General Medical Council

# External review of GMC PLAB 2019


## Professor Adrian Freeman MMedSci FRCGP


**August 2019**

## Introduction

The Professional and Linguistic Assessments Board (PLAB) test is a very high stakes examination with significant financial consequences for the candidates and major resource implications for NHS employers. The number of candidates is rising year on year. In 2018, over 7000 candidates took PLAB Part 1 and over 5000 took PLAB Part 2. The GMC has always sought to be open about its processes, seeking advice from experienced psychometricians and publishing detailed statistical reports on its website. At this stage it was felt appropriate to seek an external review which was more qualitative in nature.

The PLAB test has two parts. Part 1 is a written exam and Part 2 is a clinical exam. Both parts were reviewed using the Kane Validity framework(1).

## What is the intended use or purpose of the assessment?

This should be clearly stated and be understood by all stakeholders.

On the opening page of the website includes this statement:

*"The Professional and Linguistic Assessments Board test, or the PLAB test, helps us to make sure doctors who qualified abroad have the right knowledge and skills to practise medicine in the UK."*

Successful candidates can apply for registration with a license to practice allowing them to work as a doctor in the UK.

## Content of the tests

There is a comprehensive blueprint available to download on the website. As is commonly found, the more comprehensive the blueprint is the more complex it is to read. However, there is an additional section on the website that clearly explains how to navigate the blueprint and how to get the most out of it.

The domains arise from: the GMC *Good Medical Practice*; *Outcomes for provisionally registered doctors* and the UK Foundation Programme Curriculum. There are hyperlinks to each of these documents. The domains of the blueprint are therefore covering '*the right knowledge and skills to practise medicine in the UK'.*

Using data from READ codes and Hospital Episode Statistics ensures the content validity of the blueprint matches to the work of a junior doctor working in the UK.

The test format mirrors common practice in 'finals' for UK Medical Schools. A single best answer paper tests knowledge and an OSCE style exam tests clinical skills. The standard of the PLAB test is of a doctor entering the F2 year in the UK.

## Internal Structure

### Part 1

*Examination format*

This was established as a Single Best Answer knowledge test with 5 answer options, only one of which is correct. Choosing the correct answer generates one mark. There is no correction for guessing.

There are 180 questions and the time for the examination is 3 hours. One question per minute is an appropriate time allowance aligning with the structure of many other medical examinations.

It should be noted that for most candidates English will not be their first spoken language. Analysis of Part 1 has shown strong correlations between Part 1 scores and the scores in the International English Language Testing System (IELTS).

**Recommendation: The GMC might want to consider some other analyses of Part 1, e.g. Percentage of papers incomplete, comparison of scores at the end section of the paper, etc., to ensure that the time allowance is appropriate.**

*Item writing*

The items are written in house creating a secure question bank which has around 2000 items. Items are constructed by a group of dedicated item writers. That group is augmented

by members of the Part 1 and Part 2 panel who have question writing experience. Recognising the expanding demand, the GMC is currently recruiting more item writers.

Appointment of item writers is an open and transparent process selecting according to appropriate criteria as illustrated in this recent advertisement for item writers:

> *[Applicants should be of Consultant, General Practitioner (or equivalent senior academic status) or Specialist Trainees (ST3 and above) or equivalent Staff and Associate Specialist (SAS) Grade with knowledge and skills in the specialties required. We currently have vacancies in all specialities that are relevant to Foundation practice, with the exception of paediatrics. They should also have a knowledge and understanding of the duties and competencies expected of a doctor successfully completing Foundation Year 1.*
>
> *They should have experience of multiple choice question writing (particularly single best answer questions) in a peer review setting and/or peer-reviewed publications on assessments. Involvement with the Medical Schools Council Assessment Alliance or experience in writing questions for final year medical students would be an advantage.]*

After selection, item writers have a training day to ensure that the items are written in the correct style. Item writing is a group process with peer review of the items. All items aim to be application of knowledge rather than simple factual recall. There has been a formal style guide which is now aligning to the MSCAA (Medical Schools Council Assessment Alliance) style guide. Normal values for results are embedded within the question.

Each item is labelled to one of the three domains in the blueprint: Applying knowledge and experience to clinical practice; Good clinical care: assessment and Good clinical care: management. Of course, this labelling has limitations as application of knowledge will often cross over more than one domain.

*Standard setting*

The standard of the paper is set with an Angoff method. It is stated that ideally there should be a minimum number of 10 judges for this process. The standard setting meeting that I attended did not quite achieve that minimum but had a credible number of judges. The judges at that meeting were experienced item writers covering an appropriate range of specialities.

There was no introductory calibration discussion to align the judges at the right level and I witnessed some judgements that were mixing the '*should* and *would* get this question right' statements.

**Recommendation: The GMC could consider an appropriate calibration discussion to start each Angoff session.**

Previously the Angoff process was for each individual paper. The GMC have carried out some interesting analyses that demonstrate for these papers, the Angoff standard is equivalent if you use the historical standard setting decisions summating each individual item, rather than the paper as a whole, i.e. each item is put through the process only once. That allows more effective use of resources and will be the future methodology.

Psychometric reports are available on the GMC PLAB website for 2016/7. These are detailed analyses including reporting of facility against Blueprint areas. Generally speaking, item facility falls within 0.6 – 0.7. Cronbach Alpha is consistently around 0.9 which indicates a very reliable assessment. There is a pass rate for each test of around 70%.

In the past the panel, which meets four times a year, had a psychometric review of the tests in that quarter. Due to individual retirement, that practice was stopped.

**Recommendation: From discussion with the panel lead, it would help the function of the panel to reinstate that practice if feasible.**

## Part 2

*Exam format*

Part 2 is an OSCE exam. There are 18 stations. Each station lasts 8 minutes. The exam takes place in a dedicated test centre. The facilities of the centre allow for an appropriate mix of content of stations including role players and equipment stations including Metiman. Real patients are not used.

Until August of this year there has only been one circuit available. There are now two circuits which will run in parallel. This will obviously have an effect on logistics and that has been planned for.

Each case is marked with three domains – Data gathering, Clinical Management and Interpersonal skills. Each case is expressed as having a principal domain as a label.

Each circuit is created to sample broadly from the blueprint with six mandatory scenarios (Seriously ill patient, child, Elderly, Reproductive, Mental health and Ethical & Professional).

Previously the stations for the circuit were selected by hand. The process has now been automated through the exam management system. It ensures that there are six stations from each of the three domain labels.

There is a similar approach to Part 1 for case writing with a dedicated group of writers. There are some 350 cases in the bank.

The standard setting process is a borderline regression to provide a cut score. One standard error of the mean (SEM) is then added to achieve the pass mark. In addition, the candidate must 'pass' 11 of the 18 stations as a minimum.

Detailed psychometric analyses are presented to the Part 2 panel. Typically, reliability is between 0.7 and 0.8 with a corresponding SEM around 8. Stations are highlighted if facility is above 0.8 or below 0.5 and if discrimination is below 0.5. Highlighted stations are discussed by the panel. One assumption for a high facility is recognised as possible leakage, which is an understandable consequence of such a high stakes examination. The panel may decide to 'rest' stations under these circumstances.

**Recommendation: A more efficient solution may be to clone them which would then decrease exposure of stations and diminish the effectiveness of candidate recall.**

**Recommendation: It would be advisable to carry out some simple psychometric tests before the results are released in case a station has performed so badly that it should be suppressed.**

## Response process

### Part 1

The MCQ examination is delivered on paper. The examination is delivered in test centres in the UK and also internationally (under the auspices of the British Council).

UK test centres are at GMC offices in London, Manchester, Wales, Northern Ireland and Scotland. The London and Manchester exams are run by a private company called Invigilation Services, who have a contract which specifies secure arrangements for handling exam material. They provide invigilators at a ratio of 25 to 1. Smaller venues are run and invigilated by GMC staff.

The PLAB website gives a clear description of the style of question, with 40 sample questions. There are clear instructions about how the test will run, what candidates are expected to bring with them and what they are not allowed to have in the exam. The website also gives clear regulations, specifically with detailed descriptions of what is considered misconduct and the process and consequences of any misconduct.

Reasonable adjustments will be given for disability and examples are given on the website.

The marking of the paper is standard machine marking, outsourced to a company called SCD. The mark sheets from the international centres are both photocopied and couriered back to the UK.

### Part 2

The OSCE test is delivered in the GMC's dedicated test centre. Again, the website gives clear instructions to candidates and an example of a typical station. I witnessed a day of the Part

2 exam. It was efficiently delivered with appropriate consideration for the candidates. There were appropriate quarantine arrangements between sessions.

A doctor can apply to become a PLAB examiner if they are on the specialist or GP register or ST3 or above or a staff doctor with a minimum of 2 years' experience in that post. They must be familiar with the standards of doctors entering the F2 year. Examiners from different ethnic backgrounds are encouraged to apply.

Potential examiners submit a structured online application. If accepted the new examiners will attend a training day and observe the OSCE in action. In training they carry out group marking exercises. Appropriate emphasis is placed in training on calibrating decisions to the correct level.

Although there is no checklist marking, for each case, there are positive and negative descriptors to assist the examiners. Borderline regression has four levels – Good, Satisfactory, Borderline and Unsatisfactory. Again, training is provided for new examiners.

Examiners mark on iPad. This ensures completion and security of data. For an 18 station OSCE running twice in the day there is a considerable cognitive load on the examiners. The cognitive load of that many stations has to be balanced against the enhanced reliability resulting from appropriate sampling. From my observation, the examiners maintained concentration and appropriate marking.

For each day, there is a briefing for candidates and also (separately) for examiners. There is a senior examiner, chief investigator (CI) assigned for each day. Each station on the circuit has video link and the CI is able to track any reported difficulties (candidates, examiners, role players etc) from a video control room.

On the training day for new examiners, emphasis is placed on the importance of calibration for the correct running of the exam. This is to help the role players and to assist in consideration of the scoring. This emphasis is repeated on the day briefing for the examiners.

However, when I attended the exam, this calibration was not prominent. The sometimes poor engagement with calibration was also reported to the panel meeting. With two circuits running in parallel correct calibration becomes crucial to ensure that candidates are having a consistent experience.

**Recommendation: The assessment might be enhanced by written descriptions of what should be happening during the calibration meeting and it may be that more time should be assigned to this process.**

### Role players

The PLAB 2 uses a commercial agency (Medical Role Players (UK)) of professional actors to simulate medical roles e.g. patients, colleagues, etc. The agency aims to ensure that these

role players (RP) match as closely as possible to the requirements of the cases. The RPs receive appropriate training through the agency.

On each examining day there is a facilitator from the agency to assist in the logistics and also to provide formal quality assurance (QA). Principally this is carried out using the video surveillance system. They aim to QA all RPs at least twice. If there are persisting problems with individual RPs, they will be removed from examining. The agency provides detailed reports of the QA to the panel. The examples of the RP paperwork/script that I saw were appropriate.

## Consequences

### Part 1

Candidates receive the results six weeks after the test. They are given their overall score, the pass mark and the average marks for that test. They are also given a breakdown of marks by domain.

If they pass the test, they are then eligible to take Part 2. If they fail, they have an automatic right to a total of four attempts. If they fail four attempts, they must complete an additional 12 months of clinical training with supporting documentation from a senior doctor as supervisor before they can take a further attempt.

There is an appeal process, but only if there is clear evidence of irregularity in the conduct of the exam or there were exceptional personal circumstances which affected the candidate's performance. There must be proof of those circumstances and they should be submitted within 3 working days of the exam.

### Part 2

For the clinical exam, candidates receive the marks for each of the three domains for each station and totals for those domains across all stations. They also receive their station score and passing score for the station as well as their total score and the passing score

The examiners iPads allow some qualitative feedback selecting from a series of pre prepared statements about aspects of the candidate's performance.

Regulations about resits and appeals are the same as for Part 1.

If they pass the test, candidates can apply for registration with a license to practice.

## Correlation

There are no regular sources of correlation. However, some studies have been published looking at PLAB candidates and postgraduate specialty tests. In particular a paper by

McManus and Wakeford looked at correlations between PLAB and MRCP and MRCGP exams (2). Although they found that PLAB scores correlated well with scores in specialty exams, the PLAB candidates scored lower than UK graduates. Similarly, a study by Tiffin demonstrated similar differences looking at the Annual Review of Competence in the specialty training programs (3).

These studies were part of a thorough review of PLAB which led to some significant changes, particularly in Part 2, with longer stations, increased number of stations from 14 to 18 and change to a borderline regression standard setting mechanism.

## Summary

The PLAB test demonstrates high validity. It has developed over time, informed by best evidence in assessment practice. It is outward looking and continues to seek to develop. I have suggested a few considerations (in bold).

| | |
|---|---|
| **Recommendation** 1 | **The GMC might want to consider some other analyses of Part 1, e.g. Percentage of papers incomplete, comparison of scores at the end section of the paper, etc., to ensure that the time allowance is appropriate.** |
| 2 | **The GMC could consider an appropriate calibration discussion to start each Angoff session.** |
| 3 | **From discussion with the panel lead, it would help the function of the panel to reinstate that practice if feasible.** |
| 4 | **A more efficient solution may be to clone them which would then decrease exposure of stations and diminish the effectiveness of candidate recall.** |
| 5 | **It would be advisable to carry out some simple psychometric tests before the Part 2 results are released in case a station has performed so badly that it should be suppressed** |
| 6 | **The assessment might be enhanced by written descriptions of what should be happening during the calibration meeting and it may be that more time should be assigned to this process.** |

# References

1. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Medical Education. 2015;49(6):560-75.

2. McManus IC, Wakeford R. PLAB and UK graduates' performance on MRCP(UK) and MRCGP examinations: data linkage study. Bmj. 2014;348:g2621.

3. Tiffin PA, Illing J, Kasim AS, McLachlan JC. Annual Review of Competence Progression (ARCP) performance of doctors who passed Professional and Linguistic Assessments Board (PLAB) tests compared with UK medical graduates: national data linkage study. Bmj. 2014;348:g2622.

# The author

Professor Adrian Freeman MMedSci FRCGP, is Director of Assessments in University of Exeter Medical School.

President of European Board of Medical Assessors

Long standing examiner for MRCGP

Chair MRCGP INT accreditation panel

Educational Adviser for MRCP

Board Member of Medical Schools Assessment Alliance

Council member of Academy of Medical Educators

External examiner at many institutions nationally and internationally