Assessing the equivalence of PLAB graduates to UK graduates

A comparison of MRCP(UK) and MRCGP performance, and an evaluation of the PLAB Part 1 standard-setting process

Chris McManus: University College London
Richard Wakeford: University of Cambridge

Draft submitted 27th September 2013; minor typos etc corrected 27th October 2013

Executive Summary

1. Our primary task was to assess whether the knowledge and skills of doctors whom we call 'PLAB graduates', "continue to be equivalent" to UK graduates who have "successfully completed the first year of Foundation Programme training", by "examining any evidence of disparity between the success rates of UK medical graduates and international medical graduates in postgraduate examinations and assessments."
2. We compared the performance in recent years of PLAB graduates on the written and clinical components of the MRCP(UK) and the MRCGP with that of UK graduates. A simple comparison of overall mean scores provided strong evidence that PLAB graduates performed at a substantially lower level than UK graduates, as well as progressing more slowly.
3. A key question concerns the level at which PLAB passmarks might be set in order to result in equivalence. For this we used two different methodologies: a) finding a passmark for PLAB whereby median candidate performance equates to median UK graduate performance; and b) dividing the PLAB graduates into twelve equi-interval groups based on their performance at PLAB Part 1 and PLAB Part 2, and contrasting the performance of these groups with the performance of graduates of the different UK medical schools, in order to estimate a mark at PLAB which was equivalent.
4. Our main conclusion is that overall the assessed knowledge and skills of PLAB graduates are substantially below those of UK graduates.
5. In the absence of robust prior evidence on equivalence, we cannot comment on the extent of the 'continuation' of equivalence. Indeed, there is a deep ambiguity as to the meaning of 'equivalence', as to whether minimal standards are met, similar overall standards are achieved or distributions in UK and PLAB graduates are equivalent.
6. There are many possible reasons for the underperformance of PLAB graduates, which include:
    a. Issues with the standard setting process for PLAB Parts 1 and 2;
    b. The difficulties of questions being generated for the examination;
    c. Unlimited attempts at both Parts having been allowed;
    d. A relatively low English language performance requirement (IELTS level 7);
    e. Some candidates may have had prior sight of questions, a problem discussed as far back as the 1986 Report on PLAB.
7. In the short term, equivalence of PLAB might be attained by increasing the current pass standards by a factor of about 25-30 marks for Part 1 and 16 marks for Part 2.
8. We recognize some of the difficult implications of these conclusions for workforce planning and other issues. They result from the fact that a set of assessments with political, professional and social implications have not previously been assessed psychometrically for equivalence.
9. The current analyses were possible because of the recent, and very important, process of record linkage which the GMC has begun. Without such data, the assessment of equivalence would have been hard, although not impossible. In future years, with more data from more postgraduate assessments, equivalence might be easier to assess routinely.
10. Standard-setting of PLAB should not be taking place in a vacuum, but in conjunction with other UK bodies setting standards, including undergraduate medical schools (including the Assessment Alliance of the Medical Schools Council), the Foundation Year Programme, and Postgraduate examiners. This should probably involve sharing of questions, and crossing over of examiners between different assessments. It is also beneficial to have a wide range of opinions on standard-setting groups, including recent graduates.
11. Although it extends somewhat beyond our brief, we have also asked why the research questions asked here have not been asked previously, and we conclude that in part it may be because neither the governance of PLAB nor the organisational structure of the GMC is that of a 'research organisation' where answering such questions comes naturally. Research needs to be a routine part of the PLAB process, preferably carried out in-house as a part of service delivery.

Introduction

Issues to be addressed

The review of the Professional and Linguistic Assessments Board (PLAB) test has several objectives which may be helped by an analysis of how results at the PLAB exam relate to performance on subsequent professional assessments such as MRCP(UK) and MRCGP.

Under Theme 1, *Ensuring Standards*, the group is required,

> "To review whether the knowledge and skills demonstrated by a pass in both parts of the PLAB test continue to be equivalent to those of doctors who have successfully completed the first year of Foundation Programme training."

A related topic, under Theme 3, *Confidence*, requires the group,

> "To examine whether international medical graduates granted full registration following a successful pass in the PLAB test are more or less likely than other cohorts of doctors to experience difficulties in medical practice in the UK",

in particular by,

> "Examining any evidence of disparity between the success rates of UK medical graduates and international medical graduates in postgraduate examinations and assessments."

These separate but related issues can be assessed by examining how well PLAB predicts performance in Royal College examinations and by comparing PLAB graduates with UK graduates. In this report we will consider the following questions:

How do PLAB graduates and UK medical school graduates compare when taking MRCP(UK) and MRCGP?

This section will proceed in three different ways. Firstly the report will ask whether PLAB and UK graduates have the same mean level of performance at MRCP(UK) and MRCGP, which would be expected if they are equivalent. The rest of the section then uses two separate methods to assess equivalence between the two groups, comparing the median levels of performance, and then comparing the performance of PLAB graduates with UK graduates from different UK medical schools.

How is the standard setting for PLAB carried out?

PLAB Part 1 is the major 'gate-keeping' part of PLAB, having the higher failure rate, and so its method of standard setting is of particular interest. This section will examine both the process and some statistical data on standard setting, in order to assess whether the standard may not be set at an appropriate level. Standard-setting in PLAB Part 2 will also be briefly considered.

Although strictly beyond our brief, we will ask here also why the issue of equivalence has not properly been researched before, and how it might relate to issues of governance and that PLAB and the GMC are not a 'research organisation'.

The background to the current PLAB standard

The current standard for PLAB refers to the standard of those who pass being, "equivalent to those of doctors who have successfully completed the first year of Foundation Programme training". Previous reviews, in 1986, 1999 and 2004 have considered the level to be set for the standard. The

2004 Review [1][i] stated that the, "The standard of the test should be that of doctors successfully completing Foundation Year 1", explaining that,

> "Council has agreed that in future doctors who pass the PLAB test will be granted full registration in the same manner as UK-trained doctors who have successfully completed their pre-registration training … [i]t would be inequitable to expect UK-trained doctors and IMGs to satisfy different standards to obtain full registration. For these reasons we have concluded that the standard of the test should be that of doctors completing the end of Foundation Year 1." (Para 15).

The previous Review, in 1999 [2], set a similar standard, saying, "a pass in the PLAB test should demonstrate that the successful candidate has the ability to practise safely as an SHO in a first appointment"[ii].

The terms of reference of an earlier Review, in 1986 [3], said that the standard "related to suitability to engage safely in employment at Senior House Officer level". An Annex/Discussion Document discusses the history of PLAB, referring to a report to Council in 1972 which referred to "publicity which cast doubt on the professional or linguistic competence of individual doctors who had qualified overseas". As a result the TRAB/PLAB tests were introduced in June 1975, and included a test of professional knowledge[iii] (but not a clinical test, which was regarded as impractical). The next decade found that, "the failure of candidates was due in the main part to their lack of professional knowledge rather than difficulty in communicating in English".

Of particular interest in the Annex of the 1986 Report is a description of the pilot testing of the PLAB tests (paras 15, 35 et seq). Pilot testing in 1975 gave the MCQ paper to 87 British medical graduates who had a mean mark of 61% and an unstated number of British medical students who scored a mean of 46%. The panel therefore set the pass mark at 45%. Although this seems a sensible strategy, in practice it has several problems which are discussed further in Appendix 1, and have strong implications for any standard setting of PLAB. Although the 1975 pass mark was set at 45%, the Chair was also authorised to change the pass mark if necessary, and in the event it was actually set at 35%. Over the next year the pass rate of PLAB candidates increased, with the result that in March 1976 the pass mark was shifted to 40%.

Further pilot tests were carried out in November 1977 on 250 final year medical students in Manchester and on that basis the pass mark was raised to 45%. In July 1979 the pass mark was raised to 50%. By 1982 "the view was expressed that many candidates from overseas were now well trained in taking MCQ papers and that their examination skills did not always mirror their clinical expertise"[iv]. In May 1984 the pass mark was raised to 55%. Pass rates declined after 1980, from 43%, through 41%, 37%, 33%, 25% in the intervening years, to 22% in 1985.

[i] Superscript Roman numbers in the text refer to footnotes whereas Arabic numbers in square brackets indicate references in the bibliography.

[ii] Para 12 of the Report expanded on this, saying that, "The GMC has determined that an applicant for limited registration must provide objective evidence of competence to practise safely under supervision as a senior house officer (SHO). A pass in the PLAB test should, therefore, demonstrate the ability to practise safely as an SHO in a first appointment since this is the level at which most doctors work after obtaining limited registration for the first time. It is also the level at which UK trained doctors obtain full registration and is, therefore, an appropriate point at which to implement a broad-based test. SHOs tend to specialise increasingly as they progress."

[iii] MCQ initially in 1975, but supplemented in June 1976 by the "Medical Short Answer paper". In January 1985 a "Projected Material Examination" was added to the existing paper. There was also a Viva Voce examination.

[iv] A comment that, "on those occasions when test papers had been repeated there had been a consistent and significant increase in the pass rate", suggests that there was leakage of examination items. That had almost certainly happened for the 240 pre-recorded questions in the bank for the Comprehension of Spoken English paper. The pass rate rose from 58% in 1975, to 67% in 1976, 70% in 1978, 81% in 1980 and 87% in 1984. The Annex concludes, "There can be little doubt that the bulk of the questions have, over the years, found their way, virtually verbatim, to a number of private organisations which offer specific tuition for overseas doctors who require to pass the PLAB tests. A similar, though not as damaging situation is known to have arisen in respect of some questions used in the Multiple Choice Questions". On the basis of the latter, "policy decisions were taken by the Board in 1984 to eliminate from the Medical Short Answer papers and Parts I and II of the Written English paper any questions which had been used previously".

The 1986 Report also describes a system implemented by Council in 1975 for monitoring doctors who were granted registration after passing TRAB/PLAB, in which towards the end of the first 12 months of UK practice the doctor had to provide the names of two consultants who would provide reports on professional competence and proficiency in English. Adverse reports were received on about 2% of doctors.

This brief historical review shows that in the early days of PLAB there were attempts to assess the standard both by carrying out parallel studies in groups of UK doctors and medical students, and to follow up the performance of doctors into their practice within the UK. We know of no evidence for such activity since 1986.

How do PLAB graduates and UK medical school graduates compare at MRCP(UK) and MRCGP?

The validity of pass marks

High-stakes examinations such as PLAB have a pass mark ('cut score', 'passing score') and although it is rarely discussed in the literature, a key question concerns the validity of that pass mark. Kane [4] distinguishes clearly between a pass mark and a 'performance standard', the latter being a measure of adequate performance in the domain to which passing the assessment allows access. He then says that, "Validation … consists of a demonstration that the proposed passing score can be interpreted as representing an appropriate performance standard".

Kane distinguishes several types of validity for pass marks. '*Procedural validity*' looks at the appropriateness of the procedures used in standard-setting (and we will consider it later when looking at the details of the Angoff process used for PLAB Part 1). Kane particularly emphasises that poor procedure can cast serious doubt on the validity of a pass mark, but good procedure alone cannot validate the particular pass mark which is chosen. '*Internal validity*' of standard-setting assesses the agreement of examiners on the pass mark and hence the reliability and precision of the pass mark which is chosen. Reliability alone does not however validate a pass mark, for as Verheggen *et al* [5] have said in the context of providing normative data, it might well be, "that [although] the Angoff ratings are more reliable, they may[also] be less valid. In other words the judgements would be *consistently* off the mark" (p.210; emphasis in original).

Kane's third approach to validity looks to external criteria, particularly the '*direct, criterion-related approach'*, which essentially sees how those passing the exam perform at later tasks, whether those who pass well perform better than those who only just pass, and whether those who are only just passing are subsequently performing at an acceptable level. If they are not, then it is probable that the standard is set at the wrong level. That is the approach which is adopted in this section.

The meaning of equivalence

Although it is expected that those passing PLAB will be "be equivalent to those of doctors who have successfully completed [FY1]", the meaning of 'equivalent' has a number of philosophical and definitional subtleties.

Although little used in mainstream biomedicine and social sciences, 'equivalence testing' or 'bioequivalence testing' has been extensively used since the 1980s in clinical pharmacology, where it can make the process of acceptance of new drugs much faster and cheaper. The logic is that "whereas the purpose of a traditional hypothesis test is to determine whether two groups differ from one another, [an equivalence test] is used to determine whether two groups are sufficiently near each other to be considered equivalent" [6]. This methodology is well developed, and can be applied fairly straightforwardly to social science data [7]. Bioequivalence testing has well agreed standards for two drugs to be regarded as equivalent, typically being that the peak concentration, or some similar parameter, is within ±20% of a reference compound. Such values are more difficult to obtain in social science research [7], although they are not impossible.

Equivalence testing conventionally only looks at the mean of a test distribution in comparison with a reference distribution. The mean is often a good descriptor of a distribution, but it is not the only one that is relevant. The abilities of UK and PLAB graduates form distributions, with some graduates being excellent and others being barely acceptable. At that point it is unclear what is meant by "equivalent to ... doctors who have successfully completed FY1". Should perhaps the medians be equivalent? Alternatively, since one is dealing with qualifying examinations, should, say, all PLAB graduates be at least as good as the worst UK graduate (who is on the Register and practising?). Here we will concentrate on comparing the median performance of PLAB and UK graduates, and we will compare graduates from different UK medical schools with PLAB graduates passing at different levels.

The 'continuing equivalence' of PLAB

The Review Group was asked to assess the extent to which the standard of PLAB graduates, "*continue[s]* to be equivalent to those of doctors who have successfully completed [FY1]" (our emphasis). There are, to our knowledge, no previous assessments of PLAB equivalence which use Kane's "direct, criterion-related approach", and instead all previous evidence of continuing equivalence, with some minor exceptions in the late 1970s and early 1980s, has relied on evidence from "procedural validity" and "internal validity" of the Angoff and Borderline groups methods used in PLAB Parts 1 and 2. While important and necessary, such forms of validity cannot properly demonstrate equivalence, and that is particularly the case if those carrying them out are not also directly involved in similar, parallel processes with UK finals examinations and the formative assessment of FY1 doctors.

Assessing the equivalence of UK and PLAB graduates

The assessment of equivalence of assessments is never straightforward unless either there are two groups of individuals taking the same assessment [8] or there is cross-moderation of judgemental methods [9]. UK graduates who have "successfully completed the first year of Foundation Programme training" do not take PLAB (or indeed any other summative assessment at the end of FY1), and PLAB graduates will not have taken UK medical school finals[v]. Neither are there shared questions in PLAB and UK medical school finals (and indeed because UK medical schools run their own final examinations, different items are used in different schools[vi]. Standards may also differ between UK medical schools [10], as will be discussed later). Direct assessment of equivalence does not therefore seem possible using such methods.

An indirect method of assessing equivalence is to compare groups such as PLAB and UK graduates on some other assessment taken by both groups – an external yardstick, in effect, the procedure which Kane has referred to as a "direct, criterion-related approach". Once PLAB graduates have entered the UK medical training system they enter training programmes which are the same as those for UK graduates, with regular assessment through ARCP (Annual Review of Competence Progression), which includes various workplace based assessments (WPAs). There are also major summative assessments in the form of membership examinations for the various Royal Colleges. All doctors in the UK doctors are subject to oversight by the General Medical Council, and in some cases that results in Fitness to Practice cases. The present analysis will only consider performance in

[v] It should also be said at the beginning that since no assessment, beyond the formative and the informal, is carried out of UK graduates at the end of FY1, there is a sense in which equivalence can never be properly established. Since PLAB graduates typically and frequently join UK graduates at the beginning or the end of FY2, we will treat the equivalence standard as being either the standard of UK graduates in general on similar career tracks or, in the case of knowledge-based exams, the approximate standard of UK graduates as shown in medical school finals examinations. Neither is optimal, but should provide good indications of the appropriateness of the standard.

[vi] Attempts are being made at present for UK medical schools to use some identical items in their finals, in an initiative of the Medical Schools Council Assessment Alliance (http://www.medschools.ac.uk/MSCAA), although as yet the project is in its early stages.

examinations of two of the Royal Colleges, with ARCP and Fitness to Practice data being the subject of a separate report by a different team.

The logic of the current study is straightforward. MRCP(UK) and MRCGP are taken both by UK and PLAB graduates, and if the UK and PLAB graduates are equivalent then they should perform equally well when they take the MRCP(UK) and MRCGP examinations[vii]. The situation is made somewhat more complex as doctors choose which medical specialty to enter, and they are also selected onto training programmes such as for general practice or for core medical training (CMT)[viii]. Those taking the examinations are not therefore random or representative samples of either UK or PLAB graduates, although they are a complete sample in the present study of those taking MRCP(UK) or MRCGP in the years concerned.

Methodological details of the datasets and analyses

Background information about the two postgraduate examinations, and the creation of the databases, follows:

MRCP(UK). MRCP(UK) is run by the Royal Colleges of Physicians of London, Edinburgh and Glasgow, is in three parts, MRCP Part 1 (a 200-item best-of five (BOF) multiple choice assessment with brief clinical vignettes, which is computer-marked), MRCP Part 2 (a 270-item computer-marked BOF assessment with more complex and extensive clinical scenarios), and PACES (Practical Assessment of Clinical and Examination Skills; a modified OSCE, with eight encounters, six involving real patients, and two involving simulated patients, and two examiners at each station).

The original PACES examination[11] changed its format in 2009 to new PACES (nPACES)[12]. Separate but somewhat different marks for Communication Skills and Physical Examination Skills could be extracted from both PACES and nPACES[ix]. The MRCP(UK) Part 1 and Part 2 examination have had essentially the same structure since 2001/2 [13], although the method of standard setting for both was changed from Angoff to Statistical Equating in 2009 and 2010 respectively.

MRCGP. The MRCGP examinations, which are run by the Royal College of General Practitioners, are in two parts, the AKT (Applied Knowledge Test; a 200-item computer-displayed and computer-marked multiple choice test, with a variety of item types), and the CSA (Clinical Skills Assessment; a 13-station OSCE in the form of a simulated surgery with candidates seeing simulated patients, while being assessed by an examiner).

The AKT is typically taken during the second year of training, and the CSA is taken during the third and normally final year of training. All candidates are on UK training schemes overseen by Postgraduate Deaneries; entry to the examinations by others (e.g. foreign-based candidates) is not allowed.

The PLAB datasets: The PLAB database is extensive, and for the present purposes only a subset of the measures were used, as under.

[vii] Although it would clearly be desirable to use a wider range of Royal College examinations, data are not currently available on them (although they should be in the future), and we have therefore restricted ourselves to the two examinations for which we act as educational and psychometric advisors, MRCGP in the case of Richard Wakeford and MRCP(UK) for Chris McManus. We are very grateful to the RCGP and the Federation of Royal Colleges of Physicians for providing their consent to these analyses.

[viii] It is also possible for doctors to take MRCP(UK) (but not MRCGP) outside of a training programme, and many, for various reasons, choose to do so. MRCP(UK) holds little data on the training status of its candidates, that information only having been collected in the past few years.

[ix] Essentially these marks differentiate those stations or skills which primarily consider what can be called 'Examination Skills' (examination being in the sense of physical examination, and the interpretation of physical signs), and 'Communication Skills', principally involved in history-taking, the exploration of symptoms, the communication of information, and the handling of difficult situations including the breaking of bad news.

1. *Performance on PLAB Part 1 (knowledge assessment)*

a. PLAB Part 1 has 200 scorable items, of which a small number are removed because of problems in keying/scoring, so that in a typical examination there are about 197 scored items, for which a total score is calculated for each candidate.

b. The pass mark is set by a variant of the Angoff method, and is typically about 125, but has varied in the range 116 to 135, reflecting the adjudged relative ease or difficulty of the particular items chosen. Candidates attaining the pass mark pass the exam.

c. For present purposes it is convenient to describe performance in terms of *marks relative to the pass mark*, so that a candidate scoring zero just passes the exam, a candidate with a positive mark has passed the examination with a mark or marks to spare, and candidates with a negative mark have failed the examination. Marks relative to the pass mark are raw marks, typically out of about 197, and can be considered as percentages (approximately) by dividing by two.

d. Candidates can take the examination on multiple occasions, and do so in ever decreasing numbers. Previous analyses (of MRCP(UK)) [14] have suggested that mark at the first attempt of taking an examination is the best predictor of future performance, and therefore our main analyses consider *Mark at first attempt*. In Appendix 2 we also calculate *Mark when passing*, irrespective of the attempt, and show that while broadly similar to Mark at first attempt, it is far more problematic to interpret. A further description with illustrative data is provided in Appendix 2.

e. PLAB Part 1 has four sub-scales, described as *Context*, *Diagnosis*, *Investigations* and *Management*. Different numbers of items are used on the four sub-scales, and separate pass marks are not available for them. Performance is expressed as percentage correct on each sub-scale, and cannot take differences in item difficult into account, meaning that some variance on each sub-scale score reflects variance in difficulty of items rather than difficulty in performance of candidates. The sub-scores are only used in a few exploratory correlational analyses.

f. **In summary**: Performance on PLAB Part 1 is principally analysed in terms of *difference in raw marks from the pass mark at the first attempt*, zero and positive marks indicating a pass, and negative marks a fail.

2. *Performance on PLAB Part 2 (clinical assessment)*

a. The PLAB Part 2 exam has been described and analysed in detail in a previous report by one of us [15].

b. Candidates on a particular day see 15 OSCE stations, one of which is a pilot station, and hence marks are available from 14 stations.

c. There are four types of station, Communication Skills (C), Examination Skills (E), History-taking Skills (H) and Practical Skills (P), and since 2008 each circuit of 14 stations has had 4 Communication, Examination and History stations, and 2 Practical stations. Prior to 2008, the four station types were all included but in variable proportions.

d. There is one examiner at each station. Standardised patients do not take part in assessments.

e. Examiners have a marking scheme for each station, typically involving four to seven objectives which are rated from A to E (Excellent, Good, Adequate, Fail, Severe Fail). In addition there is an Examiner's Overall Judgment (EOJ), which is rated as Pass, Borderline or Fail.

f. The marking schemes and standard setting are relatively complex and will not be described in detail here. Suffice it to refer the reader to the previous report [15], and to say that one method was in use from 2001 until April 2008, and a second method from April 2008 onwards.

g. For the current marking scheme a (non-integer) total mark between 0 and 4 is awarded on each of the 14 stations, giving an overall mark in the range 0 to 56. Each station has its own pass mark, and to pass the exam a candidate must pass on at least 9 stations. In addition the total score across the stations must be greater than the sum of the fourteen individual station pass marks plus one standard error of measurement.

h. Pass marks are currently set by a variant on the borderline groups method, using the EOJ ratings made by the examiners, pooled across all occasions on which that station has been used.

i. Prior to April 2008, the marks were scored and a pass mark set using a somewhat different method. For the purposes of the present analysis, pre-April 2008 marks were re-scaled to make them equivalent to current marks.

j. Pass marks vary for different sets of 14 stations, and therefore all marks described here are analysed in terms of marks relative to the pass mark. A value of +5 therefore means that a candidate is 5 marks above the pass mark for that set of OSCE stations. It should be noted that the marks are not percentages, and cannot therefore be directly compared with Part 1 marks, which are either numbers correct out of 200 (approximately) or are expressed as percentage of items correct.

k. **In summary**: Performance on PLAB Part 2 is described in terms of (raw) marks above the overall pass mark which was set for the particular set of 14 OSCE stations which was used. Marks are not percentages, and cannot directly be compared with those for PLAB Part 1. As for Part 1, most of our analyses are conducted on the basis of mark at first attempt.

3. Performance in IELTS tests

a. Candidates taking PLAB must have achieved a minimum standard at IELTS (International English Language Testing System; http://www.ielts.org). IELTS has four sub-scales (Listening, Reading, Speaking and Writing) and a total score. The required level for PLAB has varied over the years, but currently is set at a score of 7 on the total score and at all four sub-scores. Candidates taking PLAB in earlier years may have had lower scores either overall or on sub-scales.

b. IELTS describes its bands as follows:

   i. Band 9: **Expert user**: has fully operational command of the language: appropriate, accurate and fluent with complete understanding.

   ii. Band 8: **Very good user**: has fully operational command of the language with only occasional unsystematic inaccuracies and inappropriacies. Misunderstandings may occur in unfamiliar situations. Handles complex detailed argumentation well.

   iii. Band 7: **Good user**: has operational command of the language, though with occasional inaccuracies, inappropriacies and misunderstandings in some situations. Generally handles complex language well and understands detailed reasoning.

   iv. Band 6: **Competent user**: has generally effective command of the language despite some inaccuracies, inappropriacies and misunderstandings. Can use and understand fairly complex language, particularly in familiar situations.

   v. Band 5: **Modest user**: has partial command of the language, coping with overall meaning in most situations, though is likely to make many mistakes. Should be able to handle basic communication in own field.

   All parts of the test and the Overall Band Score can be reported in whole and half bands, e.g.[x] 6.5, 7.0, 7.5, 8.0.

   IELTS' current advice is that, for 'linguistically demanding academic courses', scores in Bands 7.5 to 9 are 'acceptable' and that scores in Band 7 are 'probably acceptable'.

c. For convenience in some later analyses we divide candidates into three groups with a total score ≤ 7, 7.5 or ≥ 8.

d. Some candidates are exempted from requiring an IELTS pass, mainly because they can demonstrate that they have carried out their training at a medical school where the great majority of teaching is in English.

[x] http://www.ielts.org/institutions/global_recognition/setting_ielts_requirements.aspx.

Linking of the MRCP(UK) and PLAB databases

The database for the current analysis consisted initially of all 65,115 candidates who had taken at least one part of the MRCP(UK) examination between 2001 and 2012[xi]. 37329 MRCP(UK) candidates had a GMC number and therefore had at some point worked in the UK. Linkage with the PLAB database was carried out by the GMC looking for all PLAB candidates who had a GMC number in the list of those taking MRCP(UK), with 9,818 being identified overall. Of the remaining candidates, 24641 had graduated at UK medical schools, and are the group to be compared with the PLAB candidates and with whom they should be equivalent.

Linking of the MRCGP and PLAB databases

Two databases were created for the MRCGP/PLAB linkage, one for the AKT the other for the CSA. Linkage with the PLAB database was carried out by the GMC looking for all PLAB candidates who had a GMC number in the lists of those taking either or both parts of the MRCGP. There were data available on the AKT between 2008 and 2013 for a total of 22,081 candidate attempts, of which 17,395 were first attempts. Of these latter 3,160 had taken PLAB Part 1, 3,067 had taken PLAB Part 2 and 2,985 had IELTS scores reported. There were data available on the current version of the CSA (not directly comparable in terms of candidate scores with the preceding version) for candidates taking this between 2010 and 2013, a total of 11,673 candidate attempts, 8,346 of which were first attempts. Of these, 1,411 had taken PLAB Part 1, 1,388 PLAB Part 2, and 1,353 had IELTS scores reported.

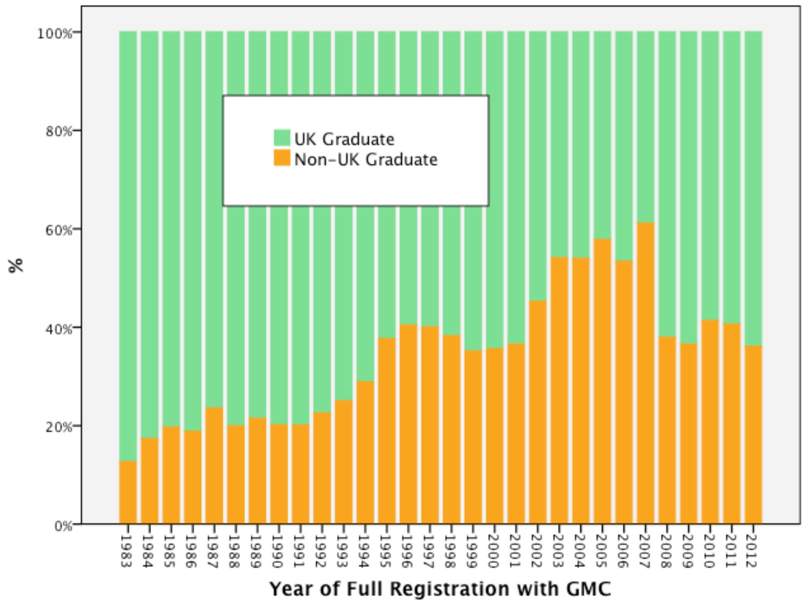
The AKT database is larger than the CSA database, principally because it covers a longer time period. Because they are separate, certain intra-MRCGP data which are available for the MRCP(UK) are not readily calculable for the MRCGP. The age of candidates is not available, and date of primary medical qualification must be used as a surrogate.

Demographics and progression of UK and PLAB groups

The numbers of candidates taking PLAB and their origins

Foreign medical graduates provide a substantial part of the UK's medical workforce. 39.2% of medical graduates currently registered with the GMC (and currently holding a licence) and who achieved full registration in the last 30 years possess non-UK primary medical qualifications (Figure 1). Many of these non-UK graduates will have qualified in the UK by means of the PLAB examinations. Take up of PLAB can be variable across years, due to such things as changes in legislation, immigration rules and NHS workforce requirements.

Figure 1: Proportions of UK and non-UK Graduates on LRMP (Aug 2013) by F



Part 1 may be sat in the UK or in British Council test centres abroad, Part 2 is only sat in the PLAB test centre in Manchester. Over the five years 2008 – 2012, 17,441 attempts were made at Part 1

[xi] Because of the nature of the sampling window, data were right and left truncated, data for some candidates early in the cohort consisting only of Part2 or PACES, and for candidates late in the process consisting only of Part 1 results.

(average, 3,488 per annum) and 9,240 attempts at Part 2 (average, 1,848 per annum). 64% of the Part 1 attempts were candidates' first ones at the examination, for Part 2 the figure was 71%.

First attempt pass rates at PLAB are approximately 45% and 71% respectively for Parts 1 and 2 but show huge variations by country of PMQ. From data available from the PLAB team, we have estimated that the ultimate pass rates for those attempting the exam are approximately 60% and 75%, respectively. Over the quinquennium 2008-2012, an average of 1,513 IMGs per annum passed the final part of the examination towards qualifying for full registration with the GMC. This compares with an annual average of 6,720 UK graduates fully registering over the same period, and represents the output of the equivalent of five or six medium-sized UK medical schools.

The main countries where PLAB Part 1 candidates' primary medical qualification had been obtained are shown in Table 1 for countries providing > 2% of the candidature. Candidates were not always nationals of the country providing their Primary Medical Qualification (PMQ) so the table also shows nationalities.

Table 1: PLAB Part 1 Examinations 2008-12 Candidates' Nationality and Country of Primary Medical Qualification: main countries (>2%)

Country	% of candidates with PMQ from Country	% of candidates with nationality of country
Bangladesh	2%	2%
China	2%	(<1%)
Egypt	4%	2%
India	17%	16%
Iraq	3%	2%
Nepal	2%	(1%)
Nigeria	12%	12%
Pakistan	28%	25%
Russian Federation	2%	(<1%)
South Africa	2%	(1%)
Sri Lanka	(1%)	2%
Sudan	9%	5%
Ukraine	2%	(<1%)
United Kingdom	n/a	13%

Comparison of UK and PLAB graduates on demographics and progression

Table 2 shows basic descriptive data on demographics and progression for PLAB and UK graduates taking MRCP(UK) and the MRCGP.

For the **MRCP(UK)**, PLAB graduates are more likely to be male and to be from ethnic minorities. UK and PLAB graduates qualify as doctors at similar ages, but PLAB graduates take MRCP(UK) later than UK graduates, not least because they have been taking PLAB Parts 1 and 2 between graduation and taking MRCP(UK) Part 1. PLAB graduates also progress more slowly through MRCP(UK) Parts 1, 2 and PACES (in large part due to having more resits – data not shown).

In PLAB graduates, age at taking PLAB Part 1 shows only a small correlation with mark at first attempt (r=-.074, n=9812, p<.001), and a similarly small correlation is found between age at first taking MRCP(UK) Part 1 and mark at first attempt at MRCP(UK) Part 1 (r=-.150, n=7823, p<.001); a similar result is found for UK graduates taking MRCP(UK) Part 1 (r= -.090, n=18532). In each case, older candidates perform somewhat less well, although the reasons for that are far from clear, not least as older candidates should overall have had more clinical experience.

For the **MRCGP**, PLAB graduates are more likely than UK graduates to be male and far more likely to be non-white. They have been, on average, qualified 5-6 years longer than their UK counterparts. PLAB graduates are far more likely to resit both the AKT and CSA than UK graduates (mean attempt number in AKT database = 1.16 for UK graduates, 1.64 for PLAB graduates p<.001; mean attempt number in CSA database = 1.12 for UK graduates, 2.17 for PLAB graduates p<.001).

Correlations between date of qualification and score on first attempts at the MRCGP components are: AKT – UK graduates r=-.023 (n=12152, p<.05), PLAB graduates r=.101 (n=3160, p<.01); and for the CSA – UK graduates r=.078 (n=5977, p<.01), PLAB graduates r=.247 (n=1388, p<.001). (Note that a positive correlation here indicates that *more recent* and thus probably younger graduates perform better.) Thus, in all groups except UK graduates taking the AKT, age correlates negatively with success.

Table 2: Demographics of candidates taking MRCP(UK) and MRCGP

	UK graduates; % or mean (SD)	PLAB graduates; % or mean (SD)	Significance
MRCP(UK) candidates			
Age at Qualification	24.85 (2.16) n=24640	24.88 (2.00) n=9798	NS
Age 1st attempt PLAB Part 1	n/a	28.66 (4.34) n=9812	-
Age 1st attempt PLAB Part 2	n/a	29.41 (4.41) n=9344	-
Age 1st attempt MRCP(UK) Part 1	26.87 (2.40) n=18532	30.34 (4.51) n=7823	p<.001
Age 1st attempt MRCP(UK) Part 2	27.71 (2.33) n=14094	31.54 (4.19) n=5133	P<.001
Age 1st attempt MRCP(UK) PACES	28.43 (2.34) n=14409	32.61 (4.11) n=4388	P<.001
Interval 1st attempts at MRCP(UK) Part1 and Part 2 (weeks)	50.0 (32.2) n=12091	101.3 (75.0) n=3947	P<.001
Interval 1st attempts at MRCP(UK) Part2 and PACES (weeks)	39.4 (23.5) n=12051	66.7 (49.7) n=4138	P<.001
Sex (% female)	56.3% n=24634	32.0% n=9802	p<.001
Ethnicity (% non-white)	41.1% n=24641	96.3% n=9804	P<.001
%UK nationals	n/a	7.8% n=9589	-
MRCGP candidates			
Year of Qualification (AKT Database)	2005.6 n=12,152	1998.7 n=3,160	p<.001
Year of Qualification (CSA Database)	2005.9 n=5,977	1999.5 n=1,388	p<.001
Sex (% female) (AKT Database)	65.2% n=12,152	44.9% n=3,160	p<.001
Sex (% female) (CSA Database)	63.8% n=5,977	45.0% n=1,388	p<.001
Ethnicity (% non-white) (AKT Database)**	32.3% n=12,152	94.2% n=3,160	p<.001
Ethnicity (% non-white) (CSA Database)**	33.1% n=5,924	94.4% n=1,381	p<.001
%UK nationals	n/a	12.0% n=3233	-

* Self-reported ethnicity from MRCP(UK) database
** In the MRCGP, candidates are constrained by training programme length (normally, 3 years) and the rules about when the components can be taken (AKT not before Year 2, CSA not before Year 3). They thus normally make their first attempt at the AKT about 6-12 months before their first attempt at the CSA.

Data on nationality is available only for the candidates taking PLAB, but as Table 2 shows, perhaps somewhat surprisingly, there is a large group of PLAB candidates who are UK nationals, about 8% (749/9589 for those taking MRCP(UK)), and 12% (n=388) for the MRCGP. The most frequent countries of graduation for the MRCP(UK) group are Pakistan (n=181), India (n=114), Nigeria (n=72), Iraq (n=56), Sudan (n=53) and Bangladesh (n=34)[xii]. The MRCPGP group is similar, 50% qualified in

[xii] These doctors are probably heterogeneous in their origins. Some may have taken up UK nationality since arriving in the UK, and some may have been educated in the UK, but, for various reasons, have chosen to study in countries such as the Czech Republic (n=23), the Russian Federation (n=14), Grenada (n=7), Romania (n=5), and Bulgaria (n=5), all of which have medical schools which advertise extensively for international students, often offering parallel courses to local ones, taught

the Indian sub-Continent, 14% in West Africa, 13% in Eastern Europe/former USSR, and 5% from various medical schools in the Caribbean[xiii]. The MRCGP candidates who were UK nationals took significantly more attempts to pass PLAB Part 1 than those who were of other nationalities(mean: 1.8 attempts vs. 1.4 attempts, p<.001). First attempt score on PLAB Part 1 was also significantly lower than for non-UK nationals, (mean: 2.98 vs 7.21, p<.001), and they also performed less well on the AKT, British (mean: -2.54 vs 4.67, p<.001), whereas on the CSA were not statistically different from non-UK nationals (mean: -3.45 vs -4.81, NS)[xiv].

Although all of these various differences in demographics and progression are not to be disputed, for the primary process of this report of assessing equivalence, they are mostly not relevant. If PLAB acts to allow only doctors who are indeed equivalent to enter UK medical training, then despite age or other demographic differences, progression of those equivalent doctors should be similar to those of UK graduates. To put it another way, this analysis is primarily about the nature of the ruler by which the performance of PLAB graduates is being measured, and the relationship of this ruler to those provided by MRCP(UK) and the MRCGP, rather than about external reasons for differences in performance of the various groups.

Correlations of PLAB results with MRCP(UK) and MRCGP results

A necessary analysis, before comparing UK and PLAB graduates in detail, is to assess the extent that PLAB results correlate with performance on MRCP(UK) and MRCGP. If PLAB is a valid assessment of skills relevant to progression during UK postgraduate training then performance on it should relate to performance on subsequent UK postgraduate assessments. Elsewhere, in longitudinal studies of UK graduates, it has been shown that there are strong continuities across performance in secondary school assessments, undergraduate medical school performance, and postgraduate examination performance in the form of MRCP(UK) [16], with a preliminary analysis suggesting that MRCGP also correlates in a similar way. This we have called the ‘Academic Backbone’.

Table 3a shows that better performance on the two parts of PLAB correlates with better performance on the various parts of MRCP(UK) and of MRCGP.

Table 3a: Correlations of performance in PLAB Parts 1 and 2 a) with performance at MRCP(UK) and b) with performance at MRCGP

	PLAB Pt1 1st attempt	**PLAB Pt2** 1st attempt
MRCP(UK)		
Part 1 1st attempt	r=.521 (p<.001; n=7823)	r=.194 (p<.001; n=7671)
Part 2 1st attempt	r=.390 (p<.001; n=5133)	r=.227 (p<.001; n=4916)
PACES 1st attempt	r=.171 (p<.001; n=4386)	r=.274 (p<.001; n=4120)
MRCGP		
AKT 1st attempt	r=.490 (p<.001; n=3160)	r=.186 (p<.001; n=3067)
CSA 1st attempt	r=.232 (p<.001; n=1411)	r=.321 (p<.001; n=1388)

in English. The largest groups of UK nationals have graduated on the Indian sub-continent. Looking at the IELTS scores of these doctors there is a specific pattern, compared with other PLAB graduates, of having higher scores on the IELTS total score (p<.001), with particularly large differences on the Speaking and Listening scores (p<.001), and no difference on the Reading and Writing scores. This suggests that these doctors have been brought up in families where English is spoken, but that they have not been in an English-speaking school system. It seems a reasonable possibility that they are from extended families which are a part of the sub-continental diaspora, which has helped their ability to comprehend spoken English. Further investigation of the large sub-group of UK nationals taking PLAB may well be warranted.

[xiii] Other data on the complete cohort of all PLAB Part 1 entrants who ever took Part 1 (including those who failed or never passed Part 20, also finds that about 12% of all first attempts at PLAB Part 1 are by British nationals.

[xiv] These results might be compatible with a group who either have failed to obtain admission to a UK medical school or are part of a diaspora, and therefore have acquired more UK cultural capital through living in the UK, and therefore perform better on PLAB Part 2 and the CSA.

There is also specificity in that the knowledge based assessment of PLAB Part 1 particularly correlates with MRCP(UK) Part 1 and MRCGP AKT, whereas the clinical assessment of PLAB Part 2 correlates better with MRCP(UK) PACES and MRCGP CSA, both of which are clinical assessments.

An important question, for later, is the extent to which the correlations found here are similar to those which would be found between, say, finals at a UK medical school, and the assessments. That cannot be assessed with the data provided in the GMC-MRCP and GMC-MRCGP linked databases. The key conclusion is that PLAB results correlate at a reasonable level with MRCP(UK) and MRCGP results, providing evidence of the predictive validity of PLAB.

For comparative purposes, table 3b shows correlations analysed in a separate study of the MRCP(UK) and MRCGP sections for those candidates who have taken both assessments. As in table 3a, it can be seen that there is specificity, knowledge-based assessments correlating highly (.673 between MRCP(UK) Part 1 and the AKT), and the clinical examinations (PACES and the CSA) correlating highly (.496). The latter correlation is particularly important as it suggests that the lowish correlation between PLAB Part 2 and both PACES (.186) and the CSA (.321) is not a reflection of poor correlation between clinical assessments in general, but is more likely explained by the low reliability of PLAB Part 2, which has been discussed elsewhere [15], and is probably in the range .55 to .71.

Table 3b: Correlations of performance at first attempts in MRCP(UK) and MRCGP by candidates who have taken both assessments

	MRCPGP AKT 1st attempt	**MRCGP CSA** 1st attempt
MRCP(UK)		
Part 1 1st attempt	r=.673 (p<.001; n=1988)	r=.348 (p<.001; n=1988)
Part 2 1st attempt	r=.600 (p<.001; n=1131)	r=.386 (p<.001; n=1131)
PACES 1st attempt	r=.471 (p<.001; n=943)	r=.496 (p<.001; n=943)
Note: These data are not part of the present PLAB dataset but are part of a separate collaborative research project between the MRCP(UK) and the MRCGP which is currently being prepared for publication.		

The role of IELTS on performance in PLAB and in MRCP(UK) and MRCGP

We have examined the extent to which poor performance in postgraduate examinations may be related to poor English language skills rather than poor medical knowledge. We will not consider it in detail here but Appendix 3 shows for the MRCGP that while poorer IELTS scores do have some impact on MRCGP performance, the majority of the effect is due to poorer medical knowledge as assessed by PLAB. Similar results are found for MRCP(UK) but are not reported here. Poorer performance at PLAB is itself related to lower IELTS scores [15], but that is a separate matter.

Equivalence of MRCP(UK) and MRCGP candidates who have taken PLAB or UK finals

If UK and PLAB candidates for MRCP(UK) or MRCGP are equivalent then the simplest of predictions is that their mean scores on the assessments should be the same. Table 4 shows very clearly that they are not. For all of the assessments, the mean marks of PLAB graduates are substantially below those of UK graduates. An idea of the size of the effect can be gained by calculating the effect size known as *Cohen's d* (the difference in the mean scores divided by the standard deviation of the reference group, which here is the UK graduates). Cohen's d is .94, .91 and 1.40 for MRCP(UK) Parts 1, 2 and PACES, and 1.01 and 1.82 respectively for MRCGP AKT and CSA. A conventional classification describes values of Cohen's d of greater than .8 as "large", and these values are undoubtedly substantial, the average effect size of 1.22 suggesting about one and a quarter standard deviations between the UK and the PLAB groups. It should also be added that conventional equivalence calculations with plausible values for the acceptable differences in performance also yield highly significant evidence for non-equivalence, mainly because of the very large sample sizes involved, and will not be considered further.

Table 4: Mean marks and SDs of UK and PLAB graduates at their first attempt at the various parts of MRCP(UK) and MRCGP

	UK graduates	PLAB graduates	Significance & Effect size
MRCP(UK)			
Part 1 1st attempt	.734 (10.12; n=18532)	-8.73 (10.75; n=7823)	P<.001; Cohen's d=.94
Part 2 1st attempt	6.406 (7.49; n=14094)	-.414 (6.906; n=5133)	P<.001; Cohen's d = .91
PACES 1st attempt*	1.148 (5.34; n=14376)	-6.337 (6.317; n=4386)	P<.001; Cohen's d=1.40
MRCGP			
AKT 1st attempt	19.02 (15.04; n=12152)	3.78 (16.23; n=3160)	p<.001; Cohen's d=1.01
CSA 1st attempt	13.44 (9.93; n=5977)	-4.61 (10.66; n=1388)	p+.001; Cohen's d=1.82

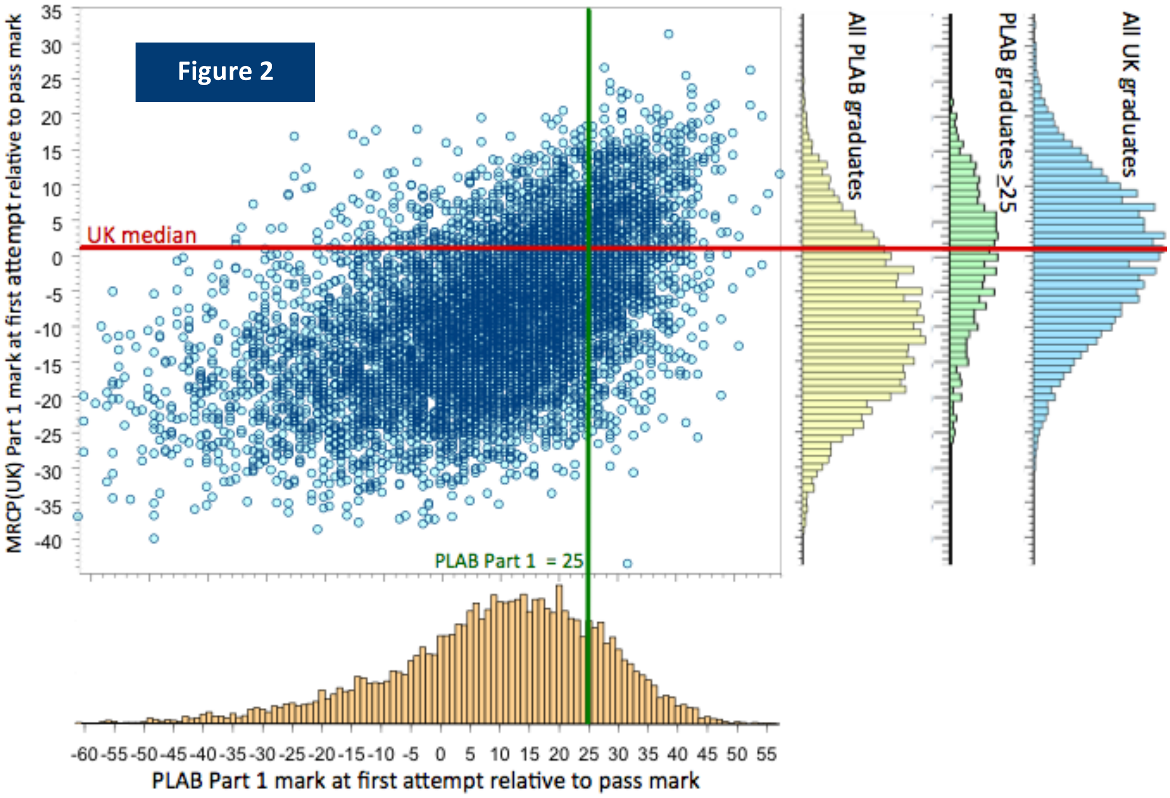
* PACES marks are scored in terms of the oldPACES scoring system

Note: Cohen's d, which is a standard measure of effect size, is calculated as the difference in means divided by the standard deviation of the UK graduates, who are treated as the reference group

The clear difference in the performance of UK and PLAB candidates, coupled with the fact that for the PLAB candidates there is a clear correlation between PLAB scores and subsequent performances in MRCP(UK) and MRCGP, forces the next question, of what pass mark there would have to be in PLAB to achieve equivalence between UK and PLAB graduates.

We will describe two separate ways of evaluating equivalence in terms of estimating what pass mark in PLAB, particularly for Part 1, would be necessary in order to achieve comparability of UK and PLAB graduates at MRCP(UK) and MRCGP. The two methods are somewhat different approaches, but together give a clearer comparison of PLAB graduates with UK graduates.

Method 1: Equating to median performance of UK graduates. A typical UK graduate taking MRCP(UK) Part 1 is at the median level of performance on that assessment, so that half of graduates perform better and half perform less well. In figure 2 the distribution of marks of UK graduates taking MRCP(UK) Part 1 is shown at the far right in blue.



On a scale relative to the pass mark of zero, their median mark is +1.03, shown as the thick horizontal red line; UK graduates are therefore slightly more likely to pass than to fail MRCP(UK) Part 1 on their first attempt). The marks of the PLAB graduates at MRCP(UK) Part 1 are shown in the pale yellow histogram, third from the right. The distribution is clearly shifted downwards relative to the UK graduates, and in fact the mark of +1.03 which is at the median for UK graduates is on the 81.1th percentile of the PLAB graduates.

The distribution of marks at first attempt on PLAB Part 1 is shown in the horizontal orange histogram at the bottom. Finding a pass mark that results in equivalence with the UK distribution requires a pass mark to be set at PLAB Part 1 which results in a distribution of MRCP(UK) Part 1 scores in PLAB graduates which has a median of +1.03, the same as that for UK graduates. That can be estimated by considering only PLAB graduates with a mark higher than some threshold, which can be adjusted until the median of those taking MRCP(UK) Part 1 is +1.03. The dark green vertical line in Figure 2 is set at a threshold ('pass mark') of +25. The MRCP(UK) Part 1 marks of all those PLAB graduates to the right of the dark green line are shown in the middle, pale green histogram at top right, and for this group the median is very close to +1.03, half being above that value and half below it. On that basis, a pass mark for PLAB1 of **+25** compared with the present pass mark would result in a group of PLAB graduates performing equivalently on MRCP(UK) Part 1 to UK graduates. Of the 7823 PLAB graduates taking MRCP(UK) Part 1, only 1409 (18.01%) are in the green distribution, and can be regarded as equivalent to the UK graduates.

A similar analysis can be carried out for MRCP(UK) Part 2 in relation to PLAB Part 1. For UK graduates the median is +6.01, a value which is at the 82.5th percentile for PLAB graduates. Adjusting the threshold for PLAB Part 1 until the PLAB graduates have a median of +6.01 requires a threshold of **+32** compared with the present PLAB1 pass mark of zero; on that basis only 516 of the 5133 PLAB graduates currently taking MRCP(UK) Part 2 can be regarded as equivalent to UK graduates.

MRCP(UK) PACES is more problematic for calculating an equivalent threshold. The UK graduates have a median mark of +2.0 at PLAB, a mark which is at the 91.5th percentile for PLAB graduates. However it is not possible to get a threshold for PLAB Part 2 which produces a median of +2.0, there simply being no candidates left. The threshold is therefore **>+18**.

The analyses for MRCGP are similar, though both the AKT and CSA produce a result: the median AKT mark for UK graduates is 21 and for PLAB graduates is 5, and the median CSA mark for UK graduates in 14 and that for PLAB graduates is -5. To achieve an equivalently performing median candidate as between UK graduates and PLAB candidates on first attempt, would require the pass mark for PLAB Part 1 to be increased by **+35** marks and that for PLAB Part 2 to be increased by **+10** marks. Each of these approaches would reduce the number of PLAB graduates in GP training greatly (PLAB1 AKT from 3160 to 106, leaving just 3%; PLAB2/CSA from 1388 to 114, a reduction to 8%).

Taken overall, with Method 1 it is possible to set a pass mark for PLAB Part 1 which results in a distribution of marks at MRCP(UK) Part 1, MRCP(UK) Part 2 and the MRCPGP AKT which has a similar median and distribution to that of UK graduates. However the pass marks required are substantially higher than at present, being 20, 32 and 35 marks higher than at present, the average increase being **+29**, so that the pass mark for PLAB Part 1 would need to be raised from a typical value of 126 (63%) to about 155 (77.5%).

We are of course aware that such changes in the pass mark would have dramatic effects upon the overall pass rate, with inevitable knock-on effects for workforce planning, etc. Those are much wider issues and cannot be considered further here.

Method 2: Comparison with performance of graduates from different UK medical schools. The second method takes a different approach. In a previous analysis of the performance of graduates of different UK medical schools at MRCP(UK) [10] there were clear and large differences in performance at MRCP between graduates of different medical schools. That result extended and developed the much earlier analysis of Wakeford *et al* [17] for MRCGP, and has been repeated in the recent analyses of the MRCGP[xv]. Similar differences between medical schools have also been reported for FRCA [18] and MRCOG [19]. The ordering across medical schools is broadly similar in all of the studies, with some variation due to sampling differences, and perhaps also differences in medical school training.

Since UK graduates differ between medical schools, the question of equivalence can also be addressed by asking at what level the performance of PLAB graduates is similar to that of those from UK medical schools[xvi].

A near equivalent which is useful for the present purpose is to take the performance PLAB graduates on the PLAB assessment and to divide them into twelve equally-spaced subgroups[xvii] according to their performance at PLAB Part 1 (or Part 2), and then compare those groups with graduates of individual UK medical schools. The key question concerns which of the twelve sub-groups produce performance equivalent to that of graduates from individual UK medical schools. The sub-groups were based on steps of five marks for PLAB Part 1 and three marks for PLAB Part 2, so that groups in each case can be directly compared with the marking scales for each assessment.

Figure 3: Mean performance at MRCP(UK) Part 1 of graduates of UK medical schools (open points) in relation to performance of PLAB graduates (red points) divided into twelve groups according to PLAB Part 1 mark at first attempt.

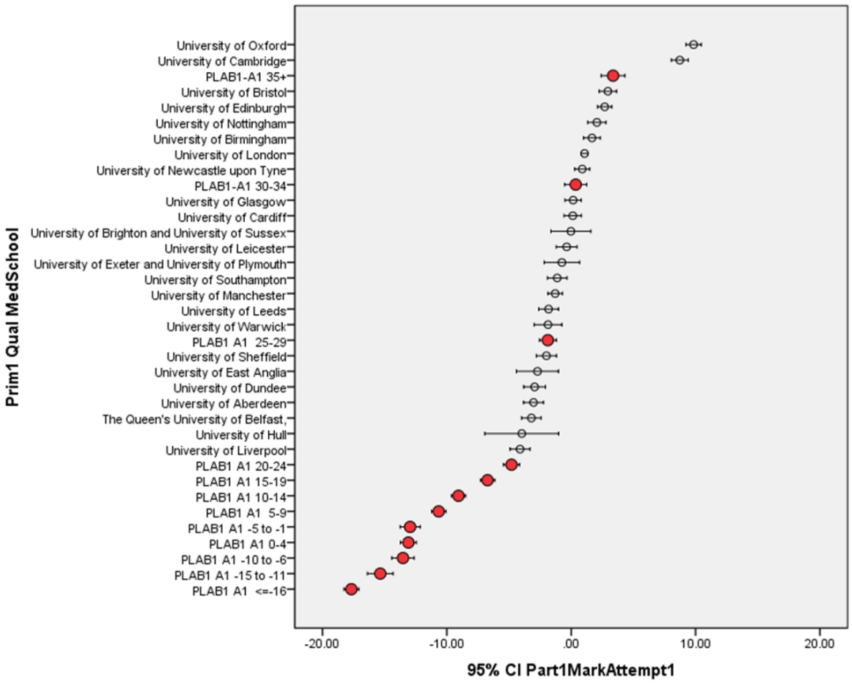


Figure 3 shows results for MRCP(UK) Part 1. The open black points show performance of graduates of UK medical schools, ranked from highest to lowest[xviii]. Differences between UK medical schools are highly significant, as can be seen from the narrowness of the 95% confidence intervals, but they are not of

[xv] See the MRCGP annual reports from 2008 onwards at http://www.rcgp.org.uk/gp-training-and-exams/mrcgp-exam-overview/mrcgp-annual-reports.aspx.

[xvi] Although it can be assumed that if there are differences in performance between UK graduates according to the medical school they attended, there will also be differences between IMGs according to the medical school they attended, with perhaps much larger variation. Although IMG medical schools are known, it is not practical or meaningful to compare IMGs by medical school they attended. Numbers of IMGs from most individual medical schools are small, and there is no reason to believe that IMGs from those schools choosing to take PLAB are a random sample of the graduates. Those taking PLAB may be the best in their year, being high flyers who wish to work in prestigious UK medical schools, they may be the weakest graduates, who could not get good jobs in their home countries, they could be taking PLAB because they are economic migrants, because they have family or other personal reasons to want to be in the UK, or they could be political refugees, fleeing their home countries out of necessity. They will also include some British school leavers, determined to study medicine but who have not managed to secure a place in a British medical school. Few of these factors are measurable and none can easily be taken into account in any analysis.

[xvii] Sub-groups based on PLAB1 and PLAB2 were originally produced as deciles, which while it has the advantage that the groups are of equal size has the disadvantages that a) the boundaries are different for the MRCP(UK) and MRCGP datasets as the samples are different, and b) the boundaries are not obviously on the scale by which PLAB1 and PLAB2 are marked. We therefore chose instead to sub-divide the data into twelve subgroups with equal numerical intervals corresponding to five marks on the PLAB1 scale and three marks on the PLAB2 scale. The majority of the groups were also of broadly similar size to the numbers of graduates from UK medical schools, making standard error bars similar and hence easy to compare.

[xviii] Medical schools have only been included if a reasonable number of recent graduates have taken MRCP(UK) or MRCGP, and hence some new medical schools are excluded as graduates have not yet taken these exams in sufficient numbers.

substantive interest here. Performance of PLAB graduates is shown by the twelve red points, which correspond to different marks on the first attempt at PLAB Part 1, scaled relative to the pass mark. The highest scoring PLAB group ("PLAB1-A1 35+", who scored 35 or more marks above the PLAB1 Pass mark) has a mean performance equivalent or better than the mean performance of graduates of all but two of the UK medical schools (Oxford and Cambridge), and is clearly achieving highly. Similarly the second and third groups ("PLAB1-A1 30-34" and PLAB1-A1 25-29" have a mean performance better than or similar to the graduates of many UK medical schools. The fourth group (PLAB1-A1 20-24"), with PLAB scores from 20-24 points above the pass mark, has a mean performance which is not distinguishable from the mean performance of the lowest performing UK medical schools, using the Ryan-Einot-Gabriel-Welsch Q post hoc test for differences between means with a (corrected) p level of .05. The eight remaining PLAB1 groups, from "PLAB1 A1 15-19" downwards all perform at a significantly lower average level than graduates of any of the UK medical schools.

Taken overall, Figure 3 suggests that the top four PLAB1 groups are equivalent to graduates from UK medical schools when taking MRCP(UK) Part 1, whereas the lower eight groups perform less well. Those results suggest that an equivalence level is within **+20 to +24** points range above the current pass mark, a value which is similar to the value of +25 calculated using the median method (above).

Similar calculations can be carried out for MRCP(UK) Part 2 and PACES, using PLAB Part 1 for MRCP(UK) Part 2, and PLAB Part 2 for MRCP(UK) PACES, and plots similar to figure are shown in Figures 4 and 5. For MRCP(UK) Part 2 (Figure 4), the mean performance of only the top two PLAB groups (30-34 and 35+) is equivalent to UK graduates, making **+30 to +34** the likely equivalence. For PACES (Figure 5), only the very top group ("PLAB2-A1 18+") has a mean performance equivalent to graduates from UK medical schools, making the equivalence **> +18**.

Figure 4: Mean performance at MRCP(UK) Part 2 of graduates of UK medical schools (open points) in relation to performance of PLAB graduates (red points) divided into twelve groups according to PLAB Part 1 mark at first attempt.

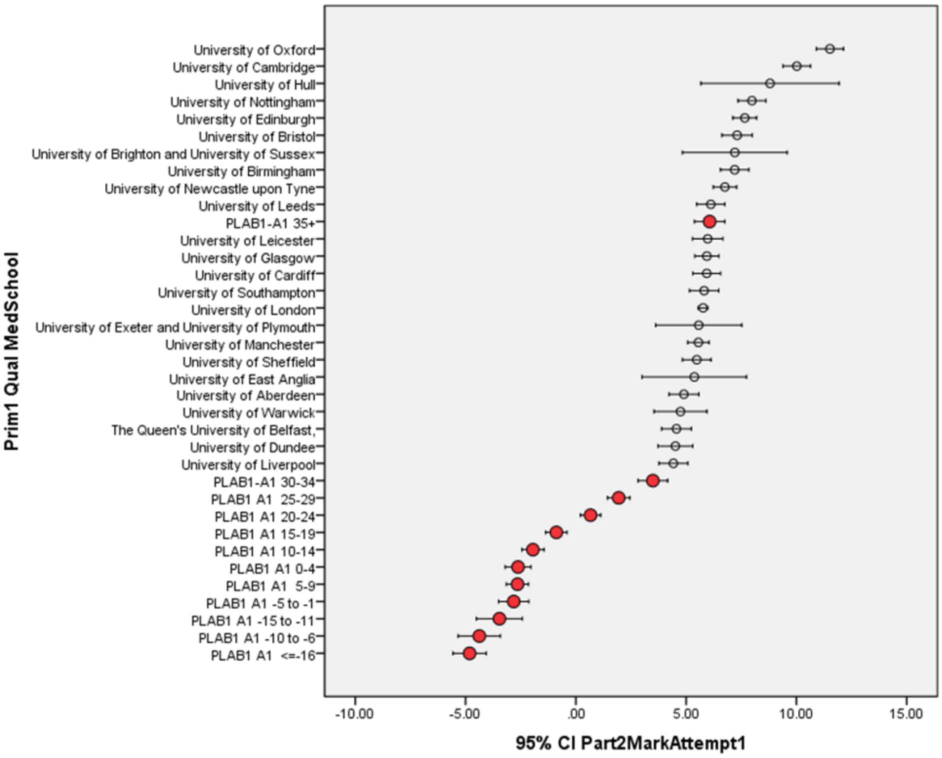
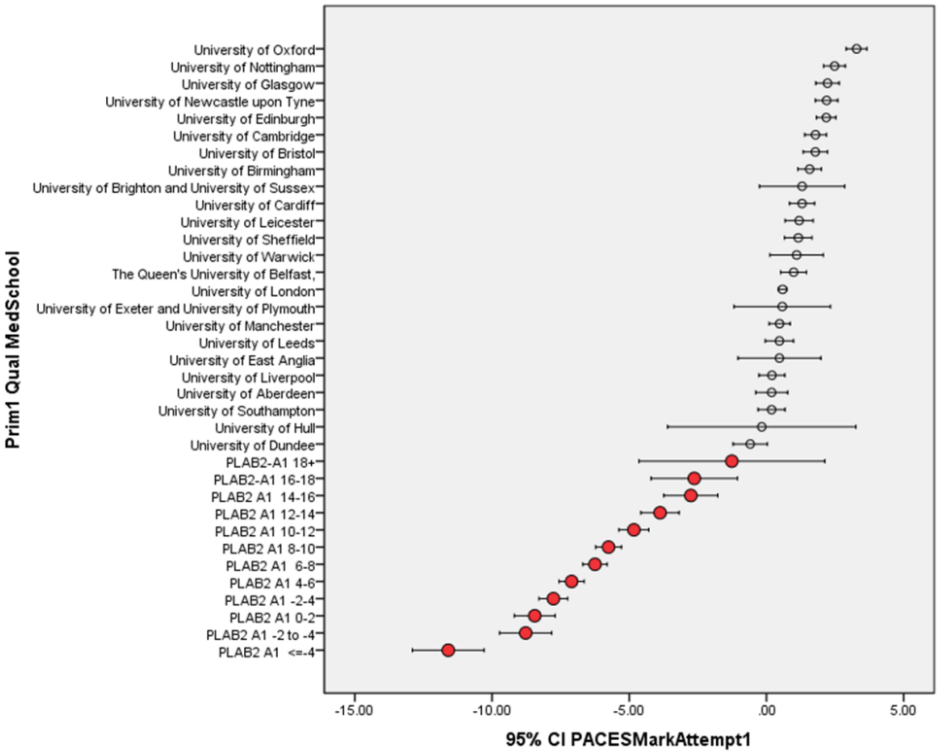


Figure 5: Mean performance at MRCP(UK) PACES of graduates of UK medical schools (open points) in relation to performance of PLAB graduates (red points) divided into twelve groups according to PLAB Part 2 mark at first attempt.



Analyses for the MRCGP are shown in Figures 6 and 7, comparing performance in the AKT with PLAB Part 1 and in the CSA with PLAB Part 2. Note that the MRCGP data sub-divide London medical schools and also include a separate group of EEA graduates, shown in green, who are not required to take PLAB.

In the AKT (Figure 6), the top PLAB group, with PLAB Part 1 scores of 35 or more above the pass mark is well within the body of the kirk, as it were. The next two groups down (PLAB Part 1 mark 25-29 and 30-34) are also clearly equivalent in performance to graduates of many UK schools. The next two groups, 20-24 and 15-19 are equivalent to graduates of the lower scoring UK medical schools, meaning that a probable equivalence is at **+15 to +19**. This method therefore suggests a lower passing standard for 'equivalence' than did Method 1, which estimated a value of +35.

Figure 6: Mean performance at the MRCGP AKT of graduates of UK medical schools (open points) in relation to performance of PLAB graduates (red points) divided into twelve groups according to PLAB Part 1 mark at first attempt. EEA graduates are shown in green.

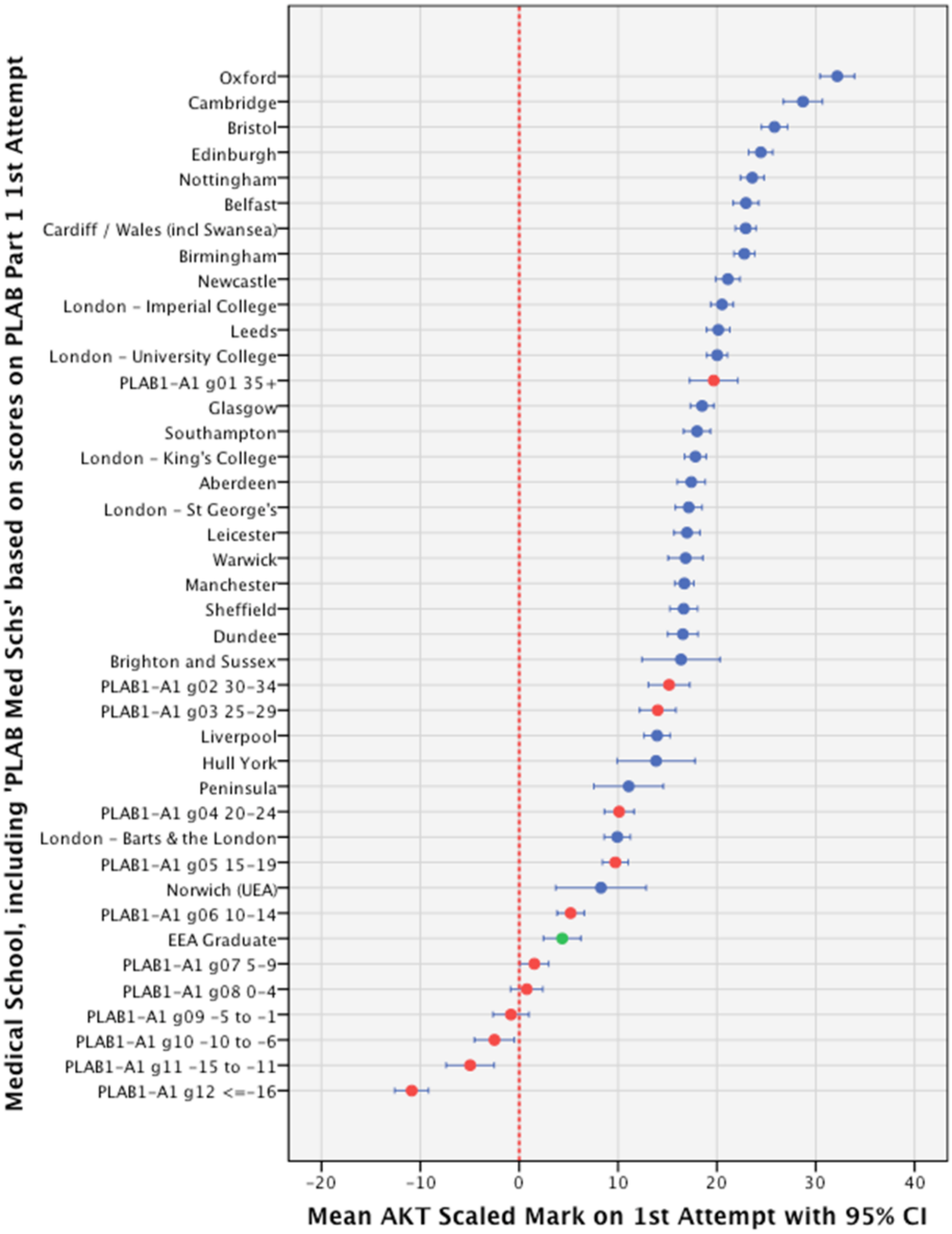
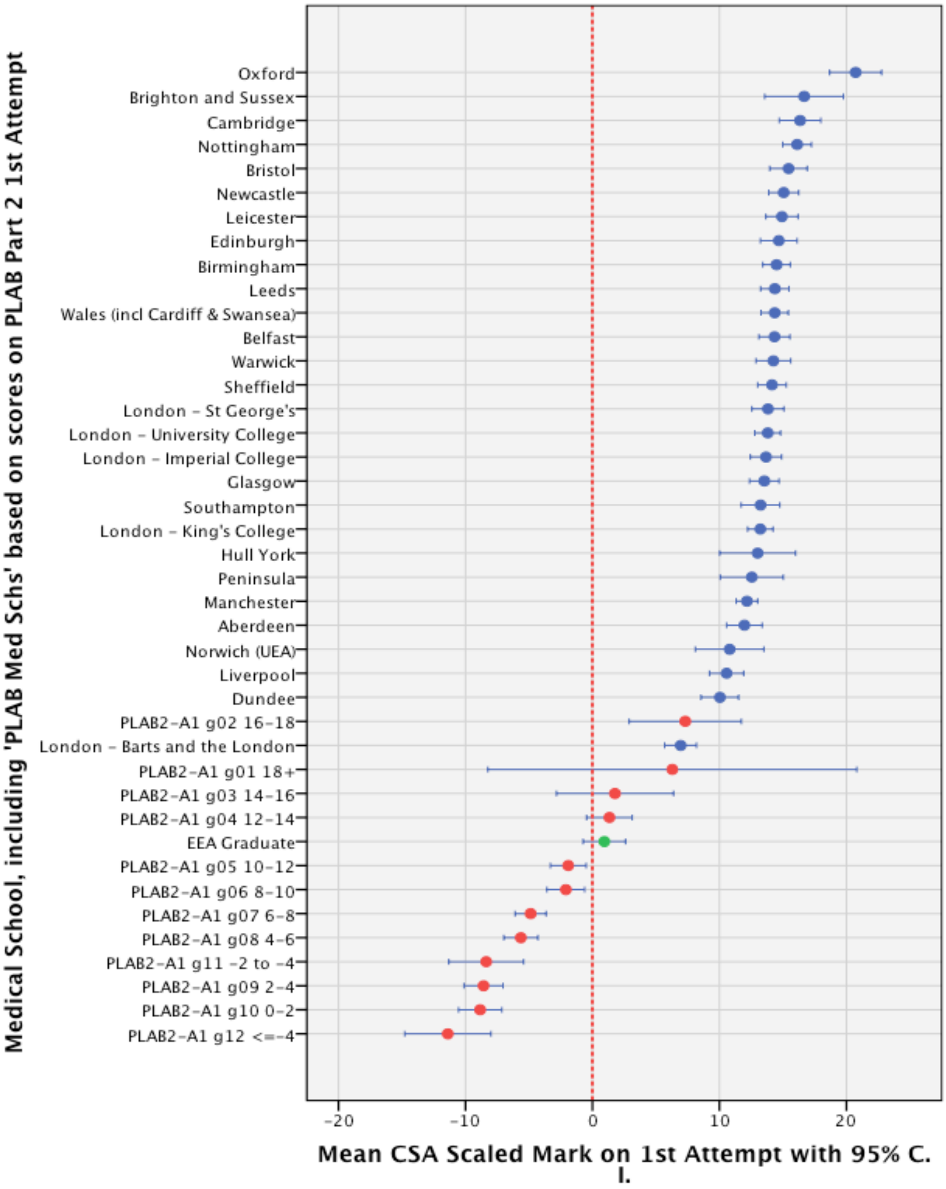


Figure 7: Mean performance at the MRCGP CSA of graduates of UK medical schools (open points) in relation to performance of PLAB graduates (red points) divided into twelve groups according to PLAB Part 2 mark at first attempt. EEA graduates are shown in green.



In the CSA (Figure 7), only the 16-18 group and the (very small) 18+ group are equivalent to the lowest scoring UK medical school, suggesting that PLAB Part 2 scores of **+16 to +18** would be necessary for 'equivalence'. This contrasts with the proposal under Method 1 that the PLAB Part 2 passing standard should be increased by only 10 marks. An increase of 16 would remove almost 99% of PLAB graduates entering the CSA.

Overall estimate of the equivalence levels of PLAB1 and PLAB2

A simple comparison of the mean performance of UK and PLAB graduates on MRCP(UK) and MRCGP makes clear that PLAB graduates are underperforming on these postgraduate examinations. Many of the PLAB candidates taking MRCP(UK) are on UK postgraduate training programmes, and all of the PLAB candidates taking MRCGP are on UK training programmes.

The two methods described here (equating medians, and comparing performance of UK medical schools with twelve groups of PLAB graduates) give somewhat different results, although the broad picture is clear enough, and can be summarised easily (Table 5):

Table 5: Estimated change in the pass mark of PLAB 1 and PLAB 2 to produce equivalence, using the two separate methods described in the text

Values are indicated as marks relative to the current pass mark and are in raw marks on the scales of the examinations themselves. Note that the marking scales for PLAB Part 1 and PLAB Part 2 are not comparable.	**PLAB Part 1**			**PLAB Part 2**	
	MRCP(UK) Part 1	MRCP(UK) Part 2	MRCGP AKT	MRCP(UK) PACES	MRCGP CSA
Method 1: Equivalence of medians	+25	+32	+35	> +18	+10
Method 2: Comparison with UK medical schools	+ 20 to +24 (+22)	+ 30 to +34 (+32)	+15 to +19 (+17)	>+18 (+18)	+16 to +18 (+17)

1. PLAB graduates based on the current pass marks for PLAB Part 1 and PLAB Part 2 underperform substantially when taking MRCP(UK) and MRCGP examinations, being about 1 standard deviation below the performance of UK graduates.

2. Estimating an equivalence level of PLAB Part 1 for the knowledge assessments of MRCP(UK) Parts 1 and 2 and MRCGP AKT suggests that a pass mark of the order of **+27** marks higher than at present would be close to equivalence (+31 based on Method 1 and +24 based on Method 2). Since the PLAB Part 1 typically has nearly 200 questions, in terms of percentage of items correct, the pass mark would need to be moved from its present level of about 63% to about 76% or so. Needless to say that would have a substantial effect on the pass rate, only about a quarter or so of current PLAB graduates achieving those levels.

3. For PLAB Part 2 both methods find that a substantially higher pass mark is needed, there being barely any level of attainment at PLAB Part 2 which is equivalent to the performance of UK graduates, so that only the very top performers appear to be equivalent to UK graduates. Averaging across the estimates, the pass mark would seem to need to rise by about **+16** marks (+14 based on Method 1 and +18 based on Method 2). Some of the problems with estimating may result either from the assessment not stretching candidates at the top end, or from the relatively low reliability of Part 2, an aspect of the assessment which has been considered elsewhere [15], and inevitably makes its predictive power less than is desirable.

How is the standard setting for PLAB carried out?

From the previous section it appears that the pass marks for both PLAB Part 1 and Part 2 may be set at too low a level since despite an intention that PLAB graduates should be equivalent to UK graduates at the end of the first Foundation Year, in practice the PLAB graduates appear to be lagging substantially behind the UK graduates in their progression through core medical and general practice training, with substantially lower performance at MRCP(UK) and MRCGP. It is therefore necessary to ask about the details of how standard setting is carried out. This report will concentrate on PLAB Part 1, since it is probably the more important 'gate-keeper', coming first and having a lower pass rate than PLAB Part 2. Of course standard setting in PLAB Part 2 also needs considering (and has in part been considered by one of us [ICM] in a separate report for the PLAB Review[15]).

Standard setting in PLAB Part 1

The PLAB Part 1 exam currently uses a variant of the Angoff standard setting procedure. Prior to September 2004 standard-setting had used a one-off Angoff in 2000 followed by linear test equating based on about 30 to 40 marker questions[xix]. The current standard setting process is described on the website thus:

> "Standard setting: The pass mark for each examination is set using the Angoff method, in which a panel of trained and experienced clinicians decide what percentage of minimally competent doctors at the appropriate stage of training would answer each question correctly. The Angoff method of standard setting is internationally recognised and it ensures that examinations are of a consistent standard over time." http://www.gmc-uk.org/doctors/plab/advice_part1.asp#14

It is an accurate description, but inevitably some of the details are missing. For providing details and data for the descriptions below we are indebted to Katharine Lang and Michael Harriman for their extensive help, and in particular providing statistics on standard setting judgements from March 2001 to June 2013. The process of PLAB Part 1 Standard Setting is currently changing somewhat, and the description that follows mainly describes the process from 2004 until 2012 (which is when the candidates who subsequently took MRCP(UK) or MRCGP would have taken the exam[xx]). The 2004-12 standard setting took place as follows:

1. The standard-setters (judges) are members of the PLAB1 Panel and come from a representative range of disciplines, their specialities consisting of Hospital Medicine (2), Accident and Emergency Medicine (3), Obstetrics and Gynaecology (2), General Practice (2), Surgery (2), Psychiatry (1) and Paediatrics (1). Not all standard-setters are present at all occasions, and there is no specific quorum, and the norm would be from 8 to 10 standard-setters[xxi]. The median number of standard-setters in the data we were provided was 10 (mean=9.74, range 7 to 12). Unsurprisingly the range of marks was significantly greater when there were more standard-setters (r=.134, p<.001).

2. Members of the PLAB1 Panel are currently appointed for four years but with the possibility of renewal. There is therefore a slow rotation of members. When new members are appointed, typically about twice a year, the group as a whole discusses the benchmark (the 'just-passing', 'minimally competent' candidate). There are also practice runs with dummy questions to allow new members to feel comfortable with the process.

[xix] The 2003 PLAB review discussed standard-setting and said: [para 36] "Part 1: When the extended matching question examination was introduced in July 2000, the standard was set using the Angoff method, in which a series of expert judges decided what percentage of minimally competent doctors at the appropriate stage of training would answer each question correctly. From these results a pass mark was derived. This standard was maintained through using a number of the same questions in two different examinations. Comparing candidates' performances on these questions in the two examinations and adjusting the pass mark accordingly ensured that the examinations were all of a similar standard." The 2003 review then recommended the use of an Angoff method in the future: [para 37] "The Angoff method of standard setting is internationally recognised and, as such, fulfils the criteria that give such systems credibility. However, there is a danger that a drift in the standard may have occurred over the years. We have, therefore, concluded that all the items in the bank should be subject to an Angoff standard setting exercise. We recognise that it would not be practical to undertake one Angoff exercise on the whole bank – it would take far too long. We recommend, therefore, that an Angoff exercise should be undertaken on each examination paper, as it is prepared. Over time, a standard will be set on every question in the bank and papers can be created, using this information, to ensure a consistent standard."

[xx] Doctors taking PLAB Part 1 between 2000 and 2004 took the version of the examination in which a single standard was set by the Angoff method in 2000, and then a form of statistical equating was used. It is unlikely that the standard was radically different from that used at present or changes in the pass rate would have been seen.

[xxi] Recently there has been a move to ask standard-setters to carry out the process individually when they happen to be visiting the GMC, perhaps for some other reason.

3. Standard-setting is usually carried out four times a year, in March, June, September and November.

4. Angoff standard-setting is carried out only for new items or for items whose wording has been changed, otherwise standards from previous outings of the questions are used. For the ten diets from March 2011 to June 2013, 887 items were standard-set, a mean of 88.7 items per occasion. Each exam would have had 200 questions, and therefore about 45% of items for each diet are new or modified and required standard-setting.

5. The standard setting process takes place in a group at the Panel meeting, and for reasons of security standard-setters are not sent the items in advance. Items are viewed, in effect, "under exam conditions". Standard-setters spend a while reading the questions sequentially and individually, for each one, making a judgement of the standard. In particular:

    a. Standard-setters are not provided with correct answers either before making their judgement or after making their judgement, and neither are they asked to indicate which they think is the correct answer. The rationale for this is that the items are thought to be relatively easy and hence all standard-setters will know the correct answer for all questions. It is also the case that most standard-setters should have seen the questions about six months previously at the 'question selection meeting', when answers are provided and standard-setters can question the keying.

    b. Standard-setting is carried out, "in terms of the likely probability that the candidate would know the correct answer to the question as it appears in the examination paper as opposed to whether the doctor could deal with the equivalent scenario in an authentic work situation. This approach takes into account whether the presentation or wording of a question affects the likelihood of the candidate answering correctly" (Email from Katharine Lang).

    c. As they read the questions, standard-setters write on the sheet their estimated standard, which is expressed as a number from 0 to 10, 5 indicating that they think 50% of just-passing candidates *would* get the answer correct, 6 that 60% would get it correct, etc. Items are best-of-five and therefore answers of 0, 1 or 2 are strongly discouraged, and indeed did not occur during 2011-13. The range of marks is therefore from 3 to 10 (30% to 100%). Of 8643 judgements, the numbers at each final level (after any adjustments) were 3: 48 (0.6%); 4: 537 (6.2%); 5: 1766 (20.4%); 6: 2663 (30.8%); 7: 2190 (25.3%); 8: 1204 (13.9%); 9: 229 (2.6%); 10: 6 (0.1%).

    d. There is no detailed record of the time standard-setters take in standard-setting the questions, but it is of the order of two hours in total, firstly for reading and judging each of the items, and then calling out judgements and discussing discrepancies. That is about a minute and a quarter per question. It is broadly similar to the time provided to candidates, who have three hours to answer 200 questions (an average of 54 seconds per question).

    e. When all standard-setters have made judgements on all questions, the standard-setters then go around the table calling out the numbers they have written down, which are collated by the Panel Secretary[xxii]. Standard-setters take it in turns to be the first to call out their judgement on each item.

    f. If standard-setters differ by 5 or more points then the item is discussed and the standard-setters at the extremes are invited to alter their responses[xxiii]. Statistics were

[xxii] There is little likelihood that examiners would change their judgement at this stage in response to other examiners' judgements since all question sheets are collected at the end of the session.

[xxiii] "We try to apply the 'egg-timer' rule: if a single item is causing a lengthy battle it is probably grounds to suppress it from the results."

only available for judgements *after* this process was carried out, when the range of standard-setter marks varied from 1 (1.1%) through 2 (18.8%), 3 (45.8%). 4 (33.1%) and 5 (1.1%) for the 887 questions. Given that marks can only be in the range 3 to 10, a discrepancy of 5 is only possible when marks are 3 to 8, 4 to 9 or 5 to 10. Since marks of 3, 9 and 10 are rare, almost all of the discrepancies are probably in the range 4 to 8. Without being given any statistics on the number of items discussed, it is probably about 5 - 10% at maximum.

6. After new questions have been standard-set, the Panel reviews the performance of items in the previous diet. Items may be rekeyed for that diet, may be withdrawn, and items may be modified for when they go into the bank again. During the review, standard-setters are informed of the correct answer and of other relevant statistics, the discussion being led by a psychometrician.

Potential problems with standard-setting

The equivalence exercise suggests that the standard currently being set may be too low given subsequent performance of PLAB graduates at MRCP(UK) and MRCGP. Although the Angoff method has been used to set the standard, and it is undoubtedly a recognised method for carrying out such tasks[20], it is known to be far from perfect, and is perhaps better described as *faute de mieux*, being less bad than many other methods, in part because its faults and problems are at least well recognised and understood, allowing some interpretation of the outcomes.

The PLAB1 Angoff method is not standard (but few implementations are[21]). Here we review various aspects of the Angoff procedure to assess whether they may be responsible for the standard being set lower than seems necessary for equivalence.

1. **Training of standard-setters on the level of the 'just-passing' (borderline) candidate.** Using an Angoff method requires that standard-setters have a clear mental model of the ability level of a borderline candidate (just-passing candidate). Induction of new members to the Panel currently does take place, and the continual rotation of panel members helps to ensure that benchmarking by existing members is refreshed. Panel members undoubtedly differ to some extent in their stringency, some setting consistently higher standards than others (e.g. Appendix 4: Table 2), and that probably relates in part to the same factors as were seen in a recent study where consultants viewed video tapes of FY1 doctors and rated them on competence [22].

2. **Standard-setters do not see items in advance**. Many users of the Angoff method send items to the standard-setters in advance for them to standard set in their own time, and examinees then email their judgements, or bring them along to the meeting. That encourages more reflection and, if necessary, the opportunity to consult literature to ascertain current good practice, to check on the accuracy of material, and so on. The argument for the current practice is based both on security and also on the timing for standard-setters corresponding to that the candidates have when they see the questions. The latter may be true, but standard-setters are making a different judgement, they may be inexperienced in the discipline, they have no opportunity to check up on information, a problem compounded perhaps by there being no correct answers provided, and little or no feedback or discussion for most items. There may also be a security risk with sending out items, but it surely could be minimised with trusted and experienced standard-setters.

3. **The process of standard-setting is relatively rushed.** The current system encourages standard-setters to carry out the standard-setting process quickly (standard-setters have to wait upon the slowest of the standard-setters), but that probably does not encourage reflection or critical consideration.

4. **Standard-setters are not provided with correct answers**. Angoff methods vary as to whether or when standard-setters are provided with correct answers. Some exam systems find that an

effective method is to provide questions without answers, require standard-setters to make their judgements in advance without correct answers, and then at the discussion phase the correct answer is provided. Perhaps unsurprisingly, when standard-setters find they have given a wrong answer (and omniscience is rare in standard-setters even at this level), they tend to modify their judgement of the correct standard. Without answers being provided at all, there is a risk that standard-setters, who mostly have been qualified a decade or more, and in some cases more than three decades, will be trading on out-of-date knowledge (and how many paediatricians or psychiatrists or general physicians or surgeons have done obstetrics since qualification?)[xxiv]. Correct answers, as standard-setters always say, act as valuable CPD, particularly in specialties other than their own.

5. **There are no 'reality checks'**. Examiners in general are frequently surprised and sometimes shocked by the discrepancy between what they think candidates know and what they actually do know. The differences are often in both directions, candidates finding some items much easier and others much harder than examiners think. Questions which are much easier are sometimes explained by current undergraduate or other courses emphasising that material. Items which are more difficult may be due to material not being in current practice, despite being in textbooks, or because of changes in disease presentation. Reality checks can be provided in several ways:

    a. **Normative data from the diet**. Although standard-setting is currently carried out prior to an exam being sat, that is not essential, and were the standard-setting to be carried out a week or so after the exam then normative data could be provided[xxv]. The role of normative data is controversial [23,24] but the risk of not using it is that standard-setters are guessing blindly in the dark.

    b. **Normative data from previous diets**. Presentation of material on questions from previous diets can act as a reality check, not least as standard-setters see that many items were much harder or much easier than anticipated during standard-setting, providing a useful standard check. In PLAB Part 1 that could readily be introduced by placing the review of previous questions *before* standard-setting for the present diet.

    Although normative data is available at some stages during the Panel's deliberations, it is not available during standard-setting itself.

6. **The distribution of standards which are set**. A strange finding in almost all uses of Angoff is that, despite a large amount of work on the part of the standard-setters, the pass mark seems always to be around 60%, and usually between 55% and 65%. Furthermore the distribution of pass marks set for individual questions is quite narrow (and for PLAB Part 1, 95% of the pass marks are between 53% and 75% (despite some questions being very easy or very difficult for just-passing candidates). Those effects can be seen in the data analysed by Clauser *et al* for the NBME examination [23] and for the MRCP(UK) examination [24][xxvi]. Likewise in a meta-analysis of 113 Angoff assessments across a range of non-medical disciplines the mean was 60.6% (SD 11.5) [25]. The reasons for the concentration of standard-setters judgements around 60% are probably complex but in part they reflect the fact that for best-of-five examinations where there is a 20% guessing rate, the mid-point from 20% to 100% is 60% (and that is also the point where

[xxiv] In one of her several very helpful emails, Katharine Lang commented on what happens if a standard-setter feels they don't know the answer to a question: "If the answer isn't obvious to a particular examiner, that would be an occasion when they might ask for verification – or they might think 'If I don't know, neither will the candidates', and score accordingly. This is the kind of situation where the examiner's speciality makes a difference and will usually provoke a discussion. For example a surgeon might say 'I have never heard of X' and an obstetrician would reply 'We see dozens of these a week and it is usually the juniors who deal with it', the rest of the group would join in the conversation and the outliers would likely adjust their scores nearer to the mean."

[xxv] Until recently standard-setting was carried out after the exam was sat, and although normative data was not provided it could have been, and was seen later in the day during review of the questions.

[xxvi] Declaration of Interest: One of us, ICM, is a co-author on the latter paper [24].

an exam has maximal discriminatory power). For standard-setters who do not like being the hawk or the dove, guessing 60% (or perhaps 50% or 70% on occasion), is a 'safe option'. Some of the 13 PLAB standard-setters are undoubtedly not using extreme judgements. The median number of more extreme judgements, defined as 3,4, 8,9, or 10, was 27%, but at the lower end, three standard-setters made only 4%, 10% and 11% of such judgements compared with the three at the high end who made 29%, 32% and 39% more extreme judgements, three times as many. There is an argument that only the latter standard-setters are carrying out the standard-setting correctly, in so far as they are generating a wide range of standards (and it can be taken for granted that a standard-setter who simply said 6 (60%) for each item would not be carrying out a worthwhile job).

7. **The questions which are set**. Little information was available to us about the questions which had been set (and hence whose standards are to be set), and we have not seen question-setting in action. If though it were the case that the questions set were too easy, and the standard-setters mostly generated standards of about 60%, then the standard for the exam as a whole would be too low. Question-writers are asked to set questions at a level suitable for the end of FY1, but it is not clear how well they can do that. Probably further investigation into the question setting would be desirable.

8. **Discussion is limited**. The PLAB Part 1 Angoff has very little discussion of items (in contrast with other exams where often each item is discussed the hawk and dove on each question being asked to comment and all standard-setters being given the option to change their judgements). Although this often has more effect on the variance (and hence the reliability) of standards which are set, the mean standard can also alter [24]. A meta-analysis across a range of non-medical disciplines found that the average passmark was set at 57.6% without discussion and at 64.9% with group discussion[25]. Group discussion also provides an important opportunity for non-specialist standard-setters to update what may be outdated information on current practice, as well as to revise and reiterate the group's perceptions of the 'just-passing' or borderline candidate. At present there is an argument that for most questions, the PLAB Part 1 standard-setters are, in some sense, blind – they generate a number, with no indication of whether they know the correct answer to the question, whether other standard-setters think the same, whether subject-area standard-setters think the same, and the reasons for any of the other decisions which are made. All that a standard-setter really finds out is that sometimes the marks they call out are similar to those of other standard-setters and sometimes they are not. That does not seem a solid basis for a robust standard-setting procedure.

9. **The use of 'would' vs 'should'**. A recurrent issue concerning Angoff methodology, which is often discussed by standard-setters but seems hardly to be discussed in the literature itself, is the use of the terms 'would' or 'should'. Angoff's original formulation, and most methodological descriptions since, have used the term 'would'. However not all UK exams use 'would', and the MRCP(UK) has used 'should' for its Part 1, Part 2 and Specialty Certificate examinations. PLAB Part 1, in common with many other exams, asks, "What percentage of minimally competent doctors at the appropriate stage of training *would* answer each question correctly?" (our emphasis). Often however, to English ears, there seems a contradiction here with the prescriptive sense in which it is felt that standards need to be set. Although it might be tempting to put the problem down to eccentric American usage[xxvii], this seems unlikely in Angoff's case since in the 1971 article [27] he uses 'would' on 12 occasions (of which six are in the footnote describing the method), and 'should' on four occasions, all of which are prescriptive.

   Part of the problem probably arises because Angoff is only describing the judgements which are made about how the "minimally acceptable person" would perform, and since those are

[xxvii] *The American Heritage Book of English Usage*[26] comments that, "Just as they ignore the traditional rules governing the use of shall and will, Americans largely ignore the traditional rules governing the use of should and would". (p.33).

predictive hypotheticals there is a sense in which 'would' is the correct usage. However, Angoff does not provide any clear definition of the minimally acceptable person.

A much clearer statement of the Angoff process is to found in the Angoff Rating Instructions for the (American) Association of Firearm and Toolmark Examiners (AFTE)[xxviii] assessments, which firstly define "Minimal Acceptable Candidate" as, "a concept that describes what the minimally acceptable candidate *should* be able to do, or *should* know, on the very first day on the job" (our emphases). The description then continues, "... subject matter experts are asked to estimate the probably that a minimally competent applicant *will* get the answer correct". With forensic clarity there is a careful separation of the definition of the minimal acceptable candidate and what they should be able to do, and the predictions about how they would behave in the assessment, given the 'shoulds' in their definition. A risk in any Angoff procedure is that the 'shoulds' in defining the borderline candidate get forgotten, and the exercise is seen instead as a mere form of precognition, a predicting of the future, irrespective of what standards should apply[xxix].

There is a danger, therefore, that conventional Angoff usage of 'would', as in PLAB Part 1, confuses and sometimes misleads standard setters. There also might be a problem if 'would' also allows an inadvertent slippage into norm-referencing, a relative-method masquerading in the guise of an absolute method, as has been suggested by Clauser *et al* [23,24].

10. **Standard-setters differ in their specific content knowledge**. Those carrying out an Angoff procedure should be experts in the content of the examination. PLAB Part 1 is, in effect, equivalent to a 'finals' examination, and therefore covers a wide range of topics, including medicine, surgery, obstetrics, psychiatry, and paediatrics. When examined in UK medical schools those subjects are typically assessed by specialists in the areas, albeit ones who are aware of the specific needs of an undergraduate rather than a specialist examination. The standard-setters for PLAB Part 1 are all experts in a similar range of specialties, but they also make standard-setting judgements on *all* of the specialties. No information is available about the extent to which the standard-setters are aware of current knowledge and standards in other disciplines than their own, but potentially there is a problem if, say, a standard-setter is an emergency medicine specialist and is assessing the standard for psychiatry items (or vice-versa). If standard-setters were to be making judgements based on their own, perhaps much out-dated, knowledge of other specialties, then they might perceive items as harder than they actually are, and then rate those items as more difficult and hence requiring a lower standard. That possibility can be assessed empirically.

    a. The standard-setters were divided into seven specialist groups (OG, Surgery, EM, GP, Psychiatry, Paediatrics, Medicine)[xxx]. The questions are divided by PLAB into 22 groups[xxxi]. *A priori*, ICM made a decision for each question area whether it could reasonably be seen as in one or more of the specialist areas (see Appendix 4: Table 1:

[xxviii] The AFTE website comments, that, "Sometimes referred to as forensic ballistics, firearm identification is the discipline of forensic science that can link a fired ammunition component (bullet, cartridge case, shotshell) to the firearm that fired it. … The Association of Firearm and Tool Mark Examiners (AFTE) is the professional organization for practitioners of firearm and/or tool mark identification" (http://www.afte.org). See www.afte.org/AssociationInfo/certification/Files/Appendix%20G.pdf for the Angoff guidelines.

[xxix] Consider the substantive difference between asking, "*How many newly qualified surgeons **should** be able to remove an appendix safely*?" and "*How many newly qualified surgeons **would** be able to remove an appendix safely?*". Any public body setting standards to ensure patient safety needs, of necessity, to consider the 'should' question.

[xxx] There are thirteen examiners in total, with very significant differences in their mean standard-setting scores, from 6.02, 6.07 and 6.11 for the three doves through to 6.36, 6.54 and 6.77 for the three most hawkish. Notably these differences are also present within specialities, the EM specialists, the OG specialists, the two physicians and the two surgeons being significantly different using post-hoc tests (although the two GPs gave similarly stringent judgements).

[xxxi] Currently there are 20 groups; http://www.gmc-uk.org/PLAB___Part_1_sampling_grid___DC2827.pdf_49191175.pdf

table 1). That is of course subjective, but it is good enough for a first approach to the problem.

b. Marks were all converted from the 0 to 10 scale used in standard setting, to a more conventional 0% to 100% value indicating the expected percentage of just passing candidates who would know the answer. These values can be doubled to give marks on the actual scale used in the PLAB Part 1 exam (which has about 200 questions).

c. The average standard setting mark was calculated for each combination of Standard-setter Specialty and Question Specialty ((see Appendix 4: table 2, and these values were then standardised so that all Standard-setter Specialties had a mean of zero and all question types had a mean of zero (see Appendix 4: Table 3).

d. Average marks were then calculated for those question types which were or were not in a specialist area, these being weighted by the actual numbers of questions in each category.

e. Questions in a standard-setter's specialist area had a weighted mean mark of +1.45% and those not in a specialist area had a mean mark of -1.17%, the difference being 2.62%, indicating that when setting a pass mark in their specialist area, standard-setters set the standard 2.62% higher than when setting in a non-specialist area.

f. Although 2.62% may not appear particularly high, it has already been mentioned that the range of mean standard setting marks is fairly low, with a standard deviation of 1.22 (in raw marks on a 0 to 10 scale), which is 12.2% on the standard percentage scale. The difference of 2.62% between the standard of specialist and non-specialist items is therefore 2.62/12.2 = 0.21 standard deviations, which is non-trivial.

g. Overall there is therefore evidence that standard-setters set a higher standard in their own specialist area than in areas in which they are not expert.

h. The results are compatible with a more sensitive analysis by Verheggen *et al* [5] where standard-setters were not only required to set a standard for exam items, but also were not provided with answers but were required to answer the items themselves. When standard-setters answered items correctly they set the standard substantially higher than when they answered items incorrectly, correctly answered items having a standard about 9.3% higher than the standard for incorrectly answered items. That is a higher value than the 2.62% described here, but it uses information on actual knowledge of correct and incorrect, rather than inferred estimates based on specialty. It seems likely that a value of 9% or so could well be attained if PLAB standard-setters were asked to answer items themselves while standard setting them.

Recent changes in the standard-setting process

PLAB Part 1 has recently introduced changes in the standard-setting process, with panel members rating questions individually, in advance of the exam, their judgements being put together in a spreadsheet, and the Panel members then coming together to discuss their judgements. When there are discrepancies of 5 or more points then the panel as a whole discusses the question, with the outliers discussing the reasons for their scores, the panel as a whole then discussing the question, and everyone then being permitted to change their scores if they wish. Clearly some of these changes are in accord with the comments above about the process. Whether they have had any substantial impact on standards, pass marks, pass rates and predictive validity, it is too early to say, not least as none of the candidates under this standard-setting procedure will have taken MRCP(UK) or MRCGP.

Standard-setting: Discussion and Conclusions

Standard-setting is one of the most difficult processes in examining. As Clauser *et al* comment [24], "Standard-setting is by its nature a policy decision; standards do not exist in nature waiting to be discovered. Although establishing examination standards is not a scientific process, in order to be defensible a standard must have procedural credibility." The question for present purposes is whether the procedure is credible, given a clear statement that the PLAB graduates should be equivalent to UK graduates at the beginning of FY2, and if not, what are the critical components of the standard-setting which reduce its potential credibility. Not least of the problems is the claim that, "The Angoff method of standard setting is internationally recognised and it ensures that examinations are of a consistent standard over time." It undoubtedly is internationally recognised[xxxii], and most medical educators have used it, not least because it has some attractive properties. Reviewers have often criticised it, Brandon [21] concluding his review by stating that,

> "Given the central place of standard setting in current American educational decision making, conclusions about the modified Angoff standard-setting method, which has been studied more than any other method, are uncomfortably tentative and narrow".

Nor does frequent use mean that the Angoff method is easy for standard-setters to use, Verheggen *et al* [5] describing the "tedious work of judging items", and Clauser *et al* [23] commenting that, "content experts struggle with making the required judgements" (and we have heard examiners describe it as "plucking numbers from thin air"). Nor does the method *ensure* a consistent standard over time, with Clauser *et al* [24] commenting that, "in the absence of performance data, content experts have a limited ability to make judgements about the relative difficulty of test items" (p.16).

Standard-setting in PLAB Part 1 is not typical of the ways in which standard-setting is generally carried out. In particular:

1. Standard-setters do not see items in advance of the standard-setting meeting and therefore cannot reflect on them or consult standard textbooks, etc.
2. The standard-setting process is fairly rushed, about 90 or so items having their standard set within two hours or so.
3. Correct answers are not provided either before, during or after standard-setting. It seems likely that standard-setters set lower standards for items for which they do not know the correct answer.
4. Normative data, as a 'reality check', are not provided at the time of standard-setting, either on current or previous items.
5. The range of standards which is set is very narrow (but in that sense, it is not substantially different from many other Angoff exercises, where standards bear little relation to normative data [23,24]).
6. As in most Angoff exercises, the pass mark always seems to be remarkably close to about 60%, a value seemingly unrelated to the difficulty of the questions. The result is that if the questions set are too easy the pass mark of the exam will be too low.
7. There is little proper opportunity for standard-setters to discuss items, only a small minority with very disparate scores being discussed. The result is that standard-setters cannot readily realise that their views are out of alignment with other standard-setters.
8. Standard-setting asks about the proportion of borderline takers who "would" get an item correct. There is a strong argument that the proper term is "should" (and we cannot help noting that in the 2009 edition of *Tomorrow's Doctors*, there are 90 occasions where 'should' is used, and none where 'would' is used). Likewise the 2012 Foundation Programme curriculum uses 'should' 102 times and would only 5 times.

[xxxii] But then in the eighteenth century, phlebotomy was internationally recognised by experts for the treatment of a wide range of diseases when, in all likelihood, it was either ineffective or positively harmful in most cases. Mere acclaim is not sufficient to justify methods, and certainly not in an age of evidence-based medicine, particularly when, as Clauser *et al* have said [24], the "literature remains less than definitive", and "there have been relatively few papers examining [the Angoff Method's] psychometric properties".

9. Standard-setters differ in their expertise, probably being expert and up-to-date with current teaching and examining standards on only some of them[xxxiii]. In non-expert areas the evidence suggests that standard-setters set lower standards than they set in their specialist areas.

In conclusion, it seems at least possible that the relatively low standard which seems to have been set for PLAB Part 1 may reflect problems with either the implementation of the Angoff process, or with the Angoff process itself.

Alternative methods of Standard-Setting

Clauser *et al* [23] provide extremely strong evidence against the acceptability of the Angoff method. They review evidence which suggests that without normative data examiners only allocate standards in a very narrow range, and the standards bear little correlation with eventual performance, but that when provided with normative data, most examiners merely adjust their standards so that they correlate almost perfectly with the normative data. In a key finding, Clauser *et al* showed that when they actually provided normative data at random, examiners still followed the normative data, rather than responding to the true difficulty of questions. The implication is that the Angoff is generally little more than a complicated way of norm-referencing. Clauser *et al* conclude their paper by asking,

> "It may be attractive to believe that the simple answer is to abandon the Angoff procedure."

They then continue, though,

> "Unfortunately , it is not clear that alternative procedures that provide accurate content-based judgements while avoiding the issues addressed [in their article] are available".

That is mostly true, although one possible alternative form of content-based judgement which they do not discuss is the technique known as *bookmarking*.

Bookmarking

Bookmarking is one of the fastest growing techniques in standard-setting at present [28,29], although it has been relatively little used in medical education [30,31]. In essence it involves printing out a booklet for standard-setters in which all of the questions used in one or several previous diets are *sorted in order of actual difficulty* (and in that sense, some limited form of normative data is provided). Examiners are asked to look through the booklet, assessing the point in the sorted items at which the questions are at the borderline where a just-passing candidate should know the answer with some particular probability (usually two-thirds). Bookmarking is usually carried out in conjunction with an item-response theory analysis, and from that a passmark may be estimated. The task is much easier and more interesting for standard-setters than the Angoff procedure, because not all items need to be looked at. Very difficult and very easy items can be quickly eliminated as the location of the borderline, with only those near the borderline needing to be read carefully and discussed. Sorting the items in order also informs standard-setters about the relative difficulty of the items, but without giving them direct information on actual percentages correct[xxxiv]. There have been few direct comparisons of bookmarking and Angoff, but of particular interest for present purposes is one in a medical context in which the standard was set at a significantly higher level with bookmarking than with Angoff [31].

[xxxiii] The criteria for being an examiner are specified as, ""Panel members must be practising clinicians of Consultant or equivalent grade with knowledge and skills in the specialty required. Applicants must be fully registered with a licence to practise and have a reasonable expectation of remaining in this position for at least five years … The panel member will have a knowledge and understanding of the duties and responsibilities of a doctor successfully completing Foundation Year 1 (FY1) and will have regular clinical contact with Foundation doctors."

[xxxiv] An additional advantage for assessments such as finals is that it is also straightforward for standard-setters to bookmark the level for, say, Honours and Distinction, a process which is exceedingly demanding if the assessment has to be Angoffed twice more in its entirety.

A potentially very useful use of bookmarking is in comparison across different examinations. It would be straightforward for a group of finals examiners and PLAB examiners each to bookmark both their own examination booklet and the other group's examination booklet to assess the extent to which the suggested standards coincide. An overall group discussion of both groups of examiners could then resolve discrepancies, etc.

Statistical equating

Whether or not bookmarking might help resolve some of the problems of the Angoff technique, Clauser *et al* have an important point when, to reiterate, they say, "it is not clear that alternative procedures that provide accurate content-based judgements while avoiding the issues addressed [in their article] are available". All content-based judgements are precisely that: judgements based on content, with all the fallibility that human judgement, even by experts, inevitably have.

And there is the rub. Angoff and other content-based procedures are probably not the answer to the difficult issues which inevitably occur with standard-setting based on assessment of content. If for PLAB the main requirement is a method, "that *ensures* a consistent standard over time", then non-content-based solutions are required. *Statistical equating* is a set of procedures which allow the standard across different diets of an examination with different items to be maintained across time under certain constraints, of which the most crucial is that there is item overlap between diets[xxxv], the 'anchor items' being used to maintain the standard of the pass mark. Nowadays the best way to carry out statistical equating is undoubtedly by using item-response theory using widely available software packages. It is technical but readily implementable, particularly with a stable assessment such as PLAB Part 1 for which large banks of previous items exist[xxxvi].

This is not the place to consider further the large literature on statistical equating and its practicalities, but it would certainly seem feasible for PLAB Part 1. Something that must be emphasised is that statistical equating cannot and should not remove professional responsibility for the content of items from the Panel. They remain responsible for the content of the exam, and need continually to check the quality of items across time, using standard statistical methods. All that has been devolved is the task, from diet to diet, of maintaining the standard of the pass mark. Setting the initial standard of the pass mark is, though, an important matter for the Panel, and regular review of the standard is also important. That process may include an Angoff assessment of the items in a diet, coupled with normative data, as well as data on the predictive validity of the assessment, and external perceptions of the appropriateness of the standard by relevant external bodies, including consultants, examiners for other examinations, employers, etc.

Other empirical approaches to assessing equivalence

Direct equating of assessments such as PLAB and medical school finals, ultimately requires that there is overlap in the content of the different assessments [8]. Merely to whet the appetite, without giving a mass of practical details, options available might include:

[xxxv] This need not necessarily mean that, say, N items in the current diet were used in the immediately previous diet, but only that N items have been used within a number of diets within a particular time window, which often excludes the previous diet or two (to prevent resit candidates encountering items they may have revised having seen them previously), and over a number of diets over, say, the three or four years before that. If there is a possibility that practice may have changed, or that items have become easier for other reasons (including leakage) then items may require removal from the pool of anchor items.

[xxxvi] We realise that to some extent we are suggesting a return to the original method of standard-setting for PLAB Part 1, which prior to September 2004 had used a one-off Angoff in 2000 followed by linear test equating based on about 30 to 40 marker questions. If PLAB Part 1 were to return to statistical equating then, as before, there would need to be a single standard-setting exercise on a recent diet, not based though entirely on an Angoff but on a range of other factors, including a Hofstee process taking into account standard-setters' and others' perceptions of the range of acceptable pass rates and pass marks, as well as information on predictive validity, etc..

1. Items from medical school finals and PLAB Part 1 could each be included in the other assessment. That is straightforward where items are 'best-of-five' as are most assessments nowadays.
2. PLAB Part 1 could include items from the bank being developed by the Medical School Council's Assessment Alliance, which are already being used in a range of medical school assessments and are best-of-five.
3. Items from MRCP(UK), MRCGP, and other Royal College Examinations could be included in PLAB, particularly when those items are somewhat too easy for postgraduate assessments but have been standard-set on the scales of those assessments by having been used live.
4. PLAB standard-setters could be invited to take part in medical school finals assessments, and vice-versa.
5. A series of bookmarking exercises could be carried out in which undergraduate, PLAB and Royal College examiners bookmark sets of questions from medical school finals, PLAB and Membership examinations to assess appropriate and comparable standards on each.

Standard-setting, research and the governance of PLAB

The paragraphs above strongly suggest that the standard of PLAB may not have been set at an appropriate level relative to the standard which the GMC has required. Whether or not that is the case, and despite it being beyond the strict remit of this report, there are important questions which perhaps need asking about the governance of PLAB and the ways in which an awareness of possible problems with PLAB might have become apparent.

The organisational structure of PLAB

Those responsible for delivering PLAB can be straightforwardly divided into two groups:

1. **GMC staff**. On a day-to-day basis the content of PLAB is organised by Michael Harriman and staff in the GMC's office in London. They analyse the results from examinations, and provide the secretariat for the committee structures underpinning the academic side of the assessments. While there can be no doubting either the professionalism or the commitment of the staff[xxxvii], nor of the quality of the results they produce regularly, year in and year out, neither is it their responsibility to instigate major changes in the structure of the assessments, which is a professional/academic responsibility.
2. **The Committee structures**. Academic and professional input into the PLAB assessments comes at five levels:
    a. **Examining and question-writing**. A large number of doctors contribute in important ways to the examination, either by writing questions for PLAB Part 1, or by being examiners for PLAB Part 2. This is time-consuming and important work which requires a dedicated commitment by those involved.
    b. **The Part 1 and Part 2 Panels**. Oversight of the academic and professional aspects of the running of the assessments, written and clinical, is under the control of the Part 1 and Part 2 Panels, who in particular are involved in question-writing and selection for Part 1 and scenario-writing for Part 2, as well as the routine work of standard-setting.
    c. **The PLA Board**. The PLA Board, of which the chairs of the Part 1 and Part 2 Panels are members, oversees academic and professional aspects of the assessments in general, including issues concerned with discipline or infringement of regulations, etc..

[xxxvii] If evidence were needed for that statement then we will just mention the regular participation of Michael Harriman, William Curnow and Katharine Lang in the informal *Sharing Good Practice* discussion days which are organised biannually by various Royal Colleges and postgraduate organisations for discussing and sharing information on quality in assessment. In particular, Michael and his colleagues organised a very successful meeting at the GMC in London in April 2011.

d. **GMC Council**. GMC Council has ultimate responsibility for the PLAB examination, with reports from the PLA Board and the Part 1 and Part 2 Panels passing to it through the Strategy and Policy Board, which reports to the Chief Executive.

e. **External Review**. At regular intervals, typically quinquennial, but for organisational reasons, somewhat longer in the present case, GMC Council instigates a formal Review of PLAB, and this document is a part of that Review process. Reviews have an external Chair and external panel members, as well as GMC and PLAB staff.

A key question for the present analyses concerns how a problem with the standard set by PLAB might be identified within these organisational structures. One possible route is through the Review, and while a possible problem has indeed been identified here, and some of the problems were much discussed in the 1986 review, it does not seem that the 1999 or 2004 reviews considered the issue, particularly to the extent of collecting external data[xxxviii].

Research within PLAB

Analyses such as those presented here are outwith the normal process of data analysis as a part of routine analyses of candidate results and the performance of individual items or stations within the assessments. Essentially they are part of *research* rather than regular *audit of service delivery*. Research uses statistical and other data collected as part of routine operations to understand underlying mechanisms and processes, the emphasis being either on deeper understanding or issues, or of addressing particular key processes or questions of theoretical, practical or political interest. The question of the equivalence of the PLAB standard clearly is of both practical and political interest[xxxix].

Research, inevitably, is carried on outside of the normal processes of service delivery, and it has a diverse range of skills and methods, often *ad hoc* and put together for the particular needs of specific projects. While it can be carried on as a part of service delivery, that can be difficult.

A key question, given the continual importance that the GMC has attached to the equivalence of PLAB, is why it has taken so long for such research to be carried out. Several different sorts of reason can be identified:

1. **Asking the question.** Research questions are only answered because someone asks them. Or more generally, asks how they might be answered in ways that are practical and possible. Experience of research makes answering research questions easier. If an organisation is not a research organisation, a theme to be returned to below, then research questions may not be asked, or if asked may not be answered or be answerable.
2. **Realising that the question could be answered.** Research, as Sir Peter Medawar said, is 'the art of the soluble', while Bismarck said that politics is 'the art of the possible'[xl]. Unless research is politically possible, and the research problems in principle are soluble, research cannot take place. Experience with research invariably helps in recognising which problems realistically have the practical possibility of being answered.
3. **Access to data**. The analyses reported here have combined data from three separate sources, PLAB, the Royal Colleges of Physicians, and the RCGP. No one organisation could have carried out the work itself. No doubt there are similar data in all of the other Royal Colleges and in other organisations which could have contributed to answering the question as well. In the present case it was the existence of the Review itself, with its clear focus on the questions of

[xxxviii] Whether there were reviews between 1986 and 1999 is unclear. Certainly none are mentioned in the 1999 review.

[xxxix] And it could hardly have escaped our notice of its relevance to the very highly-charged questions currently being asked about the performance of IMGs on the MRCGP examination, and which have resulted in the Report commissioned by the GMC from Professors Aneez Esmail and Christopher Roberts.

[xl] "Die Politik ist die Lehre vom Möglichen"

equivalence, and a will to have the question answered, which allowed the various organisations to come together.

4. **Organisation of the data**. Without the extensive work put into these data by Daniel Smith the linkages would not have been possible. And that work was possible because the GMC created his post and provided the resources and facilities, as well as the space to find answers.
5. **Experience**. Large, complex datasets are not easy to analyse, and statistics can be complicated and necessary. In the present case, both authors of this report happen to have such skills. In each case they have them because their universities, and the Royal Colleges with whom they collaborate, have helped them hone those skills.
6. **Interest**. Unless individuals at all levels have an interest, an excitement, an enthusiasm, all of which are generated by a sense of professionalism in understanding of the issues, then research will not take place.
7. **Time**. Research is not cheap. In the present case it mostly does not involve large laboratories full of expensive machines, but it does take time, and time is expensive, particularly in universities and regulatory bodies, where there are many calls on the time of researchers. For research to be done it has to be prioritised over other competing calls.

A research organisation?

Probably the main reason that this research has not been done previously, despite its importance, is that the GMC, and PLAB within it, is not a research organisation. When that is the case, research ideas are rare, the exploration of research possibilities is also rare, the realisation of research opportunities is limited, and even when those barriers are surmounted, carrying out research is difficult. The result is that research is not done. When committees have problems to solve, when questions need answering, then research is not the way that is immediately thought of. If research is seen as necessary it is tempting to outsource it like any other commodity or service, without realising that the key attributes of the researcher are their commitment, their interest, and their tacit, informed knowledge of the situations in which the questions and the data are embedded. The researchers' interests may also not be the interests of the organisation, particularly if work has been commissioned on the basis of the cheapest tender.

A particular aspect of PLAB which is very pertinent to the present concerns is psychometrics. Any modern exam requires psychometric analyses, and they underpin most aspects of the defensibility of assessments. PLAB has psychometric support, and the Panels receive sometimes lengthy reports on item statistics and other measures, although having said that, it is not clear that the Panels know what to do with such information, and neither is it clear that it is always the right information for the purposes of the Panels. Without wishing to criticise those carrying out the psychometrics, there does seem a possibility that psychometric analyses are provided somewhat routinely, with few attempts to go beyond the immediately necessary. That may in part reflect that the psychometricians involved are not a part of the organisation, have been tasked only with carrying out particular tasks, and it is neither their responsibility nor indeed their right to analyse the data further. A first step might well be to bring a psychometrician/data analyst in house, and encourage them to do proper research by funding them for a part-time research degree, working on the mass of data which is available.

We believe that the central reason that the question of the equivalence of PLAB has never properly been addressed before, or considered seriously as being answerable, is that the GMC itself is not a research organisation, and, of necessity no-one else outside would have had access to the key data. The question seems easy to answer now, mainly because of the possibility of creating a linkage between databases, using data collected for other purposes, the key which opened the door being membership of the Review group. Even so such databases require structuring in proper ways (and for instance the MRCP(UK) spent a lot of effort over several years in creating what is called the ‘history file’, which now rapidly allows questions to be answered about candidates and their progress through its assessments, and was one of the keys to the linkage here).

Between 1975 and 1986 there were the ‘green shoots’ of research into the equivalence of PLAB, with the 1986 Review reporting pilot studies of the assessment of UK medical students and doctors, and of PLAB graduates being followed up a year into their practice. It would have been entirely feasible at that time to have instigated studies based on existing epidemiological method to have followed up carefully chosen samples of PLAB graduates for comparison with UK graduates. It wouldn’t have been easy, but during that same period one of us was setting up large-scale cohort studies of UK medical students and doctors[16,32] to which cohorts of PLAB graduates could have been appended. And no doubt there have been multiple opportunities since, with various methodologies, which could have resulted in answers to the question.

Being a research organisation requires that research is embedded within a system, the system has individuals whose key role is to carry out research within the organisation, and there is the provision of infra-structure for that research. MRCP(UK) Central Office was not a research organisation in 1996 when it began to want answers to important questions about its examinations. As it became a research organisation, so it set up a Psychometrics Research Unit, with both employed staff and academic collaborators who were committed to service delivery but had time set aside for research activities which resulted both in published papers and also research for internal consumption to help direct policy. As time passed and publications began to accrue, so the organisation realised that to have a research capacity was to be ahead of the game on policy issues, in part because researchers spotted issues and problems within the data they were analysing for other purposes, and in part because it was easier to respond to questions raised by, say, regulatory authorities. Chairs and members of the various examining boards also started to ask questions and wanted questions answered with data rather than assertions, and researchers started to sit on those boards where they also realised the questions that needed asking and the sort of answers that were needed. The final step in the process of making the MRCP(UK) a research organisation was in the creation of a Medical Director, a medical academic whose primary interest was the academic and professional integrity of the examination, and who could instigate research and other processes in order to explain the past and to anticipate the future based on careful analysis of data[xli].

This section began by describing the organisational structure of PLAB and its governance, and has ended with the argument that opportunities by the GMC to assess the key question of the equivalence of PLAB might have been missed because neither the GMC nor PLAB is essentially a research organisation. It seems unlikely that any of the PLAB question writers, the PLAB standard-setters, the members of the PLAB Parts 1 and 2 Panels, or the members of PLA Board are not continually asking research questions, for they are after all doctors, and research is one of the cores of modern evidence-based medicine. What perhaps is lacking is a way of actualising and realising those ideas into research projects and outcomes which can answer the myriad of questions that any complex exam raises about itself and about its relationship to health care systems[xlii]. One partial solution might be to appoint a practising clinician who is also a researcher and who would be on all the board and panels *ex officio*, with a remit to ask questions and to help to answer them.

[xli] If one needs only one example of that, consider the continuing and difficult issue of the under-performance of both medical students and doctors from ethnic minorities and of international medical graduates. Where once data were treated as secret, now it is policy not only to publish data but to ask what are the origins of the differences which are found [33], as for example in developing new methods for identifying possible sexism or racism amongst examiners on a routine basis [34].

[xlii] An answer we heard when we asked people informally how research questions might be taken forward within PLAB was that, “there was no mechanism”.

Discussion and Conclusions

There seems little doubt from these analyses that PLAB graduates underperform in their subsequent careers in the UK, relative to UK graduates. They are not therefore equivalent under any simple definition of the term. It is therefore necessary to look at the details of the PLAB assessments to ascertain the locus of the problem.

The major gate-keeper within PLAB is the PLAB Part 1 examination, it taking place first, and having the lower pass rate, making it the immediate place to look for possible problems. In addition one of us has looked in detail elsewhere at the Part 2 examination [15].

Several potential sources of problems can be discounted quickly since:

- The blueprinting of PLAB has been examined carefully in a separate report[xliii], and appears to be fit for purpose and comparable to what is done elsewhere, in university finals, etc.
- PLAB1 has very acceptable reliability, typically of the order of .9, meaning that the examination is successful at ranking candidates into a meaningful order.
- Performance in PLAB 1 correlates well with subsequent performance on MRCP(UK) and MRCGP, suggesting that the constructs it is measuring are parallel to those assessed by the two postgraduate examinations.

PLAB Part 1 is therefore reliably putting candidates into a meaningful order which subsequently predicts other postgraduate outcomes (and UK medical school finals do a similar thing to a numerically similar extent, although getting large-scale data for assessing that is currently not straightforward). PLAB therefore seems to act as a part of what one of has called the 'Academic Backbone' that underpins secondary school, undergraduate and postgraduate performance in UK medical students and graduates [16].

The problem with PLAB is not therefore in the construct being assessed by the assessments or with its ability to put candidates into a meaningful order. Instead it has to lie elsewhere. Two major possibilities need considering, only one of which we can go into in detail.

- The setting of a standard, the drawing of a line between pass and failure, is being carried out at the wrong level, and is at an inappropriately low level
- Although the examination and the standard are appropriate, sufficient numbers of candidates are circumventing the normal procedures to ensure that they are entering UK postgraduate training with qualifications which are not an accurate representation of their true ability.

This is not an appropriate point to investigate in detail the latter possibility, but it certainly cannot be ignored, particularly given that the 1986 Review commented, as we mentioned earlier, said that,

> "the bulk of the [language test] questions have, over the years, found their way, virtually verbatim, to a number of private organisations which offer specific tuition for overseas doctors who require to pass the PLAB tests. A similar, though not as damaging situation is known to have arisen in respect of some questions used in the Multiple Choice Questions".

Three decades later, in the age of the internet and global communications, it seems unlikely that such organisations have diminished in their scope or reach, and a search of the internet rapidly finds sites offering such services. If some of the apparent passes at PLAB are the result of personation, of question leakage, or other forms of deception[xliv], then that would certainly undermine the predictive validity of the assessments.

[xliii] *Reviewing the Blueprint: Report of the Sub-group*. Documented presented to the GMC PLAB Review group, 10 October 2012.

[xliv] It is unlikely that direct cheating by copying is taking place as, for a number of years, the PLAB Part 1 assessment has used *Acinonyx* to look for undue similarities between candidates' answers to multiple-choice questions [35].

An alternative way in which candidates can be passing without in fact having sufficient knowledge or skills is by repeated attempts. Candidates failing and retaking exams repeatedly benefit from chance having been in their favour. Standards however are set on the basis that a candidate is taking an examination on a single occasion. Just as there is only a one-in-six probability of using a dice to roll a six on the first attempt, but the probability of rolling a six goes up to 66% (two-thirds) if the dice is rolled six times, so resitting candidates benefit from and capitalise on chance.

With unlimited attempts the result is that many candidates are only just passing an exam, having just crept in above the pass mark. A large proportion of successful PLAB candidates only *just* pass their relevant assessments, almost a half (47.7%) having the lowest acceptable IELTS standard (7.0), and that is after unknown numbers of attempts[xlv]. Overall, 10% of successful PLAB Part 1 candidates have taken it three or more times, and we estimate that over a third of candidates pass PLAB Part 1 with a score within two SEMs of the pass mark. Reducing the number of attempts to a maximum of six does not solve that problem, as most candidates are passing within six attempts. One possible solution is to add an ever-increasing number of standard errors of measurement to pass marks for each attempt made[xlvi]. Our analyses here have purposely used the mark at first attempt as the primary measure of PLAB Part 1 and Part 2 performance, as described in Appendix 2. PLAB marks at the first attempt are predictive of outcome, but become less and less predictive as those marks are obtained at higher and higher attempts. Those who pass exams only at a high attempt number and with a mark only just above the cutscore are unlikely to perform well later in their career progression.

Despite all of the factors described above, it is also likely that the standards for PLAB Part 1 (and Part 2) have been set inappropriately. We have described above some of the potential problems with the Angoff procedure used in Part 1, and they are sufficient to account for some of the discrepancy between the standard which is set and the standard which might need to be set if equivalence with UK graduates is required, as PLAB has always taken as its criterion for its standard.

PLAB standard-setting should also not be taking part within a closed system, hermetically sealed from the rest of the medical world. Many postgraduate medical examinations in the UK, particularly the more specialist, in effect do stand alone, and have no other examinations with which they can be directly compared. That though is not the case for PLAB, which of its very nature is a parallel assessment to that carried out in medical schools throughout the UK. Just as examining boards at secondary level work closely together to ensure that standards on assessments such as GCSEs and A-levels are comparable, using a mixture of statistical and evaluative methods [9], so PLAB and other equivalent qualifications need to be collaborating on standard-setting exercises, both judgemental, as in cross-moderation methods, and statistical, with various 'common test' methods. The GMC is surely in a position to facilitate and moderate a dialogue across PLAB, medical schools and the Royal Colleges.

We cannot finish without emphasising that any change in the pass mark for PLAB will inevitably have consequences for managing the NHS Workforce. PLAB graduates currently form a large proportion of the doctors working in the NHS and any change in their likely numbers would have inevitable consequences for service delivery. That cannot be part of this review, but we acknowledge that it is potentially problematic. Getting the standard right though is fundamental to ensuring high quality postgraduate education and training for the delivery of care of the highest quality.

[xlv] The IELTS standard of 7.0 is, incidentally, below that required to come to the UK to practice as a solicitor, the acceptable level being 7.5 (http://www.sra.org.uk/solicitors/qlts/english-requirement.page).

[xlvi] If an extreme example is needed, we are aware of one PLAB graduate who required 12 attempts to pass. Subsequently that doctor failed their first written postgraduate exam by 45 marks, failed that exam badly on a further three attempts before sitting the clinical examination and after failing that badly three times left postgraduate training. Such cases are a tragedy for the doctors involved, have untold possible consequences for the patients who have been treated, and are expensive and potentially dangerous for the employers of the doctor.

Acknowledgements

We are very grateful to many people for their help in preparing this report. Michael Harriman, Katharine Lang and William Curnow for ensuring that there are as few errors as possible in our descriptions of processes; Daniel Smith and Andy Knapton, without whom none of the complex data-merging operations would have been possible; Thomas Jones for facilitating various aspects of preparing the report; Prof Jane Dacre and the Royal Colleges of Physicians for giving permission for the analyse of MRCP(UK) data to take place, and Liliana Chis for helping in the preparing of those data; and Dr Sue Rendel and the Royal College of General Practitioners for their permission to analyse the MRCGP data.

Appendix 1: Setting a pass mark if a UK comparator group has taken PLAB

In 1975 those setting the PLAB pass mark gave the paper to a group of UK medical students in Glasgow, London, Manchester and Nottingham, "and the overall mean mark ... for British medical students [was] 46%" [3]. The report then continues, "In the light of these results the Panel agreed that the pass mark in the MCQ paper should be set at 45%", with the strong implications that it was because it was similar to the mean of the UK medical students.

Although setting the PLAB pass mark at the average mark attained by UK finalists might seem a sensible strategy, a little reflection suggests that there is something wrong about it. Since the mean and median of the UK finalists will be similar, about half of the UK finalists would have scored more than 45% and *half would have scored less*. However with a pass mark of 45% then *all* of the PLAB group had to have obtained a mark higher than 45%, making it harder for them to pass. That can be seen in Appendix1: Figure 1a, where at the top is a hypothetical distribution of marks obtained by the PLAB candidates, and below them is a hypothetical distribution of marks for the PLAB candidates. The pass mark for the PLAB candidates is the mean of the UK marks. Those passing are shown as the shaded areas, and correspond to all of the UK graduates (by definition), and the top end of the PLAB candidates. A corollary is that the median mark of the PLAB candidates is higher than that of the UK graduates.

A different approach (Appendix1: Figure 1b) would be to calculate a "UK pass mark" – the value on the test at which UK candidates would pass (and given the high pass rates of UK finalists, this would be at the lower end of the UK graduates, 95% of UK medical students passing finals at their first attempt, and almost all after resits). The pass marks for UK and PLAB candidates (PM-UK and PM-PLAB) are shown, and are the same. A problem now, though, is that despite the two groups having the same pass mark, the median of those passing *is lower in the PLAB candidates than the UK candidates* (see the figure). When those two groups go on to take further examinations, which will correlate with PLAB performance, then it is very likely to be the case that *the PLAB graduates will perform less well than the UK graduates.* And to reiterate, that is despite having passed an exam with the same pass mark.

The problem arises, here, because 'equivalent' has two separate meanings:

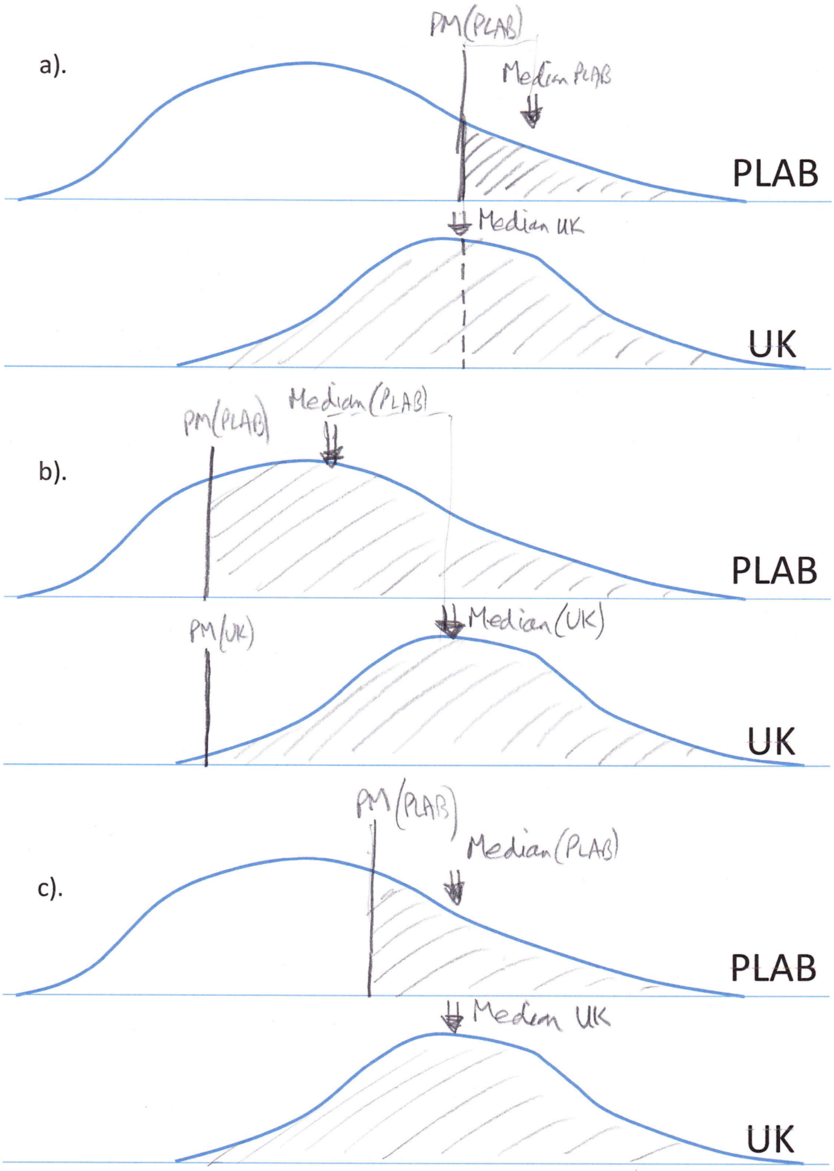
a) The pass mark for the two groups is the same (a strong meaning of equivalent);
b) The outcome of the two groups in terms of career progression will be the same.

In the final analysis these two usages are incompatible, there being no situation in which the two occur simultaneously[xlvii], and that is because one is about pass marks and the other about the medians of two differently distributed groups.

If it is of central importance that the two groups are equivalent in terms of career progression (and career progression is perhaps the best proxy for quality of service provided to the NHS), then a criterion for the PLAB pass mark is that *the medians of the passing groups should be equivalent*. In Appendix1: Figure 1c the pass mark for PLAB has been adjusted so that the median of the PLAB group is the median of the UK group. The PLAB pass mark is, though, higher than the marks attained by some of the UK graduates, which might seem unfair. The distributions of the UK and PLAB groups are also different, with the PLAB candidates being grouped at the lower end, so that while their median outcome will be equivalent to the median outcome of the UK candidates, fewer will be high achievers and more may struggle. The latter can only be avoided if a strategy is used to ensure that the entire distributions are equivalent.

[xlvii] The only exception being where the entire distributions of the UK and PLAB candidates are identical.

Appendix 1: Figure 1. Schematic representation of different methods of assessing equivalence. See text of Appendix 1 for details. PM indicates Pass Mark, and shaded areas indicate candidates who pass.



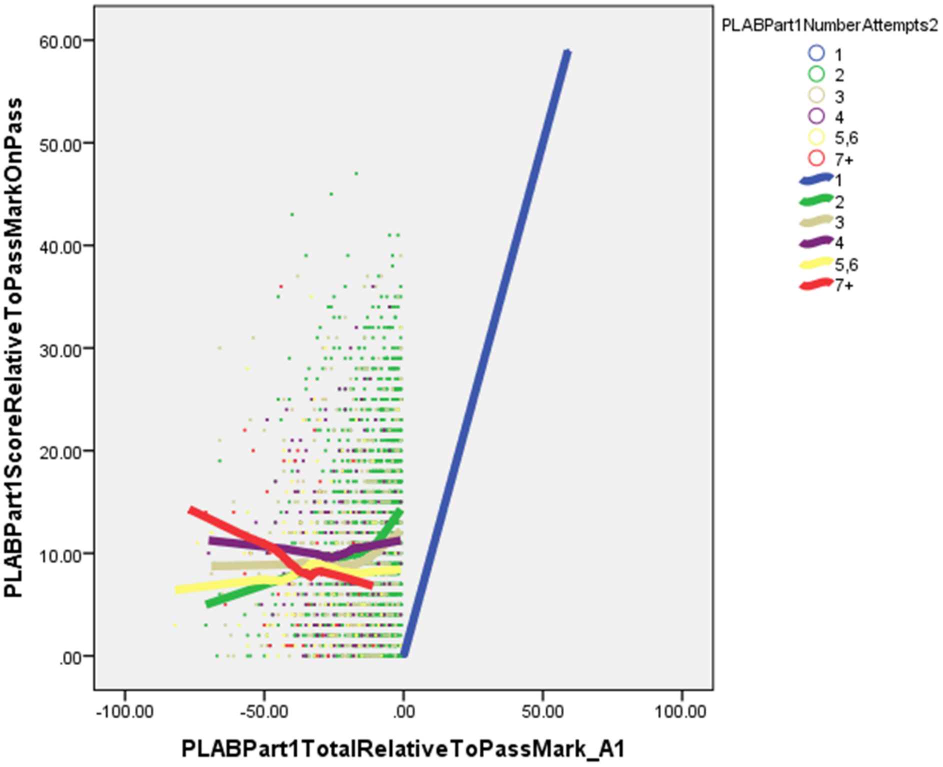
This analysis of different ways of setting pass marks for a PLAB type of examination shows how, even with direct comparative information on the performance of UK and PLAB candidates at the same level, there are still strong problems in generating equivalence of the two groups. From the point of view of the UK health service the best and most straightforward method is probably that of setting a pass mark so that the medians are the same[xlviii]. Whether that could be justified in terms of equity is a very different matter.

[xlviii] This, in effect, is what is done in our Method 1, albeit basing it the indirect outcome measures of MRCP(UK) and MRCGP performance. The different pass marks for UK and PLAB are then less obvious as they take different assessments.

Appendix 2: Comparison of Mark at first attempt with Mark on passing

Throughout this report we have used 'mark at first attempt' as the main predictor (as did Esmail and Roberts in their report and paper on the MRCGP [36] [37], published as this report was submitted). and outcome variables for the various analyses, rather than 'mark at pass'. Mark at pass may seem to have many advantages over mark at first attempt, and therefore some justification is required. Mark at pass appears to demonstrate that a candidate has acquired some pre-set standard, and therefore is competent to progress to a next stage of training, whereas in contrast mark at first attempt may seem to reflect a host of irrelevant factors, such as preparedness and so on. Elsewhere one of us has investigated in detail the performance of candidates on repeated attempts at the MRCP(UK) Part 1, Part 2 and PACES exams [14], and their inter-relationship, and here we will only provide two brief demonstrations that mark at first attempt has many advantages over mark at pass, both statistical and theoretical.

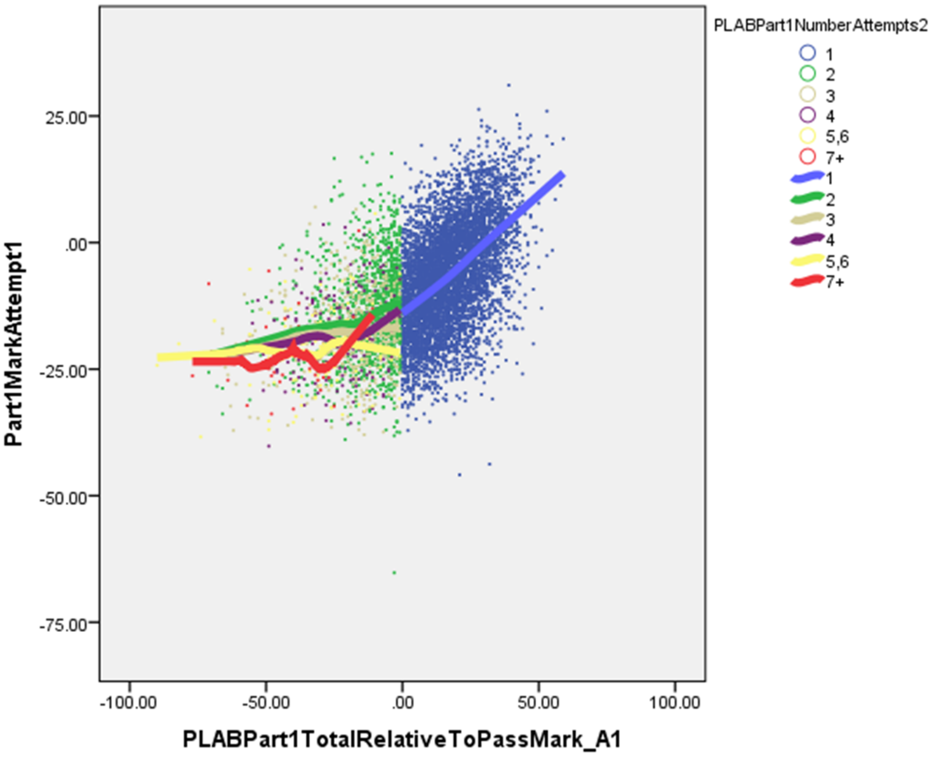
Appendix 2: Figure 1. Mark at first attempt at PLAB Part 1 (horizontal) in relation to the mark obtained when passing, in relation to the attempt on which the candidate passed Part 1. Different attempt numbers are shown by different colours (see key).



Appendix 2: Figure 1 shows what ought to be a simple plot but is in fact very complex. On the horizontal axis is plotted mark at first attempt at PLAB Part 1, and on the vertical axis is mark when passing PLAB Part 1, with the attempt at which the candidate passed being shown in different colours. The right-hand side of the figure has just a solid straight line, consisting of candidates who passed on their first attempt, and hence there is a perfect correlation between the two marks. All other candidates failed on their first attempt and hence are in the left-hand side of the graph. These candidates must have marks at first attempt which are less than zero, and marks on pass which are higher than zero. The green line, for those passing at the 2nd attempt, shows a positive upward slope, indicating that those who had done better on their first attempt and then passed at their second attempt do better at that second attempt. As attempt number increases, as shown by the different coloured lines, so there is little or no relationship between mark at first attempt and mark on pass; and indeed to a large extent, the average mark at pass bears no relation to mark at the first

attempt. Overall therefore 'mark at pass' contains two separate components, high marks for those who pass at their first attempt, and relatively low marks for those who pass at later attempts, particularly higher numbered attempts. The result is that mark at pass predicts later outcomes much less well than mark at first attempt, most marks at pass being very close to the pass/fail borderline, where there is little variation, the distribution as a whole being very skewed.

Appendix 2: Figure 2. Mark at first attempt at PLAB Part 1 (horizontal) in relation to the mark obtained at the first attempt at MRCP(UK) Part 1 (vertical), separately for those passing PLAB Part 1 at the first and later attempts, indicated by different colours (see key).

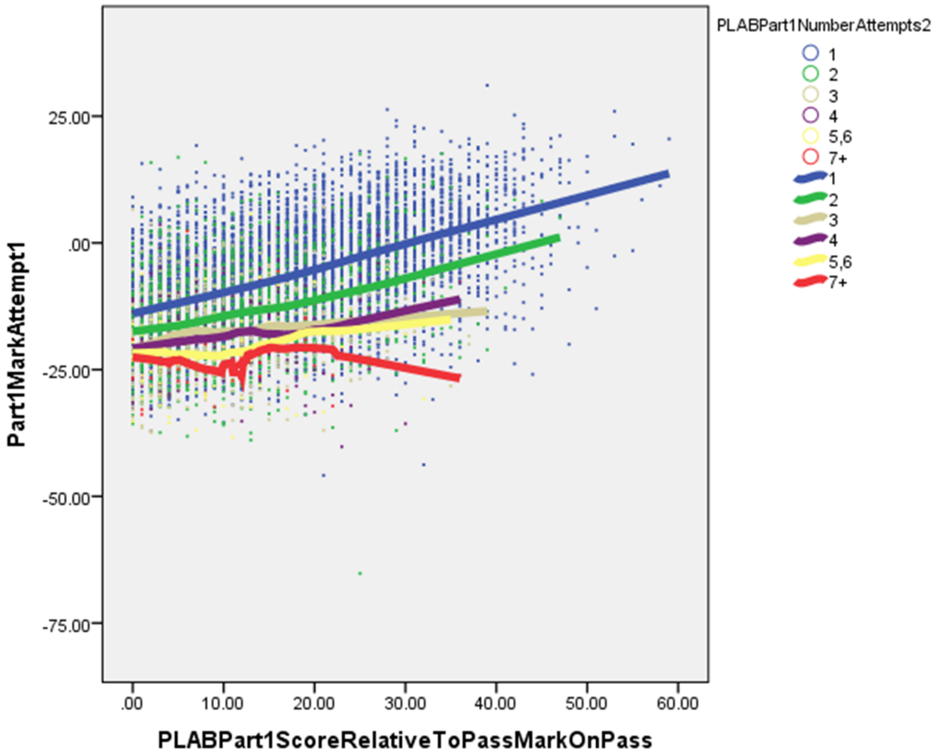


Appendix 2: Figure 2 shows a plot of mark at first attempt (horizontal) at PLAB Part 1 in relation to MRCP(UK) Part 1 mark at the first attempt (vertical; the outcome measure), with those passing at different attempts shown by different colours. There is a smooth relationship, the prediction at the lower end of those passing on the first attempt (blue line) continuing smoothly on to those passing at their second attempt (green line). And relationships between mark at first attempt and the out come measure. Mark at first attempt on PLAB Part 1 is therefore predicting MRCP(UK) Part 1 performance well

Appendix 2: Figure 3, in contrast to figure 2, shows mark at pass on PLAB Part 1 on the horizontal axis, and, as before, MRCP(UK) Part 1 on the vertical axis. Note that of course mark at pass has to be greater than zero, producing a rather strange, skewed, distribution for mark at pass. At first attempt there is a strong relationship of PLAB Part 1 mark at pass to MRCP(UK) Part 1 mark at first attempt, the slope being lower for those passing PLAB 1 on the second attempt, lower still for those passing on the third attempt, until by the fifth and higher attempts the lines are statistically flat, mark at PLAB Part 1 pass showing no prediction of the outcome measure. Notice also that the range of marks for mark at pass becomes narrower as attempt number increases.

Mark at first attempt is therefore a much better and much better behaved predictor of outcome in later assessments, and that makes it easier to interpret and to analyse, particularly given that it is approximately normally distributed, unlike mark at pass which is a skewed, truncated mixture distribution.

Appendix 2: Figure 3. Mark at pass at PLAB Part 1 (horizontal) in relation to the mark obtained at the first attempt at MRCP(UK) Part 1 (vertical), separately for those passing PLAB Part 1 at the first and later attempts, indicated by different colours (see key).



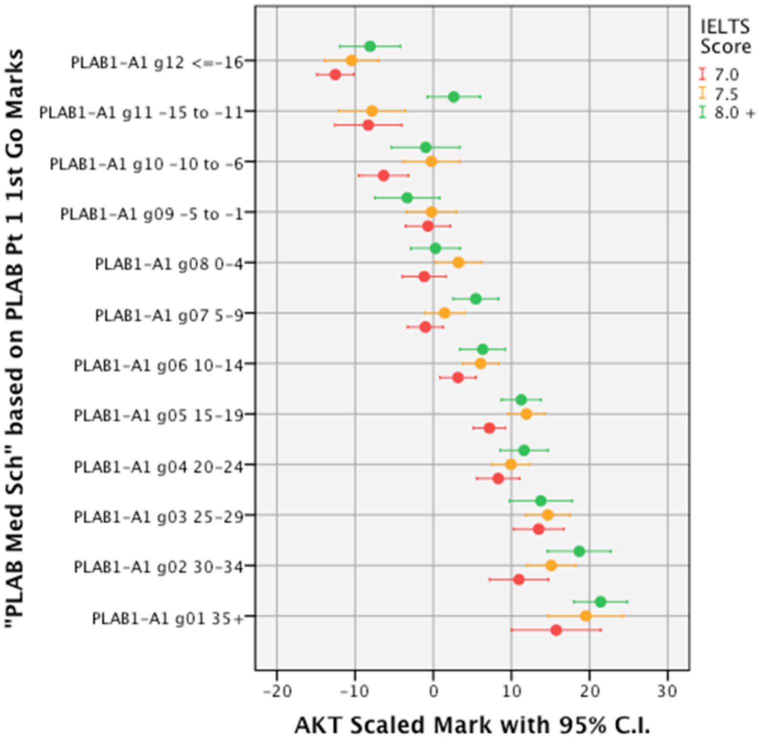
Appendix 3: IELTS, PLAB and performance at the MRCGP

PLAB candidates have mostly passed IELTS, although some are exempted. Since PLAB is an assessment carried out in English, as also are MRCP(UK) and MRCGP, an important question concerns the extent to which poor performance at later postgraduate qualifications may be mediated via problems with English. We have investigated that for both MRCP(UK) and MRCGP, but will only report here the results for MRCGP.

Few PLAB candidates had IELTS scores below 7 or over 8, and we therefore divided the candidates into three groups, ≤7, 7.5 and ≥8. The descriptions of the bands have been provided in the main text.

Appendix 3: Figure 1 shows performance at the MRCGP AKT in relation to performance at PLAB Part 1 at the first attempt and the IELTS level, the 'traffic lights' showing that at most levels of PLAB 1 performance, those with the highest IELTS scores (green) perform better than those with the lowest IELTS scores. IELTS is clearly therefore important. However a multiple regression shows that the effect of PLAB Part 1 (beta=.496) is very much stronger than the effect of IELTS (beta= .086).

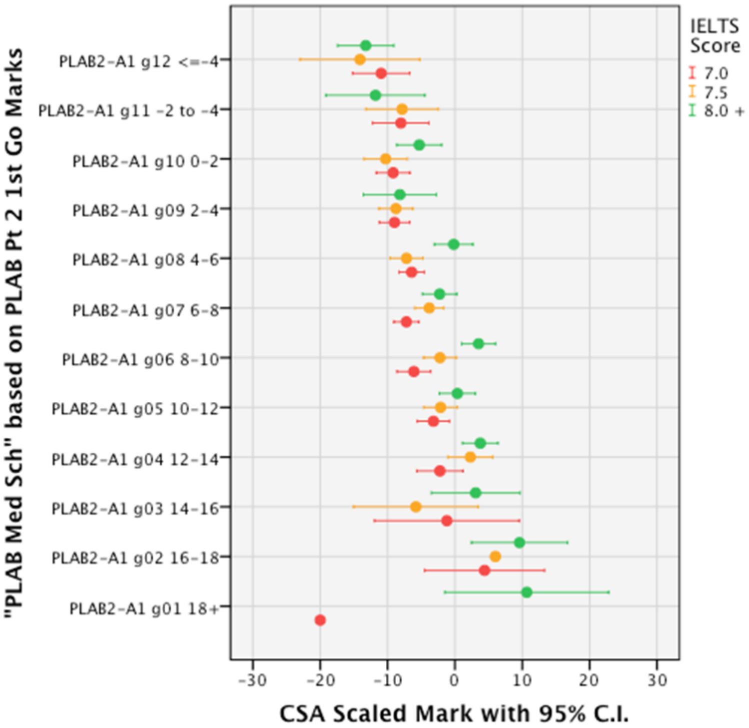
Appendix 3: Figure 1. Performance at the MRCGP AKT (horizontal) in relation to performance at PLAB Part 1 first attempt (vertical), and IELTS (red: ≤ 7, orange 7.5; green ≥ 8).



Appendix 3: Figure 2 shows a similar analysis for performance at the MRCGP CSA, broken down by PLAB Part 2 performance at first attempt and IELTS level, shown as traffic lights. The effects are somewhat less clear, in some cases due to small sample sizes. The multiple regression shows the effect of PLAB Part 2 (beta = .278) is stronger than that for IELTS (beta= .187). The lower effect of PLAB Part 2 (compared with PLAB Part 1 on the AKT) probably reflects the lower reliability of PLAB Part 2, and the larger effect of IELTS is probably due to the greater importance of language, particularly spoken language, in a clinical examination.

Taken overall, and particularly for PLAB Part 1 and the AKT, the conclusion is probably similar to that of the 1986 Review, that, "the failure of candidates was due in the main part to their lack of professional knowledge rather than difficulty in communicating in English" [3].

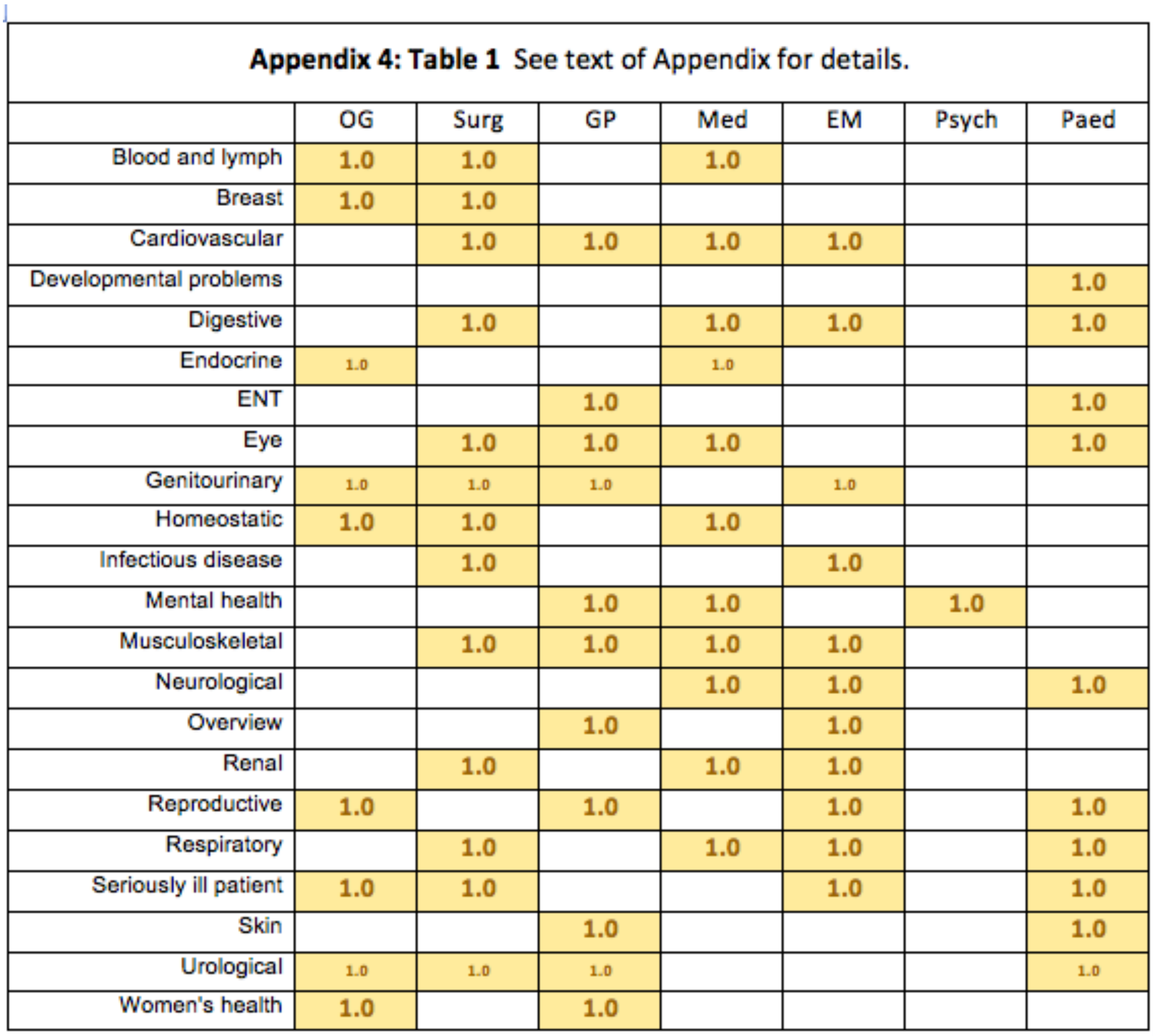
Appendix 3: Figure 2. Performance at the MRCGP CSA (horizontal) in relation to performance at PLAB Part 2 first attempt (vertical), and IELTS (red: ≤ 7, orange 7.5; green ≥ 8).



Appendix 4: Standard setting in relation to standard-setter specialty

Information was provided to us on the standards set by the 13 individual standard-setters on 887 new items set between 2011 and 2013. Not all standard-setters provided standards for all items, for various reasons. For the present analysis the standard-setters in the same specialty were grouped together. All values are expressed on a percentage scale (i.e. the raw scale of 0 to 10 multiplied by 10).

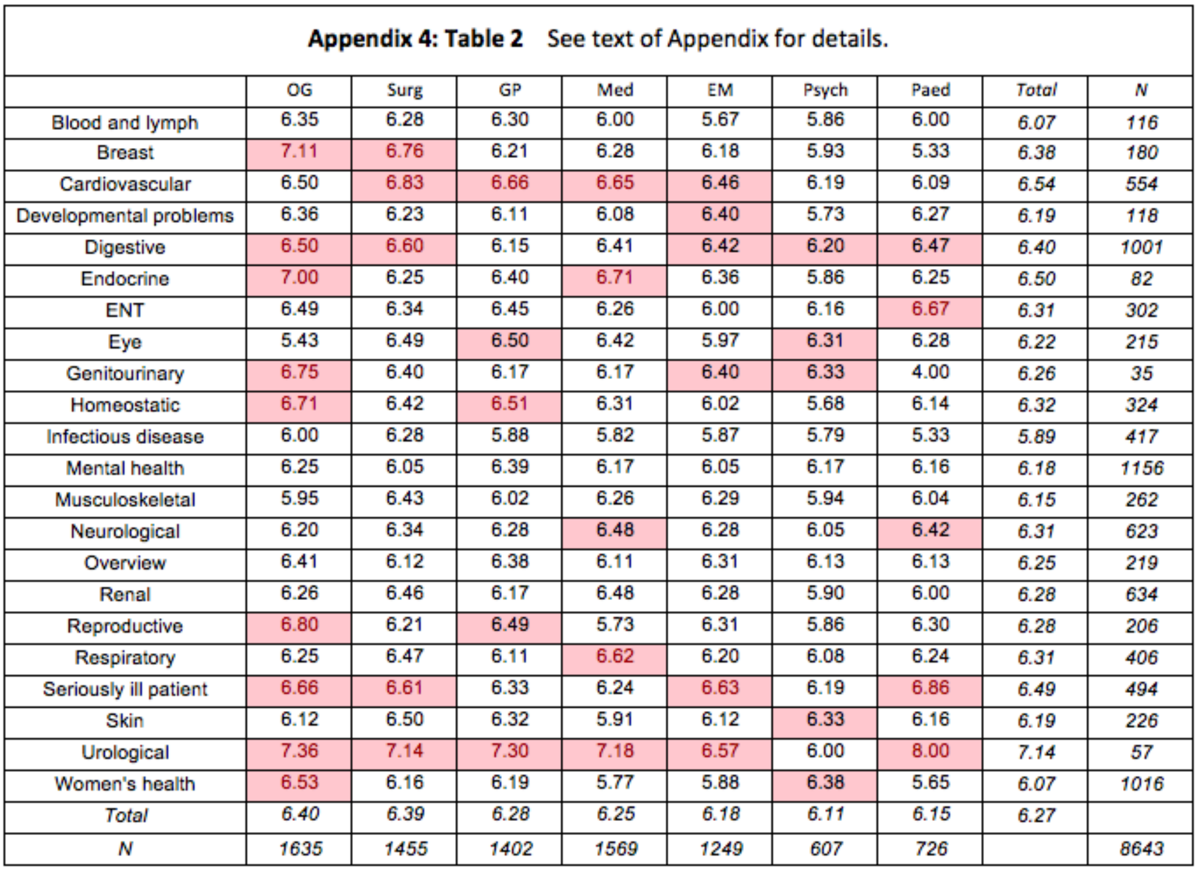
Appendix 4: Table 1 shows ICM's *a priori* classification of which question/speciality combinations are more likely to represent specialist knowledge (and it is accepted that this is fairly arbitrary), a 1 indicating the possibility of an area being with an examiner's specialist domain.



Appendix 4: Table 1 See text of Appendix for details.

	OG	Surg	GP	Med	EM	Psych	Paed
Blood and lymph	1.0	1.0		1.0			
Breast	1.0	1.0					
Cardiovascular		1.0	1.0	1.0	1.0		
Developmental problems							1.0
Digestive		1.0		1.0	1.0		1.0
Endocrine	1.0			1.0			
ENT			1.0				1.0
Eye		1.0	1.0	1.0			1.0
Genitourinary	1.0	1.0	1.0		1.0		
Homeostatic	1.0	1.0		1.0			
Infectious disease		1.0			1.0		
Mental health			1.0	1.0		1.0	
Musculoskeletal		1.0	1.0	1.0	1.0		
Neurological				1.0	1.0		1.0
Overview			1.0		1.0		
Renal		1.0		1.0	1.0		
Reproductive	1.0		1.0		1.0		1.0
Respiratory		1.0		1.0	1.0		1.0
Seriously ill patient	1.0	1.0			1.0		1.0
Skin			1.0				1.0
Urological	1.0	1.0	1.0				1.0
Women's health	1.0		1.0				

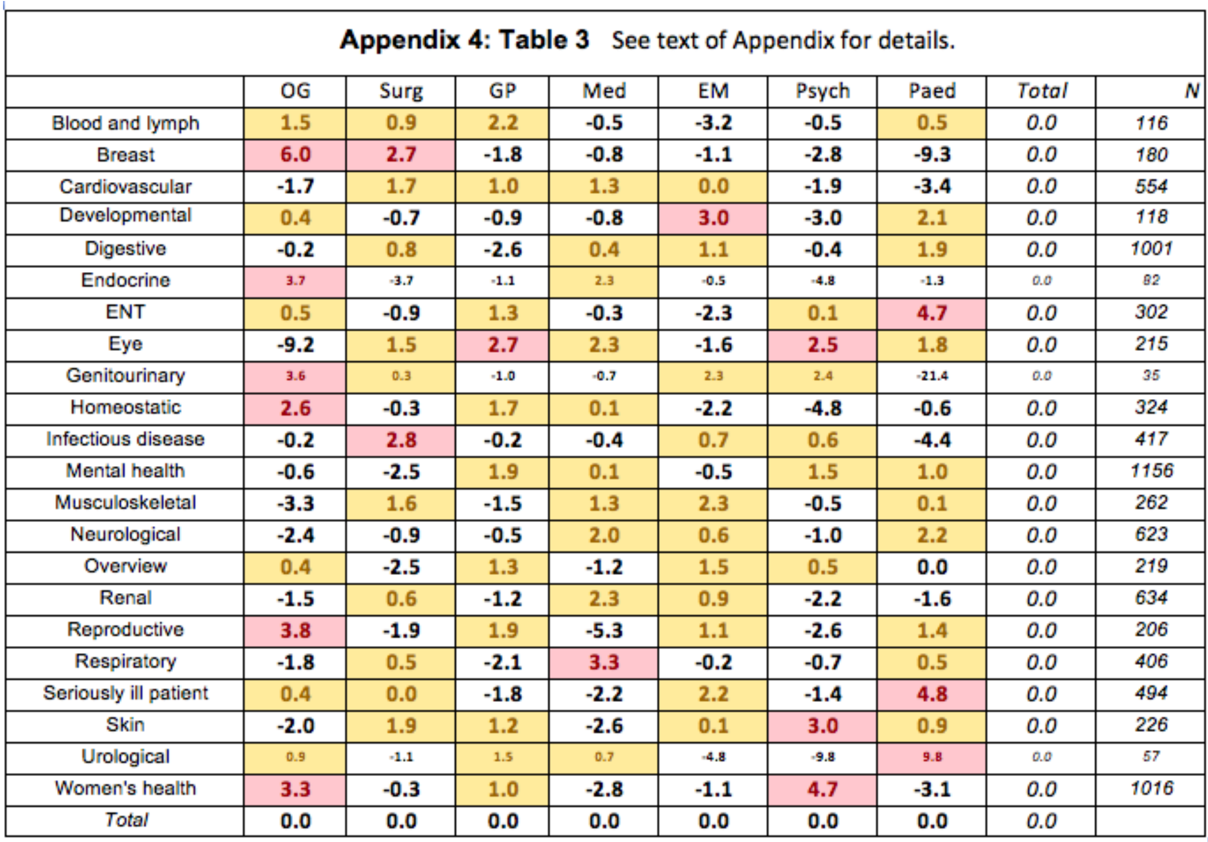
Appendix 4: Table 2 shows the mean standard for items in each category according to examiner specialty. Items with a mean above 6.5 have, arbitrarily, shown in pink. Without going into details, many of the patterns make sense, as for instance that Breast questions are rated higher by OG and Surgical specialties, or Neurological questions by Medical and Paediatric specialists. Looking at the marginal totals it is also clear that some question areas, and some examiner specialties, have higher standards than other.



Appendix 4: Table 2 See text of Appendix for details.

	OG	Surg	GP	Med	EM	Psych	Paed	*Total*	*N*
Blood and lymph	6.35	6.28	6.30	6.00	5.67	5.86	6.00	*6.07*	*116*
Breast	7.11	6.76	6.21	6.28	6.18	5.93	5.33	*6.38*	*180*
Cardiovascular	6.50	6.83	6.66	6.65	6.46	6.19	6.09	*6.54*	*554*
Developmental problems	6.36	6.23	6.11	6.08	6.40	5.73	6.27	*6.19*	*118*
Digestive	6.50	6.60	6.15	6.41	6.42	6.20	6.47	*6.40*	*1001*
Endocrine	7.00	6.25	6.40	6.71	6.36	5.86	6.25	*6.50*	*82*
ENT	6.49	6.34	6.45	6.26	6.00	6.16	6.67	*6.31*	*302*
Eye	5.43	6.49	6.50	6.42	5.97	6.31	6.28	*6.22*	*215*
Genitourinary	6.75	6.40	6.17	6.17	6.40	6.33	4.00	*6.26*	*35*
Homeostatic	6.71	6.42	6.51	6.31	6.02	5.68	6.14	*6.32*	*324*
Infectious disease	6.00	6.28	5.88	5.82	5.87	5.79	5.33	*5.89*	*417*
Mental health	6.25	6.05	6.39	6.17	6.05	6.17	6.16	*6.18*	*1156*
Musculoskeletal	5.95	6.43	6.02	6.26	6.29	5.94	6.04	*6.15*	*262*
Neurological	6.20	6.34	6.28	6.48	6.28	6.05	6.42	*6.31*	*623*
Overview	6.41	6.12	6.38	6.11	6.31	6.13	6.13	*6.25*	*219*
Renal	6.26	6.46	6.17	6.48	6.28	5.90	6.00	*6.28*	*634*
Reproductive	6.80	6.21	6.49	5.73	6.31	5.86	6.30	*6.28*	*206*
Respiratory	6.25	6.47	6.11	6.62	6.20	6.08	6.24	*6.31*	*406*
Seriously ill patient	6.66	6.61	6.33	6.24	6.63	6.19	6.86	*6.49*	*494*
Skin	6.12	6.50	6.32	5.91	6.12	6.33	6.16	*6.19*	*226*
Urological	7.36	7.14	7.30	7.18	6.57	6.00	8.00	*7.14*	*57*
Women's health	6.53	6.16	6.19	5.77	5.88	6.38	5.65	*6.07*	*1016*
Total	*6.40*	*6.39*	*6.28*	*6.25*	*6.18*	*6.11*	*6.15*	*6.27*	
N	*1635*	*1455*	*1402*	*1569*	*1249*	*607*	*726*		*8643*

Appendix 4: Table 3 shows the same data as table Appendix 4: Table 2 but with all values expressed as means relative to the column and row totals, positive values indicating that the item is set at a higher level, and negative values at a lower level. Coloured cells are positive, with pink higher than orange. Many of the higher values makes sense, although not all do so (and not all are based on the same sample sizes).



Appendix 4: Table 3 See text of Appendix for details.

	OG	Surg	GP	Med	EM	Psych	Paed	*Total*	*N*
Blood and lymph	1.5	0.9	2.2	-0.5	-3.2	-0.5	0.5	*0.0*	*116*
Breast	6.0	2.7	-1.8	-0.8	-1.1	-2.8	-9.3	*0.0*	*180*
Cardiovascular	-1.7	1.7	1.0	1.3	0.0	-1.9	-3.4	*0.0*	*554*
Developmental	0.4	-0.7	-0.9	-0.8	3.0	-3.0	2.1	*0.0*	*118*
Digestive	-0.2	0.8	-2.6	0.4	1.1	-0.4	1.9	*0.0*	*1001*
Endocrine	3.7	-3.7	-1.1	2.3	-0.5	-4.8	-1.3	*0.0*	*82*
ENT	0.5	-0.9	1.3	-0.3	-2.3	0.1	4.7	*0.0*	*302*
Eye	-9.2	1.5	2.7	2.3	-1.6	2.5	1.8	*0.0*	*215*
Genitourinary	3.6	0.3	-1.0	-0.7	2.3	2.4	-21.4	*0.0*	*35*
Homeostatic	2.6	-0.3	1.7	0.1	-2.2	-4.8	-0.6	*0.0*	*324*
Infectious disease	-0.2	2.8	-0.2	-0.4	0.7	0.6	-4.4	*0.0*	*417*
Mental health	-0.6	-2.5	1.9	0.1	-0.5	1.5	1.0	*0.0*	*1156*
Musculoskeletal	-3.3	1.6	-1.5	1.3	2.3	-0.5	0.1	*0.0*	*262*
Neurological	-2.4	-0.9	-0.5	2.0	0.6	-1.0	2.2	*0.0*	*623*
Overview	0.4	-2.5	1.3	-1.2	1.5	0.5	0.0	*0.0*	*219*
Renal	-1.5	0.6	-1.2	2.3	0.9	-2.2	-1.6	*0.0*	*634*
Reproductive	3.8	-1.9	1.9	-5.3	1.1	-2.6	1.4	*0.0*	*206*
Respiratory	-1.8	0.5	-2.1	3.3	-0.2	-0.7	0.5	*0.0*	*406*
Seriously ill patient	0.4	0.0	-1.8	-2.2	2.2	-1.4	4.8	*0.0*	*494*
Skin	-2.0	1.9	1.2	-2.6	0.1	3.0	0.9	*0.0*	*226*
Urological	0.9	-1.1	1.5	0.7	-4.8	-9.8	9.8	*0.0*	*57*
Women's health	3.3	-0.3	1.0	-2.8	-1.1	4.7	-3.1	*0.0*	*1016*
Total	0.0	0.0	0.0	0.0	0.0	0.0	0.0	*0.0*	

The overall effect of being specialist was found by finding the mean relative standard set for areas in the specialist area (a value of +1.45%) and those not in the specialist area (a value of -1.17%), giving a difference of 2.62%.

Bibliography

Reference List

1. *PLAB Test Review Group: Final Report [July 2004; Chair: The Lord Patel]*. London: General Medical Council; 2004.

2. General Medical Council: *Steering Group on the review of the PLAB test: Report: November 1999*. London: General Medical Council; 1999.

3. General Medical Council: *Report of the Working Party on the PLAB tests [January 1986; Chair Sir David Innes Williams]*. London: General Medical Council; 1986.

4. Kane M: **Validating the performance standards associated with passing scores.** *Review of Educational Research* 1994, **64:** 425-461.

5. Verheggen MM, Muijtens AMM, Van Os J, Schuwirth LWT: **Is an Angoff standard an indication of minimal competence of examinees of of judges?** *Advances in Health Sciences Education* 2008, **13:** 211.

6. Rogers JL, Howard KI, Vessey JT: **Using significance tests to evaluate equivalence between two experimental groups.** *Psychol Bull* 1993, **113:** 553-565.

7. Cleland J, Dowell J, McLachlan J, Nicholson S, Patterson F: *Identifying best practice in the selection of medical students (literature review and interview survey)*. London: General Medical Council (http://www.gmc-uk.org/Identifying_best_practice_in_the_selection_of_medical_students.pdf_51119804.pdf); 2012.

8. Kolen MJ, Brennan RL: *Test equating: Methods and practices*. New York: Springer-Verlag; 1995.

9. Newton P, Baird J-A, Goldstein H, Patrick H, Tymms P: *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority; 2007.

10. McManus IC, Elder AT, De Champlain A, Dacre JE, Mollon J, Chis L: **Graduates of different UK medical schools show substantial differences in performance on MRCP(UK) Part 1, Part 2 and PACES examinations.** *BMC Medicine* 2008, **6:5:** http://www.biomedcentral.com/1741-7015/6/5.

11. Anonymous: **PACES: Practical Assessment of Clinical Examination Skills. The new MRCP(UK) clinical examination.** *J R Coll Physicians Lond* 2000, **34:** 57-60.

12. Elder A, McAlpine L, Bateman N, Dacre J, Kopelman P, McManus IC: **Changing PACES: developments ot the examination in 2009.** *Clinical Medicine* 2011, **11:** 231-234.

13. Tighe J, McManus IC, Dewhurst NG, Chis L, Mucklow J: **The Standard Error of Measurement is a more appropriate measure of quality in postgraduate medical assessments than is reliability: An analysis of MRCP(UK) written examinations.** *BMC Medical Education (www biomedccentral com/1472-6920/10/40)* 2010, **10:** 40.

14. McManus IC, Ludka K: **Resitting a high-stakes postgraduate medical examination on multiple occasions: nonlinear multilevel modelling of performance in the MRCP(UK) examinations.** *BMC Medicine* 2012, **10:** 60.

15. McManus IC: *The marking and standard setting of PLAB Part 2*. London: Report to the PLAB Review, April 2013; 2013.

16. McManus IC, Woolf K, Dacre J, Paice E, Dewberry C: **The academic backbone: Longitudinal continuities in educational achievement from secondary school and medical school to MRCP(UK) and the Specialist Register in UK medical students and doctors.** *BMC Medicine* 2013, **Submitted**.

17. Wakeford R, Foulkes J, McManus IC, Southgate L: **MRCGP pass rate by medical school and region of postgraduate training.** *Brit Med J* 1993, **307:** 542-543.

18. Bowhay AR, Watmough SD: **An evaluation of the performance in the UK Royal College of Anaesthetists primary examination by UK medical school and gender.** *BMC Medical Education* 2009, **9**.

19. Rushd S, Landau AB, Khan JA, Allgar V, Lindow SW: **An analysis of the performance of UK medical graduates in the MRCOG Part 1 and Part 2 written examinations.** *Postgraduate Medical Journal* 2012, **88:** 249-254.

20. Plake BS, Cizek GJ: **Variations on a theme: The modified Angoff, extended Angoff, and Yes/No standard setting Methods.** In *Setting Performance Standards: Foundations, Methods and Innovations*. Edited by Cizek GJ. New York: Routledge; 2012:181-199.

21. Brandon PR: **Conclusions about frequently studied modified Angoff standard-setting topics.** *Applied Measurement in Education* 2004, **17:** 59-88.

22. Yeates P, O'Neill P, Mann K, Eva K: **Seeing the same thing differently: Mechanisms that contribute to assessor differences in directly-observed performance assessments.** *Advances in Health Sciences Education* 2013, **18:** 325-341.

23. Clauser BE, Mee J, Baldwin SG, Margolis MJ, Dillon GF: **Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study.** *JEM* 2009, **46:** 390-407.

24. Clauser BE, Harik P, Margolis MJ, McManus IC, Mollon J, Chis L *et al.*: **An Empirical Examination of the Impact of Group Discussion and Examinee Performance Information on Judgments Made in the Angoff Standard-Setting Procedure.** *Applied Measurement in Education* 2009, **22:** 1-21.

25. Hurtz GM, Auerbach MA: **A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus.** *EPM* 2003, **63:** 584-601.

26. Anonymous: *The American Heritage Book of English Usage: A practical and authoritative guide to contemporary English*. Boston: Houghton Mifflin; 1996.

27. Angoff WH: **Scales, norms, and equivalent scores.** In *Educational Measurement*. 2 edition. Edited by Thorndike RL. Washington, DC: American Council on Education; 1971:508-600.

28. Cizek GJ, (editor): *Setting performance standards: Foundations, methods and innovations (Second edition)*. New York: Routledge; 2012.

29. Lewis DM, Mitzel HC, Mercado RL, Schulz EM: **The bookmark standard setting procedure.** In *Setting Performance Standards: Foundations, Methods and Innovations*. Edited by Cizek GJ. New York: Routledge; 2012:225-253.

30. Schneid SD, Kingston PA, Apperson AJ: **Simplified bookmark method for local medical school examinations.** *Med Educ* 2011, **45:** 533-534.

31. Lypson ML, Downing SM, Gruppen LD, Yudkowsky R: **Applying the Bookmark method to medical education: Standard setting for an aseptic technique station.** *Med Teach* 2013, **35:** 581-585.

32. McManus IC, Richards P: **An audit of admission to medical school: 1. Acceptances and rejects.** *Brit Med J* 1984, **289:** 1201-1204.

33. Dewhurst NG, McManus IC, Mollon J, Dacre JE, Vale JA: **Performance in the MRCP(UK) Examination 2003-4: Analysis of pass rates of UK graduates in the Clinical Examination in relation to self-reported ethnicity and gender.** *BMC Medicine* 2007, **5:8:** www.biomedcentral.com/1741-7015/5/8/abstract- doi:10.1186/1741-7015-5-8.

34. McManus IC, Elder AT, Dacre J: **Investigating possible ethnicity and sex bias in clinical examiners: An analysis of data from the MRCP(UK) PACES and nPACES examinations.** *BMC Medical Education* 2013, **in press**.

35. McManus IC, Lissauer T, Williams SE: **Detecting cheating in written medical examinations by statistical analysis of similarity of answers: pilot study.** *Brit Med J* 2005, **330:** 1064-1066.

36. Esmail A, Roberts C: *Independent review of the Membership of the Royal College of General Practitioners (MRCGP) examination [published 26th September 2013]*. London: General Medical Council (http://www.gmc-uk.org/MRCGP_Final_Report__18th_September_2013.pdf_53516840.pdf); 2013.

37. Esmail A, Roberts C: **Academic performance of ethnic minority candidates and discrimination in the MRCGP examinations between 2010 and 2012: analysis of data.** *Brit Med J* 2013, **347**.