Work Psychology Group
Thinking differently

Exploring the Relationship between General Practice Selection Scores and MRCGP Examination Performance

Professor Fiona Patterson
Dr Maire Kerrin
Helen Baron
Safiatu Lopes

September 2015 Final Report

Contents

1 Introduction

1.1 As many people's first point of contact with the NHS, around 90% of patient interaction is with primary care services, in particular General Practice (GP) [1]. Selecting the right doctors into GP training is imperative due to both the importance of the work and the particular combination of knowledge and skills it requires. There is a high level of investment in training, estimated at £488,730 per UK trainee achieving GP Certificate of Completion of Training (CCT) [2] and a standard of practice, assessed by the MRCGP is required to qualify. Most trainees progress through training well and meet or exceed the required standard at first attempt, but there is a small minority of trainees who consistently struggle to meet the criteria for independent practice.

1.2 Since 2009/10, there has been a steady decline in numbers of those who apply for training in General Practice, as well as the number of qualified GPs that are actively practising [3]. The need to significantly increase the number of practicing GPs highlights the critical importance of maximising the number of trainees who qualify. Identifying early those trainees who are likely to struggle to meet the required standard and therefore require targeted support in training could promote a higher qualification rate. As well as improving retention during training, such intervention has the potential to reduce training costs (a recent estimate suggested there was a cost of £40,000 per doctor for a typical six month extension to training [4]).

1.3 A number of large scale meta-analytic studies exploring various assessments used in medical education, both at undergraduate and postgraduate level, have found differences in candidate performance according to ethnicity [5] [6] [7]. Specifically, Black and Minority Ethnic (BME) candidates have repeatedly been found to perform less well than comparable White candidates. This effect is significantly more apparent for those who have received training outside of the United Kingdom – EEA (European Economic Area) or IMG (International Medical Graduate). Identifying the specific causes of these differences continues to be problematic since candidates' ethnicity is strongly confounded with place of medical qualification (PMQ), with a majority of IMGs coming from ethnic minorities [7].

1.4 While the issue of differential attainment in medical assessment is not limited to the UK [8], it is a concern to ensure that trainees from BME backgrounds are not being unfairly discriminated against when it comes to their prospects and progression in the medical training pathway. Where there is differential attainment at assessments in the pathway, it is important to understand whether this reflects real differences in levels of performance or is an artefact of the assessment design or implementation. In particular, in terms of assessment design and evaluation, issues of validity are closely allied to issues of fairness. 'Validity' refers to how well a test measures what it is purported to measure. If a test is a valid measure of capability, then differences in test scores (between individuals or different groups of doctors) reflect real differences in performance, whereas if a test has little or no relationship with performance, differences in test scores will not consistently be reflected in outcomes and consequently

decisions based on such assessment results may not be fair. It is therefore important to establish the validity of the assessments used throughout training, as well as during the selection process, in order to ensure that issues of fairness have been addressed.

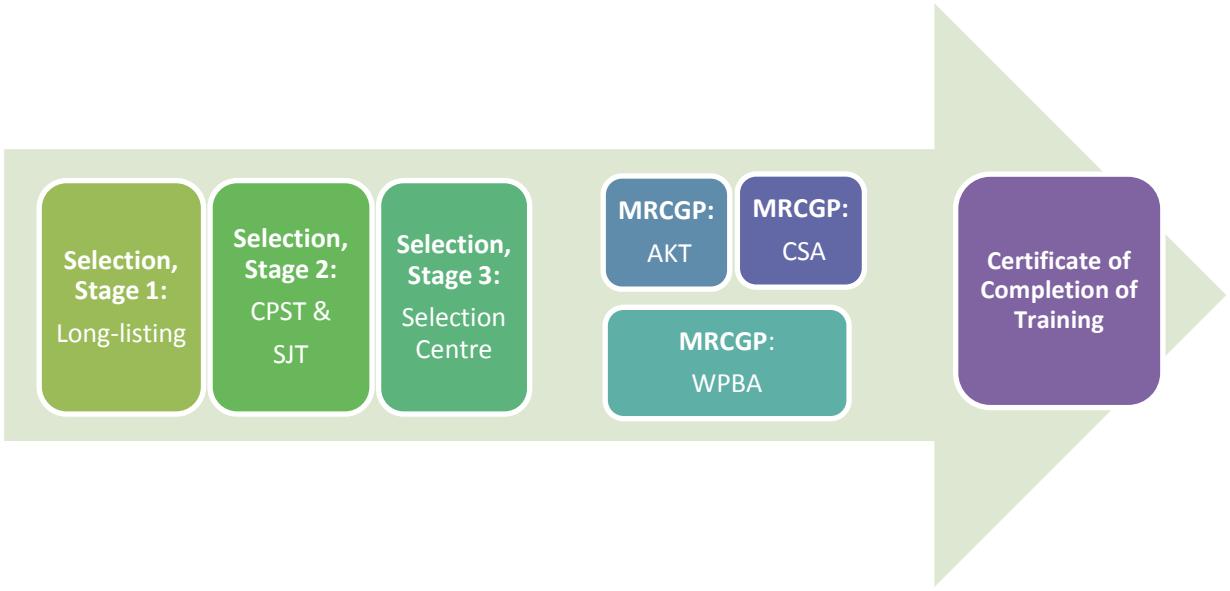
1.5 Effective assessments can not only ensure that doctors with the appropriate capabilities are selected into training, there is also potential to identify those who may find elements of the training more challenging and would benefit from additional support to enable them to best develop their skills. Tailoring the training to the needs of trainees could help ensure better success rates, increasing the overall number of doctors qualifying as GPs.

1.6 This study explores differential attainment in GP specialty training assessments between White and BME trainees, also taking into account PMQ, linking progression back to performance at selection, and examining the validity of each method of assessment.

1.7 The UK GP selection process

1.7.1 In order to apply for GP specialty training, prospective candidates must undergo a three stage, competency-based selection assessments (see **Fig. 1**), which is administered by the respective Local Education and Training Boards (LETBs) and Postgraduate Deaneries, and nationally coordinated by the National Recruitment Office (NRO). Candidates are scored at each stage of the process against core attributes identified as important for training in general practice (e.g. empathy, communication, clinical expertise, etc.). An aggregated summary score is then produced for each individual. The GP selection process is based on multi-source, multi-method job analysis [9] [10] and has been shown to be predictive of in-training performance [11].

Figure 1. General Practice Selection and MRCGP Assessment Pathway



Note: MRCGP = Membership of the Royal College of General Practitioners; CPST = Clinical Problem Solving Test; SJT = Situational Judgement Test; WPBA = Workplace Based Assessment; AKT = Applied Knowledge Test; CSA = Clinical Skills Assessment

1.7.2 *Stage 1* of the selection process requires candidates to first submit an application to up to four LETBs or Postgraduate Deaneries, with a longlisting process in place to ensure that eligibility criteria are met by each candidate. For IMGs – with the exception of those who have graduated from the European Economic Area (EEA) – the pathway to enter GP specialty training is likely to differ, with the registration route for Stage 1 being regulated by the General Medical Council (GMC). For example, most IMG candidates are required to achieve a minimum overall standard in the International English Language Testing System (IELTS) test. The IELTS consists of four components – Listening, Reading, Writing and Speaking. Candidates receive scores on a Band Score from 1.0 (Non User) to 9.0 (Expert User) for each component. Individual scores are then averaged and rounded to produce an Overall Band Score. The minimum requirement for the overall and individual component scores was raised in October 2010 to be 7.0 for the overall score, as well as each component (prior to this date and also within the period of study, the requirement was an overall score of 7.0, with a minimum of 7.0 in the speaking component and 6.0 in the other three components). IMG candidates who have not undergone Foundation Training must also sit the Professional and Linguistic Assessments Board (PLAB), which consists of two parts. The PLAB 1 is a computer-based written examination consisting of single best answer questions that assess clinical knowledge, whilst PLAB 2 is comprised of an Objective Structured Clinical Examination (OSCE), which takes the form of 14 clinical scenarios or 'stations' as well as a rest station and one or two pilot stations.

1.7.3 In *Stage 2*, all eligible UK, EEA and IMG candidates undertake a computer-delivered assessment focused on a Clinical Problem Solving Test (CPST) and a Situational Judgement Test (SJT; also referred to as the Professional Dilemmas Test or PDT). The CPST measures ability to apply clinical knowledge in a relevant context and to make clinical decisions in everyday practice. It aims to test synthesis and evaluation of medical knowledge. In the SJT, candidates are presented with a set of hypothetical work-relevant scenarios and asked to make judgements about possible responses. Following best practice, SJT scenarios are based on a thorough analysis of the job role to determine the key attributes and behaviours associated with competent performance in the role [11] [12] [13] The SJT is designed to measure empathy and sensitivity, professional integrity and coping with pressure.

1.7.4 In *Stage 3*, which is the final stage of selection, successful candidates are then invited to take part in a selection centre (based on a multi-trait, multi-method assessment approach) comprising a 30 minute written prioritisation exercise and three doctor-patient simulations using role actors as patients[1]. The selection centre is a test of aptitude for training in GP (not an OSCE assessment of competence), and should be compared to other interview processes for specialty training (e.g. three station interviews used by the RCP and Surgery CT1, for example). Once the Selection Centre (SC) is completed, allocation to LETB/Deanery takes place based on overall rank and trainee preferences.

[1] The GP selection centre structure changed in 2010. Prior to this it also comprised of a group simulation.

1.8 GP end-of-training assessment process

1.8.1 Once a candidate has been accepted into GP specialty training, their progress and performance is measured through an integrated MRCGP (Membership of the Royal College of General Practitioners) assessment system (see **Fig. 1**). This is devised of three separate components: an Applied Knowledge Test (AKT), a Clinical Skills Assessment (CSA) and Workplace Based Assessment (WPBA). Each of these test different competences using different assessment methods, which together cover the full spectrum of knowledge, skills, behaviours and attitudes defined by the GP curriculum.

1.8.2 Satisfactory completion of the MRCGP is a compulsory element of the curriculum in order to attain CCT. The number of attempts at each component of the MRCGP examination has been restricted to four attempts since 2010.

1.9 Context and rationale

1.9.1 The current study aims to identify predictors of poor trainee progression, in particular, likelihood of failing either the AKT or the CSA examination at first attempt, in order to establish appropriate interventions for improving success rates. As a large proportion of candidates who fail the MRCGP examinations at first attempt are IMGs, it is therefore important to examine IMGs differential attainment in further detail (assessing the value add of IELTS and PLAB scores, in addition to performance at selection) in order to improve the prediction of those who are likely to struggle early on in training.

1.9.2 In 2013, an independent review of the MRCGP licensure examination was published by Esmail & Roberts [14]. In their review, they found evidence of significant differences in failure rates between different groups of trainees in the CSA examination. In particular, BME trainees were found to significantly underperform compared to their White counterparts.

1.9.3 While Esmail & Roberts were able to explore the relationship of this outcome with previous assessment and entry routes to registration (i.e. PLAB and IELTS performance), detailed specialty recruitment data were not available at the time. Within the report, Esmail & Roberts recommended that "*Data from the selection scores of doctors recruited into general practice and held by the NRO should be integrated with CSA outcome data so that we can better understand the relationship between attainment at this level and CSA outcome"* (p. 20)

1.9.4 The current study expands on this previous work, by examining differences in the relationship between performance at selection and subsequent MRCGP assessments. While Esmail and Roberts were particularly interested in the relationship with ethnicity and PMQ, with a focus on the CSA, the overarching aims of this project were to; (i) explore the extent to which it is possible to identify trainees who are likely to struggle in training at an early stage, and whether there are any indicators within the current selection tests that can help

identify specific targeted support within training to accelerate 'time to competence', and (ii) better understand the *causes of differential attainment* on both the AKT and CSA.

1.9.5 Evidence of a positive correlation between selection data and the MRCGP examinations would lend support to the selection assessments being used to identify broad areas were trainees would benefit from support at the outset of training and hence optimise trainees' time to competence.

2 Method

2.1 Design and approach

2.1.1 The objective of the study is to identify variables that will identify candidates who are at risk of failing to reach the required standard of practice as measured by the AKT and the CSA and therefore the focus of the analysis is those who fail the assessments at their first sitting. The analysis aims to not only investigate the validity of each of the stages of assessment in isolation, but look beyond this to the incremental validity (added value) of using them together. The incremental validity is the degree to which prediction is improved by adding another selection measure.

2.1.2 We examined the following relationships in the full group of GP trainees (i.e. both UK, EEA and IMG trainees):

i. Performance on the GP selection tests (CPST and SJT)[2], (with and without controlling for age and sex) as predictors of AKT and CSA examination scores.

ii. Incremental validity of Selection Centre (SC)[3] scores over CPST and SJT scores, as a predictor of AKT and CSA examination scores. This analysis indicates the value added by the selection centre assessment process.

iii. Incremental validity of PMQ and BME status on the prediction of both the AKT and CSA examination scores. This analysis shows to what extent the differences in performance on the AKT and CSA by PMQ and Ethnicity can be accounted for by differences in performance at selection. Both the direct effects and interaction effects are considered.

2.1.3 For applicants who completed the IELTS and PLAB assessments, to assist in identifying candidates at risk of failing, the following analyses were conducted:

[2] Performance which resulted in a successful offer was used. During the period of study candidates were only allowed to sit the selection tests once within an application year, either in Round 1 or Round 2 of recruitment, but could resit again in the following year (which may still be within the study period).

[3] Performance which resulted in a successful offer was used. During the period of study candidates were allowed to resit the SC multiple times.

iv. Performance in the GP selection tests (CPST and SJT)[2], (with and without controlling for age and sex) as predictors of AKT and CSA examination scores.

v. Incremental validity of SC scores[4] over CPST and SJT scores, as a predictor of AKT and CSA examination scores.

vi. Incremental validity of PLAB assessments over all GP selection data (CPST, SJT and SC), as predictors of AKT and CSA examination scores.

vii. Incremental validity of IELTS assessments over all GP selection data (CPST, SJT and SC) and PLAB scores, as predictors of AKT and CSA examination scores.

2.2 Data matching

2.2.1 Data were compiled by the GMC and supplied to Work Psychology Group as an anonymised data file for analysis. A more detailed account of the data matching and data cleaning process is contained in a report held by the GMC.

2.2.2 Selection data for n=19,233 trainees between 2007 and 2012 were provided by the GP National Recruitment Office (NRO). Some discrepancies were found in the NRO data files. Data validation procedures were positive for n=16,996 cases. The largest source of non-validated cases came from the 2007 applicants, with less than 50% validating. Therefore those assessed in 2007 were dropped from the analysis. 98% of the remaining cases passed the case validation process.

2.2.3 MRCGP performance data for the years 2008 to midway through 2013 were provided by the Royal College of GPs (RCGP). AKT scores were matched for n=10,041 cases using trainees' GMC numbers. Over 90% of trainees from 2008 and 2009 had matched AKT scores, but the proportion was lower for later years as trainees may not have yet attempted the AKT. In addition, the AKT data file contained n=466 cases that did not match any of the NRO cases. They may have been trainees from earlier training cohorts. These individuals tended to have made more AKT attempts and have lower AKT scores. CSA scores were matched for n=7,340 cases. This reduction compared to the AKT reflects the fact that CSA is taken during ST3. There were n=343 cases in the CSA files that did not match any of the NRO cases. As for the AKT, these individuals tended to have had more attempts and lower scores on their first attempt at the CSA.

2.2.4 Trainees' PLAB and IELTS scores were provided by the GMC from their own records and matched in for those trainees who had taken the MRCGP examinations. Approximately 88% of the IMG trainees and 20% of the EEA trainees had matched scores. The GMC also provided demographic data from their records.

[4] Performance which resulted in a successful offer was used. During the period of study candidates were allowed to resit the SC multiple times.

2.2.5 Trainees may have sat selection and MRCGP assessments multiple times. For the purpose of the analyses reported here the following data were used:

a. Individual standardised CPST and SJT scores were used from the successful (last) attempt that resulted in the offer of a training place. This means that some individuals who first applied in 2007 could be included in the study if they were selected in a later year when there were no issues with data matching. Therefore after dropping trainees who were successful in 2007, the sample available for analysis consisted of n=10,028 trainees with AKT scores of whom n=2,179 had matched scores for the PLAB and IELTS assessments. There were n=7,333 trainees with CSA scores of whom n=1,623 had matched scores for the PLAB and IELTS assessments.

b. The selection centre (SC) score was used as an additional predictor. During the period of study (2007 to 2012), the calculation of the total SC scores varied, sometimes including a weighting of one or both of the CPST and SJT scores. For equivalence, a reweighted SC score (excluding CPST and SJT scores) was calculated and used in analyses for those years where one or both of the CPST and SJT scores were included as part of the total SC scores.

c. Personal information (including demographic data and PMQ).

d. Score from the first AKT and CSA sitting, expressed as a deviation from the passing score for the test form.

e. PLAB 1 and PLAB 2 scores, expressed as a deviation from the passing score for the test form for those trainees who had sat the assessment. The number of attempts before passing was also used as an indicator of the difficulty the candidate had in meeting the PLAB standard.

f. IELTS band scores for the overall and each of the four components.

2.3 Research methodology

2.3.1 SPSS Version 22.0 was used to conduct the analyses.

2.3.2 The research methodology used hierarchal linear regression analyses in order to establish the predictive validity of GP selection scores for AKT and CSA performance. Consistent with previous research, in all instances, data on first examination attempt was used. Both main and interaction effects were examined, and separate analyses looked at the moderating impact of English as a Foreign Language (ENFL) (IELTS scores), as well as the PLAB assessments on performance for those applicants who had taken these tests, predominantly IMG candidates. Some restriction of range analyses were used to better understand the differences in correlations between the full sample and specific subgroups.

3 Results

3.1 Sample

3.1.1 The sample consists of all trainees who entered training between 2008 and 2012, for whom validated selection scores could be matched to available AKT, CSA, IELTS and PLAB scores. There were different proportions of the total samples for each analysis because fewer trainees had completed the CSA than the AKT, and only IMGs had PLAB and IELTS scores.

3.1.2 **Table 1.1** below shows the sample sizes for the whole sample (n=10,028 for AKT and n=7,333 for CSA). **Table 1.2** compares two parts of the sample: the UK graduates (n=7,353 for AKT and n=5,323 for CSA) and the IMGs who had PLAB and IELTS scores available (n=2,179 for AKT and n=1,623 for CSA). The IMG group therefore comprises approximately a fifth of the whole sample and has markedly lower scores than the group as a whole on those assessments where a comparison was possible. The score difference is between 0.70 and 1 standard deviation (SD) in size compared to the total group scores. The difference between this group and the UK Graduates is larger with all the test score differences above 1 standard deviation. Only the selection centre is a little below this. Differences of this order of magnitude are highly statistically significant in these large samples.

3.1.3 The difference in scores between the full group and IMG group who took the PLAB and IELTS for the CPS are closer to 0.7 of a SD whereas the differences for the SJT are towards 1 SD. This mirrors the differences for the AKT which are towards the lower end whereas the difference for the CSA is nearer to a full SD.

Table 1.1 Sample Breakdown and Mean Scores

	Selection scores and AKT		Selection scores and CSA	
	Mean	SD	Mean	SD
N	10,028		7,333	
Age	29.8	4.9	29.8	4.9
% Female	59%		56%	
% Male	41%		44%	
% UK trained	73%		73%	
% EEA	24%		25%	
% IMG	3%		3%	
CPST	262.0	33.3	262.6	33.0
SJT	264.4	30.5	265.6	30.0
SC	81.5	7.3	81.3	7.3
AKT *(respective to pass mark)*	13.3	18.1		
CSA *(respective to pass mark)*			8.5	12.9

Table 1.2 Sample Breakdown and Mean Scores: UK trained versus PLAB and IELTS candidates

	Selection scores and AKT, UK Graduates		Selection scores, PLAB, IELTS and AKT		Selection scores and CSA, UK Graduates		Selection scores, PLAB, IELTS and CSA	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
N	7,353		2,179		5,323		1,623	
Age	28.1	3.7	34.2	4.6	28.1	3.7	34.1	4.5
% Female	59%		46%		56%		40%	
% Male	41%		54%		44%		60%	
CPST	271.0	28.0	236.8	33.7	271.2	28.1	239.7	33.6
SJT	274.7	24.4	234.3	26.7	276.0	23.8	236.5	26.4
SC	83.3	6.9	76.5	6.1	83.2	6.9	76.4	6.1
AKT *(respective to pass mark)*	18.1	15.6	-0.3	17.7				
CSA *(respective to pass mark)*					13.4	9.9	-4.8	10.5
IELTS			7.4	0.6			7.4	0.5
PLAB1			14.3	10.3			14.6	10.2
% pass PLAB1 first time			63%				65%	
No of attempts to pass PLAB1			1.7	1.2			1.6	1.2
PLAB2			7.3	3.5			7.4	3.5
% pass PLAB2 first time			79%				79%	
No of attempts to pass PLAB2			1.3	0.52			1.2	0.51

3.2 Descriptives

3.2.1 **Table 2.1** shows the correlations between each of the predictor variables (the CPST, SJT, SC, PLAB 1, PLAB 2, and IELTS) and outcome variables (AKT and CSA). Correlations provide an indication of the association between different variables; if two variables correlate positively and significantly, then high scores in one are associated with high scores in another. Generally, within selection research and practice a correlation of approximately r=0.10 is considered weak, r=0.30 is considered moderate, and r=0.50 and above is considered strong [15]. However, research has consistently shown that even correlations in the low range between selection and criterion data can have significant utility in a selection context.

3.2.2 All correlations were found to be positive. The strongest correlate of the AKT performance is the CPST score (r=0.73, p<0.001). This is a very strong correlation, and although a strong relationship was expected since both assessments are written knowledge tests, it is higher

than might be anticipated. Although the tests have some similarities there will typically be a substantial time gap between the two tests and therefore somewhat lower values were anticipated. It might be expected that the PLAB 1 which is also similar in format to the AKT would show a similar correlation, but while still substantial, it is much lower (r=0.37, p<0.001). This could be due to the restriction of the sample for the PLAB correlations (mostly taken by only IMG trainees). As a comparison, the correlation between the AKT and CPST was calculated for the same restricted sample (n=2,179) and presented in **Table 2.2.** While a little reduced (r=0.60, p<0.001), this is still substantially larger than the PLAB 1 and AKT correlation.

3.2.3 The correlation between the SC score (without any CPS or SJT components) with the AKT was smaller, but still statistically significant (r=0.31, p<0.001) in the full sample. The correlation reduces in the two part samples to 0.11 for the PLAB group and 0.18 for the UK trained group. Both of these values remain statistically significant (p<0.001) although small.

3.2.4 The correlates of the CSA were smaller but still substantial. The strongest predictor in the full sample was the SJT (r=0.54, p<0.001). The correlations in the sub samples are lower at 0.26 for the PLAB group and 0.29 for UK graduates. These values are both statistically significant (p<0.001) and for the PLAB group the SJT is still by a small margin the best predictor of the CSA scores, however for the UK graduates the CPST has a higher correlation with the CSA scores (r=0.37, p<0.001). The format of the SJT is very different from the CSA; the SJT is a low fidelity written exercise and the CSA consists of observed practical simulations and is thus, high fidelity. The CPST also has a strong correlation with the CSA in the full group (r=0.49, p<0.001). The SC, which is more similar in format to the CSA, is also significantly correlated with the CSA (r=0.42, p<0.001) in the full group. The correlation in the PLAB group is somewhat lower although still statistically significant (r=0.17, p<0.001). The PLAB 2 correlates are a little higher (r=0.24, p<0.001) with the CSA in this group, perhaps because the two measures have a very similar format. The SJT correlates higher than the PLAB 2 with the CSA, even in the restricted sample that completed all three tests (n=1,623, r=0.26, p<0.001).

Table 2.1. Correlations between Predictor and Outcome Variables

	SJT	SC	PLAB 1	PLAB 2	IELTS	AKT	CSA
CPST	.48** n=10,028	.31** n=10,028	.31** n=2,179	.12** n=2,179	.08** n=2,179	.73** n=10,028	.49** n=7,333
SJT		.32** n=10,028	.11** n=2,179	.19** n=2,179	.23** n=2,179	.47** n=10,028	.54** n=7,333
SC			.08** n=2,179	.15** n=2,179	.14** n=2,179	.31** n=10,028	.42** n=7,333
PLAB 1				.15** n=2,179	.10** n=2,179	.37** n=2,179	.18** n=1,623
PLAB 2					.16** n=2,179	.14** n=2,179	.24** n=1,623
IELTS						.13** n=2,179	.25** n=1,623
AKT							.55** n=7,326

Note: ** Significant at the 0.01 level (2-tailed).

Table 2.2. Correlations between Predictor and Outcome Variables for PLAB candidates (above diagonal) and UK Graduates (below diagonal)

	CPST	SJT	SC	PLAB 1	PLAB 2	IELTS	AKT	CSA
CPST		.28** n=2,179	.09** n=2,179	.31** n=2,179	.12** n=2,179	.08** n=2,179	.60** n=2,179	.23** n=1,623
SJT	.32** n=7,353		.09** n=2,179	.11** n=2,179	.19** n=2,179	.23** n=2,179	.27** n=2,179	.26** n=1,623
SC	.19** n=7,353	.13** n=7,353		.08** n=2,179	.15** n=2,179	.14** n=2,179	.11** n=2,179	.17** n=1,623
PLAB 1					.15** n=2,179	.10** n=2,179	.37** n=2,179	.18** n=1,623
PLAB 2						.16** n=2,179	.14** n=2,179	.24** n=1,623
IELTS							.13** n=2,179	.25** n=1,623
AKT	.70** n=7,353	.29** n=7,353	.18** n=7,353					.36** n=1,623
CSA	.37** n=5,323	.29** n=5,323	.25** n=5,323				.42** n=5,317	

Note: ** Significant at the 0.01 level (2-tailed).

3.2.5 Corrections for restriction of range ((Case 2c) [16] were applied to the correlations for the UK and PLAB groups to see to what extent the within group reduction in the correlations was caused by the restriction of variance in the sub-samples. The results should be treated with care as the correction formulae only provide an estimate of the unrestricted values. The results are shown in **Table 2.3**.

3.2.6 The correlations all increased, many to values similar to those in the unrestricted group, showing that the restriction of variance was a clear factor in the lower correlations. For the UK group, all the correlations rose to values similar to those in the full population, suggesting that for this group the drop in correlation was due to the restriction of variance. For the PLAB group, although the correlations rose substantially, not all reached the unrestricted values. The correlation between the SC and the CSA did, but the corrected CPST-AKT correlation and the SJT-CSA correlations for this group were a little lower than their full group values. The raw CPST-AKT correlation was already high at 0.60 and only rose to 0.63. The raw SJT-CSA correlation for this group was only 0.26, which was well below the full group value of 0.54. The corrected value of 0.46 was much closer. Overall this suggests that measurement is not quite as effective for the PLAB group. This could be the result of slightly lower reliability in either the selection measures (CPST, SJT and SC) or the criterion measures (AKT and CSA) or both for this group. This could have numerous causes. For example, if even a minority of the PLAB group struggled to understand the assessment tasks due to language difficulties, their results would be less accurate and this could lower the test reliability and validity across the group.

3.2.7 The selection assessments (CPST, SJT and SC) are therefore individually significant predictors of performance on the MRCGP assessments, particularly across the full sample. These correlations make the selection assessments good predictors of outcome performance and replicate the validity study results previously found using 2007 data [12]. The within sample prediction is lower and this is partly, if not wholly, accounted for by the restriction of range.

Table 2.3. Correlations between Predictor and Outcome Variables for PLAB candidates (above diagonal) and UK Graduates (below diagonal) corrected for restriction of range

	CPST	SJT	SC	AKT	CSA
CPST				.63** n=2,179	.52** n=1,623
SJT				.33** n=2,179	.46** n=1,623
SC				.20** n=2,179	.43** n=1,623
AKT	.78** n=7,353	.50** n=7,353	.31** n=7,353		
CSA	.58** n=5,323	.57** n=5,323	.39** n=5,323		

Note: ** Significant at the 0.01 level (2-tailed).

3.2.8 **Table 3** provides an overview of predictor and outcome scores by ethnic background and PMQ. A total of six groups were explored: White and BME IMGs, White and BME EEA trained, and White and BME UK trained.

3.2.9 As outlined in **Table 3**, there are clear differences in the pattern of results for the different groups. Compared to the mean for the total trainee group (as outlined in **Table 1.1**) there are differences between the assessment scores of more than one standard deviation and these large differences in mean scores will have a significant impact on pass rates for the different groups. The White UK group consistently perform the best on all of the assessments they take. Only the BME IMG group is large enough to provide a robust estimate of group differences with the White UK group among the non UK-qualified doctors and the differences are very similar for the selection tests and the MRCGP exams. The CPS and AKT show a difference of about 1.5 SDs. The SJT and CSA have an even bigger difference. With differences of this magnitude it is not necessary to carry out a significance test to be sure of the difference. Trainees qualified outside the UK tending to be older by four to seven years on average.

Table 3. Descriptives for CSA, AKT, IELTS and PLAB Examinations based on Ethnicity

Variables		Ethnicity and PMQ region					
		White UK	**BME UK**	**White EEA**	**BME EEA**	**White IMG**	**BME IMG**
Age (years) at application to training	Mean	28.2	27.8	32.6	32.9	35.5	34.6
	95% CI	28.1 – 28.3	27.7 – 27.9	31.9 – 33.2	32.0 – 33.8	34.6 – 36.4	34.4 – 34.8
	SD	3.9	3.0	4.5	4.8	5.5	4.9
	n	4,869	2,402	187	104	138	2,224
Male	N (%)	1,554 (31.9%)	1,031 (42.9%)	68 (36.4%)	58 (55.8%)	56 (40.6%)	1,257 (56.5%)
Female	N (%)	3,315 (68.1%)	1,371 (57.1%)	119 (63.6%)	46 (44.2%)	82 (59.4%)	967 (43.5%)
GP Selection Results							
CPST	Mean	275.9	261.5	248.2	223.5	240.0	236.5
	95% CI	275.2 – 276.6	260.3 – 262.7	243.3 – 253.0	216.8 – 230.0	234.1 – 245.8	235.1 – 237.9
	SD	25.9	29.3	33.3	34.0	34.5	34.0
	n	4,869	2,402	187	104	138	2,224
SJT	Mean	279.5	265.0	255.6	238.7	240.8	234.3
	95% CI	278.9 – 280.1	264.0 – 266.0	251.7 – 259.4	233.1 – 244.2	235.6 – 245.9	233.2 – 235.4
	SD	22.4	25.1	26.8	28.8	30.3	26.8
	n	4,869	2,402	187	104	138	2,224
SC	Mean	83.9	82.1	78.2	78.7	77.3	76.3
	95% CI	83.7 – 84.1	81.8 – 82.4	77.2 – 79.1	77.3 – 80.1	76.2 – 78.4	76.0 – 76.5
	SD	6.8	7.0	6.5	7.1	6.5	6.0
	n	4,869	2,402	187	104	138	2,224

Variables		Ethnicity and PMQ region					
		White UK	BME UK	White EEA	BME EEA	White IMG	BME IMG
AKT First Attempt (relative to pass mark)							
AKT total score	Mean	21.1	12.0	9.4	-10.8	-0.2	0.1
	95% CI	20.7 – 21.5	11.3 – 12.7	6.6 – 12.1	-14.9 – -6 .7	-3.3 – 2.8	-0.7 – 0.8
	SD	14.2	16.4	18.9	21.1	18.2	17.7
	n	4,869	2,402	187	104	138	2,224
Pass	%	92%	78%	74%	35%	55%	54%
CSA First Attempt (relative to pass mark)							
CSA total score	Mean	15.4	9.4	2.0	-3.9	-0.4	-5.1
	95% CI	15.1 – 15.7	8.9 – 9.9	0.0 – 3.9	-6.7 – 1.2	-2.8 – 2.0	-5.6 – -4.6
	SD	9.0	10.3	11.7	11.8	12.7	10.6
	n	3,505	1,770	142	73	111	1,670
Pass	%	95%	83%	65%	33%	55%	35%
PLAB Scores (Total score relative to pass mark)							
PLAB 1	Mean	n/a	n/a	18.0	11.9	14.8	14.2
	95% CI			13.5 – 22.4	9.3 – 14.5	12.5 – 7.1	13.7 – 14.6
	SD			11.4	7.7	11.8	10.2
	n			26	34	103	1,997
PLAB 2	Mean	n/a	n/a	7.6	7.1	8.4	7.3
	95% CI			6.1 – 9.1	6.0 – 8.2	7.6 – 9.2	7.1 – 7.5
	SD			3.9	3.3	4.1	3.5
	n			26	34	103	1,997
IELTS							
Overall	Mean	n/a	n/a	7.4	7.4	7.3	7.4
	95% CI			7.2 – 7.6	7.2 – 7.6	7.2 – 7.4	7.4 – 7.4
	SD			0.5	0.5	0.4	0.6
	n			26	34	103	1,997
Reading	Mean	n/a	n/a	7.2	7.0	7.2	7.2
	95% CI			6.9 – 7.5	6.8 – 7.2	7.1 – 7.3	7.2 – 7.2
	SD			0.8	0.7	0.7	0.8
	n			26	34	103	1,997
Speaking	Mean	n/a	n/a	7.7	7.8	7.4	7.5
	95% CI			7.4 – 8.0	7.5 – 8.1	7.3 – 7.5	7.5 – 7.5
	SD			0.7	0.8	0.6	0.7
	n			26	34	103	1,997
Understanding	Mean	n/a	n/a	7.3	7.5	7.3	7.5
	95% CI			7.0 – 7.6	7.3 – 7.7	7.2 – 7.4	7.5 – 7.5
	SD			0.7	0.7	0.6	0.8

Variables		Ethnicity and PMQ region					
		White UK	BME UK	White EEA	BME EEA	White IMG	BME IMG
	n			26	34	103	1,997
Writing	Mean 95% CI SD n	n/a	n/a	7.0 6.7 – 7.3 0.7 26	6.9 6.7 – 7.1 0.7 34	6.7 6.6 – 6.8 0.7 103	7.1 7.1 – 7.1 0.8 1,997

Note: IMG = International Medical Graduate; BME = Black and Minority Ethnic; EEA = European Economic Area; CI = Confidence Interval around the mean; SD = Standard Deviation; n = sample size.

3.3 Borderline performance in GP selection tests

3.3.1 Scores on the GP selection tests are converted to standardised scores, which have a mean of 250 and a standard deviation of 40. Candidates' scores are then reported in four bands. Band 1 is below the minimum required standard. The minimum standard is determined from time to time using an Angoff exercise. During the period of the study it ranged from 148 to 194 on the standardised scale. Candidates whose scores fall into Band 1 for either the CPST or SJT are not allowed to proceed to the next stage of selection (i.e. selection centre). Band 2 is just above the minimum required standard (scores that fall between the minimum standard and half a standard deviation below the mean – 230), while Band 3 is well above the minimum required standard (scores that fall between half a standard deviation below the mean and a standard deviation above the mean – 231 to 290) and Band 4 is the highest level (a score of 291 and above).

3.3.2 For the purposes of analysis, trainees whose scores fell into Band 2 for either the CPST or the SJT were classified as 'borderline'. Those with scores in Bands 3 and 4 for both tests were classified as having a 'clear' pass.

3.3.3 **Table 4** below shows the pass rate for different groups broken down by selection test scores into the borderline and clear passing groups.

Table 4. MRCGP Examination Pass rates by CPST and SJT scores, Ethnicity and PMQ

	GP Selection Performance (CPST & SJT)	
	Borderline at selection % n=2,573	**Clear pass at selection %** n=7,455
AKT Pass first time		
All candidates n=10,028	**46.2%**	**90.2%**
White UK trained n=4,869	62.6%	95.1%
BME UK trained n=2,402	47.5%	87.0%
White IMG trained n=138	38.6%	72.1%
BME IMG trained n=2,224	42.6%	74.3%
White EEA trained n=187	50.7%	88.1%
BME EEA trained n=104	21.9%	64.5%
CSA Pass first time		
All candidates n=7,333	**46.4%**	**86.7%**
White UK trained n=3,505	84.9%	96.5%
BME UK trained n=1,770	68.6%	87.0%
White IMG trained n=111	43.6%	66.1%
BME IMG trained n=1,670	27.2%	46.1%
White EEA trained n=142	54.9%	70.3%
BME EEA trained n=73	26.0%	47.8%

3.3.4 In summary, less than half of the borderline group are likely to pass both MRCGP assessments first time (46% for both the CSA and AKT). Conversely, the vast majority of trainees who attain a clear pass in both the CPST and SJT (Band 3 or Band 4) will pass both examinations at first attempt (90.2% for the AKT and 86.7% for the CSA). However, when the GP trainee population is broken down in terms of ethnicity and PMQ, there are apparent differences between groups. The White UK trained group who attain borderline scores in the

selection tests have better pass rates in the MRCGP exams, compared to other groups who attain borderline scores (62.6% for the AKT and 84.9% for the CSA). The BME IMG group has a particularly low pass rate on the CSA with an overall pass rate of 34.9%. Even BME IMGs with a clear pass on the selection tests do less well than White UK trained candidates with borderline scores (46.1% vs 84.9%). The next section investigates further variables to better understand the drivers for underperformance in IMG graduates.

3.4 Analysis of the added predictive power of different predictor variables

3.4.1 Hierarchal regression analyses were carried out for the full group of GP trainees (i.e. both UK and IMG trainees):

i. Performance on the GP selection tests (CPST and SJT), controlling for the effects of age and sex, as predictors of AKT and CSA examination scores.

ii. Incremental validity of SC scores over CPST and SJT scores, as a predictor of AKT and CSA examination scores.

iii. Incremental validity of PMQ and BME status on the prediction of both the AKT and CSA examination scores;

3.4.2 **Table 5** shows the results of the regression analysis to identify predictors of the AKT and CSA scores. The R-squared (R^2) change is reported for each level of the analysis, and is an indication of how much additional variance, in terms of a percentage total, in the AKT and CSA scores is explained by adding each new set of variables (i.e. how much variance is explained by age and sex alone, then by adding the CPS and SJT to age and sex, then by adding the SC, and so on). Put another way, this indicates the degree of incremental predictive validity that is contributed by each new predictor. The adjusted R^2 for the first stage in each analysis is provided in parenthesis if different. These values differed little, if at all, from the unadjusted value in samples of the size analysed here.

3.4.3 Analyses were run controlling for the possible effects of age and sex and then additional analyses were run omitting these variables. It can be desirable to control for variables such as age and sex as an initial step in the regression analysis so that differences between groups in their make-up does not affect the estimates of variance explained by different variables. In the current case, IMG trainees tended to be older and to be predominantly male, whereas UK trained trainees tend to be younger and predominantly female. However, because of this confound between age and sex and place of medical training, it is possible that the results after the control is applied, underestimate the explanatory power of the different assessments. Furthermore, practically it would not be appropriate to use trainees' age and sex when making decisions about whether to select them. Therefore, the model which controls for these variables cannot be used in practice, e.g. to identify candidates likely to fail the MRCGP assessments, even if it is potentially useful in understanding the relationships between variables from a more theoretical perspective. For this reason the analysis was carried out both with, and without, the control for age and sex.

3.4.4 Place of medical training and BME status was coded using five dummy variables where each individual received a score of one if they belonged to the group and zero if not. While it is possible to model these groups using two variables, PMQ and BME, this assumes that the meaning of the BME variable is the same for each region of training. However, this is not necessarily the case. For example, White candidates from the UK are almost entirely primary language speakers of English, whereas those from the EEA speak a variety of languages; 15% of BME IMG applicants are from African backgrounds but only 5% of UK trained IMG applicants come from this background; 7% of UK trained BME applicants are of Chinese backgrounds but less than 1% of IMG BME applicants are from this background. For this reason the six groups created by considering PMQ and BME status were treated as independent.

3.4.5 As well as the direct effect of these variables, the interaction terms (i.e. the product of the dummy variable with each of the selection tests) were also included in order to include all the variance due to grouping. Significance for main effect for these variables (i.e. PMQ and BME status) means that a particular group consistently performs better, or worse, than expected from their selection scores compared to the rest of the sample. Significance in the interaction term means that the relationship between the predictor (selection test performance) and outcome (AKT or CSA performance) is different for members of the group (e.g. BME IMG) compared to the remainder of the sample (i.e. not BME IMG). This would mean that the tests are better predictors for one group than the other.

3.4.6 The table also shows the semi-partial correlation (sr^2) for each variable in the equation at each stage, to indicate the individual contribution to variance explained. Higher values indicate variables that are stronger predictors of the outcome.

Table 5. Predicting MRCGP Examination Scores from Selection Performance and Demographics

Variables added to the equation at each level		AKT n= 10,028		CSA n=7,333	
		No control for age and sex	Controlling for age and sex	No control for age and sex	Controlling for age and sex
0. Age and sex	R^2 change %		**15.5%*** **		**28.5%*** **
	F statistic	n/a	920	n/a	1,463
	df		*2, 10,025*		*2, 7,330*
1. CPST & SJT	R^2 change %	**55.3%*** **	**41.1%*** **	**36.7%*** **	**16.8%*** **
	F statistic	6,211	4,745	2,120	1,124
	df	*2, 10,025*	*2, 10,023*	*2, 7,330*	*2, 7,328*
CPST	sr^2	33.8%***	29.7%***	7.6%***	4.2%***
SJT	sr^2	1.6%***	0.8%***	12.3%***	6.5%***
2. Selection Centre (SC)	R^2 change %	**0.4%*** **	**0.2%*** **	**4.3%*** **	**2.7%*** **
	F statistic	84	49	529	377
	df	*1, 10,024*	*1, 10,022*	*1, 7,329*	*1, 7,327*
CPST	sr^2	31.1%***	28.2%***	5.3%***	3.1%***
SJT	sr^2	1.3%***	0.7%***	8.9%***	5.0%***
SC	sr^2	0.4%***	0.2%***	4.2%***	2.7%***
3. PMQ and BME/White	R^2 change %	**1.3%*** **	**1.1%*** **	**9.1%*** **	**5.3%*** **
	F statistic	20	18	88	55
	df	*15, 10,009*	*15, 10,007*	*15, 7,314*	*15, 7,312*
CPST	sr^2	12.5%***	11.8%***	1.6%***	1.1%***
SJT	sr^2	0.1%***	<0.1%***	0.4%***	0.3%***
SC	sr^2	0.2%***	0.1%***	1.4%***	1.1%***
BME UK	sr^2	n.s.	n.s.	<0.1%*	<0.1%**
White EEA	sr^2	n.s.	n.s.	n.s.	<0.1%*
BME EEA	sr^2	<0.1%***	<0.1%***	n.s.	n.s.
White IMG	sr^2	n.s.	n.s.	<0.1%**	<0.1%**
BME IMG	sr^2	n.s.	n.s.	<0.1%**	<0.1%*
*BME UK*CPS*	sr^2	n.s.	n.s.	n.s.	n.s.
*White EEA*CPS*	sr^2	n.s.	n.s.	n.s.	n.s.
*BME EEA*CPS*	sr^2	n.s.	n.s.	n.s.	n.s.
*White IMG*CPS*	sr^2	<0.1%**	<0.1%**	n.s.	n.s.
*BME IMG*CPS*	sr^2	0.2%***	0.2%***	0.2%***	0.1%***
*BME UK*SJT*	sr^2	n.s.	n.s.	n.s.	<0.1%*
*White EEA*SJT*	sr^2	n.s.	n.s.	n.s.	<0.1%*
*BME EEA*SJT*	sr^2	<0.1%*	n.s.	n.s.	n.s.
*White IMG*SJT*	sr^2	<0.1%***	<0.1%***	n.s.	n.s.
*BME IMG*SJT*	sr^2	<0.1%**	<0.1%*	<0.1%**	<0.1%*

Note: *df* = degrees of freedom; sr^2 = semi-partial correlation; *n.s.*= not significant; *** Significant at the 0.001 level (2-tailed). ** Significant at the 0.01 level (2-tailed). * Significant at the 0.05 level (2-tailed).

3.4.7 The best predictor of performance on both the AKT and CSA is found to be the GP selection tests (CPST and SJT). Alone, they are able to account for 55.3% of the variance in AKT scores and 36.7% of the variance in CSA scores when age and sex are not controlled for. The SC scores have a smaller, but statistically significant incremental predictive value (0.4% for the AKT and 4.3% for the CSA) over and above the selection tests. However, when PMQ and BME status are added in the last step of the analysis, they also significantly improve the prediction of AKT and CSA scores, as they have incremental validity over the selection tests and the SC, with the effect being largest for the CSA (1.3% for the AKT and 9.1% for the CSA). Therefore, it seems that the difference in performance on AKT and CSA is not accounted for entirely by their differences on the selection tests.

3.4.8 A preliminary regression analysis considered the variance accounted for by the ethnicity and PMQ variables before the assessment data was considered. The R^2 change percentage was 23.2% (F=606, *df*=5, 10,022, *p*<0.001) and 11.1% (F=303, *df*=5, 10,020, *p*<0.001) without controlling for age and sex for the AKT and CSA respectively. With controls, for the AKT this was 40.9% (F=1012, *df*=5, 7,327, *p*<0.001) and for the CSA this was 18.0% (F=495, *df*=5, 7,325, *p*<0.001). Thus, the difference in performance on the NRO assessments accounts for a great deal of the group difference in the MRCGP performance.

3.4.9 The largest effects for PMQ and BME status are seen for the BME IMG group. The interaction term with the BME IMG group and the CPST is the largest effect with the interaction with the BME IMG group and the SJT also showing significant effects although smaller, for both outcome variables (AKT and CSA). These significant interaction effects mean that the relationship between the CPST and the SJT, and the outcome variables may be different for the BME IMG group compared to the rest of the sample. In this case, the correlation between the selection tests and the MRCGP assessments is lower for the BME IMG group than for the full sample (CPST and AKT *r*=.73 for full sample and *r*=.60 for the BME IMG group; for the SJT and CSA the values are *r*=.54 and *r*=.36 respectively). The selection tests are still good predictors for this group, but not quite as good as for the UK trained group. Restriction of variance in the scores has already been shown to be one cause of this drop. Another contribution to the drop could be a proportion of the IMG group having difficultly engaging with the selection tests; perhaps because of language difficulties or because of unfamiliarity with the context of medicine in the UK, or differences in the syllabus of study or approach to medicine in their place of training. There is also a main effect for the IMG groups in predicting CSA scores but this is difficult to interpret in the presence of the interaction.

3.4.10 When age and gender are entered into the equation first, they explain a substantial amount of variance. This seems to be particularly at the expense of the predictive power of the selection tests rather than the other variables, although this effect could also be due to a degree of confounding.

3.4.11 In summary the selection assessments are good predictors of later success in the MRCGP assessments. This means that the selection data could be used to identify trainees at

selection who may struggle to reach the standard required for the MRCGP exams and to be more likely to fail at first sitting.

3.4.12 The results of this set of analyses suggested a further investigation of IMG candidates, in order to identify potential indicators of why they perform less well on the AKT and CSA examinations. A further series of regression analyses was therefore undertaken to see whether PLAB or IELTs results are more predictive of examination performance for this group. The regression analyses were therefore repeated, with only those trainees who completed the IELTS and PLAB assessments, in order to explore the following:

i. Performance in the GP selection tests (CPST and SJT), controlling for age and sex as predictors of AKT and CSA examination scores.

ii. Incremental validity of SC scores over CPST and SJT scores, as a predictor of AKT and CSA examination scores.

iii. Incremental validity of PLAB assessments over all GP selection data (CPST, SJT and SC), as predictors of AKT and CSA examination scores.

iv. Incremental validity of IELTS assessments over all GP selection data (CPST, SJT and SC) and PLAB scores, as predictors of AKT and CSA examination scores.

3.4.13 **Tables 6.1 and 6.2** shows that the power of the CPST and SJT in predicting AKT and CSA performance in the IMG group is substantially reduced compared to the full sample. As such, the GP selection tests are still good predictors of performance, but are not as strong predictors as was found in the whole sample. The contribution of the SC in predicting AKT and CSA scores is also reduced, but is still statistically significant for all four analyses. Restriction of range in the sample can be one cause of a reduction in predictive power and comparing **Tables 2.1-2.3** shows that this is likely to be the case here. While **Table 3** shows that the IMG groups have at least as high variance for the different assessments as the other groups and often larger, when compared to the values for the total sample, the variances are generally somewhat reduced. For this group, the total variance in AKT scores explained by the selection assessments is 38% (or 40% taking into account age and gender data), compared with 56% (57%) for the full sample. For the CSA, the figures are 11% (23% taking into account age and gender) for this group, compared to 41% (48%) for all groups together. The impact of the PLAB and IELTS need to be compared to these figures.

3.4.14 **Table 6.1** reports regression analyses using only the scores from the test sitting where the candidate passed. **Table 6.2** shows the results adding in variables for the number of sittings the candidate required to reach a passing score on the two parts of the PLAB. This variable will provide some indication of how difficult the candidate found it to reach the required PLAB standard. For some candidates, multiple sittings may be required because of a deficit in the required knowledge or ability to apply it. For others, there may be an issue with English language skills which is affecting their ability to demonstrate their competence. Both analyses are shown because it may not be practical or appropriate to collect accurate information regarding the number of PLAB sittings within the selection context. In this case,

Table 6.1 might better reflect the practical possibility of predicting performance using the PLAB.

Table 6.1 Predicting MRCGP Examination Scores from Selection performance, PLAB and IELTS scores

Variables added to the equation at each level		AKT n=2,179		CSA n=1,623	
		No control for age and sex	Controlling for age and sex	No control for age and sex	Controlling for age and sex
0. Age and sex	R^2 change %		**5.5%*** **		**16.2%*** **
	F statistic	n/a	63	n/a	156
	df		*2, 2,176*		*2, 1,620*
1. CPST & SJT	R^2 change %	**37.3%*** **	**34.0%*** **	**9.4%*** **	**5.5%*** **
	F statistic	646	611	83	56
	df	*2, 2,176*	*2, 2,174*	*2, 1,620*	*2, 1,618*
CPST	sr^2	30.0%***	28.7%***	2.6%***	1.6%***
SJT	sr^2	1.1%***	0.7%***	4.2%***	2.5%***
2. Selection Centre (SC)	R^2 change %	**0.2%** **	**0.1%***	**1.9%*** **	**1.0%*** **
	F statistic	8	4	35	20
	df	*1, 2,175*	*1, 2,173*	*1, 1,619*	*1, 1,617*
CPST	sr^2	29.6%***	28.4%***	2.3%***	1.5%***
SJT	sr^2	1.0%***	0.6%***	3.8%***	2.3%***
SC	sr^2	0.2%**	0.1%*	1.9%***	1.0%***
3. PLAB 1 and PLAB 2	R^2 change %	**3.8 %*** **	**3.5%*** **	**3.4%*** **	**2.1%*** **
	F statistic	69	66	32	22
	df	*2, 2,173*	*2, 2,171*	*2, 1,617*	*2, 1,615*
CPST	sr^2	19.8%***	20.8%***	1.3%***	0.9%***
SJT	sr^2	0.4%***	0.5%***	2.7%***	1.6%***
SC	sr^2	0.1%**	0.1%*	1.3%***	0.7%***
PLAB 1	sr^2	3.2***	3.4%***	0.8%***	0.4%**
PLAB 2	sr^2	n.s.	n.s.	2.2%***	1.4%***
4. IELTS	R^2 change %	**2.3%*** **	**2.0%*** **	**3.0%*** **	**1.4%*** **
	F statistic	22	20	14	7
	df	*4, 2,169*	*4, 2,167*	*4, 1,613*	*4, 1,611*
CPST	sr^2	19.8%***	19.3%***	1.4%***	1.0%***
SJT	sr^2	0.4%***	0.3%**	1.4%***	0.9%***
SC	sr^2	0.1%**	n.s.	0.9%***	0.5%***
PLAB 1	sr^2	3.2%***	3.1%***	0.5%***	0.3%**
PLAB 2	sr^2	n.s.	n.s.	1.8%***	1.1%***
Reading	sr^2	1.8%***	1.6%***	0.4%**	0.3%*
Speaking	sr^2	0.7%***	0.6%***	n.s.	0.2%*

Understanding	sr^2	n.s.	n.s.	1.4%***	0.4%**
Writing	sr^2	n.s.	n.s.	n.s.	n.s.

Note: *df* = degrees of freedom; sr^2 = semi-partial correlation; *n.s.* = not significant; * Significant at the 0.05 level (2-tailed), ** Significant at the 0.01 level (2-tailed), *** Significant at the 0.001 level (2-tailed).

Table 6.2 Predicting MRCGP Examination Scores from Selection Performance, PLAB attempts and IELTS scores

Variables added to the equation at each level		AKT n=2,179		CSA n=1,623	
		No control for age and sex	Controlling for age and sex	No control for age and sex	Controlling for age and sex
2. Age and sex	R^2 change %		**5.5%*** **		**16.2%*** **
	F statistic	n/a	63	n/a	156
	df		*2, 2,176*		*2, 1,620*
3. CPST & SJT	R^2 change %	**37.3%*** **	**34.0%*** **	**9.4%*** **	**5.5%*** **
	F statistic	646	611	84	56
	df	*2, 2,176*	*2, 2,174*	*2, 1,620*	*2, 1,618*
CPST	sr^2	30.0%***	28.7%***	2.6%***	1.6%***
SJT	sr^2	1.1%***	0.7%***	4.2%***	2.5%***
2. Selection Centre (SC)	R^2 change %	**0.2%****	**0.1%***	**1.9%*** **	**1.0%*** **
	F statistic	8	4	35	21
	df	*1, 2,175*	*1, 2,173*	*1, 1,619*	*1, 1,617*
CPST	sr^2	29.6%***	28.4%***	2.3%***	1.5%***
SJT	sr^2	1.0%***	0.6%***	3.8%***	2.3%***
SC	sr^2	0.2%**	0.1%*	1.9%***	1.0%***
3. PLAB 1 and PLAB 2	R^2 change %	**8.8 %*** **	**7.8%*** **	**5.9%*** **	**3.5%*** **
	F statistic	89	81	29	19
	df	*4, 2,171*	*4, 2,169*	*2, 1,615*	*4, 1,613*
CPST	sr^2	15.8%***	15.7%***	0.7%***	0.6%***
SJT	sr^2	0.7%***	0.5%***	2.4%***	1.5%***
SC	sr^2	n.s.	n.s.	1.0%***	0.6%***
PLAB 1	sr^2	2.6***	2.5%***	0.4%***	0.2%*
No of attempts to pass PLAB 1	sr^2	4.9***	4.3%***	1.1%***	0.5%**
PLAB 2	sr^2	n.s.	n.s.	2.3%***	1.5%***
No of attempts to pass PLAB 2	sr^2	n.s.	n.s.	1.3%***	0.9%***
5. IELTS	R^2 change %	**1.6%*** **	**1.5%*** **	**2.6%*** **	**1.3%*** **
	F statistic	17	15	13	7
	df	*4, 2,167*	*4, 2,165*	*4, 1,611*	*4, 1,609*
CPST	sr^2	15.0%***	14.8%***	0.9%***	0.7%***
SJT	sr^2	0.4%***	0.3%**	1.4%***	0.9%***
SC	sr^2	0.1%*	n.s.	0.7%***	0.4%**
PLAB 1	sr^2	2.4%***	2.3%***	0.2%*	n.s.
No of attempts to pass PLAB 1	sr^2	4.2%***	3.8%***	1.1%***	0.5%**
PLAB 2	sr^2	n.s.	n.s.	1.8%***	1.2%***
No of attempts to pass PLAB 2	sr^2	n.s.	n.s.	1.0%***	0.7%**
Reading	sr^2	1.3%***	1.2%***	0.2%*	n.s.

Speaking	sr^2	0.5%***	0.4%***	0.2%*	0.2%*
Understanding	sr^2	n.s.	n.s.	1.3%***	0.4%**
Writing	sr^2	n.s.	n.s.	0.2%*	n.s.

Note: *df* = degrees of freedom; sr^2 = semi-partial correlation; *n.s.* = not significant; * Significant at the 0.05 level (2-tailed), ** Significant at the 0.01 level (2-tailed), *** Significant at the 0.001 level (2-tailed).

3.4.15 Both the PLAB and the IELTS scores are significant predictors of performance. PLAB 1 has a greater contribution to predicting the AKT (sr^2=3.2%; sr^2=3.4% when age and gender are controlled) and PLAB 2 for predicting the CSA (sr^2=2.2%; sr^2=1.4% when age and gender are controlled). **Table 6.2** shows that when the number of attempts is added, this variable substantially increases the proportion of variance explained by the PLAB as a whole. The number of attempts at PLAB 1 is particularly predictive of AKT scores (sr^2=4.9%; sr^2=4.3% when age and gender are controlled). Both the number of attempts of PLAB1 and PLAB2 help predict CSA performance (sr^2=1.1% and 1.3% respectively; sr^2=0.5% and 0.9% respectively when age and gender are controlled). Those who struggle to pass the PLAB assessments may be weaker candidates. Indeed some of those who pass after multiple attempts may barely meet the standard.

3.4.16 Of the IELTS component scores, the 'Understanding' component was the best predictor for the CSA (sr^2=1.4%; sr^2=0.4% when age and gender are controlled), whereas 'Reading' was the best predictor for the AKT (sr^2=1.8%; sr^2=1.6% when age and gender are controlled). The pattern is similar when the number of attempts to pass PLAB is included in the equation but the overall contribution of the IELTS reduces slightly from between 1.4%-3.0% of additional variance to between 1.3% to 2.6% of additional variance explained. This supports the hypothesis that poor language skills may cause some candidates to require several attempts to pass the PLAB.

3.4.17 Review of the pass rate by IELTS scores showed that trainees with an IELTS 'Understanding' level of 7.5 or below had a 32% pass rate on first attempt at the CSA, whereas trainees with a level of 8.0 or above had a 45% pass rate. The pass rate goes above 50% if only those with a level 8.5 or above on 'Understanding' are considered. A similar comparison was made for the AKT first time pass rate and the IELTS 'Reading' score. Those scoring 7.5 or below had a 49% pass rate, whereas those scoring 7.5 or above on the IELTS 'Reading' component had a 68% pass rate. These results suggest that English language skills have a substantial impact on success in training.

3.4.18 These results suggest that by adding the PLAB and IELTs scores to the GP selection test scores, it may be possible to refine the identification of borderline trainees for the IMG group that have sat these assessments.

3.5 Improving the prediction of MRCGP pass rates

3.5.1 Regression analysis can also be used to calculate predicted scores on the outcome variable, and thus we are able to predict outcome variables with the inclusion of PLAB and IELTS

scores as predictor variables, in addition to the original selection test scores. In this case, the outcome variables are the AKT and CSA, expressed as the number of points deviated from the pass score for the exam version trainees sat. Trainees predicted a negative score would be expected to fail at the first sitting and those predicted a positive score would be more likely to pass. Analysis was therefore undertaken to identify the pass rates using this approach for the group as a whole (i.e. those who have PLAB and IELTS scores).

3.5.2 **Table 7** shows the relative identification of passing rates for the group as a whole. The prediction of outcomes is clearly improved for this predominantly IMG group by using PLAB and IELTS scores. The improvement is most marked for the AKT. The proportion passing in the predicted fail group for the AKT reduces from 42% to 31%, whereas the pass rate for those predicted to pass increases by over 3%. For the CSA, using the PLAB and IELTS, results in more trainees being predicted to fail than the borderline method currently used. The overall pass rate for this group is 36%, so more than half fail first time. Just under a third (32%) of those predicted to fail actually passed the CSA first time. The biggest improvement is in those predicted to pass; 61% actually did (compared to the overall pass rate for the sample of 36%).

3.5.3 These exact results should be treated with caution due to the limited sample sizes of the IMG groups. Generally, it would be desirable to cross-validate with a different sample, to ensure that the prediction equation generated by using this particular sample also applies to a more general population.

Table 7. MRCGP Pass Rates by Predicted Score by level of Selection Test Scores

	Predicted Fail		Predicted Pass		All trainees
	Original Borderline	Using PLAB and IELTS to predict	Original Borderline	Using PLAB and IELTs to predict	
Proportion Passing AKT first time					
All trainees					
Percentage passing	**46.2%**	n/a	**90.2%**	n/a	**78.9%**
N Passing/total group	1,189 / 2,573		6,726 / 7,455		7,915 / 10,028
Trainees with PLAB and IELTS					
Percentage passing	**41.8%**	**31.2%**	**74.3%**	**77.8%**	**53.9%**
N Passing/total group	571 / 1,367	349 / 1,118	603 / 812	825 / 1,061	1,174 / 2,179
Proportion Passing CSA first time					
All trainees					
Percentage passing	**46.4%**	n/a	**86.7%**	n/a	**76.7%**
N Passing/total group	844 / 1,819		4,783 / 5,514		5,627 / 7,333
Trainees with PLAB and IELTS					
Percentage passing	**28.2%**	**32.1%**	**47.1%**	**61.4%**	**36.0%**
N Passing/total group	270 / 957	452 / 1,408	314 / 666	132 / 215	584 / 1,623

3.5.4 **Tables 8 to 13** repeat the previous information in an alternate format showing the proportion of correct and incorrect classifications. A correct classification is where a trainee is predicted to pass first time and does so or when someone is predicted to fail first time and does so. For example, **Table 8** shows the predictions and outcomes for the full group for the AKT. The overall pass rate is 79% - so if everyone was predicted a pass, 79% of the predictions would be correct. Table 8 shows the results of using the previously defined borderline group on the NRO tests to predict failures with all others predicted to pass. 90% of those predicted to pass do so and 54% of those predicted to fail do so. Overall 81% of predictions are correct, so this is a small improvement in prediction, and in particularly it correctly identifies 65% of those who go on to fail.

Table 8. AKT predicted (NRO selection tests) versus actual Pass and Fails: *All trainees*

		Predicted Fail	Predicted Pass	Total
Actual Fail first time	**%**	**13.8%**	**7.3%**	**21.1%**
	n	1,384	729	2,113
Actual Pass first time	**%**	**11.9%**	**67.1%**	**78.9%**
	n	1,189	6,726	7,915
Total	**%**	**25.7%**	**74.3%**	**100.0%**
	n	2,573	7,455	10,028

Table 9. AKT predicted (NRO selection tests) versus actual Pass and Fails: *PLAB and IELTS trainees only*

		Predicted Fail	Predicted Pass	Total
Actual Fail first time	**%**	**36.5%**	**9.6%**	**46.1%**
	n	796	209	1,005
Actual Pass first time	**%**	**26.2%**	**27.7%**	**53.9%**
	n	571	603	1,174
Total	**%**	**62.7%**	**37.3%**	**100.0%**
	n	1,367	812	2,179

Table 10. AKT predicted (NRO selection tests, PLAB and IELTS) versus actual Pass and Fails: *PLAB and IELTS trainees only*

		Predicted Fail	Predicted Pass	Total
Actual Fail first time	**%**	**35.3%**	**10.8%**	**46.1%**
	n	769	236	1,005
Actual Pass first time	**%**	**16.0%**	**37.9%**	**53.9%**
	n	349	825	1,174
Total	**%**	**51.3%**	**48.9%**	**100.0%**
	n	1,118	1,061	2,179

3.5.5 For those with IELTS and PLAB scores only **Table 9** shows that borderline scores on the GP selection tests provide a correct classification rate of 64% compared to an overall pass rate of 54% for this group. Introducing the PLAB and IELTS data (**Table 10**) improves this correct classification rate to 73%.

Table 11. CSA predicted (NRO selection tests) versus actual Pass and Fails: *All trainees*

		Predicted Fail	Predicted Pass	Total
Actual Fail first time	**%**	**13.3%**	**10.0%**	**23.3%**
	n	975	731	1,706
Actual Pass first time	**%**	**11.5%**	**65.2%**	**76.7%**
	n	844	4,783	5,627
Total	**%**	**24.8%**	**75.2%**	**100.0%**
	n	1,819	5,514	7,333

Table 12. CSA predicted (NRO selection tests) versus actual Pass and Fails: *PLAB and IELTS trainees only*

		Predicted Fail	Predicted Pass	Total
Actual Fail first time	**%**	**42.3%**	**21.7%**	**64.0%**
	n	687	352	1,039
Actual Pass first time	**%**	**16.6%**	**19.3%**	**36.0%**
	n	270	314	584
Total	**%**	**59.0%**	**41.0%**	**100.0%**
	n	957	666	1,623

Table 13. CSA predicted (NRO selection tests, PLAB and IELTS) versus actual Pass and Fails: *PLAB and IELTS trainees only*

		Predicted Fail	Predicted Pass	Total
Actual Fail first time	**%**	**58.9%**	**5.1%**	**64.0%**
	n	956	83	1,039
Actual Pass first time	**%**	**27.8%**	**8.1%**	**36.0%**
	n	452	132	584
Total	**%**	**86.8%**	**13.2%**	**100.0%**
	n	1,408	215	1,623

3.5.6 For the CSA in the full sample, **Table 11** shows that using the borderline scores on the GP selection tests results in a correct classification rate of 79% compared to an overall pass rate of 77%. More than half the first time fails occur in the borderline group although this constitutes only 25% of the whole group. For those with IELTS and PLAB scores, **Table 12**

shows that borderline scores on the GP selection tests provide a correct classification rate of 62% compared to an overall fail rate of 64%. It also identifies 66% of those who go on to fail their first sitting of the CSA. Introducing the PLAB and IELTS data (**Table 13**) improves the correct classification rate to 67%.

3.5.7 In both cases, more candidates with PLAB and IELTS scores are predicted an initial fail result than to pass on their first attempt. Those identified as likely to fail at first sitting might benefit from a more intensive intervention earlier in their training. Currently, it may require a trainee to fail an MRCGP exam sitting in order to be identified as in need of further support. However, this is likely to be well into the training process when sections of the course may need to be repeated or extended.

3.6 Examining the impact of alternative PLAB and IELTS cut scores

3.6.1 While using the regression equation can provide an optimum classification, it does not take into account the desired numbers in different categories and it is a relatively complex algorithm to implement. An alternative approach would be to set cut scores on the different assessments.

3.6.2 For example, assigning those who score less than 10 points above the pass score on either PLAB 1 or PLAB 2 would create a PLAB borderline group. Other work has suggested that the PLAB passing score is set too low [15]. Since the PLAB passing score is set at entry level to the Foundation Year Two (FY2), the current standard is likely to underestimate requirements for specialist training entry after FY2.

3.6.3 Those with IELTS Understanding and Reading scores (the IELTS elements with the strongest relationship with examination performance) of less than 7.5 are assigned to an IELTS borderline group in a similar manner. A count of the number of borderline indicators (NRO, PLAB and IELTS) for each individual provides a combined indicator of potential difficulties. **Tables 14 to 19** show the impact of these cut scores on identifying candidates likely to fail. For the purpose of the model, those who had a score of two or three on the combined indicator are marked as likely to fail.

Table 14. AKT predicted (PLAB scores < 10) versus actual Pass and Fails

		Predicted Fail	Predicted Pass	Total
Actual Fail first time	%	**49.6%**	**27.5%**	**46.1%**
	n	910	95	1,005
Actual Pass first time	%	**50.4%**	**72.5%**	**53.9%**
	n	924	250	1,174
Total	%	**100.0%**	**100.0%**	**100.0%**
	n	1,834	345	2,179

Table 15. AKT predicted (IELTS < 7.5) versus actual Pass and Fails

		Predicted Fail	Predicted Pass	Total
Actual Fail first time	%	52.0%	33.6%	46.1%
	n	770	235	1,005
Actual Pass first time	%	48.0%	66.4%	53.9%
	n	710	464	1,174
Total	%	100.0%	100.0%	100.0%
	n	1,480	699	2,179

Table 16. AKT predicted (2 or 3 borderline scores) versus actual Pass and Fails

		Predicted Fail	Predicted Pass	Total
Actual Fail first time	%	52.9%	23.0%	46.1%
	n	892	113	1,005
Actual Pass first time	%	47.1%	77.0%	53.9%
	n	795	379	1,174
Total	%	100.0%	100.0%	100.0%
	n	1,687	492	2,179

Table 17. CSA predicted (PLAB scores < 10) versus actual Pass and Fails

		Predicted Fail	Predicted Pass	Total
Actual Fail first time	%	67.1%	48.9%	64.0%
	n	906	133	1,039
Actual Pass first time	%	32.9%	51.1%	36.0%
	n	445	139	584
Total	%	100.0%	100.0%	100.0%
	n	1,351	272	1,623

Table 18. CSA predicted (IELTS < 7.5) versus actual Pass and Fails

		Predicted Fail	Predicted Pass	Total
Actual Fail first time	%	**69.9%**	**51.4%**	**64.0%**
	n	775	264	1,039
Actual Pass first time	%	**30.1%**	**48.6%**	**36.0%**
	n	334	250	584
Total	%	**100.0%**	**100.0%**	**100.0%**
	n	1,109	514	1,623

Table 19. CSA predicted (2 or 3 borderline scores) versus actual Pass and Fails

		Predicted Fail	Predicted Pass	Total
Actual Fail first time	%	**70.5%**	**44.0%**	**64.0%**
	n	864	175	1,039
Actual Pass first time	%	**29.5%**	**56.0%**	**36.0%**
	n	361	223	584
Total	%	**100.0%**	**100.0%**	**100.0%**
	n	1,225	398	1,623

3.6.4 For the AKT, the PLAB cut score of 10 gives a correct classification rate of 53%, IELTS 64% and the combined cut score 58%. For the CSA, the classification success rates are 64%, 63% and 66%.

3.6.5 If those who are identified by the indicator, rather than being rejected are referred for additional support, it is likely that the pass rate will rise. For example, if 50% of those in the predicted fail group who actually fail could be supported to pass first time, first time pass rate on the AKT could rise from 54% for the group taking the PLAB to perhaps 74%. For the CSA, the pass rate for this group might rise from 36% to 62%. This would have a substantial impact on the pass rate for BME trainees since they are substantially overrepresented in the IMG group.

3.7. Indicators of likelihood to resit the MRCGP examinations

3.7.1. **Table 20** shows the number of exam attempts made in the time period studied broken down by the number of borderline flags (i.e. borderline performance in both the CPST and SJT, or PLAB and IELTS assessments). It should be noted that because some candidates with a single attempt in the data set might have sat the examination more once before the time period of the study that these results should be interpreted with caution. In addition, trainees with a single fail in the data set may go on to take the exam several times before passing.

Table 20. No of attempts at MRCGP exams broken down number of Borderline Flags

No of borderline Flags		All trainees		Trainees with PLAB and IELTS Scores	
		No of AKT attempts between 2007 and 2013	No of CSA attempts between 2007 and 2013	No of AKT attempts between 2007 and 2013	No of CSA attempts between 2007 and 2013
0	Mean	1.1	1.1	1.3	1.6
	n	6,721	4,907	105	81
	SD	0.37	0.39	0.77	0.87
1	Mean	1.6	1.5	1.34	1.6
	n	1,587	1,177	387	317
	SD	0.93	0.86	0.71	0.94
2	Mean	1.6	2.1	1.6	2.1
	n	792	595	767	575
	SD	0.92	1.20	0.90	1.19
3	Mean	2.0	2.2	2.0	2.2
	n	928	654	920	650
	SD	1.10	1.14	1.10	1.14
Total	Mean	1.7	1.8	1.7	2.0
	n	3,307	2,426	2,074	1,542
	SD	0.99	1.08	1.00	1.14

3.7.2 Candidates with no flags have just over one attempt on average – so nearly all pass first time but a few need to take one or both of the exams a second time. However, the average number of sittings for the group with three flags is around two – so these candidates predominantly need to resit and perhaps more than once. The effect is statistically significant for both analyses using a one-way ANOVA (F=710, df=3, 10,024, p<0.001, Eta Squared= 0.17 for the AKT and F=794, df=3, 7,329, p<0.001, Eta Squared= 0.25 for the CSA).

3.7.3 An alternative method to reduce the failure rate in the MRCGP exams would be to increase the cut score on the NRO selection tests. For example, if those with borderline scores on the NRO assessments were rejected at the selection stage, the pass rate on the AKT would rise to 90% instead of the current 79%. Similarly for the CPS, increasing the cut score to above the borderline band would increase the pass rate from 77% to 87%. However, this improvement would be at the expense of a loss of approximately 25% of candidates who would then not be invited to attend the selection centre. In particular, there would be a disproportionate reduction in the numbers of IMG and BME trainees, with around 40% of BME candidates in the borderline band and 60% of IMG trainees. Thus, raising the bar in

terms of performance required in the GP selection tests would substantially reduce the pool of trainees at a time when more, rather the fewer GPs are needed and in addition doing so would have a disproportionate impact on some ethnic groups.

3.8 Comparisons by Nationality

3.8.1 There are a number of potential reasons why IMG candidates are performing less well in both performance and outcome variables. Trainees may have received training which is lacking in areas which are covered by the assessments. The internationally trained applicants may not be a representative sample of local students. If predominantly those who cannot get work locally apply for placements in the UK, this group will tend to perform less well on average than students from their country that did find employment. As discussed earlier, applicants from some countries may be hampered in their performance by a lower fluency in English than the assessments were designed for. Another possible explanation is some kind of cultural bias within the tests themselves. This section compares the performance of candidates from different countries and profiles those from particular countries where numbers allow. If the differences are due predominantly to reasons such as differential self-selection to apply in different countries or differential English fluency then applicants from different regions might be expected to show differences in levels of performance, whereas if the assessments suffer from a UK bias in their content and delivery, candidates from different cultures might perform equally as badly.

3.8.2 **Table 21** provides background information for the sample broken down by continent of training for those areas with more than 100 trainees in the data set. It shows that British applicants are younger than those from other countries. This may be because the decision to move occurs later, perhaps after local alternatives have been ruled out. From all three continents, applicants are over 30 years old on average whereas British graduates are below 30. The higher age may be due to longer courses of study or may reflect time spent on an unsuccessful local search for a training place. African and Asian applicants are more likely to be male whereas those from Europe and the UK are more likely to be female.

3.8.3 Comparisons of performance on the tests for different groups shows UK trained individuals performing substantially better than those from elsewhere on all the indicators. Other European applicants perform better than the African and Asian groups but not as well as the UK group. Trainees from Asia perform a little better than those from Africa on the more knowledge based CPST ($t=3.05$, $df=562$ $p<0.01$), whereas those from Africa perform marginally better on the more practice based assessments, namely the SJT, SC and CSA ($t=2.6$, 2.4 and 2.1; $df=2,047$, 2,047 and 1,521; $p<0.01$, $p<0.05$ $p<0.05$ respectively).

3.8.4 While the performance of European and UK groups on the PLAB and IELTS is not substantially different from the African and Asian groups, it should be remembered that only a very small proportion of the former groups take these tests so that the results do not reflect the general level of performance of these groups.

Table 21. Descriptives for CSA, AKT, IELTS and PLAB Examinations broken down by region of training

Variables		Ethnicity and PMQ region			
		Africa	**Asia**	**Europe**	**United Kingdom**
Age (years) at application to training	Mean	34.7	33.7	31.3	28.4
	95% CI	34.1-35.2	33.5-33.9	30.8-31.8	28.3-28.5
	SD	5.6	4.7	5.0	4.0
	n	393	1656	447	7295
Male	N (%)	212 (53.9%)	903 (54.5%)	158 (35.3%)	2,678 (36.7%)
Female	N (%)	181 (46.1%)	753 (45.5%)	289 (64.7%)	4,617 (63.3%)
GP Selection Results					
CPST	Mean	234.1	240.3	251.7	269.2
	95% CI	230.6– 237.7	238.6-241.9	248.6 – 254.8	268.5 – 269.8
	SD	36.3	33.6	33.6	29.8
	n	393	1,656	447	7,295
SJT	Mean	239.9	235.9	258.7	272.8
	95% CI	237.2 – 242.7	234.5– 237.2	256.0 – 261.4	272.2 – 273.4
	SD	27.6	27.5	29.3	26.3
	n	393	1,656	447	7,295
SC	Mean	77.4	76.6	79.7	83.1
	95% CI	76.8-78.1	76.3– 76.9	79.0 – 80.3	82.9-83.2
	SD	6.6	6.1	7.0	7.0
	n	393	1,656	447	7,295
AKT First Attempt (relative to pass mark)					
AKT total score	Mean	1.3	1.8	9.7	17.0
	95% CI	-0.5-3.2	1.0– 2.7	8.0 – 11.4	16.6-17.4
	SD	18.7	17.5	18.3	16.7
	n	393	1,656	447	7,295
Pass	%	59%	58%	75%	85%
CSA First Attempt (relative to pass mark)					
CSA total score	Mean	-3.0	-4.5	5.6	12.6
	95% CI	-0.5-3.2	-5.1- -3.9	4.3 – 7.0	12.3-12.9
	SD	12.0	10.8	12.5	10.7
	n	284	1,239	331	5,272
Pass	%	40%	37%	71%	89%
PLAB Scores (Total score relative to pass mark)					
PLAB 1	Mean	16.6	13.6	14.7	13.5
	95% CI	15.3-18.0	13.1-14.1	11.2-18.2	12.5-14.6
	SD	11.6	9.8	15.3	10.4

Variables		Ethnicity and PMQ region			
		Africa	Asia	Europe	United Kingdom
	n	298	1,405	74	365
PLAB 2	Mean	7.3	7.0	8.5	8.0
	95% CI	6.9-7.7	6.9-7.2	7.7-9.4	7.6-8.4
	SD	3.5	3.3	3.6	3.9
	n	302	1,421	69	371
IELTS					
Overall	Mean	7.6	7.3	7.3	7.4
	95% CI	7.6-7.7	7.3-7.4	7.2 – 7.4	7.4-7.5
	SD	0.5	0.4	0.41	0.77
	n	296	1,406	71	362
Reading	Mean	7.5	7.2	7.1	7.2
	95% CI	7.4-7.6	7.1-7.2	7.0– 7.3	7.1-7.3
	SD	0.85	0.77	0.75	0.96
	n	296	1,406	71	362
Speaking	Mean	7.9	7.4	7.5	7.7
	95% CI	7.8-8.0	7.3-7.4	7.3 – 7.6	7.6-7.8
	SD	0.7	0.58	0.58	0.97
	n	296	1,406	71	362
Understanding	Mean	7.6	7.4	7.3	7.5
	95% CI	7.5-7.7	7.4-7.5	7.2 – 7.5	7.4-7.6
	SD	0.8	0.73	0.63	0.97
	n	296	1,406	71	362
Writing	Mean	7.4	7.0	6.9	7.1
	95% CI	7.3-7.5	7.0-7.1	6.8 – 7.0	7.0-7.1
	SD	0.8	0.67	0.60	0.94
	n	296	1,406	71	362

3.8.5 **Tables 22 and 23** show the regression results for the different groups. For all groups the CPST is the strongest predictor of performance on the AKT with only marginally reduced prediction despite the restriction of range. Although overall predictive power for the CSA is lower, the SJT predicts CSA scores on a par with the SJT. The SC also shows more prediction in these analyses than for the AKT and it accounts for between 1% and 2% of the variance in CSA scores after the effect of the SJT has been taken into account.

Table 22. Predicting MRCGP AKT Examination Scores by group

Variables added to the equation at each level		Africa n=393	Asia n=1,656	Europe n=447	UK n=7,295
0. Age and sex	R^2 change %	**12.2%*** **	**5.5%*** **	**11.4%*** **	**9%*** **
	F statistic	27	48	28	343
	df	*2, 390*	*2, 1,653*	*2, 444*	*2, 7,292*
1. CPST & SJT	R^2 change %	**33.5%*** **	**33.3%*** **	**42.9%*** **	**46.2%*** **
	F statistic	119	450	207	3726
	df	*2, 388*	*2, 1,651*	*2, 442*	*2, 7,290*
CPST	sr^2	30.1%***	26.9%***	27.0%***	35.9%***
SJT	sr^2	n.s.	0.7%***	1.3%***	0.9%***
2. Selection Centre (SC)	R^2 change %	**n.s.**	**0.3%** **	**n.s.**	**0.2%*** **
	F statistic		9		38
	df		*1, 387*		*1, 7,289*
CPST	sr^2	30.0%***	26.1%***	26.5%***	34.1%***
SJT	sr^2	n.s.	0.7%***	1.2%**	0.8%***
SC	sr^2	n.s.	0.3%**	n.s.	0.2%***

Note: *Significant at the 0.05 level (2-tailed); ** Significant at the 0.01 level (2-tailed); ***significant at the 0.001 level (2-tailed).

Table 23. Predicting MRCGP CSA Examination Scores by group

Variables added to the equation at each level		Africa n=284	Asia n=1,239	Europe n=331	UK n=5,279
0. Age and sex	R^2 change %	**29.3%*** **	**17.9%*** **	**27.9%*** **	**13.4%*** **
	F statistic	58	134	5	407
	df	*2, 284*	*2, 1,236*	*2, 328*	*2, 5,276*
1. CPST & SJT	R^2 change %	**10.9%*** **	**6.4%*** **	**13.6%*** **	**16.3%*** **
	F statistic	25	52	38	609
	df	*2, 279*	*2, 1,234*	*2, 326*	*2, 5,274*
CPST	sr^2	4.9%***	1.2%***	3.1%***	6.4%***
SJT	sr^2	3.1%***	3.3%***	4.8%***	4.7%***
2. Selection Centre (SC)	R^2 change %	**1.7%** **	**0.8%** **	**1.1%**	**2.5%*** **
	F statistic	8	12	6	198
	df	*1, 278*	*1, 1,233*	*1, 325*	*1, 5,273*
CPST	sr^2	4.2%***	1.0%***	2.7%***	5.0%***
SJT	sr^2	2.7%***	3.2%***	3.7%**	3.9%***
SC	sr^2	1.7%**	0.8%**	1.1%*	2.6%***

Note: *Significant at the 0.05 level (2-tailed); ** Significant at the 0.01 level (2-tailed); ***significant at the 0.001 level (2-tailed).

4 Discussion & Implications

4.1 Summary

4.1.1 The analysis here confirms the findings from previous studies. The strong correlations between selection test scores and the MRCGP exams (CPST and AKT, *r*=0.73; SJT and CSA, *r*=0.54) make the tests useful tools for predicting earlier in training trainees at risk of failing, so that supportive action can be taken if desired. The correlations are lower within more homogenous subgroups. We looked at UK graduates and those who qualified via the PLAB and IELTS. However, when the correlations for these two groups are corrected from restriction of range, they are similar to the full sample values; the correlation between CPST and AKT performance ranges from *r*=0.63 for PLAB candidates and *r*=.78 for UK graduates, while the correlation between the SJT and CSA ranges from *r*=0.46 for PLAB candidates and *r*=0.57 for UK graduates.

4.1.2 On its own, the Selection Centre shows moderate correlations with the AKT (*r*=0.31) and CSA (*r*=0.42) in the full sample. They again drop within the two subgroups but when restriction of range corrections are applied the correlations between the SC and the CSA remains relatively stable and range from *r*=0.43 for PLAB (predominately IMG) candidates and r=0.39 for UK graduates. In combination with the CPST and SJT through, it has some, but not large incremental validity. The effect size is smaller for the PLAB group - but as a proportional increase in variance explained it is higher. However, it is important to acknowledge that the SC is the only face-to-face encounter throughout the selection process, and the assessment of values and behaviour also requires direct observation of behaviours (which is what the SC does, and the CPST and SJT cannot). It should also be noted that the SC is not an OSCE examination and should be compared with other interview processes used for selection purposes. The face validity of any selection process should also not be undervalued – it is of significant importance, not only to Local Education and Training Boards (LETBs), but also for candidates to have the opportunity to show how they exemplify qualities necessary for successful performance in General Practice. The face validity of any selection process is also linked to candidate perceptions of fairness, which has important implications. In particular, the SC here focuses on attributes centred on empathy and communication, which from job analyses of general practice have shown to be of high priority for anyone entering a career in General Practice [9] [10], and these attributes are more appropriately assessed through a face-to-face process, where behaviour in a simulated consultation can be observed.

4.1.3 There are substantial differences in performance between UK trained and IMG groups both on the selection tests for GP training and on the MRCGP examinations with lower scorers on both coming predominantly from the IMG group. The mean difference is about one standard deviation for the CPST and AKT and somewhat less for the SJT and CSA. The results of the PLAB and IELTS tests that these candidates complete were examined to see if they were able to provide additional information regarding risk of failure. These analyses showed that while

the selection assessments remained by far the best predictor of later performance, both the PLAB and IELTS provided incremental validity over the selection assessments alone.

4.1.4 The further breakdown by continent of training for non UK graduates showed similar patterns in terms of the predictive power of the tests, however, there were some small differences, most likely reflecting differences in the self-selection of candidates from the different regions wanting to continue their training in the UK.

4.1.5 Defining a borderline group with reference to the selection tests only, enabled some improvement in prediction for the IMG group. The borderline group were identified using the current banding structure for selection test scores with the lowest passing band defined as borderline. The original pass rate was 54%, for the whole group, but rose to 74% in those not identified as borderline. Using the PLAB and IELTS scores with a regression approach resulted in a larger group identified as likely to pass with a higher pass rate of nearly 78%. As well as groups with a high pass rate, the remaining candidates are identified as at high risk of failing with less than a third passing either of the MRCGP assessments at first sitting.

4.1.6 If a borderline group were to be identified for additional support, consideration would need to be given to the resources available to provide such support. Intensive input from tutors would be difficult to provide for large numbers, but a mentoring approach with students who are more advanced in their training supporting those at the early stages might be rolled out to large number of students. Specific (rather than regression based) modelling for the PLAB and IELTS showed that a 10 point margin on the PLAB and a minimum of 7.5 on the two IELTS scores most related to MRCGP performance (Reading and Understanding) could identify potential success quite well, although not as well as the full regression analysis. Selecting those who had two or more of these flags identified about two thirds of the IMG group.

4.1.7 By using the GP selection data (CPST, SJT and SC) it has been possible to accurately identify trainees who are likely to struggle to meet the standards required by the MRCGP at the outset of their training. This would allow targeted interventions to be offered at the outset of training, where early intervention is likely to be preferable (and both more effective and more efficient). The group with the most difficulty are those trained outside of the UK (i.e. IMGs). This group are assessed with the PLAB and IELTS as well as the usual selection process. By using these scores, in addition to the selection tests (CPST and SJT), it would be possible to identify trainees who would potentially benefit from additional support early on in training. The assessments themselves could provide diagnostic information regarding where help is needed, to inform specific training interventions to accelerate trainees' time to competence.

4.1.8 It is acknowledged that if the GP NRO wishes to access PLAB data, the most practical method would be to collect the data from applicants directly via the ORIEL system when they are submitting their application of GP recruitment. At present, trainees are provided with their PLAB 1 scores in all instances, but only receive their PLAB 2 scores if they have failed to achieve the minimum standard. Hence, it is noted that the GMC does not currently provide

PLAB 2 scores to passing candidates, but this may need to be reviewed in light of the current recommendations. Additionally, while the number of attempts required to pass the PLAB 1 and PLAB 2 appears to be more predictive of performance in licensing exams, it may not be appropriate or practical to request this additional information for candidates at point of selection, given that unlimited attempts at PLAB are allowed.

4.2 Strengths and limitations

4.2.1 Our study goes beyond the Esmail & Roberts report [14] by providing a more detailed analysis which includes selection data. The outcomes of this study can contribute to our understanding, not only where there may be group differences, but where the key issues may lie with group differences (differential attainment). The use of continuous data (rather than dichotomous data) allows for more robust predictions and conclusions to be made about the value add of each of the measures in the process. Consequently, we believe these findings contribute to a better understanding on the theoretical underpinnings of differential attainment.

4.2.2 While it might not be surprising to find that the CPST and PLAB 1 are significant predictors of future performance in a knowledge test (i.e. AKT), our study also found that the SJT was a significant predictor of performance in the CSA. The theory underlying SJTs suggests they measure implicit trait policies (ITPs) and general experience (and, depending on job level, specific job knowledge) [17]. ITPs can be described as an individual's judgement about the relative cost/benefits of expressing certain traits/behaviours in certain situations. In this way, SJTs can be said to measure the procedural awareness about what is effective behaviour in a given situation. Given that the CSA is designed to measure domains such as data gathering, clinical management and interpersonal skills, it is reasonable to assume that there should be overlap in terms of what the SJT and CSA assess.

4.2.3 Finally, this study shows that concerns already identified during performance at selection, are likely to persist throughout training and emerge as poor performance in MRCGP examinations. Candidates that exhibit borderline performance at selection would therefore likely benefit from early intervention and support systems.

4.2.4 There exist a number of limitations within this study that need to be acknowledged in interpreting the results. As trainees are allowed up to a maximum of four sittings of either the CSA or AKT, it is important to note that the first attempt at either examination may not in each case be a true reflection of their typical performance. However, other approaches such as using the average of either examination grade are equally, if not more, problematic. Subsequent scores are usually higher, but only those who failed initially will have more than one score, therefore this approach could distort the sample and create additional restriction of range in the data, therefore it is more meaningful to use first attempt data, as other researchers have done previously [12]. Future research could explore other dependent variables, such as number of attempts required to pass.

4.2.5 Between 2008 and 2013 there have been changes to the GP selection process, as well as with the AKT and CSA assessments. For example, the selection centre was restructured in

2010 so that it no longer included a group exercise. The weighting of the CPST and SJT in the final selection score has increased since 2008. Moreover, IELTS requirements changed over the period from a requirement of a minimum score of 7.0 overall to a minimum score of 7.0 on all components. Lastly, in terms of the MRCGP assessments, a change in the standard setting method was introduced in 2010 that took account of the pass-fail borderline. These various changes may have had an impact on the results but it is important to note that in general the methods used at selection and in the MRCGP have remained broadly stable over this time.

4.2.6 Another limitation of the present study is that there is no information about the training interventions and programmes on which GP trainees were placed. For example, it was not possible to ascertain whether borderline candidates systematically receive poorer quality training than their counterparts, which theoretically could account for some of the effects reported here. This could be the case if weaker trainees are assigned to less desirable training posts where the learning potential is restricted, for example. The present dataset does not allow for researchers to analyse the impact for any training interventions or differences in training programmes, which are also likely to influence outcomes. Further profiling and analysis of actual training interventions in each programme would be beneficial to understand the 'value-added' by various education and training interventions.

4.2.7 Finally, it is important to note that due to the restriction in sample sizes for certain demographic groups (see **Table 3**), some of the results should be interpreted with caution. In particular, while White UK, BME UK and BME IMG are robust groups (n>1,000), White EEA, White IMG, and BME EEA have far more limited sample sizes (n<300).

4.3 Comparison with existing literature

4.3.1 This study expands on previous research carried out by Esmail & Roberts [14], Patterson et al. [12], and McManus & Wakeford [18].

4.3.2 Esmail & Roberts [14] found significant differences in MRCGP examination performance between white UK graduates and other candidate groups, even after controlling for age and sex. Their more striking finding was that BME graduates were more likely to fail the CSA at their first attempt than their white UK colleagues, even if they had been trained in the UK (failure rate 17% vs. 4.5%). Previous analyses were conducted using logistic regression analyses on dichotomous "Pass" or "Fail" outcome variables. The current study used linear regression analyses, as information on the number of points by which candidates had either passed or failed was available.

4.3.3 Patterson et al. [12] found the CPST and the SJT correlate with the AKT at *r*=0.73 and 0.43, respectively (uncorrected[5]), and the CSA at *r*=0.38 and 0.43, respectively (uncorrected), and the selection centre was found to correlate with the CSA at *r*=0.32. Our most recent findings,

[5] The range of scores is restricted in the sample since there is an absence of data for those who were either unsuccessful in short-listing, long-listing or rejected from the selection process.

which are based on a larger dataset, are comparable with and replicate previous validity study results. The CPST and SJT correlate 0.73 and 0.46 with the AKT, again uncorrected for restriction of range. The correlations for the CSA are 0.49 and 0.54, respectively, which is slightly stronger than previously found. Similarly, scores from the selection centre correlate at r=0.42 (uncorrected) with the CSA. All these results, including previously reported findings, are statistically significant at p<0.001.

4.3.4 McManus & Wakeford [18] found the PLAB 1 correlated with the AKT at r=0.49 (p<0.001) and PLAB 2 correlated with the CSA at r=0.32 (p<0.001). In addition, they found that, in terms of PLAB performance, those who had higher IELTS scores were also likely to perform better in the AKT and CSA. The correlations in the current data set are a little lower, but still showing a substantial relationship: PLAB 1 and AKT correlate r=0.38 (p<0.001) and PLAB 2 and the CSA correlate at r=0.24 (p<0.001). The total IELTS score had a similar relations with the CSA with a correlation of r=0.25 (p<0.001).

4.3.5 In addition to the study carried out by McManus & Wakeford [18], other studies have found similar trends in terms of language skills and their impact on assessment performance [19]. In particular, researchers found that candidates who perform poorly in the CSA examination are more likely to have difficulties explaining matters to patients. Such candidates are also more likely to encounter misunderstandings in the consultation and to have more difficulty repairing misunderstandings [19]. The interpersonal skills domain of the CSA was "particularly problematic". Patterson et al. (2013) note that deficiencies in language skills are likely in some way to affect a trainee's ability to perform to a satisfactory level within their training. In particular, a trainee might show "good comprehension and make accurate diagnoses, but the difficulty lies more with the nuances and phraseology that are specific to the UK context" (p. 336) [20].

4.4 Implications for practice

4.4.1 The results of this study demonstrate that the GP selection tests, as well as PLAB and IELTS performance, have potential to accurately identify trainees who are likely to struggle in the GP specialty training programme. Based on our findings, we propose the following recommendations:

a. Candidates who are deemed to be 'borderline' at selection are significantly more likely to fail the AKT and the CSA on the first attempt. One method of addressing this would be to increase the CPST and SJT cut scores, however, this would significantly impact on the number of GPs entering training which is clearly undesirable at this point in time, as there is currently a decline in trainees applying to GP specialty training. This option may be revisited in future, however, if GP recruitment uptake is seen to improve. Currently, a more suitable option would be to allow applicants with higher selection test scores to be 'fast tracked' into training, without having to attend the selection centre. The selection centre, in addition to being used as a selection tool, could be used to identify 'borderline' candidates and inform the level

of support that may be required upon entering training. Selection test performance may enhance the ability to identify broad areas such as knowledge, soft skills and language, which may require support based on CPST and SJT performance. This will significantly accelerate the time to competence for trainees that may be deemed borderline at point of selection. The selection assessments can thus be used as a mechanism for *early identification and pick up*, which will increase the potential for borderline trainees (identified at selection) to pass the MRCGP first time.

b. Since language skills are associated with training progression, it might be necessary to consider increasing IELTS entry requirements. Alternatively, candidates with scores that fall between 7.0 and 8.0 could be immediately identified as requiring additional language support upon entering training (again using the assessments not just as selection hurdles but also as diagnostic profiling tools at point of selection), however, as this would potentially incur significant financial costs per trainee, it may be more practical to reconsider increasing the entry requirements and using the IELTS as a selection tool.

c. The PLAB examinations and respective pass marks are designed to be set at an equivalent standard to UK graduates (at the end of Foundation Year 1), however, there is currently little evidence to justify its current level, with PLAB graduates performing substantially less well at MRCGP than UK graduates. Data from previous research and reviews of PLAB [7] [18], along with our findings, indicate that the standard for PLAB maybe set too low if equivalent progression by PLAB graduates to UK graduates is expected and required. The standard for PLAB may therefore require reconsideration. However, it is noted that the current purpose of the PLAB test is not to identify whether candidates have the potential to achieve equivalent outcomes as UK graduates in postgraduate medical education and training or through medical career pathways. It is designed to test candidates' ability to practise medicine at the level expected at the end of F1 training [21].

References

1. The King's Fund. General Practice in England: An overview. Briefing, September 2009. http://www.kingsfund.org.uk/sites/files/kf/general-practice-in-england-overview-sarah-gregory-kings-fund-september-2009.pdf (accessed 20/04/15).

2. BMA Unit Costs of Health and Social Care 2011, PSSRU, University of Kent. http://bma.org.uk/-media/Files/Word%20files/News%20views%20analysis/pressbriefing_cost%20of%20training%20doctors.doc (accessed 12/10/12).

3. Rimmer, A. One in eight GP training posts vacant, despite unprecedented third round of recruitment. BMJ Careers 24 Oct 2014. http://careers.bmj.com/careers/advice/view-article.html?id=20019782 (accessed 20/03/15).

4. Irish B, Carr A, Sowden D, Douglas N, Patterson F. Recruitment into specialty training in the UK. BMJ Careers. 12 Jan 2011. http://careers.bmj.com/careers/advice/view-article.html?id=20001789#ref26 (accessed 23/04/15).

5. Woolf K, Potts HWW, McManus IC. The relationship between ethnicity and academic performance in UK-trained doctors and medical students: a systematic review and meta-analysis. Brit Med J 2011, 342.

6. Menzies L, Minson S, Brightwell A, Davies-Muir A, Long A, Fertleman C. An evaluation of demographic factors affecting performance in a paediatric membership multiple-choice examination. Jan 2015. Postgrad Med J 2015;91:72-76.

7. Wakeford R, Denney ML, Ludka-Stempien K, Dacre J, McManus C. Cross-comparison of MRCGP & MRCP(UK) in a database linkage study of 2,284 candidates taking both examinations: assessment of validity and differential performance by ethnicity. BMC Medical Education 2015, 15:1.

8. Spike N, Hays RB: Analysis by training status of performance in the certification examination for Australian family doctors. Med Educ 1999, 33:612-5

9. Patterson F, Ferguson E, Lane P, et al. A competency model for general practice: implications for selection and development. Br J Gen Pract. 2000; 50:188–193.

10. Patterson F, Tavabie A, Denney M, et al. A new competency model for general practice: Implications for selection, training and careers. Br J Gen Pract. 2013; 63: 249–250.

11. Patterson, F., Carr, V., Zibarras, L., Burr, B., Berkin, L., Plint, S., Irish, B., Gregory, S., 2009. New machine-marked tests for selection into core medical training: evidence from two validation studies. Clin. Med. (Northfield. Il). 9, 417–420.

12. Patterson F, Lievens F, Kerrin M, et al. The predictive validity of selection for entry into postgraduate training in general practice: evidence from three longitudinal studies. Br J Gen Pract Nov 2013; 63 (616) e734-e741.

13. Lievens, F., Peeters, H., Schollaert, E., 2008. Situational judgment tests: a review of recent research. Pers. Rev. 37, 426–441.

14. Esmail A, Roberts C. Academic performance of ethnic minority candidates and discrimination in the MRCGP examinations between 2010 and 2012: analysis of data. BMJ. 2013; 347:f5662.

15. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd Edition. New York: Academic Press; 1988.

16. Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. Journal of Applied Psychology, 85, 112–118.

17. Motowidlo SJ, Beier ME. Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. Journal of Applied Psychology, 2010; 95(2):321-333.

18. McManus IC, Wakeford R. PLAB and UK graduates' performance on MRCP (UK) and MRCGP examinations: data linkage study. BMJ : British Medical Journal. 2014; 348: g2621.

19. Roberts C, Atkins S, Hawthorne K. Performance features in clinical skills assessment: linguistic and cultural factors in the membership of the Royal College of General Practitioners examination. www.kcl.ac.uk/sspp/departments/education/research/ldc/publications/MRCGPling/MRCGP (accessed online 17/09/15).

20. Patterson, F., Knight, A., Stewart, F., and MacLeod, S. (2013). How best to assist struggling trainees? Developing an evidence-based framework to guide support interventions. Education for Primary Care, 24, 330-339.

21. General Medical Council. Review of the GMC's PLAB test: final report. Published September 2014. http://www.gmc-uk.org/PLAB_review_final.pdf_57946943.pdf (accessed online 09/06/15).