the arbitrary nature of standard setting is generally accepted, it is possible to challenge this view, at least in principle. If a stably administered test corresponding to the domain of desired knowledge generates an ordered series of outcomes, and if performance in practice can be graded in some way (even if only dichotomised into satisfactory and unsatisfactory performance), then it is in principle possible to compare the test outcome with the practice outcome, and set a standard corresponding more or less to the dividing line. In general, this idea is referred to as ‘prospective validity’ (see Appendix C, the Glossary), but it would be even more appropriate to refer to it as ‘retrospective validation’. Since there is clear evidence (see 4.3) that there is indeed a relationship between test performance and performance in practice, this is not an impossible notion, although of course the practical difficulties are huge; not least the observation that tests rarely remain stable over the time scales required to explore future practice.

Two major categories of standard setting are generally employed: norm referencing, where outcomes are related to a referenced population, and criterion referencing, based on some absolute standard of knowledge or performance (See Appendix C, the Glossary, for fuller discussions). A taxonomy of standard setting methods is provided in Figure 10. The details of these methods are provided in the Glossary (Appendix).

Figure 10 (McLachlan, from post graduate teaching material)



Norm referencing is usually described as not in widespread use for medical professional assessments. One reason for this is that candidate performance may vary significantly from year to year (not only by chance, but due to systematic changes in educational practice philosophy and opportunity at earlier levels of training), and therefore changes in the standard of candidates passing the assessment may vary correspondingly. That norm referencing alone is not appropriate in these high stakes professional contexts is shown by a study by McManus et al (2005)[cix] which used test equating in the form of marker questions to explore absolute standards in the MRCP (UK) Part 1 examination from 1985 to 2002[cx]. This demonstrated real changes in performance over this period. (While the sign of this change is not important to the general argument, out of interest it suggested a decline in performance). Cohen-Schotanus (1999)[cxi] also identifies significant variations year on year for medical student performance.

However, Richter Lagha et al (2012)[cxii] suggest, perhaps surprisingly, that norm referencing (with a cut off at 1 or 2 SD or SEM below the mean representing a fail) is still common in US medical school OSCEs, due to the difficulty of standard setting.

Furthermore, in a published commentary, Van der Vleuten makes a defence of norm referencing in standard setting in undergraduate assessment, noting its low costs, and the lower degree of variability between students than between assessment diets (Van der Vleuten 2010)[cxiii].

One appropriate use of norm referencing is when there is an expectation that multiple tests will be set repeatedly as in progress testing (McHarg et al, 2005)[cxiv]. Here, there is evidence that consistent performance below a set norm value corresponds to performance below the criterion of competence.

In general, however, criterion referencing is preferred. This is despite the eloquent contempt which Glass (1978)[cxv] poured on the very idea of criterion referencing ("One can justifiably and fruitfully ask" he wrote of such attempts, " 'What manner of discourse are these persons engaging in?' "). Despite these observations, included for the sake of representing all sides of the argument, we are confident that a high stakes professional assessment such as PLAB, where patient safety is the intended outcome, should employ the best evidenced criterion referenced methods available, despite the formidable difficulties involved. This is the generally accepted view in the field.

As shown in Figure 10, criterion referenced methods can be divided into judgements on test items, and judgements on test takers. The former can be made in advance of test administration, and are therefore often described as prospective: the latter are made after test administration, and are retrospective. In addition, there are a number of 'compromise' methods. These can be compromises between norm and criterion referenced methods, or between item and candidate methods, depending on how they are used.

Prospective methods are frequently used with knowledge tests, and retrospective methods for skills tests, though there is no absolute requirement that this be the case, and there are many exceptions. But supporting reasons are that knowledge tests are normally of Crossed Design (i.e. all candidates undertake all items), environmentally stable (each administration is much like every other administration), items are often independent of each other[cxvi] (Clauser and Clymna, 1994) and (where computer marked), assessor variability is less important. All this accords well with prospective methods. Conversely, skills tests may be undertaken at a number of simultaneous circuits, so each candidate receives a slightly different experience, environmentally variable (depending on factors such as location and time of day), may have linking constructs (such as communication skills), and have considerable assessor variability. Here, retrospective methods, where the standard is 'negotiated' through direct observation, offer advantages. Statistical methods are available to test the independence of items in skills tests, but they generally require rather large samples (Yen, 1993)[cxvii].

Angoff methods (see Appendix C, Glossary) are perhaps the most familiar version of prospective methods based on test items. However, reliability and validity may improve if information on the consequences of the Angoff decisions is subsequently provided to the assessors (See 'modified Angoff'), and this introduces a retrospective element.

Three main retrospective methods are employed with skills tests such as OSCEs. These are Borderlines Groups, Contrasting Groups, and Borderline Regression, and they have much in common. All require that a score based on a checklist is calculated from observation, but the assessors also give a grade, based on a global judgement of overall performance. The Glossary (Appendix C) describes how cut scores are calculated from these two kinds of information. Perhaps surprisingly, global judgements can prove to be more reliable than ratings (Regehr et al, 1998)[cxviii].

In comparing these, Contrasting Groups generally gives higher cut scores than Borderline groups, and hence lower pass rates (See appendix E). Where the numbers of candidates rated as Borderline is small, both methods may be harder to implement. Borderline Regression is easier to calculate even when group sizes are small, and appears to have a higher precision than the other two methods, and is therefore particularly suitable in small scale tests, but in a number of instances, gives lower cut scores and higher pass rates than the other methods. This evidently varies with circusmtances, however.

A number of recent studies have studied these methods on a comparative basis, though, sadly, the choices of tests for comparison has been unsystematic.

Kaufman et al (2000)[cxix] compared Angoff, Borderline Groups, and 'relative' methods in standard setting for OSCEs. 'Relative methods' included two normative approaches – a cut score 1.96.SEM below the mean, and one set at 60% of the 95th percentile rank score – the Cohen Method q.v.), and a 'holistic' approach – the University cut score of 60%. The Angoff and Borderline methods were preferred on the basis that they gave similar results to each other, while the other approaches gave highly divergent outcomes in terms of cut scores and pass rates. The Angoff variant involving discussion amongst judges after initial setting was preferred in this paper. The Borderline Groups method was noted as being simpler and less expensive that the Angoff method.

Downing et al (2003)[cxx] compared Nedelsky, Hofstee, 'Direct Borderline', and Ebel methods for written examinations. Here, 'Direct Borderline' is used to describe an Angoff approach, in which judges determine whether a borderline student would pass this question on a 'yes/no' basis. The authors point out that this described in the text of Angoff's original work on standard setting[cxxi] – the currently used method of estimating proportions of borderline students passing a question is described in a footnote) The Hofstee and Ebel approaches had the lowest cut scores. Nedelsky and 'Direct Borderline' methods gave similar outcomes, so the authors concluded that the latter is workable as an approach to standard setting.

Cusimano and Rothman (2003)[cxxii] compared Angoff, Ebel and Hofstee methods in standard setting for OSCEs. The Hofstee approach gave 'more realistic' cut off scores and better reliability indices than the Angoff or Ebel approaches, and these authors therefore recommended its employment. However, 'more realistic' in this context meant 'more similar to previous outcomes' rather than of demonstrated validity. The pass rates were lower with Hofstee as opposed to the other two methods. Reliability was the same for both Angoff approaches.

Kramer et al (2003)[cxxiii] compared Modified Angoff methods with Borderline Regression with regard to OSCEs: Borderline Regression gave higher reliability as assessed by generalizability theory but also higher pass rates than either of the Angoff methods used. Of the latter, the method incorporating feedback on performance gave a lower cut score and higher pass rate. Oddly, in this study, trainees outperformed GPs.

Wood et al (2006)[cxxiv] compared Borderline Regression with a Borderline Group approach and preferred the former on the basis that it gave smaller confidence intervals. They describe this observation as indicating that Borderline regression is "more accurate" than the Borderline group approach. However 'accuracy' carries an implication of validity, which was not explored in their study. Rather, they were looking at consistency. They did note that the overall cut score was lower and the overall pass rate was higher with the Borderline Regression method. In this particular study, the number of 'Borderline' candidates was as few as 12 at one particular station.

Downing et al (2006)[cxxv] compared five different standard setting methods (Angoff, Ebel, Hofstee, Borderline Group, and Contrasting Groups) for OSCEs. Noting that the different methods produced different cut scores, they commented that there is no "gold standard" in standard setting, and the best that can be obtained are defensible methods and outcomes. Different standard-setting methods produce different passing scores. The key to defensible standards lies in the choice of credible judges and in the use of a systematic approach to collecting their judgments. Ultimately, they suggest, all standards are policy decisions.

Boursicot et al (2007)[cxxvi] demonstrated significant differences between the outcomes of OSCE standard setting methods (Borderline Group, Borderline Regression, and Angoff) at three UK medical schools. However, these differences refer more to the schools involved than to the methods, suggesting that it is difficult to compare different methods across different environments, due to the different interpretations that may be placed on the Global Judgement categories by examiners of different backgrounds and experience.

Schoonheim-Klein et al (2009)[cxxvii] compared Angoff and Borderline Regression in a dental OSCE, and further explored various compensatory models. Borderline Regression was preferred, on the basis of higher reliability, but it is worth noting that it once again gave a lower cut score and a much higher pass rate, raising issues about validity in addition to reliability performance. The authors explore this possibility through calculation of a weighted Loss Function, using calculations of sensitivity and specificity based on staff estimates. The weighting represents the relative costs of False Positive and False Negative outcomes. The authors conclude that this Loss Function is at a minimum with a combination of Borderline Regression and Partial Compensation. However, this does not seem in accord with their own Table 6, in which the loss function is smaller for both Angoff methods used whenever the weighting reaches 15. Moreover, in the real world, the cost of a false negative is mitigated if there is a possibility to re-sit: the cost of a false positive is not. Since the pass rate is very high with the Borderline Regression approach (passing 93% of incompetent students and 98% of competent students) it is not surprising that false negatives are initially very low.

Jalili et al (2011)[cxxviii] compared the usual 'modified Angoff' method with what they describe as a 'three level Angoff', which is the Direct Angoff yes/no method with the addition of a 'don't know' option, with regard to a 14 station OSCE administered to undergraduate students and concluded that the three level Angoff gave lower agreement between judges and wider confidence intervals. The standard modified Angoff procedure also offered much more 'realistic' cut scores in the authors' view – i.e. corresponding to their expectations.

Clauser and Clyman (1994[cxxix]) point out that the original version of Contrasting Groups identified two groups a priori, masters and non-masters, making it prospective with regard to test takers. The difficulty in performing this task gave it low Acceptability. A variant of this approach was described by Livingston and Zieky (1982)[cxxx] to centre on the classification of actual student performance by categories, making it retrospective with regard to test takers, in the now familiar form. They conclude that the Contrasting Groups method is appropriate for setting pass fail standards.

5.2.2 Item Response Theory

Item Response Theory (IRT – see under **Reliability** in the Glossary, Appendix C) is an approach which can be used both in standard setting and in the development of a sophisticated understanding of the performance of both items and candidates. Analysis of individual items by these means can lead to opportunities for test equating (see 5.2.3), and for further approaches to standard setting (Bakhta et al 2005 [cxxxi]; Bond 2003[cxxxii]; Bond and Fox, 2007[cxxxiii]).

5.2.3 Test Equating

It is possible to test-equate assessments, so that the standard remains constant, as is done in RCP (UK) assessments. Nungester et al (1991)[cxxxiv] describe the NBME approach in 1991, where test equating was used with an anchor test which had been standard set as the key. Standard setting for the anchor test appears to have been by Angoff methods.

However, as the Nungester et al (1991) paper indicates, test equating does not solve the problem of setting the cut score on the equated exam. If historical data are used, this may be viewed as a kind of norm referencing, even if there is a rationale behind selecting the cut score. De Champlain argues against this approach, on just these grounds (De Champlain, 2010)[cxxxv].

5.2.4 Simulated patient rating

It is possible to imagine including patient or simulated patient scores in the OSCE assessment process, in addition to the expert rater, and there is some evidence that this can bring about small but real improvements to the psychometric properties of an assessment (see discussion in Homer and Pell, 2009)[cxxxvi]. In the particular circumstances of the PLAB test, this approach would need to be fully tested (using generalisability theory), in order to determine the magnitude of rater-candidate interactions. We comment elsewhere on a possible role for lay observers of the process. Simulated patient rating process is also employed in the certification of IMGs in the Netherlands (Sonderen et al, 2009[x]).

5.2.5 Summary

It is clear that there is no single universally optimum method of standard setting (see also Woehr, Arthur and Fehrmann, 1991[cxxxvii]; Norcini and Shea , 1991[cxxxviii]; De Champlain, 2004[cxxxix]). The current methods used in PLAB are well studied in the literature, show acceptable properties, and there is no consistently better method under all circumstances. A valuable additional source of information is the use of Item Response Theory which can lead to a deeper understanding of the performance of items and candidates and opens the way to further opportunities in standard setting in the future. A significant hazard is the risk of false positives. Addition of 1 Standard Error of Measurement, as is currently practiced in Part 2 of PLAB, and hence increasing the cut score, significantly reduces the likelihood of false positives, at the cost of a slightly raised number of false negatives. However, the possibility of resitting the test helps alleviate the consequences of being a false negative. Although we know of no calculation that determines 1 SEM as the optimum number for this purpose, it represents a reasonable compromise between the competing factors.

Recommendation: Pending strategic review, the GMC should retain Angoff and Borderline Groups standard setting methods, since both are well recognised and supported by evidence, and the *addition of 1 SEM to the cut score for Part 2, as this helps reduce false positives in the crucial Skills element.*

Recommendation: We recommend the collection of IRT data for GMC PLAB Part 1 data in particular, to expand the range of options available to the GMC subsequently.

5.3 Marking and confidence intervals

iii. Marking / scoring systems (with an emphasis on the statistical tools, techniques or models used to set thresholds - and the acceptable confidence intervals around these - and validate results).

5.3.1 Validity and Reliability

Measuring 'reliability '*per se* is a complex issue. See the Glossary, Appendix C, for a further discussion of various approaches.

It will be noted that frequently, 'reliability' figures quoted refer to Cronbach's co-efficient Alpha. While this is a standard practice, it presents several difficulties. One is that Alpha is a measure of consistency, and therefore assumes that only one construct is under study. While this is an interesting debate in its own right as far as some kind of 'generalised medical intelligence' is concerned, it is certainly possible to hypothesise an assessment which covers more than one construct, and therefore displays a low alpha, although each of the constructs might have a higher alpha if measured separately. The second issue is that, as Tighe et al (2010)[cxl] cogently point out, reliability is a property both of the assessment and the candidates, not just the assessment alone. Reliability as conventionally measured increases as the range of scores (as a surrogate for the range of abilities of candidates) increases. This is because, where a wide range of scores is observed, the variance in candidate ability becomes larger in comparison to the error variance[cxli] (McManus et al, 2003). Tighe et al (2010)[cxlii] argue cogently for the use of the Standard Error of Measurement as a more stable measure of reliability, particularly where the ability range and number of candidates are both small. Indeed Cronbach (2004)[cxliii] himself argued for the use of the Standard Error of Measurement and arising confidence intervals as a measure of precision around scores as a better and more accurate statistical interpretation for reliability.

Equally, since Cronbach's alpha is derived from Classical Measurement Theory (qv), it treats all sources of error together, where Generalisability Theory allows consideration of a range of sources of possible error.

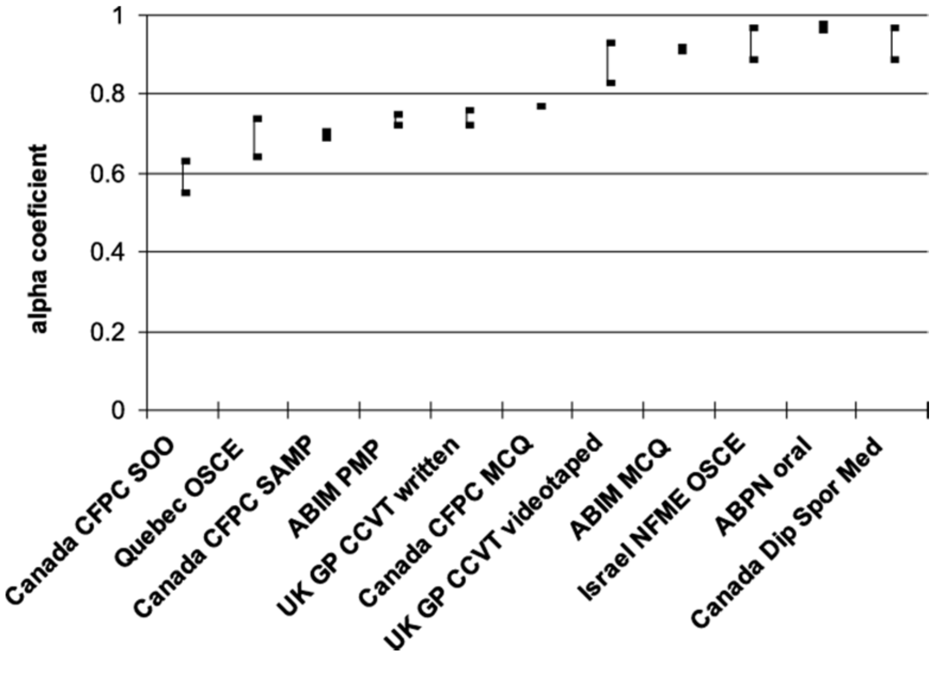
Considerable confusion exists around the relationship between validity and reliability. It is often said, for instance, that an assessment cannot be valid if it is not reliable. The reality is more nuanced. It is possible, even desirable to initially consider validity and reliability as formally separate constructs. Let us hypothesise an assessment for which we have a gold standard outcome in the form of good predictive validity data, and a relationship between the predictor assessment variable and the outcome variable which is at least ordinal and at best linear or curvilinear, *and* where the outcome variable can be dichotomised into *adequate* and *inadequate* performance. And since we are hypothesising, we can initially assume that the reliability of the predictor variable is 1. Now, let us consider what happens as the reliability decreases. What changes is our ability to determine what the value of the predictive validity is, not the existence of the validity as a construct . But this is a highly idealised case. We could say that validity and reliability are formally separate but functionally intertwined.

Frequently, however, we are interested in the *inferences* that can be drawn from the assessment outcomes, and here we can say that the less reliable a test is, the less we can conclude about the validity. So it is not that an unreliable test is not *valid*, it is that we cannot say whether an unreliable test is valid or not. An analogy can be drawn between calculations of the correlation coefficient between two variables, and the error bars on each of them. The size of the error sets an upper bound on the value of the correlation coefficient that can be calculated, but it does not tell you that there is no correlation.

Confidence intervals are derived from estimates of the reliability of the assessment (for instance, the Standard Error of Measurement is a function of the reliability), and therefore the same arguments apply.

In a systematic review, Hutchinson et al (2002)[cxliv]explored the reliability of a variety of post graduate certification examinations. They identified 55 relevant papers, and found that a variety of methods were used to determine reliability, including inter-rater reliability (17 papers), internal consistency (12 papers) and examination stability (11 papers). Their Figure 2 is shown below as our Figure 11 for the Alpha statistic, and shows just how variable the outcomes can be.

Figure 11 (from Hutchinson et al, 2002)



Interestingly, standard setting is not mentioned in this literature review.

We note that the GMC currently reports SEM routinely for Part 1 of PLAB and we commend this practice.

5.3.2 How many stations in an OSCE?

> iv. The optimum number of stations that should be included in OSCEs and other comparable practical assessments.

In a systematic review of the reliability of OSCEs, Brannick et al (2011) considered both alpha values

(a measure of the internal consistency of performance of an assessment) and generalizability from a large number of studies. When meta-analytic methods were used, mean alpha values across stations was 0.66 (95% CI 0.62 to 0.70)[cxlv]. When plotted against the number of stations (their Fig 2, our Figure 12) a modest generic increase was observed with number of stations (and is the nature of the mathematical formula for alpha that this should generally be the case). Generalisability also increased with the number of stations. However, in this case, the differences between studies are the truly significant finding: as the authors point out "some studies report a reliability of >0.80 with < 10 stations, and others report a reliability of < 0.80 with > 20 stations. This is explained by saying that a 'good' OSCE with well designed stations and well trained examiners is likely to be reliable, while the converse is true. The correct answer to a question on the optimum number of stations in PLAB Part 2 can only be derived from the GMC data for the GMC OSCE, through a Decision (D) study following from a Generalisability (G) study of GMC OSCEs over the years. The choice of the acceptable level of reliability is again a matter of arbitrary choice, but values below 0.7 are generally considered inadequate. An example of such a D study by a Royal College is shown in Figure 13.

Figure 12 (from Brannick et al 2012)

This Figure shows that reliability (measured either by Alpha or by generalizability) increases with the number of stations, but there is huge variability in the values obtained.

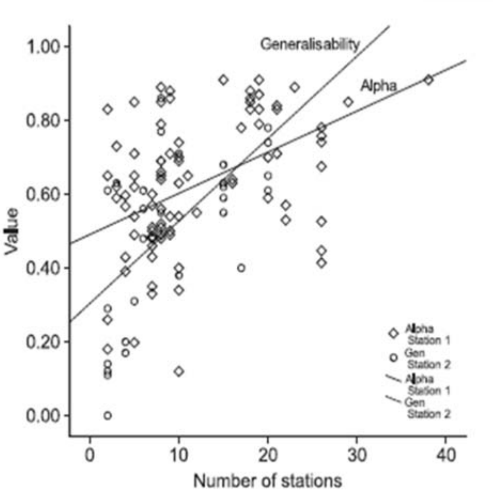
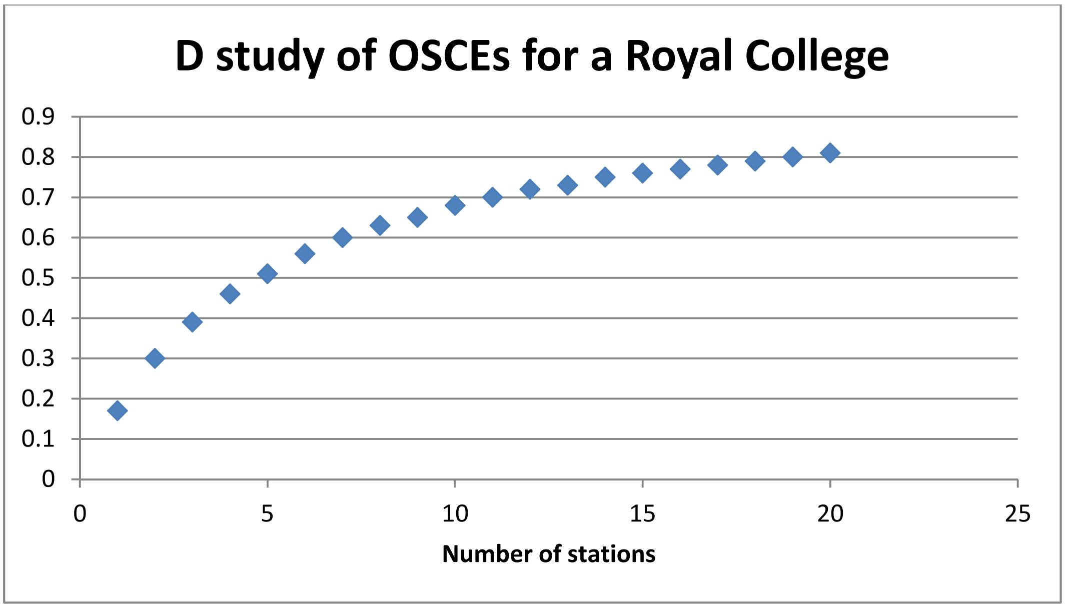


Figure 2 Scatterplot of alpha and generalisability coefficients by number of stations with unweighted regression lines

Figure 13 (data provided by one of the Royal Colleges)



However, station number is not the only way of looking at this issue. If taking each station independently is described as a 'vertical' approach, then it is also possible to imagine a 'horizontal' approach, as observed in the PACES OSCE for the Royal College of Physicians[cxlvi]. Here, 'themes' are integrated across stations (as is also the case with the RCGP CSA exam).

Recommendation: that the GMC conduct a Generalisability Theory Analysis of PLAB Part 2 on a routine basis, and subsequently a Decision Study, to calculate the number of stations appropriate for this particular assessment.

5.4 Compensation Models

v.The optimum number of stations that candidates should pass in OSCE's or other comparable practical assessments

There are a variety of approaches that can be taken to compensation in one aspect of performance by performance in another. Currently, the GMC employs a partial compensation model based on the number of stations passed. Failure in four stations results in an overall fail, even if the aggregated score exceeds the cut score. We will call this rather common conjunctive approach a 'profile criterion', and while it seems intuitively sensible, we know of no underpinning evidence to support it in detail. A different approach is to consider 'horizontal' domains of knowledge rather than 'vertical' stations. For example, "Communication Skills" might be viewed as a key attribute, and is likely to be present as a component in a number of stations. Instead of assessing each station separately, it is possible to aggregate the scores for Communication Skills across all the stations in which it is present, and determine the cut score for this attribute. The candidates are then required to 'pass each attribute separately. The PACES component of the RCP (UK) membership tests operates on this basis. However, a 'critical action' approach (see Glossary) based on this principle found low reliability of this approach (Richter Lagha 2012)[cxlvii]. As evidence of performance of this approach in PACES unfolds, it would be valuable to keep it under review as a possibility for PLAB Part 2.

Reece et al (2008)[cxlviii] used confirmatory factor analysis to compare a three domain model ('examination skills', 'communication skills' and 'history taking skills') with a one domain full compensation model. The 3 factor model offered significantly better fit to the data. Interestingly, 23% of students failed if no compensation was allowed between domains, while only 4% of students failed in the 'full compensation' model, suggesting the latter might result in higher numbers of false positives. Their data suggests that stations should be re-designed to optimise performance on a domain based model. Schoonheim-Klein et al (2009)[cxlix] also present data to show that partial compensation models outperform full compensation models as measured by the weighted Loss Function (which integrates Sensitivity and Specificity). Against this approach, it is a general finding that individual OSCE stations have a high degree of case specificity, even for common domains. Case specificity (in this particular context) is the phenomenon that good performance in, say, one OSCE station is frequently not matched with good performance in all OSCE stations. area (e.g. Communication Skills, as measured in one OSCE station) may not correlate with Similarly, a study of a 'critical action' or 'critical approach' standard setting method (q.v.) which takes a similar domain based approach (but obtains the domains through a Delphi style consultation between clinicians)[cl] (Richter Lagha et al 2012), found it performed even more poorly than norm referencing in terms of reliability, with the critical action approach offering G of 0.20 over six stations.

It might be argued by those unfamiliar with assessment good practice, that all stations should be passed or even that the cut score should be 100%. This is not the case. In order to set such high standards every station would be required to have perfect validity, and perfect reliability, and this is an impossible requirement. Brian Jolly (1999)[cli] notes that on one occasion where he was involved, an expert speciality group 'determined that a passing performance on the test as a whole should entail passing every station. In the pilot, all the examiners took all the stations. Not one passed'.

5.4.1. Conclusions

A full compensation model (i.e. without a conjunctive approach) is undesirable. The GMC's current policy restricts full compensation in a way that is desirable, but we know of no evidence on how many stations should be required to be passed. This might be susceptible to retrospective analysis from the GMC's own data.

5.5 Exam Conditions

vi.PLAB Part 1 candidates undertake the written examination (in the UK and overseas) on paper scripts in an examination hall under test conditions. Is there evidence that this is not best practice and that other methods of conducting written examinations are better practice, for example by using electronic examination tools?

As Steven Downing stated: “Written or computer based assessment is the most appropriate modality to test cognitive knowledge” and “Cognitive knowledge is best assessed using written test forms” (Downing , 2002))[clii].

It is possible to cheat under written examination conditions in an examination space, even for supposedly ethical post graduate doctors (McManus et al, 2005) [cliii] , and this may be an argument against their use. Again, the GMC may have extractable data on the influence of location on candidate scores which would be well worth exploring. However, it is not clear that alternative assessment methods would offer greater examination security.

5.6 Other Approaches That Might Be Tried

5.6.1 Situational Judgement Tests

Situational Judgement Tests (SJTs) are a measurement method designed to assess judgement in work-relevant situations. Typically, they present challenging situations likely to be encountered at work, and require candidates to make judgements about possible (pre-defined) responses. This may be in the form of ranking the possible responses, or of selecting the best responses from those available (e.g. ‘best three responses of a choice of 5’). These are then scored against a pre-determined key. In general the situations focus on non-academic/professional attributes such as integrity, empathy, resilience, or team involvement. In an upcoming systematic review (personal communication Fiona Patterson; in press in Medical Education[cliv]), the authors conclude that “SJTs are a cost-efficient methodology compared with high-fidelity assessments of non-academic attributes, such as those used in objective structured clinical examinations. In general, SJTs are found to demonstrate less adverse impact than IQ tests and are positively received by candidates”.

In the particular circumstances of the PLAB test, it is known that IMGs face particular challenges around cultural integration, rather than skills. It therefore seems plausible that use of appropriately written SJTs, focussed on such issues as confidentiality, asking for help and team interactions with Allied Health Professionals, would be able to explore such areas in advance of entering practice. SJTs of this kind could be delivered either within an extended Part 1 of PLAB, or as a stand-alone test.

Koczwara et al (2012)[clv] show that Situational Judgement Tests measure procedural knowledge as well as declarative knowledge, and are the single best predictors of performance on extended tests at a Selection Centre (which are in turn good predictors of performance in clinical practice. These authors also suggest that a test of verbal, numerical and diagrammatic reasoning (the Swift Analysis Aptitude Test) has predictive validity with regard to Selection Centre decisions, and corresponded to a clinical problem solving test approach.

Recommendation: We recommend that the GMC consider exploring the use of Situational Judgements tests as part of the PLAB process.

5.6.2. Computer Assisted and Adaptive Testing

(See Glossary, Appendix C, for these terms)

Computer Assisted testing enables a wider range of assessment approaches to be used, compared to written forms. For instance, video or audio clips can be used. If Item Response Theory (q.v.) analysis has been employed, then tests can be equated, and if a standard has been appropriately set, candidates can be given immediate feedback on their performance. Computer assisted testing also enables a wide range of analytic data (such as Differential Item Functioning, q.v.) to be gathered automatically. Against this, its use requires the availability of a considerable number of computer stations in a wide variety of locations, and this is not always feasible.

Computer Adaptive Testing (CAT) allows a reduction in the total number of questions required for a given level of reliability, and has been implemented by the Australian Medical Council for IMGs. However, it can be argued that the great strength of Adaptive testing is that it rapidly segregates candidates into a number of different ability bands. But in testing for licensure, only two bands are required: competent and not competent. Use of CAT in this context is likely to draw very heavily on borderline difficulty, high discrimination items, posing challenges to the question bank.

For these reasons, we do not wish to make an unequivocal recommendation in favour either of Computer Assisted or Adaptive Testing: rather, the GMC should keep this possibility under review, and observe the outcomes of the Australian experience.

Recommendation: The GMC should keep under review the possibility of the use of Computer Assisted Testing, Computer Adaptive Testing and Test Equating for PLAB Part 1, particularly in the light of the Australian Medical Council's approach.

5.7 Exam Preparation

b. Is there any evidence on best practice in terms of preparing candidates for either written examinations or practical assessments? For example, publishing question banks and other information to assist candidate preparation.

Test anxiety is a significant factor in many testing environments, and is generally viewed as construct irrelevant (Ergene, 2003)[clvi]. It may have differential effects by gender and social status (Zeidner, 1990)[clvii]. It may well be alleviated by skills deficit training (such as test taking strategies and study skills) and by interventions, such as relaxation techniques) aimed at reducing stress in this context (Dendate & Diener, 1986)[clviii]. However, test taking strategies etc are frequently promoted as part of commercial packages aimed at professional exam sitters (see section on test support). It has been suggested (Van Iddeking et al, 2011)[clix] that re-sitting might reduce test anxiety, and therefore provide a construct relevant value. On the other hand, Matton et al (2009)[clx] argue that test anxiety is, essentially by definition, construct irrelevant.

Some professional bodies (such as the United Kingdom Clinical Aptitude Test, UKCAT) make official practice tests available on line, so that candidates may become familiar with test format etc. However, this depends on the availability of a sufficient number of test items in the bank to avoid test familiarity (See 5.1), and given the lack of evidence on whether or not this is helpful in the PLAB context, we cannot make a recommendation on this point.

5.8 Feedback

> c. Candidates are informed if they have passed or failed the PLAB test: they are not given a breakdown of their marks and there is no appeal against the outcome. Is there any evidence on best practice in terms of providing a right of appeal for unsuccessful candidates in written examinations and practical assessments? This should include specific procedures for individuals with protected characteristics.

"Organisational Justice" relates to the relationship between an employee or student (or applicant) and an organisation, and how their judgements on justice affect their subsequent feelings, attitudes and behaviours on organisational justice (Greenberg, 1987)[clxi]. It is often subdivided into distributive, procedural, and interactional justice.

"Distributive Justice" obtains when the outcomes of decision and distribution of resources are perceived as fair and equally applied (Adams, 1965)[clxii]. "Procedural Justice" relates to the fairness of processes which lead to outcomes. A procedurally just process is one in which valid and reliable methods are used with lack of bias (Leventhal, 1980)[clxiii]". "Interactional Justice" relates to the consequences of decisions, and benefits from the provision of explanations for decisions, and sensitive delivery of the decision itself. See also Bies and Moag (1986)[clxiv] and Greenberg (1990)[clxv].

These views of justice generate strong emotional responses in those who feel aggrieved. From an interactional justice point of view, there is much to be said for providing more detailed feedback to failing candidates.

Although we know of no direct evidence relating to the impact of organisational justice issues on performance or likelihood of legal challenge, in the context of employment Terpstra et al (2000)[clxvi] studied litigation arising from a variety of selection methods. They found that unstructured interviews were the most likely to lead to subsequent litigation, followed by tests of cognitive ability and tests of physical ability. Conversely, structured interviews, work samples, assessment centres and personality tests were significantly under-represented in litigation as compared to their frequency of use. The authors also considered (on much smaller samples) the outcome of litigation, and found that the litigant was successful most often in the case of unstructured interviews, followed by physical ability tests, cognitive ability tests and work samples. Structured interviews and assessment centres survived all of the challenges that had been mounted against them (admittedly a small number). Similar conclusions were drawn by Posthuma et al (2002[clxvii]), who concluded that structured interviews were defensible in law. This suggests that courts at least prefer robust and objective measures to subjective ones.

Recommendation: We recommend giving a more detailed breakdown of performance to PLAB OSCE examinees to enable them to improve. The model for this is the Royal College of General Practitioners (RCGP) Clinical Skills Assessment (CSA) feedback (as described in Appendix A).

Recommendation: We also recommend the GMC publish on its website the further data which this Report suggests is gathered, to demonstrate a culture of transparency to patients and to PLAB examinees, and to improve examinees understanding of, and expectations of PLAB.

5.9 Exam conditions

> d. Is there any evidence on best practice in terms of the design and conduct of examinations/assessment for candidates with learning difficulties (including dyslexia).

Guidance from the UK National Dyslexia Association (personal communication) is as follows.

MCQs may pose particular challenges to dyslexics by their very nature. Dyslexics may also find difficulty in tracking from one sheet to another, and in retaining information from one page to another. Boxes for selection may also be too small. Exam papers are best printed on cream vellum, using a sans serif font such as Arial. Clarity of syntax is particularly important. Coloured overlays may benefit some dyslexics, and the option of bringing their own overlays should be offered to such candidates.

6. Discussion, and Further Recommendations Emerging from the Project

6.1 Discussion

Despite the unusual nature of PLAB, the GMC faces a familiar task for any examination body, which is how to uphold standards for a group (here, the public) while behaving equitably towards individuals (here, PLAB examinees). Recognising these issues might focus the necessary process of continual review of PLAB in two key areas (1) securing the safety of the community as patients, and (2) the nature of the GMC’s duties towards PLAB examinees for whom it has a role in controlling access to UK postgraduate medical education.

However, in this balance, the preponderance must lie towards patient safety. In the Introduction, we raised the question of whether or not PLAB was fit for purpose, and noted that international medical graduates who have passed PLAB still appear to be over represented in disciplinary proceedings and NCAS reports. As we indicated, this does not necessarily indicate that PLAB or the IMGs are at fault here: it could be that current NHS structures fail to support IMGs in an adequate manner. However, we believe that it is essential that the relationship between performance in PLAB tests and later clinical performance be explored. To this end we have recommended that the GMC commission research into the relationship between PLAB test performance and later clinical practice (see 4.3.2), but we also wish to recommend that the GMC explore the reasons why IMGs are over-represented in NCAS reports and GMC data. Formally:

Recommendation: We recommend the GMC examines through further research why some international medical graduates perform poorly, as seen in the National Clinical Assessment Service (NCAS) and GMC data, despite having passed PLAB.

One approach to dealing with this issue is to award interim registration of some kind as is the case in Australia, where the AMC is piloting Work-Place Based Assessment in four states (See Appendix A, Table A1). This seems to us to deserve consideration. The kinds of Work-Placed Based Assessment that might readily be employed ion the UK include the tools of Multi-Source patient and Colleague feedback, Mini-clinical examination, Direct Observation of Practical Procedures and Case Based Discussion currently or recently employed in the Foundation Assessment programme. These have a the merit of familiarity in the UK context and currently existing usage for relevant comparator groups. A strategic view of assessment in medical work is provided by Norcini (2003)[clxviii].

Recommendation: 'Interim' Registration followed by Workplace Based Assessment (which could include patient feedback) for a defined period as part of PLAB assessment. There could be a process of linking this with PLAB Part 1 and 2 results in a portfolio for overall assessment before grant of Full Registration.

In similar vein, some countries (such as the Netherlands – see Sonderen et al, 2009[x]) have targeted further training which may depend on the outcome of the performance tests, and this approach seems to have merits too.

Recommendation: We recommend consideration of a model for targetting bespoke further training as in the College of Emergency Medicine (CEM) Membership Examination (MCE) and the Netherlands.

We have only been dealing with two of the three components of PLAB. The third is use of the IELTS score, under separate study. However, in the light of evidence that 'English as a first language' is significantly associated with lower scores in ECFMG, then we believe that there is a significant task in integrating these findings with those in our Report, and exploring the relationship between IELTS scores and performance on PLAB Parts 1 and 2.

Recommendation: We recommend the GMC statistically analyse the relationship between score performance in PLAB Part 1 and 2 and IELTS score performance.

In the following Decision Matrix (Table 8), we summarise some of the possible approaches we have discussed above, from the perspectives of patients and candidates.

Table 8

Proposal	Rationale and models	Strengths	Weaknesses
1. Housekeeping Recommendations			
1) Pending a strategic review retain Angoff and Borderline Groups standard setting methods	Both methods are well recognised and supported by evidence, and the addition of 1 SEM helps reduce false positives. This is the current PLAB model	**For patient safety:** As a short-term measure this is an evidence-based approach and its retention would allow cohort study data to be gathered. **For PLAB examinee** Certainty	**For patient safety:** Other methods may be more discriminating (fewer false positives) **For PLAB examinee:** Other methods may be more precise (smaller borderline group)

2) Calculate Differential Item Functioning	To explore why any individual test items perform poorly	**For patient safety:** Improved evidence can inform design of PLAB in the future. Poorly-formulated questions could be removed **For PLAB examinee:** Questions showing evidence of bias could be improved or eliminated.	**For patient safety:** Time needed to achieve results F**or PLAB examinee:** None
3) Conduct an analysis of existing PLAB results in the following three areas: 1) conduct an Item Response Theory analysis of PLAB results 2) analyse and report the Standard Error of Measurement (SEM) routinely as a measure of reliability for PLAB	1) to allow (a) Test Equating or (b) a Computer Adaptive Testing approach if desired or (c) test optimisation near the cut score 1) the SEM, in addition to Cronbach's Alpha, would enable the GMC to demonstrate its commitment to robust	**For patient safety:** This would also allow the test to be optimised near the cut score (De Champlain, 2010). **For PLAB examinee:** Calculating the SEM would improve information on reliability	**For patient safety:** Time needed to achieve results of analysis **For PLAB examinee:** None

3) conduct a Generalisability analysis of PLAB Part 2 on a routine basis, and a Decision Study	testing 3) to calculate the number of stations appropriate for PLAB, to demonstrate an evidence-based approach.	A Generalisability analysis could improve use of resources	
4) Collate and analyse further data relating to PLAB candidates in the following four areas: 1) Data relating to the time interval between candidates' attempts 2) Demographic data (which should include a section for voluntary completion of information on all protected characteristics) on candidates 3) data on the number of attempts at PLAB, and the time period since the last	1) to identify attrition or accumulation of knowledge in candidates 2) to focus the GMC's approach to equality and diversity for PLAB candidates 3) to enable ongoing analysis of the link between candidate performance and PLAB	**For patient safety:** Improved evidence on the impact of time-delay, demographics, number of attempts and level of experience can inform design of PLAB in the future **For PLAB examinee:** Improved evidence on the impact of time-delay, demographics, number of attempts and level of experience can inform examinees in the future	**For patient safety:** Time needed to achieve results of analysis **For PLAB examinee:** None

attempt 4) data on the current level of experience when taking PLAB	4) to enable better assessment of whether candidates are in fact well-matched to the Foundation Year 1 level of the exam		
5) Match PLAB Part 1 and Part 2 data with IELTS data	To enable an integrated approach to the GMC assessment of PLAB	**For patient safety:** Improved robustness of GMC assessment of language skills **For PLAB examinee:** Greater confidence in language skills	**For patient safety:** None **For PLAB examinee:** Challenge to the relevance of IELTS to medical practice
6) Publish on the GMC website the data gathered resulting from these recommendations	To demonstrate a culture of transparency in keeping with current policy of seeking to rely on an evidence-base Modelled on RCGP (**7**), NRO for GP Training (**4**), UK Foundation Programme Office (**2**)	**For patient safety:** Patients would be fully informed about the reasons for the design of the exam process **For PLAB examinee:** Examinees would have improved understanding and expectations of PLAB and be better able to direct their efforts appropriately	**For patient safety:** None **For PLAB examinee:** None

2. Strategic Recommendations			
1) Consider the following strategic approaches to examinations which could be piloted alongside the GMC existing approach and/or developed in the future:			
Limiting the number of attempts 4, with a compulsory period of further development required before further	National and International models limiting the number of attempts	**For patient safety:** Reducing the risk of false positives **For PLAB examinee:** Avoiding wasting money and time on re-sits for which the candidate is not prepared	**For patient safety:** Potentially limiting the availability of doctors **For PLAB examinee:** Not having the opportunity to sit ad lib: possible delays in entering practice
2) Computer Assisted and/or Adaptive Testing	Computer adaptive testing is modelled on the Australian	**For patient safety:**	**For patient safety:**

and Test Equating	Medical Council Ltd CAT exam (**56**)	Wider range of items (e.g. audio) in test **For PLAB examinee:** Immediate feedback, chance to show knowledge of visually based materials	Computer based testing poses different risks for exam security than paper based tests **For PLAB examinee:** Difficulty in accessing computer based test centres
Situational Judgement Tests as part of PLAB Part 1	Situational Judgement Tests are modelled on the NRO for GP Training (**4**) and UK Foundation Programme Office (**2**)	**For patient safety:** Demonstrated validity in many test settings. Chance to explore scenarios not based solely on cognitive knowledge, but based on the UK health care system **For PLAB examinee:** Chance to show knowledge of understanding of complex scenarios	**For patient safety:** none **For PLAB examinee:** Difficulty in developing a universal ‘job scheme’ against which to derive the Test questions. May add to total exam burden

Consider Workplace Based Assessment (which could include patient feedback) as part of PLAB assessment	There could be a process of linking this with PLAB Part 1 and 2 results in a portfolio for overall assessment before grant of full registration. The model for this is the Australian Medical Council Ltd (**56**) which is currently piloting WBA in 4 states	**For patient safety:** On the job assessment will have taken place over a period of time **For PLAB examinee:** A chance to show their skills in the role a doctor in the UK health care system	**For patient safety:** Assessments may not be correctly completed by clinicians as they are designed to be **For PLAB examinee:** A period of uncertainty after PLAB pass
3. Best Practice Recommendations From Other Professional bodies			
Give further consideration to the following key approaches in particular:			
Giving a more detailed breakdown of performance to PLAB examinees	RCGP (**7**)	**For patient safety:** Examinees can focus their efforts to improve and possibly qualify faster **For PLAB examinee:** 1)Examinees can focus their efforts	**For patient safety:** Feedback should not allow gaming from candidates

		to improve 2)Encouragement for examinees to learn 3)Robust exam design 4) Sense of organisational justice	
Adopting a critical incident or red light rule	Bar	**For patient safety:** Allows critical incidents to receive due attention **For PLAB examinee:** None	**For patient safety:** Low reliability of incident reporting and assessment **For PLAB examinee:** Low reliability of incident reporting and assessment
Building in an exceptional circumstances clause	RCGP Applied Knowledge Test, and CSA exam	**For patient safety:** Allows good doctors with difficult circumstances to have a chance to enter practice in the UK **For PLAB examinee:** Organisational justice	**For patient safety:** Possibility of gaming the system **For PLAB examinee:** None

4. Recommendations for key areas for further research

Examine through further research the following aspects	The evidence-gap in our review has been apparent in many critical areas		
Why some international medical graduates perform so poorly by The National Clinical Assessment Service (NCAS) and GMC data, despite having passed PLAB		**For patient safety:** Identifying organisational changes (e.g. support, induction) which help IMGs practice better). **For PLAB examinee:** Identifying organisational changes (e.g. support, induction) which help IMGs practice better).	**For patient safety:** Delaying registration of good doctors. **For PLAB examinee:** Blame culture, focussed on IMGs
Qualitative research to understand why some candidates re-sit PLAB so many times, how they interpret their failures, and how they respond to them.		**For patient safety:** Might find ways of promoting personal reflection on the part of those failing **For PLAB examinee:** Targeted guidance on how best to	**For patient safety:** None **For PLAB examinee:** None

		focus their studies.	
5. Further Recommendations			
Consider offering clarification of GMC January 2012 Guidance	The guidance is leading a number of Royal Colleges to amend their regulations for all their examinations prematurely in the light of this report	**For patient safety:** Colleges need to design their examinations to fit with their training programmes to ensure robust examination systems **For PLAB examinee:** Important insights from this report could inform future examination design to benefit examinees	**For patient safety:** None **For PLAB examinee:** None
Review the gap as regards registration of international medical graduates in other EEA countries who then come to practice in the UK	Setting up a working group to liaise with other EEA organisations on this issue, and seeking advice on whether the way the UK Bar Standards Board deals with this issue for barristers coming to practice in the UK might be lawfully adopted for medicine are two suggested approaches	**For patient safety:** This is a serious gap for patient safety **For PLAB examinee:**Assessment of all doctors in the UK before independent practice would be fairer to those who do take PLAB and to all doctors in UK practice	**For patient safety:** None **For PLAB examinee:** None

7 References

[i] Cohen J. (1988) *Statistical power analysis for the behavioural sciences*. 2 ed. Lawrence Earlbaum Associates, Hillsdale, NJ.

[ii] RAND Report International Comparison of Ten Medical Regulatory Systems (2009)

http://www.rand.org/pubs/technical_reports/TR691.html Accessed 01/08/12

[iii]Tooke J. (2009) Aspiring to Excellence http://www.mmcinquiry.org.uk/MMC_FINAL_REPORT_REVD_4jan.pdf

[iv] van der Vleuten C. (2005) Assessing Professional Competence. *Medical Education;* 39: 314

[v] McManus IC, Ludka K. (2012) Re-sitting a high-stakes postgraduate medical examination on multiple occasions: nonlinear multilevel modelling of performance in the MRCP(UK) examinations. *BMC Medicine;* 10:60: 1-19

[vi] McManus IC, Ludka K. (2012) op. cit.

[vii] Burke S, Davison E, et al. (2005) The Involvement of lay people in selection to general practice training schemes. *Education for Primary Care;* 16: 450-7

[viii] Sonderen MJ, Denessen E, ten Cate O, Splinter TAW, Postma CT. (2009) The clinical skills assessment for international medical graduates in The Netherlands. *Medical Teacher;* 31:e533-e538 p. e533

[ix] http://www.gmc-uk.org/doctors/plab/part2_examination_regulations_may_2008.asp, accessed 05/09/12.

[x] Schuwirth L, Van der Vleuten C (2010) How to design a useful test: the principles of assessment. In *Understanding Medical Education*, (Ed Swanwick T) Wiley-Blackwell.

[xi]Sonderen (2009) ibid.

[xii] Ricketts C. (2010) A new look at re-sits: are they simply a second chance? *Assessment & Evaluation in Higher Education;* 35: 351-356

[xiii] Matton N, Vautier S, Raufaste E. (2009) Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. *Intelligence;* 37**:** 412-21

[xiv] Kulik JA, Kulik CC, Bangert RL. (1984) Effects of practice on aptitude and achievement test scores. *American Educational Research Journal;* 21*:* 435–447

[xv] Geving AM, Webb S et al. (2005) Opportunities for repeat testing: Practice doesn't always make perfect. *Applied H.R.M. Research;* 2:47-56

[xvi] Raymond MR, Neustel S, Anderson D. (2009) Same-Form Retest Effects on Credentialling Examinations. *National Council on Measurement in Education: Issues and Practice;* 2009: 19-27

[xvii] Boulet JR, McKinley DW, Whelan GP, Hambleton RK. (2003) The effect of task exposure on repeat candidate scores in a high-stakes standardized patient assessment. *Teaching and Learning in Medicine;* 15:227-232

[xviii] Swygert KA, Balog MS, Jobe A. (2010) The Impact of Repeat Information on Examinee Performance for a Large-Scale Standardized-Patient Examination. *Academic Medicine;* 89:1506-1510

[xix] Boulet JR, McKinley DW, Whelan GP, Hambleton RK. (2003) Op cit.

[xx] Hausknecht JP, Trevor CO, Farr JL. (2002) Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology;* 87**:** 243-254

[xxi] Raymond MR, Neustel S, Anderson D. (2007) Retest effects on identical and parallel forms in certification and licensure testing. *Personnel Psychology;* 60: 367–396

[xxii] Schleicher DJ, Van Iddekinge CH, Morgeson FP, et al. (2010) If At First You Don't Succeed, Try, Try Again: Understanding Race, Age, and Gender Differences in Retesting Score Improvement. *Journal of Applied Psychology;* 95**:** 603-617

[xxiii] Reeve CL, Lam H. (2005) The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence*; 33**:** 535-549

[xxiv] Matton N, Vautier S, Raufaste E. (2009) Situational effects may account for gain scores in cognitive ability testing: a longitudinal SEM approach. *Intelligence;* 37: 421-421

[xxv] Hausknecht JP, Trevor CO, Farr JL. (2002) op. cit.

[xxvi] United States Department of Labor. (1970) *Manual for the USES General Aptitude Test Battery*. Washington, D.C.: U.S. Department of Labor.

[xxvii] Lievens F, Buyse T, Sackett PR. (2005) Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology;* 58**:** 981-1007

[xxviii] Hausknecht JP, Trevor CO, Farr JL. (2002) op. cit.

[xxix] McManus IC, Lockwood DNJ. (1992) Does performance improve when candidates resit a post-graduate examination? *Medical Education;* 26: 157-162

[xxx] Bandaranayake RC, Buzzard AJ. (1993) The probability of passing at resits in the part 1 fellowship examination. *Australian and New Zealand Journal of Surgery;* 63: 723-726

[xxxi] Cohen-Schotanus J. (1999) Student assessment and examination rules. *Medical Teacher;* 21(3): 318-21

[xxxii] McManus IC, Ludka K. (2012) Re-sitting a high-stakes postgraduate medical examination on multiple occasions: nonlinear multilevel modelling of performance in the MRCP(UK) examinations. *BMC Medicine;* 10:60: 1-19

[xxxiii] Millman J. (1989) If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher;* 18: 5-9

[xxxiv] Matton N, Vautier SP, Raufaste (2011) Test-Specificity of the Advantage of Retaking Cognitive Ability Tests. *International Journal of Selection and Assessment;* 19: 11-7

[xxxv] Pell G, Fuller R, Homer M, Roberts T. (2012) Is short-term remediation after OSCE failure sustained? A retrospective analysis of the longitudinal attainment of underperforming students in OSCE assessments. *Medical Teacher;* 34: 146-150

[xxxvi] Hays RB. (2012) Remediation and re-assessment in undergraduate medical school examinations. *Medical Teacher*; 34: 91-2

[xxxvii] Hays RB, Sen Gupta TK, Veitch J. (2008) The practical value of the Standard Error of Measurement in borderline pass/fail decisions. *Medical Education*; 42: 810–815

[xxxviii] Bandaranayake RC, Buzzard AJ. (1993) The probability of passing at resits in the part 1 fellowship examination. *Australian and New Zealand Journal of Surgery*; 63: 723-726

[xxxix] Clauser BE, Margolis MJ, Case SM. (2006) Testing for licensure and certification in the professions. In: Brennan RL, editor. Educational Measurement. 4. Westport, CT: Praeger Publishers; pp. 701–731

[xl] Raymond MR, Kahraman N, Swygert KA, Balog KP. (2011) Evaluating Construct Equivalence and Criterion-Related Validity for Repeat Examinees on a Standardized Patient Examination. *Academic Medicine;* 86: 1253-1259

[xli] Lievens F, Buyse T, Sackett PR. (2005) Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology;* 58**:** 981-1007

[xlii] Pell G, Fuller R, Homer M, Roberts T. (2012) Is short-term remediation after OSCE failure sustained? A retrospective analysis of the longitudinal attainment of underperforming students in OSCE assessments. *Medical Teacher;* 34: 146-150

[xliii] Tighe J, McManus IC, Dewhurst NG, Chis L, Mucklow J. (2010) The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP(UK) examinations. *BMC Medical Education*; 10: 40

[xliv] Clauser BE, Nungster RJ. (2001) Classification accuracy for tests that allow retakes. *Academic Medicine*; 76: S108-S110

[xlv] Millman J. (1989) If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher;* 18: 5-9

[xlvi] Clauser BE, Margolis MJ, Case SM. (2006) Testing for licensure and certification in the professions. In *Educational Measurement* 4 ed. Ed Brennan RL. ACE/Praeger Series on Higher Education

[xlvii] Millman J. (1989) op. cit.

[xlviii] Millman J. (1989) op. cit.

[xlix] Juul D, Loewy EH. (1988) The selection of critical errors. Paper presented at the annual meeting of the National Council for Measurement in Education, New Orleans, LA. Cited in Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher;* 18 (6):5-9

[l] Millman J. (1989) op. cit.

[li] Hamdy H, Prasad K, Anderson MB, Scherpbier A, Williams R, Zwierstra R, Cuddihy H. (2006) BEME systematic Review: predictive values of measurements obtained in medical schools and future performance in medical practice. *Medical Teacher;* 28: 103-106

[lii] Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. (1989) Predictive validity of certification by the American Board of Internal Medicine. *Ann Intern Med*; 110:719-26.

[liii] Tamblyn R, Abrahamowicz M, Brailovsky P, et al. (1998) Association between licensing examination scores and resource use and quality of care in primary care practice. *JAMA;* 280: 989–96

Tamblyn R, Abrahamowicz M, Dauphinee D, et al. (2007) Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA* 298: 993–1001

Tamblyn R, Abrahamowicz M, Dauphinee WD, Hanley JA, Norcin J, Girard N, Grand'Maison P, Brailovsky C (2002) Association between licensure examination scores and practice in primary care. *JAMA;* 288: 3019-3026

[liv] Beard JD. (2005) Setting standards for the assessment of operative competence. *European Journal of Vascular and Endovascular Surgery;* 30(2): 215-218

[lv] Holmboe ES, Wang Y, Meehan T, et al. (2008) Association between maintenance of certification examination scores and quality of care for medicare beneficiaries. *Archives of Internal Medicine;* 168: 1396-403

[lvi] Wenghofer E, Klass D, Abrahamowicz M, Dauphinee D, Jacques A, Smee S, Blackmore D, Winslade N, Reidel K, Bartman I, Tamblyn R. (2009) Doctor scores on national qualifying examinations predict quality of care in future practice. *Medical Education*; 43: 1166-1173

[lvii] Mitchell C, Bhat S, Herbert A, Baker P. (2011) Workplace-based assessment of junior doctors: do scores predict training difficulties? *Medical Education;* 45: 1190-1198

[lviii] Hess BJ, Weng W, Lynn LA, Holmboe ES, Lipner RS. (2011) Setting a fair performance standard for physicians' quality of patient care. *Journal of General Internal Medicine;* 26(5): 467-473

[lix] Wakeford R. (2012) International medical graduates' relative underperformance in the MRCGP AKT and CSA examinations. *Education for Primary Care;* 23: 148-152

[lx] Southgate L, Campbell M, Cox J, Foulkes J, Jolly B, McRorie P, Tombleson P. (2001) The General Medical Council's Performance Procedures: The development and implementation of tests of competence with examples from general practice. *Medical Education, Supplement;* 35: 20-28

[lxi] Hausknecht JP, Trevor CO, Farr JL. (2002) Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology*; 87:243-54

[lxii] Choudry KC, Fletcher RH, Soumerai SB. (2005) Systematic Review: The Relationship between Clinical Experience and Quality of Health Care. *Annals of Internal Medicine;* 142:260-273

[lxiii] Norcini JJ, Kimball HR, Lipner RS. **(**2000) Certification and specialization: do they matter in the outcome of acute myocardial infarction? *Academic Medicine;* 75: 1193-8

[lxiv] Papadakis MA, Arnold GK, Blank LL, Holmboe ES, Lipner RS. (2008) Performance during Internal Medicine Residency Training and Subsequent Disciplinary Action by State Licensing Boards. *Annals of Internal Medicine;* 148:869-876

[lxv] McLachlan JC. (2010) Measuring conscientiousness and professionalism in undergraduate medical students. *The Clinical Teacher;* 7:37-40

[lxvi] Veloski JJ, Callahan CA, Xu G, Hojat M, Nash DB. (2000) Prediction of Students' Performances on Licensing Examinations Using Age, Race, Sex, Undergraduate GPAs, and MCAT Scores. *Academic Medicine;* 75: S28-S30

[lxvii] Ferguson E, James D, Madeley L. (2002) Factors associated with success in medical schools: systematic review of the literature. *British Medical Journal*; 324: 952-7

[lxviii] Wiskin CMD, Allan TF, Skelton JR. (2004) Gender as a variable in the assessment of final year degree-level communication skills. *Medical Education*; 38: 129-137

[lxix] Boulet JR, McKinley DW. (2005) Investigating gender-related construct-irrelevant components of scores on the written assessment exercise of a high-stakes certification assessment. *Advances in Health Sciences Education;* 10: 53-63

[lxx] Dewhurst N, McManus C, Mollon, J, Dace E, Vale E. (2007) Performance in the MRCP (UK) Examination 2003-4: analysis of pass rates of UK graduates in relation to self-declared ethnicity and gender. *BMC Medicine;* 5:8

[lxxi] White MT, Welch K. (2012) Does gender predict performance of novices undergoing Fundamentals of Laparoscopic Surgery (FLS) training? *American Journal of Surgery*; 203: 397-400

[lxxii] GMC PLAB Part 2 data (2009-2011) provided by Suzanne Chamberlain, psychometrician and statistician for the Part 2 exam, contracted to provide services to the GMC.

[lxxiii] Dawson B, Iwamoto CI, Ross LP, Nungester RJ, Swanson DB, Volle RL. (1994) Performance on the National Board of Medical Examiners Part I Examination by Men and Women of Different Race and Ethnicity *JAMA;* 272: 674-679

[lxxiv] McManus IC, Elder AT, de Champlain A, Dacre JE, Mollon J, Chis L. (2008) Graduates of different UK medical schools show substantial differences in performance on MRCP(UK) Part 1, Part 2 and PACES examinations. *BMC Medicine;* 6:5

[lxxv] Bowhay AR, Watmough SD. (2009) An evaluation of the performance in the UK Royal College of Anaesthetists primary examination by UK medical school and gender. *BMC Medical Education;* 9:38

[lxxvi] Ferguson E, James D, Madeley L. (2002) op. cit.

[lxxvii] McManus IC, Richards P, Winder BC, Sproston KA. (1996) Final examination performance of medical students from ethnic minorities. *Medical Education*, 30:195-200.

[lxxviii] Moore RA, Rodenbaugh EJ. (2002) The unkindest cut of all: are international medical school graduates subjected to discrimination by general surgery residency programs? *Current Surgery*; 59: 228-236

[lxxix] Wass V, Roberts C, Hoogenboom R, Jones R, van der Vleuten CPM. (2003) Effect of ethnicity on performance in final objective structured clinical examination: qualitative and quantitative study. *British Medical Journal*; 326: 800-803

[lxxx] Van Zanten M, Boulet JR, Mckinley DW. (2003) Correlates of Performance of the ECFMG Clinical Skills Assessment: Influences of Candidate Characteristics on Performance. *Academic Medicine;* 78 : S72-S74

[lxxxi] Dewhurst N, McManus IC, Mollon, J, Dace E, Vale E. (2007) Performance in the MRCP (UK) Examination 2003-4: analysis of pass rates of UK graduates in relation to self-declared ethnicity and gender. *BMC Medicine;* 5:8

[lxxxii] Woolf K, Potts HW, McManus IC. (2011) Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. *British Medical Journal*; 342: d901

[lxxxiii] Wass V, Roberts C, Hoogenboom R, Jones R ,Van der Vleuten CPM. (2003) op. cit.

[lxxxiv] Yates J, James D. (2006) Predicting the “strugglers”: case-control study of students at Nottingham University Medical School. *British Medical Journal*; 332: 1009-1012

[lxxxv] Hofmeister M, Lockyer J, Crutcher R. (2009) The multiple mini interview for selection of international medical graduates into family medicine residency education. *Medical Education*; 43: 573-579

[lxxxvi] GMC PLAB Part 1 data provided by Suzanne Chamberlain.

[lxxxvii] Miles TR. (2004) Some problems in determining the prevalence of dyslexia. *Electronic Journal of Research in Educational Psychology*; 2 (2): 5-12

[lxxxviii] (BDA) British Dyslexia Association (2011) Accessed on 30/10/11 via: http://www.bdadyslexia.org.uk/about-us.html

[lxxxix] Shrewsbury D. (2011) State of Play: Supporting students with specific learning difficulties. *Medical Teacher*; 33: 254-257

[xc] Shrewsbury D. (2011) ibid.

[xci] Gibson S, Leinster S. (2011) How do medical students with dyslexia perform in extended matching questions, short answer questions and observed structured clinical examinations? *Advances in Health Science Education;* 16: 395-404

[xcii] Ricketts C, Brice J, Coombes L. (2010) Are multiple choice tests fair to medical students with specific learning disabilities? *Advances in Health Sciences Education;* 15: 265-275

[xciii] Gibson S, Leinster S. (2011) How do medical students with dyslexia perform in extended matching questions, short answer questions and observed structured clinical examinations? *Advances in Health Science Education;* 16: 395-404

[xciv] McKendree J, Snowling MJ. (2011) Examination results of medical students with dyslexia. *Medical Education;* 45: 176-182

[xcv] Hough LM, Oswald FL, Ployhart RE. (2001) Determinants, Detection and Amelioration of Adverse Impact in Personnel Selection Procedures: Issues, Evidence and Lessons Learned. *International Journal of Selection and Assessment;* 9:152 – 194

[xcvi] Gibson S, Leinster S. (2011) How do medical students with dyslexia perform in extended matching questions, short answer questions and observed structured clinical examinations? *Advances in Health Science Education;* 16: 395-404

[xcvii] Van Iddekinge CH, Morgeson FP, Schleicher DJ et al. (2011) Can I Retake It? Exploring Subgroup Differences and Criterion-Related Validity in Promotion Retesting**.** *Journal of Applied Psychology;* 96**:** 941-955.

[xcviii] Schleicher DJ, Van Iddekinge CH, Morgeson FP et al. (2010) If At First You Don't Succeed, Try, Try Again: Understanding Race, Age, and Gender Differences in Retesting Score Improvement. *Journal of Applied Psychology*; 95**:** 603-617

[xcix] Wisher RA, Sabol MA, Ellis JA. (1999) *Retention of Military Knowledge and Skills.*Arlington, VA: US Army Research Institute for Social and Behavioral Research

[c] Smith KK, Gilcreast D, Pierce K. (2008) Evaluation of staff's retention of ACLS and BLS skills. *Resuscitation;* 78: 59-65

[ci] Kim JW, Koubek RJ, Ritter FE. (2007) Investigation of procedural skills degradation from different modalities. In *Proceedings of the 8th International Conference on Cognitive Modeling*. 255-260. Oxford, UK: Taylor & Francis/Psychology Press

[cii] I am indebted to Professor Lambert Schuwirth for the immediately following comments.

[ciii] http://www.onexamination.com/general-medicine/plab?gclid=CMr9u-3n_68CFalmtAodODQ4FQ Accessed 01/08/12

[civ] http://www.examdoctor.co.uk/exam/plab-part-1/?gclid=CPDRlcTp_68CFQ8htAodZGH Accessed 01/08/12

[cv] http://www.aippg.com/plab-uk/EMQs/default.htm Accessed 01/08/12

[cvi] http://www.lancs.ac.uk/fass/projects/examreform/ Accessed 01/08/12

[cvii] Berk RA. (1986) A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*; 56: 137-172

[cviii] National Council for Measurement in Education (1999) *Standard 1.7: Standards for educational and psychological testing, 1999* http://www.apa.org/science/programs/testing/standards.aspx Accessed 01/08/12

[cix] McManus IC, Mollon J, Duke OL, Vale JA. (2005) Changes in standard of candidates taking the MRCP(UK) Part 1 examination, 1985 to 2002: Analysis of marker questions. *BMC Medicine;* 3:13

[cx] McManus IC, Mollon J, Duke OL, Vale JA. (2005) ibid.

[cxi] Cohen-Schotanus J. (1999) Student assessment and examination rules. *Medical Teacher*; 21(3): 318-21

[cxii] Richter Lagha RA, Boscardin CK, May W, Fung C-C. (2012) A Comparison of two standard-setting approaches in high stakes clinical performance performance assessments using Generalizability theory. *Academic Medicine;* 87(8):1-6

[cxiii] van der Vleuten C. (2010) Setting and maintaining standards in multiple choice examinations: Guide supplement 37.1 – Viewpoint. *Medical Teacher*; 32: 174-176

[cxiv] McHarg J, McLachlan JC, Bradley P, Searle J, Ricketts C, Chamberlain S. (2005) Assessment of progress tests. *Medical Education*; 39:221-227

[cxv] Glass GV. (1978) Standards and Criteria. *Journal of Educational Measurement*; 15: 237-261

[cxvi] Clauser BE, Clyman SG. (1994) A contrasting-groups approach to standard setting for performance assessments of clinical skills. *Academic Medicine*; 69(10 Suppl): S42-44

[cxvii] Yen WM. (1993) Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement*; 30: 187-213

[cxviii] Rehgehr G, MacRae H, Reznick R, Szalay D. (1998) Comparing the psychometric properties of check lists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine*; 73: 993-997

[cxix] Kaufman DM, Mann KV, Muijtjens AMM, van der Vleuten CPM. (2000) A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Academic Medicine; 75:*267–71

[cxx] Downing SM, Lieska NG, Raible MD. (2003) Establishing passing standards for classroom achievement tests in Medical Education: a comparative study of four methods. *Academic Medicine;* 78 Suppl 10. S85-S87

[cxxi] Angoff WH. (1971) Scales, norms and equivalent scores. In Thorndyke RL (ed) *Educational Measurement*, 2nd ed. American Council on Education; Washington DC. pp 508-600

[cxxii] Cusimano MD, Rothman AI. (2003) The effect of incorporating normative data into a criterion-referenced standard setting in medical education. *Academic Medicine;* 78 (Suppl10): 88–90

[cxxiii] Kramer A, Muijtjens A, Jansen K, Dusman H, Tan L, van der Vleuten C. (2003) Comparison of a rational and an empirical standard-setting procedure for an objective structured clinical examination. *Medical Education*; 2: 132–9

[cxxiv] Wood TJ, Humphrey-Murto SM, Norman GR. (2006) Standard setting in a small scale OSCE: a comparison of the modified borderline-group method and the borderline regression method. *Advances in Health Sciences Education: Theory and Practice;* 11: 115–122

[cxxv] Downing SM, Tekian A, Yudkowsky R. (2006) Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teaching and Learning in Medicine;* 18: 50–7

[cxxvi] Boursicot KAM, Roberts TE, Pell G. (2007) Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical Education;* 41(11): 1024-1031

[cxxvii] Schoonheim-Klein M, Muijtjens A, Habets L, Manogue M, van der Vleuten C, van der Velden U. (2009) Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard-setting method. *European Journal of Dental Education;* 13: 162–71

[cxxviii] Jalili M, Hejri SM, Norcini JJ. (2011) Comparison of two methods of standard setting: the performance of the three-level Angoff method. *Medical Education;* 45: 1199-1208

[cxxix] Clauser BE, Clyman SG. (1994) A contrasting-groups approach to standard setting for performance assessments of clinical skills. *Academic Medicine*; 69(10 Suppl): S42-44

[cxxx] Livingston SA, Zieky MJ. (1982) *Passing Scores*. Princeton, New Jersey: Educational Testing Service

[cxxxi] Bhakta B, Tennant A, Horton M, Lawton G, Andrich D. (2005) Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. *BMC Medical Education*; 5:9

[cxxxii] Bond TG. (2003) Validity and assessment: A Rasch measurement perspective. *Metodologia de las Ciencias del Comportamiento;* 5(2): 179–194

[cxxxiii] Bond TG, Fox CM. (2007) *Applying the Rasch Model. Fundamental measurement in the human sciences*. 2nd ed. Lawrence Erlbaum Associates, Inc. New Jersey

[cxxxiv] Nungester RJ, Dillon GF, Swanson DB, Orr NA, Powell RD. (1991) Standard-setting plans for the NBME comprehensive Part I and Part II examinations. *Academic Medicine;* 66(8): 429-433

[cxxxv] De Champlain AF. (2010) Setting and maintaining standards in multiple-choice examinations: guide supplement 37.2 - viewpoint. *Medical Teacher;* 32(5): 436-437

[cxxxvi] Homer M, Pell G. (2009) The impact of the inclusion of simulated patient ratings on the reliability of OSCE assessments under the borderline regression method. *Medical Teacher;* 31(5): 420-425

[cxxxvii] Woehr DJ, Arthur W, Fehrmann ML. (1991) An empirical comparison of cut off score methods for content –related and criterion-related validity settings. *Educational and Psychological Measurement*, 51: 1029-1037.

[cxxxviii] Norcini JJ, Shea J (1992) The reproducibility of standards over groups and occasions. *Appl. Meas Educ* 5: 63-72.

[cxxxix] De Champlain (2004) Ensruring that the competent are truly competent: an overview of common methods and procedures used to set standards on high stakes examinations. *Journal of Veterinary Medical Education*, 31:61-5

[cxl] Tighe J, McManus IC, Dewhurst NG, Chis L, Mucklow J. (2010) The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP(UK) examinations. *BMC Medical Education*; 10:40

[cxli] McManus IC, Mooney-Somers J, Dacre JE, Vale JA. (2003) Reliability of the MRCP (UK) Part 1 Examination. *Medical Education*; 37: 609-611

[cxlii] Tighe J, McManus IC, Dewhurst NG, Chis L, Mucklow J. (2010) The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP(UK) examinations, *BMC Medical Education*; 10:40

[cxliii] Cronbach L, Shavelson RJ. (2004) My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement;* 64(3): 391-418

[cxliv] Hutchinson L, Aitken P, Hayes T. (2002) Are medical postgraduate certification processes valid? A systematic review of the published evidence. *Medical Education;* 36: 73-91

[cxlv] Brannick MT, Erol-Korkmaz HT, Prewett M. (2011) A systematic review of the reliability of objective structured clinical examination records. *Medical Education;* 45: 1181-1189

[cxlvi] Elder A, McManus IC, McAlpine L, Dacre J. (2011) What skills are tested in the new PACES examination? *Annals of the Academy of Medicine, Singapore;* 40: 119-125

[cxlvii] Richter Lagha RA, Boscardin CK, May W, Fung C-C. (2012) A Comparison of two standard-setting approaches in high stakes clinical performance assessments using Generalizability theory. *Academic Medicine;* 87(8): 1077-1082.

[cxlviii] Reece A, Chung EMK, Gardiner RM, Williams SE. (2008) Competency domains in an undergraduate Objective Structured Clinical Examination: their impact on compensatory standard setting. *Medical Education*; 42(6): 600-606

[cxlix] Schoonheim-Klein M*,* Muijtjens A*,* Habets L*,* Manogue M*,* van der Vleuten C*,* van der Velden U. (2009) Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard-setting method*. European Journal of Dental Education;* 13*:*162–71

[cl] Richter Lagha RA, Boscardin CK, May W, Fung C-C. (2012) A Comparison of two standard-setting approaches in high stakes clinical performance assessments using Generalizability theory. *Academic Medicine;* 87(8): 1077-1082.

[cli] Jolly B. (1999) Setting standards for tomorrow's doctors. *Medical Education*; 33(11): 792-793

[clii] Downing SM. (2002) Assessment of knowledge with written test forms. In Norman GR, van der Vleuten CPM, Newble DI (eds) *International Handbook of Medical Education Research*. Dordrecht: Kluwer Academic Publishers. pp 647-672

[cliii] McManus IC, Lissauer T, Williams SE. (2005) Detecting cheating in written medical examinations by statistical analysis of similarity of answers: a pilot study. *British Medical Journal*; 330: 1064-1066

[cliv] Patterson F, Ashworth V, Zibarras L, Coan P, Kerrinz M, O'Neill P. (2012) Evaluations of situational judgement tests to assess non-academic attributes in selection. *Medical Education* (in press)

[clv] Koczwara A, Patterson F, Zibarras L, Kerrin M, Irish B, Wilkinson M. (2012) Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. *Medical Education;* 46: 399-408

[clvi] Ergene T. (2003) Effective interventions on test anxiety. A meta analysis. *School Psychology International;* 24: 313-328

[clvii] Zeidner M. (1990) Does test anxiety bias scholastic aptitude test performance by gender and social group? *Journal of Personality Assessment*; 55: 145-160

[clviii] Dendato KM, Diener D. (1986) Effectiveness of cognitive/relaxation therapy and study skills training in reducing self reported anxiety and improving the academic performance of test anxious students. *Journal of Counseling Psychology;* 33: 131-135

[clix] Van Iddekinge CH, Morgeson FP, Schleicher DJ, Campion MA. (2011) Can I Retake It? Exploring Subgroup Differences and Criterion-Related Validity in Promotion Retesting**.** *Journal of Applied Psychology;* 96**:** 941-955

[clx] Matton N, Vautier S, Raufaste E. (2009) Situational effects may account for gain scores in cognitive ability testing: a longitudinal SEM approach. *Intelligence;* 37: 421-421

[clxi] Greenberg J. (1987) A taxonomy of organizational justice theories. *Academy of Management Review*; 12:9-22

[clxii] Adams JS. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 267-299). New York: Academic Press

[clxiii] Leventhal GS. (1980) What should be done with equity theory? New approaches to the study of fairness in social relationship. In K. Gergen, M. Greenberg, & R. Willis (Eds.), *Social exchange: Advances in theory and research* (pp. 27-55). New York: Plenum Press

[clxiv] Bies RJ, Moag JF. (1986). Interactional justice: Communication criteria of fairness. In R.J. Lewicki, B. H. Sheppard, & M. H. Bazerman (Eds.) *Research on negotiations in organizations* (Vol. 1, pp. 43–55).Greenwich, CT: JAI Press

[clxv] Greenberg, J. (1990). Organizational justice: yesterday, today, and tomorrow. *Journal of Management;* 16: 399-432

[clxvi] Terpstra DE, Kethley RB, Foley RT, Limpaphayom WT. (2000). The nature of litigation surrounding five screening devices. Public Personnel Management, **29**, 43-54.

[clxvii] Posthuma RA, Morgeson FP, Campion MA. (2002) Beyond employment interview validity: a comprehensive narrative review of recent research trends over time. *Personnel Psychology* **55**: 1-81.

[clxviii] Norcini JJ (2003) Work Based Assessment BMJ 326: 753-755.