Evaluation of GP Specialty Selection

Davison, Ian; McManus, Chris; Taylor, Celia

Evaluation of GP Specialty Selection

January 2016

Evaluation of GP Specialty Selection

Ian Davison

Chris McManus

Celia Taylor

- January 2016 -

UNIVERSITY OF BIRMINGHAM

UCL

THE UNIVERSITY OF WARWICK

Acknowledgements.

The authors would like to thank numerous stakeholders and others who have inputted into this evaluation:

- Specialty selection candidates who completed the survey
- Our Advisory Committee: Sheona MacLeod (Chair), Lisa Johnsen, Bob Kirk, Ged Byrne, Selena Gray, Sarah Parsons, Graham Rutt and Salman Waqar
- Members of HEE who kept us on track and helped us obtain the data required for this evaluation, particularly Wendy Reid, Jonathan Howes and Clare Kennedy
- Derek Gallen, Keith Gardiner, Claire Loughrey and Stuart Irvine who gave us important perspectives from Wales, Northern Ireland and Scotland
- Moya Kelly, Roger Price, Richard Jones, Kelly Chambers and Kalpesh Thankey for data and information about the GP national recruitment office processes
- Fiona Patterson and Máire Kerrin for providing evidence from the Work Psychology Group
- Representatives from LETBs and Scotland, Northern Ireland and Wales who responded to our requests for costs data or questions regarding resource use and extension costs.

Data have been provided by HEE, the General Medical Council (via Daniel Smith), the Royal College of General Practitioners and UK Foundation Programme Office, without which, this work would not have been possible.

This independent evaluation was commissioned and funded by Health Education England (HEE).

Finally, we greatly appreciate Ashley Cook's skills in editing and improving the presentation of this report.

The views expressed, here, are our own and are not necessarily shared by those who have participated in the study.

Contents.

Executive Summary

Executive summary.

INTRODUCTION

Chapter 1 outlines the evaluation brief and how we have addressed it. Important political and historical contexts are noted; it also provides an overview of the datasets, questions asked, limitations and it explains why the analysis chapters are so complex, detailed and long.

We believe that this is one of the first analyses of medical training which uses data collected across the whole of training, from data on undergraduate performance (FPAS)[1], foundation performance (ARCP), selection into GP Selection Training, as well as data on applications overall to other specialties (Oriel), a questionnaire study of applicant motivations, performance during training (ARCP), performance at MRCGP (AKT and CSA), entry onto the GP Register, and Fitness to Practice problems on the GMC LRMP. In addition we have reviewed historical datasets to gain a wider view of GP selection and training. The resulting 'mosaic' provides, we hope, a broad picture of the issues involved in GP choice, selection and training, and in conjunction with economic data, will allow an analysis of cost-effectiveness.

This 'Big Picture' approach does mean that some of the recommendations are long-term in nature and beyond the remit of those involved in GP selection. Therefore, this report is split into:

- Part 1: The GP Selection System
- Part 2: Broader Recruitment and Selection Perspectives

Our conclusions and recommendations, synthesising what we have found within each of these two parts, are provided in the final chapter.

The late arrival of datasets and the complexity of the analyses mean that time has been extremely tight; we apologise for errors and omissions that have resulted.

SYNTHESIS AND RECOMMENDATIONS TAKEN FROM CHAPTER 11

1. When reflecting on the results reported here, policy makers will need to weigh up competing concerns. We are aware that some suggested changes are beyond the remit of those involved in specialty training and that some of these recommendations have already been discussed and may have been implemented.

2. Although there is considerable uncertainty, the historical trends suggest the **current annual recruitment target of 3250 GP trainees for England** and the previous 50% target (whatever denominator is used) **will be very challenging to meet** unless the number of specialty training posts is substantially reduced. Given so many doctors take several years to enter GP, l**ong-term systemic changes may be more effective than quick fixes.** These are considered in Part 2, alongside comparison with other specialties and data issues.

Part 1: The GP selection system

3. This evaluation can be viewed in terms of Utility, consisting of: reliability, validity, educational impact, acceptability of the method to the stakeholders and cost. We have looked closely at **reliability, validity and cost-effectiveness** but not educational impact and acceptability.

[1] Given the large number of acronyms and abbreviations, a list of these is provided at the end of this report.

4. For the health service overall, selection costs are insignificant; even a trainee who requires a 12 month extension is worth training; unfilled posts and trainees who fail are huge costs to the system. **The number of unfilled posts is the main driver of cost-effectiveness when considering a ten-year time horizon.** A secondary driver is the predictive validity of the selection processes, so 'Stage 2 only' is better than 'combining CPST, SJT and Stage 3 scores' (we call this 'Equal weight to all'), which in turn is better than only using 'Stage 3', although the differences are marginal compared with that between each of these approaches and the current (2015), 'baseline' approach.

5. In Stage 1, around 7% of GP candidates were rejected for immigration issues or not providing evidence for Foundation competence. About half of those who re-applied for Round 2 were offered a place.

6. We estimated alternate forms reliability to be 0.73 for CPST, 0.57 for SJT, and 0.73 for the Stage 2 total. Currently, the Stage 2 assessments have a relative standard i.e. not equivalent between years. **Using Band scores is much cruder than using the continuous scores and is less reliable.**

7. **Stage 3 has low reliability**, as expected given the low number of stations (alternate forms reliability averaged about 0.50). Correcting for restriction in range and using multiple imputation, the Stage 2 selection assessments are better predictors of AKT and CSA than is the Stage 3 Selection Centre assessment. **Stage 3 scores account only for about 3 or 4% of additional variance in CSA scores after taking Stage 2 into account.**

8. Abolishing Stage 3 could save around £3 million (which equates to the service provision of seven trainees) and result in approximately 43 more trainees completing GP training within 3 years, 22 fewer needing extensions, and 21 fewer not entering the GP Register within 5 years. **However, the real debate concerns the acceptability and educational impact of losing Stage 3.**

Recommendations: Part 1 - The GP selection system

General:

R.1: If a '50%' target is to be pursued then it probably should reflect the percentage of UK graduates who do not go straight from Foundation into Specialty training (currently 48% and rising), and/or who never enter either the GP or Specialist Registers (historically 15% and possibly rising). Similarly, as the '3250' target is for England only, the targets for Scotland, Wales and Northern Ireland should also be published[2]. As GP selection is UK-wide, it would be helpful if such announcements are agreed and coordinated between the four countries.

R.2: Investment in improving the validity and reliability of GP and specialty selection would likely, in terms of the whole UK healthcare system, be highly cost-effective in the long-term.

Stage 1:

R.3: Seek to reduce Stage 1 rejections and encourage those who are rejected to re-apply. In 2016, candidates are being given longer to provide evidence of Foundation competency and for IMGs to complete their application for sponsorship so a move to addressing this recommendation is already underway.

Stage 2:

R.4: Use statistical equating of the CPST and SJT scores across years e.g. by Rasch modelling; this makes Stage 2 an absolute form of assessment.

R.5: Use continuous Stage 2 scores rather than Bands.

[2] Historical 'places' or 'vacancies' are available at https://gprecruitment.hee.nhs.uk/Resource-Bank (accessed 07/01/2016), but we do not know if future targets are agreed or published.

R.6: Select to Stage 3 using the total CPST + SJT score rather than a cut-off mark for each; however, we have not investigated this in detail e.g. the optimal way to weight SJT compared with CPST.

R.7: Increase the length of the Stage 2 assessments, particularly SJT, to increase reliability.

R.8: In terms of transparency, it would make sense to separate development and administration of the selection system from its evaluation (this also applies to Stage 3).

Stage 3 and offers of training places:

R.9: Use a combination of CPST, SJT and the Stage 3 total scores rather than the algorithm to decide which candidates should be offered places. If this is done, weight the Stage 2 CPST and SJT continuous scores more heavily than the Stage 3 score.

R.10: Withdraw the moderation procedure. If uncomfortable with this without further evidence, undertake sociolinguistic and additional statistical analyses of moderation to enable a more comprehensive assessment to be made of what it adds, and whether it does so consistently.

R.11: Investigate the possibility of increasing the number of stations to increase reliability and predictive validity. Removing the moderation session would mean about a third more time is available. There may also be potential to reduce the length of the scenarios. Any such changes would need to be designed carefully.

R.12: The generalizability analysis suggests that simulators and cases contribute roughly similar amounts of variance (error) as assessors; so:

a) Consider ways to reduce differences between simulators and cases, as much as between assessors e.g. more simulator training and feedback on their 'hawkishness' or 'doveishness'; ideally cases would be piloted, but perhaps more realistic is consideration of the difficulty of previous cases when developing new ones (for which Rasch analysis provides a useful methodology).

b) Design the Selection Centres so that variances due to assessors, simulators and cases can be partitioned (distinguished) i.e. keep careful records of which assessor, simulator and case are seen by each candidate, and ensure each assessor and simulator experiences at least two simulators/ assessors and cases e.g. rather than confounding assessors with simulators by always pairing them in the same way.

R 13: Avoid using Cronbach's alpha as a measure of reliability. Generalisability analyses are well developed and should be used routinely within and between selection centres. Techniques such as the EM algorithm and alternative forms reliability can also be used .

R.14: Investigate why candidates (particularly those who are graduates of UK medical schools) are failing Stage 3 (and Stage 2) and consider enhancing medical school or Foundation training to help applicants address the areas of weakness identified. Publish the results to help prospective candidates understand what is required of them and consider using videos of poor, acceptable and excellent candidates in example scenarios to help applicants prepare for the Selection Centre.

R.15: Consider whether the greater reliability, validity and cost-effectiveness of abolishing Stage 3 are beneficial, given potential losses in terms of educational impact and acceptability.

Part 2: Broader recruitment and selection perspectives

9. We are delighted by the breadth and depth of data made available for this study. However, on many occasions, the data were unclear or inconsistent. To really understand the processes at work, routine collection, **checking/ cleaning, linkage, analysis and archiving** of such data are of utmost importance.

10. Our work in Part 2 examined how the GP Selection process is part of a much wider system and therefore our results are contextualised by considering these wider issues relating to junior doctors' career choices and the interplay of GP with concurrent recruitment to other specialties.

11. Although the remit of this report is GP selection, we have argued strongly that the ideal perspective is from **'cradle to grave'.** Attitudes favouring hospital specialties over general practice are evident before medical school, but become more prevalent during medical school. Favourable GP experiences can encourage positive attitudes towards general practice. Substantial numbers of doctors only enter GP training after time abroad or in another specialty.

12. When candidates applied to GP and another specialty, a higher proportion were considered appointable by GP than the other specialty, suggesting **'the bar' may be a little lower in GP**.

13. Tackling the fall in numbers of non-UK candidates and their low success rate are areas where relatively 'quick wins' may be possible, perhaps by providing more opportunities for these candidates to become acquainted with the NHS and the expected consultation styles.

14. We advocate much stronger data management and linking, coupled with better use of statistical techniques, particularly as doctors progress through selection and training into practice. Predictive validity for patient care outcome measures is the ultimate aim for the assessment of any selection system.

Recommendations: Part 2 - Broader recruitment and selection perspectives

R.16: Take a long-term strategic approach to obtaining, checking, using and storing data related to recruitment and training. Ideally this is in all specialties from the beginning of medical school to retirement. This is now underway with UKMED, but it will take many years before it will have sufficient longitudinal data to address many of the issues in this report; consequently, 'bottom-up' improvements in data retention are also desirable.

R.17: UKMED will also inevitably be limited in the measures it obtains of medical students and their interests and intentions, and separate methods of collecting such data routinely need to be developed.

R.18: Positive GP experiences at medical school are important influences on applying for GP, particularly as they help students become aware that GP suits their personality. It is possible that improving these experiences and providing positive careers advice towards general practice may increase long-term uptake of GP.

R.19: Address the problems that hinder Year 1 Foundation doctors from gaining GP experience.

R.20: Consider ways to encourage and facilitate applications from those who have been working abroad or in another specialty.

CHAPTER EXECUTIVE SUMMARIES

PART 1: THE GP SELECTION SYSTEM S

Chapter 2: Progression through GP selection

This chapter outlines some of the key processes involved in GP selection.

1. Candidates pass through three Selection Stages, Stage 1 (Administrative Checks), Stage 2 (CPST and SJT), and Stage 3 (Selection Centre), and may fail (be rejected) at each stage.

2. **Between 2009 and 2015, round 1 applications for UK graduates rose from 3503 in 2009 to 4318 in 2013, and then down to 3696 in 2015; non-UK graduate application have halved in that time.**

3. Throughout that time, **around 61% of UK graduate applications have led to an offer being accepted,** with about 22% failing at Stages 1, 2 and 3, and 18% withdrawing. **For non-UK graduates, the acceptance rate is much lower at about 24%**, and perhaps falling during this period; about 71% fail at Stages 1, 2 and 3 and just 6% withdraw.

4. The process of developing and testing Stage 2 questions is outlined.

5. CPST and SJT scores are put into 4 Bands; a candidate in either Band 1 is not invited to Stage 3.

6. The Band 2 threshold was raised in 2011 meaning about 10% of candidates were excluded from Stage 3, instead of the previous 5%.

7. However, in 2014 round 1, **5% of candidates (334/6688) met the criterion of both CPST and SJT Band 2 or higher but were not invited to Stage 3.**

8. **Stage 3 face validity was thought to be very high.**

9. There are four stations in Stage 3: three OSCE-style role plays and a written exercise. One assessor at each station awards marks for 3 or 4 competencies on a 1 to 4 scale (equating to something like: little, limited, satisfactory and strong evidence).

10. **These Stage 3 scores and the SJT Band score are converted to outcomes in a complex way: we have traced 29 branches of this algorithm.** For each of the four competencies, the mean score is calculated. The number of competencies with a mean score of 3 or greater is the major factor in determining the initial outcome. However, the number of scores of 1 and any concerns raised by the assessors are also involved in the decision making process. This initial outcome usually moves up with an SJT Band 4, stays the same with Band 3, and moves down with SJT Band 2. Sometimes whether there are 3 stations with 3 or 4 scores of 3 or 4 also affects the outcome.

11. This algorithm leads to one of the following outcomes: 1= demonstrated i.e. offered a post; 2= Review likely; 3= Review unclear; and 4=Not demonstrated i.e. rejected. Those with 'Review likely' or 'Review unclear' are discussed at a moderation session, with 'Review likely' more likely to be offered a post; so the final outcome is either 'Demonstrated' or 'Not demonstrated'.

12. **In 2014 round 1, 874/1173 (75%) candidates reviewed at moderation were offered places.**

13. **It may be better to use a total score and no moderation instead of the current complex decision tree followed by moderation**. At the end of Chapter 3, we return to this issue when we have considered the distinctiveness of the four competencies.

Chapter 3: Reliability and Generalizability

Reliability, generalizability and precision of measurement are fundamental to any assessment measure.

1. Reliability is not only of **educational importance**, low reliability meaning that the wrong candidates pass or fail, but also has **economic consequences**.

2. **Costs** arise when weak candidates falsely pass, incurring expense for additional training, lost investment due to drop-out, or indirect costs for inappropriate patient care, both financial and in terms of mortality and morbidity. When good candidates fail then overall standards in the specialty decline, and there is a reputational cost to the specialty as high quality candidates are lost[3].

3. Reliability is often assessed with Cronbach's alpha and generalizability but, as Brennan emphasizes, they give **misleadingly high coefficients** when based on assessments only at a single occasion with a single set of items.

4. The AERA/APA/NCME 2014 Standards recommend the use of alternate forms reliability, which Brennan calls the coefficient of stability and equivalence (r_{se}). Brennan emphasizes that calculation of r_{se} is in principle simple and robust, being the correlation of marks across the same candidates **in two separate full-length assessments with different questions or different cases/ assessors/ simulators.**

5. The data from GP selection over the seven years (2009-2015) **allows the calculation of r_{se}**, because many candidates sit the assessments in different years. The **EM algorithm** accounts for the inevitable restriction of range.

6. Generalizability analysis can also calculate an alternate forms reliability, which again can be corrected for range restriction, and gives results which are very similar to the EM algorithm.

7. Although previous analyses have suggested that stage 2 has an alpha of about 0.88 for CPS and 0.81 for SJT, our analyses find that **r_{se} is 0.73 for CPS, 0.57 for SJT, and 0.73 for Stage 2 overall.** Reliabilities of band scores are inevitably lower because of being quantised.

8. The reliability of Stage 3 has been quoted as being as high as 0.87 and 0.89, but probably reflects an inappropriate use of Cronbach's alpha. **Between 2011 and 2015, Cronbach's alpha for the 4 independent stations in Stage 3 has averaged 0.62, but fell to 0.55 in 2015.**

9. Generalizability has previously been estimated by Wakeford and Jolly at 0.64, based on a single test occasion; our similar analysis produced a mean of 0.59.

10. Our alternate forms reliability/generalizability estimates finds a **reliability coefficient for Stage 3, estimated from several sources, of the order of 0.5 for Stage 3.** These estimates are lower in 2014 and 2015, but we are unsure of the reasons for this decline in reliability.

11. Delays in retaking between one and six years give remarkably similar reliability coefficients, indicating that candidates' performance over time is remarkably consistent; it is primarily the unreliability of Stage 3 that leads to candidates obtaining different retake scores relative to their peers.

12. Stage 3 has four separate **sub-scales** (ES, CS, CT&PS and PI) assessed on four different **stations** (A: Patient, B: Relative/Carer, C: Co-professional and W: Written). Over and above the total score, there is only a small amount of specific variance attributable to ES, and its reliability is very low (r_{se}= 0.15). There is no detectable specific variance due to CS, CT&PS or PI, or the individual stations. **The different sub-scales and different types of station probably contribute little that is unique to the overall Stage 3 score.**

[3] Costs related to under-recruitment are considered in Chapters 6 and 7.

13. Item-response theory analyses of the 2015 data using FACETS suggest that **assessors, simulators and cases all differ in difficulty (stringency, hawkishness).** These effects inevitably decrease the reliability of Stage 3, so all three of these sources of error need careful consideration.

Chapter 4: Outcome measures and their relationship to selection measures

1. **Outcome measures** for the study consisted of **ARCP performance**, entry to the **GP Register** of the LRMP, **performance at MRCGP**, and indicators of **Fitness to Practice Issues** with the GMC.

2. **Selection measures** consisted of the **CPST and SJT scores for Stage 2, the total score for Stage 3**, and its four component scores of ES, CS, CT&PS and PI. In addition, the **FPAS EPM and FPAS SJT** scores, taken two to three years before Stage 2, were also included as selection measures since they were prior to GP training.

3. The datasets are complex, with multiple selection measures for many applicants. Analyses for the present chapter are therefore at the level of **applicants** rather than **applications**. The next chapter will primarily look at applications.

4. The main sampling frame consists of **33,520 uniquely identifiable applicants** from **48,200 applications** for seven applicant cohorts from 2009 to 2015. Applicants on two or more occasions were credited with the highest marks they attained at Stage 2 or Stage 3.

5. Seven **selection endpoints** were described, from Rejected at Stage 1 through to Offer made and Accepted. Failing candidates had lower Stage 2 and Stage 3 scores (unsurprisingly). Candidates rejecting offers after Stage 3 had equivalent Stage 2 scores to those accepting them.

6. FPAS scores were available for some applicants in the later cohorts. There was a small tendency for selected applicants to have higher EPM and SJT scores, but Ns were low and the effect was small. There was a tendency for candidates applying for GP selection to have somewhat lower FPAS scores than for UK medical graduates in general.

7. The FPAS EPM decile score showed a good correlation with Stage 2 CPST (0.63 corrected for missingness) and a somewhat smaller correlation with Stage 2 SJT (0.47) and a lower correlation still with Stage 3 total (0.33). FPAS SJT correlated equally with Stage 2 CPST (0.46) and SJT (0.48), and had a somewhat lower correlation with Stage 3 total (0.34). Given the problems with FPAS EPM in particular, which is within-school, these results suggest that **for UK graduates many features of the scores in Stage 2 are replicating those already obtained by FPAS at graduation.**

8. The remainder of the chapter considers the various outcome measures, as well as measures of selection into training, and their association with each other and in particular with performance on the Stage 2 and Stage 3 selection measures.

9. The **primary outcome measure was entry to the GP Register**, up until the end of August 2015. **ARCP data** are available from 2010 to 2014, and therefore only the 2009 and 2010 selection cohorts are likely to have complete data (i.e. four years of ARCP), although later cohorts can contribute data, particularly for ST1 and ST2. ARCP records are sometimes incomplete or inconsistent, but nevertheless provide a reasonable picture of the rate of progression and of problems. **MRCGP results** were available for all trainees from 2009 to August 2015. Linkage in all cases was via the GMC number. FtP issues were analysed from a cumulative LRMP file until the end of August 2015.

10. Of 48,200 applications from 2009 to 2015, the selection files suggested that 21,979 applications were accepted. Some applications were from the same doctors in different rounds or years, and some applicants were not identifiable, not having a GMC number. For the present chapter, **the main sample consists of 21842 doctors who had accepted offers for GP training.**

11. Of **6598 applicants who were not offered training places** between 2009 and 2012, 6359 had no record of training but **25 (0.4%) were on the GP Register.** An additional 2**39 doctors had ARCP GP records**, despite being recorded as not accepting offers, and **219 (9.6%) had entered the GP Register**. A small group of doctors, about 2% of all trainees, therefore appear to enter training despite formally not having accepted offers.

12. By linking across the files the rates of various outcomes could be assessed. Overall, **about 2.5% of doctors who have accepted training places do not continue into training.** Dates of ARCP GP records suggest that about **5% of trainees accepting places begin their training in the training year after they have been accepted**.

13. About **75% of applicants are accepted into training in the first year of application to GP training, 15% the year after, about 5% two years later, and 5% or so from three to six years later.** Candidates not accepted in their first year of application had lower Stage 2, Stage 3 and FPAS EPM scores.

14. For the major outcomes of ARCP, MRCGP, entering the GP Register, and having Fitness to Practice issues, we provide descriptive statistics, and also look at the relationship to GP selection measures and other outcome measures.

15. ARCP records are quite messy, with various inconsistencies between years in particular.

16. The **major outcome for ARCP** was a **Modified Tiffin Scale (mTiffin)**, which has five levels. For trainees in ST3, the proportions in the groups were **1: Satisfactory Progression (69%), 2: Insufficient Evidence (12%), 3: Target training (3%), 4: Extended training time (12%) and 5: Left programme [i.e. ARCP Outcome 4 or resigned] (5%).** In ST1 and ST2, about 0.7% and 1.2% of trainees required Extended Training time, and an additional 0.5% and 0.3% left the programme.

17. **Higher mTiffin Scores were associated with progressively lower Stage 2 and Stage 3 scores.** However the selection measures showed **no differences between those achieving 1: Satisfactory Progression and II: Insufficient Evidence**.

18. Not all GP trainees are in full-time training, and ARCP provides some information on those who are Less than Full-time (LTFT) or who are Out of Programme (OOP). However it does not provide the percentage of time less than full-time or the destination of trainees who are OOP.

19. Rates of LTFT and OOP can only be estimated accurately in the 2009 and 2010 cohorts. About 15% of trainees are LTFT at some point in their training, about 14% are OOP at some point, and **about 23% of trainees are either LTFT or OOP at some point in their training**.

20. **Trainees who are LTFT have lower Stage 2 and Stage 3 selection scores**, although OOP trainees show no differences at Stages 2 and 3.

21. Entering the GP Register was assessed at 27th August 2015, and therefore includes a majority of those who entered the Register at the end of the 2014-15 training year. Trainees in the 2009, 2010, 2011 and 2012 cohorts were **followed up for 6, 5, 4 and 3 years after the beginning of training**, and data are available for those trainees who entered the GP Register within that time period.

22. Overall about **50% of trainees enter the GP register within 3 years, 71% within four years, 78% within 5 years and 81 percent within 6 years, so that about 19% of trainees are not on the Register within six years of starting training**.

23. **63% of trainees without LTFT or OOP are on the GP Register in three years, and 85% within six years**[4].

24. For those without LTFT or OOP, **trainees with mTiffin Scores of 3, 4 or 5 were later to enter the Register.** However an mTiffin Score of 2 in ST3 did not predict later time of Register entry compared with mTiffin Group 1. Despite

[4] Our calculations suggest only 68% of those who are either LTFT or OOP are on the GP register after 6 years fulltime equivalent training; however, we are not confident that the records are sufficiently accurate, particularly regarding maternity leave.

gaining an mTiffin Score 5 (ARCP outcome 4 or resigned), 20% of such trainees entered the GP Register: we do not know the specific reasons for this, but ARCP outcomes are far from complete and consistent.

25. **mTiffin scores of 1, 2 or 3 in ST1 or ST2 did not differ in their ability to predict entry to the GP Register,** although the very small group with an mTiffin Score 4 in ST1 or ST2 (and an mTiffin 1 or 2 in ST3 and not LTFT or OOP) were less likely to get onto the Register. There were too few trainees to analyse for mTiffin Score 5 in ST1 or ST2.

26. Overall, **mTiffin Score 2 (insufficient evidence) shows little evidence of providing differential prediction of outcome compared with mTiffin Score 1 (Satisfactory progression)**.

27. Performance at MRCGP AKT and CSA were assessed at the first attempt at each examination, as the literature suggests that first attempts have the best statistical and predictive properties.

28. Trainees can **first take AKT in ST2, with a higher proportion in recent years taking it at the first opportunity. CSA is first taken in ST3, with some trainees apparently leaving it late in ST3**, leaving them fewer opportunities to be on the Register within three years should they fail.

29. MRCGP marks were expressed in relation to the pass mark at each assessment day. AKT and CSA marks are both predicted by Stage 2 and Stage 3 marks. However it seems clear that **the Stage 2 selection assessments are better predictors of AKT and CSA than is the Stage 3 selection centre assessment. CPST is a much better predictor of AKT than is the SJT, whereas the SJT and the CPST are equally good predictors of CSA**.

30. Fitness to Practice measures were assessed in terms of Erasure, Suspension, Conditions, Undertakings or Warnings (ESCUW) at any time during a doctor's career (and hence some may have been before GP selection or training). **Trainees with FtP issues had lower scores on the Stage 2 and Stage 3 selection measures**.

31. Taken overall there is little doubt that the **Stage 2 and Stage 3 selection measures provide predictive validity of poorer performance in ARCP, at MRCGP, and in terms of FtP**. The Stage 2 scores generally have higher correlations with outcomes than do the Stage 3 scores. At Stage 2 the CPST and SJT scores have differential prediction, CPST, the knowledge test, being better at predicting AKT, whereas the SJT contributes equally with CPST to predicting CSA performance.

Chapter 5: Assessing the predictive validity of selection and the consequences of various selection methods using multiple imputation

1. **Selection data are always limited** because candidates with lower selection scores are not accepted. Accepted candidates who go into training, and eventually have outcome measures, therefore have a higher mean and a lower range of scores than candidates in general. This problem is known as **range restriction**.

2. **Validity of selection**, the correlation between selection measures and outcome measures, **needs to be estimated in candidates.** Estimation only in the selected group, the trainees, under-estimates the true effectiveness of selection. However of necessity the performance of non-selected candidates is not known and therefore has to be estimated.

3. **Range restriction can be seen as a problem of missing data**, the outcomes for non-selected candidates not being known. The modern statistical technique of **multiple imputation** can be used to estimate or 'fill-in' the missing outcomes for candidates who are not selected.

4. The imputed datasets can be used for **estimating the consequences of different methods of selection**, such as random selection (which albeit unrealistic is a useful conceptual model for assessing the benefits of selection at all), the omission of Stage 2 or Stage 3, and the use of specific cutoffs defined in advance in comparison to models which select until all posts are filled.

5. Imputation allows the consequences of selection methods to be examined across a range of outcome variables, including Stage 3 scores, MRCGP examinations, date of entry to the GP Register, ARCP Outcome 4, or being LTFT or OOP.

6. Fully conditional multiple imputation with linear modelling was carried out on ten separate occasions, and estimates of parameters compared across the imputations to assess the robustness of the imputation process.

7. Imputed values were checked for plausibility by assessing graphically their relationship to Stage 2 and Stage 3 scores. In particular the assumption of linearity and extrapolation outside the data range seemed acceptable.

8. Although imputation was carried out for the selection years 2009 to 2015, modelling of different selection models outcomes was restricted to the years 2011 to 2014 when selection was the most stable.

9. The imputed datasets have 'filled-in' values for all GP candidates on all of the key measures, whatever the endpoint at which they left either GP selection or GP training. **The imputed datasets therefore allow virtual selection using different selection processes.** Overall eleven different selection models are compared.

10. The **Baseline Model consisted of the current selection process**, and all other selection models were compared with it.

11. Models differ both in the relative proportions of trainees entering the Register after four years of FTE training, and also in terms of the absolute numbers being trained and these are not always consistent. A model with a lower proportion of successful trainees can nevertheless result in more trainees on the GP Register if it is a lower proportion of a higher number.

12. Random selection provides a 'ground truth' model, in which no selection takes place and all available training places are filled. **Random selection is always worse than other selection models in terms of the proportion of successful outcomes**, but it can be better if many places otherwise remain unfilled.

13. The selection models are also compared in terms of attainment at MRCGP AKT and CSA, and the likelihood of ARCP outcome 4s. **Selection models also differ in their potential for adverse impact** in terms of the demographic mix in trainees in terms of place of primary medical qualification, ethnicity and sex.

14. The various selection models assess the effect of selecting entirely on Stage 2, entirely on Stage 3 and using hybrids of the various methods.

15. Selection on Stage 2 total score is effective at increasing rates of entry to the GP Register after four years of training, and also results in lower ARCP Outcome 4 rates. There is however a shift in trainee demographics towards UK graduates and non-BME trainees.

16. Selection on Stage 3 alone has relatively little difference from the Baseline Model, and in particular does not increase MRCGP CSA scores.

17. Hybrid models tend to result in outcomes which are middling in their impact.

18. The modelling of selection suggests that: **strong selection tends to result in leaving training places unfilled; selection is better when selection measures are more reliable; and because selection scores covary, differential effects on AKT and CSA scores are relatively small, most candidates being relatively good or relatively poor at all measures.**

19. All selection measures inevitably have their limitations. Many of those who do not enter the GP Register within four years probably do so because a composite of factors such as poor motivation, lack of interest, competing career attractions, or simply the 'stuff' of everyday life, perhaps involving illness, personal relationships, family

problems, domestic responsibilities, or a host of other events that could never be predicted by Stage 2 and Stage 3 scores.

20. The multiply imputed datasets allow an assessment of **the incremental validity of Stage 3 scores after taking Stage 2 into account.** Our analysis concurs with the concurrent analysis of Patterson et al (2015) which found that Stage 3 scores contribute almost no additional prediction of MRCGP AKT scores over and above the Stage 2 scores. **Stage 3 scores account only for about 3-4% of additional variance in CSA scores after taking Stage 2 into account.** Whether that additional prediction is cost-effective is considered in Chapter 7.

21. The multiply imputed datasets allow an analysis of the incremental predictive validity of Stage2 and Stage 3 scores for entry to the GP Register within four years, or an ARCP Outcome 4, after taking MRCGP scores into account. There is no evidence of additional variance in the outcomes after taking MRCGP into account, and hence **the effect of Stage 2 and Stage 3 on GP Register entry or ARCP outcome 4 is mediated entirely via MRCGP results**.

Chapter 6: Costs of GP selection and training

Costs to the health system of GP selection (based on the 2014 process) and training are estimated in this chapter.

1. The costs of selection and training for GP trainees are estimated, but the perspective adopted (a virtual provider of all NHS GP services across the UK) and assumptions made mean that these costs may not be comparable to those given elsewhere e.g. regarding other specialties where consultant and applicant time may not be included.

2. Cost components relating to the selection of GP trainees incorporate costs incurred by GPNRO, the Health Education organisations of the devolved nations and LETBs, as well as the value of the health care forgone when applicants and assessors take paid leave from NHS positions to participate in selection events.

3. Estimates of these cost components, in 2014/15 prices, are provided and applied to the 2014 selection cohort of 5,992 candidates (from which 3,090 trainees were appointed).

4. **The total 2014 selection cost was estimated at £4.90M** (approximately £800 per candidate and £1,550 per post filled). This cost was made up of: nationally-incurred costs, including Stage 2 test fees (£896,000), LETB Stage 3 'on the day' costs (£2,290,000), assessor training (£466,000) and candidate time (£1,150,000).

5. A simplified set of five potential training consequences of the GP selection process are then outlined: GP Registration achieved in three years FTE, and with 6, 12 and 24 month extensions and not achieved within five years FTE training time. The final outcome of selection is that a post remains unfilled.

6. Each of these selection training consequences is then valued, considering a ten year time horizon from the beginning of training: **in the 'best case' outcome a trainee would qualify in three years at a cost of approximately £210,000** and then provide seven years' service as a GP.

7. **Training extensions of 6, 12 and 24 months added approximately £64,000, £131,000 and £258,000 to the 'best case' cost of training, per trainee. A trainee completing three years of training but not obtaining GP Registration incurs a total cost of approximately £561,000**.

8. **One unfilled training post results in NHS replacement costs and lost health care provision valued at approximately £415,000** with a replacement hospital doctor (for the 18 month placement during training) at 100% FTE. **In 2014, 481 posts were unfilled, resulting in a total ten year cost of £200m.**

9. **Over ten years, financial savings would be made by recruiting a trainee who requires an extension of around 12 months or less prior to passing compared to leaving a post unfilled.** However trainees who do not obtain GP

Registration are always more expensive than unfilled posts, so it is important to consider trainee quality if more "marginal" trainees are to be appointed (e.g. by lowering the standard required to determine 'competence' in selection).

Chapter 7: Economic evaluation of GP selection

This chapter reports a cost-effectiveness analysis of the GP selection process, extending the analysis of costs reported in the previous chapter.

1. The primary measure of effectiveness used is the number of trainees obtaining GP Registration within four years FTE training time. Two 'target recruitment' totals are used: 3,250 and 3,750 trainees per year.

2. The costs considered are: (1) selection costs, (2) costs of unfilled training posts and (3) additional training and service delivery costs that arise when a trainee does not obtain GP Registration within three years FTE.

3. We compare the cost-effectiveness of the 2015 approach to selection (the 'baseline' approach) against six alternative approaches to selection: random selection, reducing the Stage 2 cut-score from 181 on each test back to its pre-2013 value of 166, selecting on combined Stage 2 scores only (with no cut-score), selecting on Stage 3 total scores only (eliminating moderation and with no cut-score), a by-pass of Stage 3 (automatic offers) for those with a combined score of at least 575 at Stage 2 and giving equal weight to each of the Stage 2 tests and Stage 3 total scores (eliminating moderation, maintaining 181 on each test as the Stage 2 cut-score but with no Stage 3 cut-score).

4. We use an incremental approach to analysis, estimating the additional costs and number of GP Registrations with alternative approached to selection when compared to the baseline approach.

5. Data for applications in the four years from 2011 to 2014 are included in the analysis. However the results are based on a process of multiple imputation of missing data which uses data for all applications with GMC numbers and Stage 2 scores from 2009 to 2015. The imputation was undertaken ten times so that the uncertainty arising from having to use imputed data could be ascertained. The results reported here are means across the ten imputations.

6. A mean of **3,014 trainees were selected per year in the baseline approach; while all 3,250/3,750 posts were filled in the random, Stage 2 only and Stage 3 only approaches.** The equal weight to all approach filled all 3,250 posts and almost all of the 3,750. All alternative approaches would increase the number of GP Registrations in four years or less compared with baseline. **The maximum increase in the number of GP Registrations was 237 with 3,250 posts and 550 with 3,750 posts, both with the Stage 2 only approach**.

7. The increase in the number of GP Registrations was primarily due to the increase in the number of posts filled with all alternative models (which subsequently has a significant impact on cost-effectiveness due to the high cost of not filling posts).

8. All other approaches except random selection were also more effective than baseline in producing GP Registrations within four years FTE: only with random selection of 3,014 trainees would the number of GP Registrations within four years FTE have been lower than baseline. Nevertheless, the low variation in the percentage of trainees obtaining GP Registration within four years FTE between alternative approaches to selection increases the relative importance of filling posts in determining cost-effectiveness. For example, with 3,250 posts, these percentages are 76.8% with baseline and ranged from 73.9% with random selection to 78.6% with Stage 2 only.

9. **The baseline approach had an estimated annual selection cost of £4.3m for 3,250 posts and £5.0m for 3,750 posts. Using the old Stage 2 cut score increased costs slightly; with all other alternative approaches reducing selection costs. Random selection cost £0.3m and using Stage 2 only £1.3m for both numbers of posts.**

10. With 3,250 posts, **the average ten year cost per GP Registration in four years FTE or less was £442,000 with baseline selection and, for the alternative approaches, ranged from £383,000 with Stage 2 only to £438,000 with the Stage 3 by-pass. With 3,750 posts, the average costs were £532,000 with baseline and ranged from £404,000 with Stage 2 only to £527,000 with Stage 3 by-pass across the alternative approaches**.

11. **All alternatives except random selection of 3,250 trainees were less costly AND produced more GP Registrations than baseline (i.e. dominated over baseline)**. It is likely that even random selection of 3,250 trainees would be more cost-effective than baseline if the value of care provided by a qualified GP in the seven years following qualification (estimated using their salary) is used to define the willingness to pay for a qualified GP.

12. **The most cost-effective approach to selection with both 3,250 and 3,750 posts would be to use Stage 2 scores only and to fill all posts based on a ranking process (i.e. with no cut-score).** Other approaches with similar levels of cost-effectiveness were Equal weight to all and Stage 3 only.

13. Considering whether all 3,750 posts should be filled is beyond the scope of this evaluation. With all alternative approaches except random selection that increase the fill rate beyond 3,250 posts, the 'marginal 500' trainees have worse outcomes than the first 3,250 trainees selected. However, **compared to leaving the 500 posts unfilled, increasing recruitment is cost-saving over ten years for all approaches.** If an acceptable threshold probability of obtaining GP Registration within four years FTE was identified by those responsible for training it would be possible to estimate appropriate cut-scores for Stage 2 and/or Stage 3 scores and the subsequent impact on the number of posts filled.

14. **Using Stage 2 scores only to select 3,250 trainees would not increase the number of appointed trainees who do not obtain GP Registration compared with baseline selection of 3,014 but would do so by around 150 per year with 3,750 trainees. However the number of extensions to training would increase (by around 40 per year/3,250 posts and by 220/year with 3,750 posts).** Such additional extensions may have non-economic costs, such as a negative impact on morale and patient safety that have not been included in this evaluation but need to be recognised when making decisions about changing the selection process.

PART 2: BROADER RECRUITMENT AND SELECTION PERSPECTIVES

Chapter 8: The numbers of GPs and influences before and during medical school on choosing to specialize in General Practice

This chapter draws on several sources of data to consider the historical context of the numbers of doctors entering general practice since the 1970s.

1. The DH has proposed that "[I]n future **at least half of doctors going into specialty training will be training as GPs.**" (Department of Health, 2008, p.15, para 36, our emphasis) although that is proving hard to achieve. This chapter looks at influences on becoming a GP, historically, and in relation to background, training in medical school, and differences between medical schools.

2. Data are available from the Medical Register, over the last half-century, on the proportions of graduates entering general practice. Various studies since 1966 have looked at attitudes towards GP as a specialty career, from the Todd Report, the UK Medical Careers Research Group (MCRG) studies (started by Parkhouse), three large cohort studies of 1981, 1986 and 1991 medical school applicants, and smaller later studies of 16-year olds and 11-year olds.

3. **The percentage of doctors at the end of the Foundation Programme moving straight into specialty training (including GP) has fallen from 83% in 2010 to 52% in 2015.**

4. Historically, about 15% of UK graduates did not enter either the GP or Specialist Registers. Therefore it is important that the DH target of 50% is interpreted correctly; as the proportion of those entering training and not of those graduating.

5. About 46% of medical school graduates from 1974-1987 went on to the GP Register, which would have met the DoH target of 50% of those in any specialty becoming GPs.

6. The **proportion of graduates entering the GP Register dropped from 46% to about 36% between 1987 and 1991.** That drop is shown clearly in the career intentions of final year students in the three cohort studies. However the cohort studies suggest this is not mirrored by a drop in interest in GP at entry to medical school over the same time period and may be due to a differential increase in other specialties' training posts.

7. Non-UK graduates eligible to work in the UK are less likely to be on the Specialist Register than UK graduates, and **much less likely to be on the GP Register**.

8. The introduction of the GP Register in 2006 means that the time between graduating and going onto the Register can be followed in more detail. Of doctors who eventually were on the GP Register **only half entered it within 5 years of qualifying**. The remainder entered the register up to ten years after graduating. Many doctors are therefore 'late converts' to General Practice, as the MCRG studies have also shown.

9. For the selection rounds for 2009 to 2015 it is possible to look at the time from graduation to applying for GP training. For UK doctors graduating from 2007 to 2009, **over 50% had applied for GP selection by the 2015 round**, with a sizeable proportion applying up to 5 years after graduation. If all of that 50% of applicants started training as GPs then the DH target would be exceeded.

10. In the cohorts of UK doctors applying for GP training, who graduated from 2007 to 2013, the proportions applying in their second year after graduation (F2) **dropped continuously from 38.6%, to 24.1% of graduates.** Some evidence suggests that doctors may just be taking 'a gap year', but it is also possible that there is a sustained reduction in applications to general practice.

11. Data from ARCP records for 2010 to 2014 provide an indication of flows between different specialty training programmes. For those doctors with different ARCPs in different specialties, **almost eight times as many are likely to flow into GP as out of it.** Losses are mainly to Psychiatry and Medicine. **Transfers to GP are not only from the traditional specialties of Medicine, Paediatrics and Psychiatry, but also occur in large numbers from Surgery, Anaesthetics, Obstetrics & Gynaecology and Acute Care Common Stem (ACCS).**

12. Intentions to become a GP at application to medical school are predictive of eventually entering the GP Register. The prediction is not particularly strong at entry, with low specificity and sensitivity, but improves through medical school.

13. Longitudinal, causal path modelling of the cohort studies data separates out various influences on becoming a GP. Apart from the three cohort studies of 1986/7, 1991/2 and 1996/7 graduates, **there are almost no data available on specialty preferences in medical school applicants**, and even fewer data before that. At a policy level, **data urgently need collecting on those applying to and entering medical school, as well as on students in secondary and even primary school.** In the absence of such data, much theorising on policy cannot be evidence based.

14. **Some doctors think of careers in medicine at an early age**. The median age of first thinking of becoming a doctor is about 12 (quartiles 12 to 14) and the median age of definitely deciding to become a doctor is 16 (quartiles 14 to 16). Even **at the age of 11 general practice is a less attractive career option than hospital medicine**.

15. The General Practice experience in medical school has an influence on considering becoming a GP. **Students who report that GP teaching is interesting and potentially useful are more likely to wish to become GPs**, even after taking into account prior interest in a GP career.

16. Women are more likely to become GPs. However in the cohort studies, **female applicants to medical school are no different from males in their wish to become GPs**. By Year 3 however women are already more interested in a career as a GP, they report that GP teaching is more interesting and more useful, and they report further increases in interest in a GP career in the Final Year and after the first Foundation year, with a further increase in the proportion eventually entering the GP Register. **Women [men] appear to be encouraged [discouraged] towards GP by multiple independent events through medical school and in postgraduate training**. The data cannot however exclude the possibility that women are discouraged from careers in hospital medicine by multiple independent events.

17. Compared with students who have no medical relatives, students who have a relative who is a GP show no increased likelihood of wishing to be a GP. However students who have one or more medical relatives but none are GPs, show a lower rate of wishing to be a GP. It seems that **doctors who are not GPs somehow put off family members who are applying for medical school from considering General Practice as a career**.

18. It seems indisputable that **GPs are more likely to graduate from some medical schools rather than others**. The *reasons* for the differences in **'GP-productivity'** are however another matter entirely, and have not been well investigated. In research terms there are inevitably problems of statistical power, particularly as the newer medical schools have had relatively few graduates.

19. Relative differences in GP-productivity between the 'older' medical schools (pre-2000) are stable from 1974 through to the recent UKFPO exit surveys of F2 doctors. In the **data for GP applicants since 2009**, as well as in other data which includes the newer medical schools, there **continues to be evidence of differences in GP productivity between medical schools.**

20. The proportion of GPs produced by the older medical schools is **independently related to interest in general practice of students at entry, and (negatively) to higher levels of academic attainment at entry.** For whatever reasons, applicants who are interested in GP are more likely to apply to (and get into) medical schools that produce more GPs.

21. Analysis at medical school level did not identify any effect of medical school on influencing students' level of interest in GP once differences in levels of interest in GP at entry had been taken into account.

22. GP selection data since 2009 shows that **students from schools with high GP-productivity do <u>less</u> well at the CPST and SJT Stage 2 tests** and they **do not do any better at the Stage 3 Selection Centre measures**. Such results undermine many simplistic explanations that GP-producing schools are providing their students with particular skills and expertise which will be appropriate for General Practice, and which will benefit those students in the GP selection process.

Chapter 9: Specialty Training Applicant Questionnaire: what predicts whether doctors apply for specialty training in general practice?

This chapter analyses responses to a questionnaire sent to all round 1 applicants for 2015 CT1/ST1 specialty training posts.

1. A questionnaire was emailed to all 11,782 round 1 applicants for 2015 CT1/ST1 specialty training posts. There were **3,838 responses (response rate, 33%)**. 1,748 (46%) applied to general practice, of whom 1202 (31%) had GP as their first choice specialty.

2. Applicants took an average of eight and a half minutes completing the questionnaire suggesting good engagement.

3. The questionnaire provided information on preferences for programmes and LETBs applied to, as well as reasons for those choices, which are not available elsewhere.

4. The data highlighted some inconsistencies between the GPNRO and Oriel datasets.

5. Unsuccessful applicants, males and, to a certain extent, non-EU trained and Asian/Asian British applicants were less likely to respond to the questionnaire and are thus under-represented in the results.

6. 69% of GP applicants indicated that it was their first choice specialty. **There is some evidence that GP is a popular 'back-up' specialty**.

7. Questions were asked of all candidates about the reasons for their interest or lack of interest in General Practice.

8. Those applying to GP as their first choice compared to other specialties are more likely to do so to choose where they want to work and for work-life balance, and are much less likely to do so because of prestige and high competitiveness.

9. **Those applying to GP as their first choice have had less experience in GP than those applying to other specialties have experience in their chosen specialty; this is particularly true of non-UK graduates**. This may explain why those applying to GP are less likely to give "positive experience in clinical posting in specialty" as a reason for choosing their specialty.

10. Those applying to GP as their first choice specialty are less likely to receive advice from every source compared to applicants to other specialties; **lack of positive careers advice could be reducing the size of the GP applicant pool**.

11. **GP applicants are more likely to consider location as more important than specialty than other applicants.** Overall, 70% of applicants (76% for GP applicants) stated that their current LETB was their first choice LETB. However, LETBs are less likely to be chosen on account of their general or training reputations amongst those applying to GP as their first choice compared with those applying to other specialties, with proportions stating these reasons of 38% vs. 52% for general reputation and 26% vs 44% for training reputation.

12. Only two applicants were offered an interview in a LETB that was not their LETB of application. 85 (2.8%) applicants were offered a post in a different LETB; of these, only 44 (51.8%) accepted their offer, compared to 83.9% of those offered a post in the same LETB. Although statistically significant, the number of GP trainees 'lost' through not meeting LETB preferences is low given the small numbers.

13. **Multiple regression was used to carry out a path analysis of interest in GP as a training programme**. To do this, doctors' interest in GP were categorized into four groups: *No interest in GP* (41%); *Considered GP but had not applied* (14%); *Applied for GP but not as a first choice* (14%); and *Applied for GP as first choice* (31%).

14. The major predictors of interest in GP were *Believing that one's personality was suited to GP*, and *Work-Life balance*. There were smaller effects of *The patient care that I could provide*, and *Intellectual Challenge*.

15. *Positive experiences of GP while working (presumably mostly at Foundation level) or at Medical School* had small direct effects on choice, but larger indirect effects by driving an awareness that one's personality was suited to GP.

16. Media images of GP had a significant but small negative influence on interest in GP.

17. **Being male, white, and having a later year of graduation all had negative influences on interest in GP**.

18. Doctors with greater experience working in other specialties had less interest in GP as a specialty.

19. **A simple measure of whether a doctor is likely to apply for GP as a first-choice specialty could be calculated based on the answers to six questions:**

    a) My personality is suited to GP

    b) I would like to provide patient care in GP

    c) Work-life balance is important

    d) I find GP is intellectually stimulating

    e) I had a positive experience of GP at medical school

    f) I had a positive experience of working in GP

    In terms of system changes, work-life balance and positive experiences of GP, both at medical school and Foundation training may be the most amenable to change.

20. Free text comments, analysed across respondents with differing degrees of interest in GP, provide a different perspective on what makes GP attractive or unattractive. **A lack of interest in GP was the most-cited reason for not considering it, but perceptions of the nature of the work of a GP (e.g. short consultations and a lonely working environment), uncertainty over future contract changes (including an erosion of work-life balance) and the portrayal of GP in the media were also highlighted as putting respondents off applying.**

Chapter 10: Applications to Specialty Training 2015

This chapter is based on the Oriel dataset which contains information about applications to UK specialty training in 2015. Inconsistences in these data mean that the results should be interpreted cautiously.

1. Based on the Oriel applications dataset, the population of applicants to specialty training in 2015, including GP are described. We identified some likely missing data and inconsistencies when comparing the Oriel data for GP applicants with data provided by GPNRO; a further difficulty is that Oriel does not distinguish between longlisting (Stage 1) and shortlisting (Stage 2).

2. Most applicants only applied to one specialty: 73% in round 1 and 79% in round 2.

3. GP had the lowest number of applicants per post (1.3) and almost the lowest fill rate (69%) across all ST1/CT1 specialties. In general, the specialties with the highest number of posts were less competitive and had a lower fill rate.

4. Of the 4837 who applied to GP in round 1, 51% (2477) accepted offers, and 24% withdrew (13% before interview and 11% after offer made). The remaining 25% either were not shortlisted (8%), not appointable (12%) or appointable but still no offer made (4%). To increase accepted offers requires encouraging more applications, persuading candidates not to withdraw and make offers to everyone who is appointable.

5. Of UK graduates, in GP round 1, 316 (9%) were interviewed but deemed unappointable and a further 201 (6%) were appointable but not offered a place; 51% (160/316) of those deemed unappointable and 51% (103/201) appointable but not offered a place, accepted an offer in GP round 2 or another specialty.

6. 83% of those receiving a GP offer in round 1 accepted it, compared with 79% for all other specialties combined.

7. For those who applied to GP in both rounds 1 and 2, the likelihood of being offered a place in round 2 was 66% if they had attended Stage 3 in round 1 and 48% if they had not been shortlisted in round 1. Re-applications should therefore be encouraged, as there is a fairly high probability of success.

8. 454 candidates chose between GP and non-GP offers in round 1: 281 (62%) chose the non-GP specialty and 173 (38%) GP. The percentage of applicants accepting their GP offer ranged from 12.5% when paired with Clinical Radiology to 57% when paired with Obstetrics and Gynaecology. Fewer applicants received multiple offers (including GP) in round 2; of those accepting one offer, 14/30 accepted their GP offer (47%).

9. When considering those who applied to GP and another specialty, across all specialties, 85% of applicants considered appointable by GP and/or another specialty were considered appointable by GP and 79% by the other specialty. This very crude analysis suggests that overall 'the bar' is a little lower in GP than other specialties i.e. it is slightly easier to be deemed appointable in GP. Across all specialty comparisons, the overall level of agreement regarding appointability, as measured using Kappa, was 0.40 (45% of the maximum Kappa possible given differences in the relative frequencies of appointable/non-appointable between specialties; this is considered fair to moderate agreement).

10. Based on GPNRO data, of 360 (7%) candidates rejected at Stage 1, the most common reasons for this rejection were visa/immigration issues (230, 64%) and not providing satisfactory evidence of foundation competence (115, 32%).

~ This page is intentionally left blank ~

Chapter 1

Introduction to Evaluation of GP Specialty Selection

Chapter 1.

Introduction to Evaluation of GP Specialty Selection

> *"...it is impossible to separate selection from training. A good selection scheme can be ruined by a bad system of training. A bad selection scheme can, within limits, be bolstered up by a good system of training. [...] In this country the only case I know of a thoroughly validated selection procedure from first to last was one in which selection and training were treated as a single problem"*
>
> - Sir Frederic Bartlett, British Medical Journal (1946)

The comment of Sir Frederic Bartlett, the great Cambridge psychologist, that selection and training need treating as a single entity, was probably based on experience working with the RAF during the Second World War, on the selection and training of aircrews. Seventy years on there are still almost no large-scale studies in UK postgraduate medical education where selection and training have been analysed together. Indeed one could go further and argue that even the combination of GP selection and training should not be treated in isolation as GP trainees have already been selected into and trained at medical school, and selected into and trained during the Foundation Years, and will (hopefully) be subsequently selected into practices, and continuing professional development means that training continues throughout the working lives of GPs. While this study cannot quite adopt a 'cradle to the grave' perspective, we believe that we can in this report consider aspects of both medical school and Foundation training as precursors of GP selection and training in a way that has not yet been properly attempted. That was possible because those commissioning the research made access available to the multiple and various sets of data which describe the progression of GPs through training.

The present study makes possible, perhaps for the first time, the linking together of multiple large scale databases on the training of GPs, including FPAS[1] measures of medical school performance, ARCP measures during Foundation, a range of selection measures collected during selection into General Practice and other specialties, ARCP performance during GP training, performance in the MRCGP examinations, entry onto the GP Register, as well as problems with Fitness to Practice. Those databases have not been straightforward to link together, but the linkage has produced results which, we believe, are not available anywhere else. A particular problem was the unanticipated delays in some of the data being provided, perhaps inevitably given the complexities of management and ownership of the various datasets, and the result was that this report has been delayed somewhat beyond its original target date. We believe that the results justify the delay. Those delays did allow us time to investigate a number of other issues of relevance to those commissioning the research, and in particular we have explored GP recruitment and training since the 1960s, using various data sources beginning with the 1968 Royal Commission on Medical Education (the Todd Report), and looking for emerging patterns, trends and influences on doctors choosing to become general practitioners. There is currently a crisis in recruiting sufficient GPs, and it is at such times that it can become particularly helpful to stand back and take a longer, wider overview of the issues, as well as delving deeply into the specifics of current practice. We believe the current report does both. Part 1 focusses on the GP selection system, whilst Part 2 considers broader perspectives.

In 2015, 48% of doctors did not go straight into either GP or Specialty training after Foundation training and significant numbers transfer later in their careers into GP from other specialties (see Chapter 8). Therefore, it is also important to consider transfers between specialties, as well as the international dimension both in terms of doctors entering and leaving the UK. Post-training issues such as GPs working part-time and/or leaving the profession before retirement age are key concerns in terms of workforce planning; however, they are of course beyond the scope of this evaluation. Nevertheless the databases linked together in the current analyses provide a launch pad from which to address a range of future questions as the doctors in the various cohorts

[1] We apologise for the large number of acronyms and abbreviations: please refer to the List of Abbreviations at the end of the report.

move into practices, and become the trainers and leaders of the next generations of GPs. Longitudinal data become more valuable the longer they are collected, the influences of the past upon the present become clearer, and the effects of policy decisions can be analysed more accurately, particularly if clinical outcomes of patients can be incorporated into the studies.

1.1 REPORT BRIEF

On 18th December 2014, an Invitation to Tender was announced by HEE asking for expressions of interest "preferably no later than 17:00 on 19 December 2014"; the urgency indicating some of the current pressures upon GP recruitment. The research brief stated that "The evaluation will assist HEE to ensure that [the] GP recruitment and selection process is robust and able to support the growth in GP recruitment required by the HEE mandate… The evaluation must include quality and value for money assessments of recruitment processes including utility, validity and the costs of remediation if poorer candidates are appointed with recommendations for future action." Appendix 1.1 contains the 'Plain Language Summary' of our agreed proposal.

Given the tight timetable, work commenced before contracts were signed and the exact work to be undertaken shifted somewhat due to discussions with our Advisory Group (and other stakeholders) and the availability of data. Informal discussions revealed an ever widening set of questions about the nature and effectiveness of GP selection which were being asked by the Advisory Group and others. The result is that some aspects of the report have become much more detailed than originally envisaged, new approaches have been introduced, and some analyses have not been undertaken. It should be noted that on the agreed timeline, HEE would provide the required data by February 2015. Some arrived then, but most did not due to issues that HEE faced over data governance across multiple organisations and nations (as indicated in the Acknowledgements, we are most grateful for numerous individuals and organisations for helping to overcome these issues). Most crucially, the summative assessment scores were provided in August 2015, which is the main reason for delays to this report. On 16th October 2015, we finally received FPAS (Foundation Programme Application System) data regarding final year medical students, and we are pleased to have included some analyses of these as they provide an important perspective on the broader picture of undergraduate and postgraduate training.

In comparing this current report with the 'Plain Language Summary' (Appendix 1.1) agreed with the Advisory Group in March 2015, the following points should be noted:

- Overall we couched the evaluation in terms of a Utility framework: reliability, validity, educational impact, acceptability of the method to the stakeholders, and cost (van der Vleuten, 1996, van der Vleuten and Schuwirth, 2005). It is worth noting here "the overriding message the model was intended to convey was that choosing an assessment method inevitably entails compromises" (van der Vleuten and Schuwirth, 2005, p310). Our endeavour in this report is to enable policy makers to have a clearer grasp of the compromises, particularly between predictive validity, educational impact and cost. This overall aim will be returned to in the discussion (Chapter 11).

- Work Package 1: score analyses. Reliability is the focus of Chapter 3 and Chapters 4 and 5 are concerned with predictive validity. Our analyses have been far more detailed than originally envisaged, primarily due to using alternate forms reliability in Chapter 3 and use of the Expectation Maximisation (EM) algorithm in Chapters 4 and 5 to provide more robust validity estimates. However, we have not fully investigated differences between LETBs and various groups of trainees. Passmark reliability estimates have been included in our reliability analysis. As the project progressed, it was decided that ARCP and being accepted onto the GP register were also important outcomes. In addition, using prior attainment from medical school can greatly enhance the analytic power of this investigation: hence FPAS data were sought (but arrived too late to be fully utilised).

- Work package 2: Cost and economic evaluation. Costs are considered in Chapter 6 which includes those related to the selection system, training (including extensions) and unfilled GP posts. The economic evaluation in Chapter 7 draws upon the EM modelling described in Chapter 5 to consider the likely costs per qualified GP of various approaches to selection, compared with the current system as the baseline.

- Work package 3: Survey data. This is the focus of Chapter 9, and the major focus is on attitudes towards and against GP training for all respondents, including those who applied to other specialties. The resulting path analysis investigated reasons for and against applying to general practice. Results from other studies are drawn on in Chapter 8, which considers medical school and other influences on the decision to become a GP; however we did not undertake a full review of the literature on the acceptability of selection processes for specialty training.

- Work package 4: Synthesis. This we endeavour to do in the discussion (Chapter 11) where we pull together the key findings and recommend potential courses of action.

We are delighted to have been involved in such an innovative, important, timely and ambitious project. Despite serious problems with the quality of some of the data, we are excited by the potential for longitudinal research of this nature to address complex and important research questions. We believe that this is one of the first analyses of medical training which uses data collected across the whole of training, from data on undergraduate performance (FPAS), foundation performance (ARCP), selection into GP Selection Training (Stage 2 and Stage 3), as well as data on applications overall to other specialties (Oriel), a questionnaire study of applicant motivations, performance during training (ARCP), performance at MRCGP (AKT and CSA), entry onto the GP Register, and Fitness to Practice problems on the GMC LRMP. In addition we have reviewed historical datasets to gain a wider view of GP selection and training. The resulting 'mosaic' provides, we hope, a broad picture of the issues involved in GP choice, selection and training, which in conjunction with economic data, allows an analysis of cost-effectiveness.

1.2 HISTORICAL AND POLITICAL CONTEXTS

Here, we outline some of the context within which this evaluation was conducted. General Practice has a history of employing innovative approaches to selection, so it is important to be mindful of changes that have been implemented in the last two decades. A historical perspective is required to understand the data upon which this report is based: to track trainees from graduation from medical school, working through Foundation then GP training requires a minimum of five years, and often considerably longer. We start with the political context, both in terms of the need for more GP trainees and the different pressures in different parts of the UK.

1.2.1 Political context

A major impetus for this study is the view that more GPs need to be trained. Failure to recruit sufficient GP trainees has been widely reported as a "looming crisis" in the medical press and national newspapers[2]. The mandate to Health Education England in May 2013 included "Ensuring that 50% of specialty trainees choose to enter GP specialty training" as a longer term objective (Department of Health, 2013, p17). In 2014, this mandate was worded, "HEE will ensure that 50% of trainees completing Foundation level training enter GP training programmes by 2016" (Department of Health, 2015, p17). As well as setting a date, this changed wording focused on UK trained doctors. The precise targets to be attained by GP recruitment in the UK have though continued to remain unclear, and as the Centre for Workforce Intelligence put it in 2014, "although the Government's desire for a significant increase in GP training numbers is clear, the magnitude … of the increase is open to interpretation" (p.21). In Chapter 8, we give a detailed consideration of recruitment trends and include two important caveats regarding this 50% target: first, that not all doctors enter GP or specialty training; and second that increasing numbers of doctors are delaying entry (including those training in another specialty directly after Foundation before entering General Practice), so that it could be several years before we know whether the target is (eventually) met.

An awareness of a continuing shortfall in the numbers of GPs being trained, with about 2,700 GP trainees being recruited annually, resulted in the establishment of the GP Taskforce, chaired by Dr Simon Plint, which reported in 2014. "The GP Taskforce was established by Medical Education England (MEE) and the Department of Health (DH) to recommend how the system could achieve the longstanding workforce target for 3,250 trainees to enter GP training in England each year by 2015" (Plint

[2] For example in The Pulse http://www.pulsetoday.co.uk/your-practice/practice-topics/education/training-bosses-given-two-years-to-boost-gp-trainee-numbers/20006602.fullarticle and http://www.telegraph.co.uk/news/health/news/11517019/One-in-3-trainee-GP-posts-are-empty-amid-warnings-of-crisis-shortage.html

2014 p3). The taskforce also reported further issues concerning increases in part-time working, significant numbers of GPs approaching retirement age and women in their 30s leaving the profession early. They also noted the extra demands being placed on general practice and concluded, "there is a GP workforce crisis which must be addressed immediately…" (op. cit. p6). Nevertheless, the Migration Advisory Committee felt the GP shortage was insufficiently severe to include General Practice on the Shortage Occupation List as "any shortage of GPs can be addressed by changing the incentive structure such that the GP route becomes more attractive relative to the hospital consultant route" (Migration Advisory Committee, 2015, p2)[3].

Despite this political landscape suggesting a sense of urgency, discussions with our Advisory Committee and several other stakeholders gave us a strong message that we should be academically rigorous, avoid knee-jerk conclusions, and consider options regardless of whether they are quick-fixes or longer-term solutions.

From discussion with stakeholders, we understand that the 'recruitment crisis' applies much more strongly to some parts of the four countries within the UK than others. We have endeavoured to take a UK perspective throughout, but acknowledge that some parts of this report may not appear relevant to all parts of the UK. Indeed the mandate and taskforce referenced above were focused on England. The target recruitment for England in 2016 is 3,250 trainees. With no significant concurrent increases in post numbers for Northern Ireland, Wales or Scotland, this amounts to approximately 3,750 posts available across the UK . Our modelling suggested an average of 3,014 posts filled per year from 2011 to 2014 across the UK, when the number of posts available across the UK averaged 3,250 per year. Therefore the England target of 3,250 (and 3,750 posts UK-wide) means that filling these posts will be particularly challenging. In Chapter 5, we apply the average number of posts available in the UK for 2011 to 2014, 3,250, when we consider the consequences of using alternative approaches to selection that enable more (or all) posts to be filled, although the methods we describe for evaluating selection methods in this chapter could be applied to any targets. In Chapter 7 (the economic evaluation), both the 3,250 and 3,750 targets are considered.

Like Simon Plint[5] and his team, we hope that the analyses and recommendations in this evaluation are useful to GP educators throughout the UK.

1.2.2 Historical perspective of recruitment to GP training in UK

> *"Those who cannot remember the past are condemned to repeat it"*
>
> - George Santayana

This section draws upon the reflections of Roger Price, based on his experience of GP selection since 1987.

Prior to about 2000, GP selection was undertaken in local regions of UK and was primarily practice or programme led. Such processes did not have explicit governance, or evaluation by external bodies or candidates. How appointment decisions were made was neither transparent nor explicit; there were no competency-based processes. The interview panel would have read the application which included a CV and biography. Questions asked were often about GP practice of which the candidate may well have had no experience. The approach used by these panels could vary between interviews and interviewers, and certainly between different regions.

Roger was not aware of any training for panel members, other than for basic Equality and Diversity; nor was there calibration of panel members. Trainees were subsequently appointed to a 2-3 year rotation by the local Vocational Training Scheme (VTS) programme, or a group of such schemes acting together in a deanery. An individual training practice might appoint a doctor who only required a single year in practice as a trainee, usually without any external input to the process: this stopped when the Gold Guide was introduced and the Certificate of Completion of Training (CCT) was required.

[3] The Plint report advocates such changes, too.

[4] Posts available for 2016 are at https://gprecruitment.hee.nhs.uk/Recruitment.

[5] In the forward to the taskforce report, Simon Plint wrote: "The statistics and recommendations in this report refer to England, unless specifically stated otherwise, but we hope that the recommendations will be useful for all UK nations" (Plint 2014, p3).

Since the late 1990s, GP training has undertaken a long process of becoming more explicit and systematic. For example, Pat Lane from Sheffield initiated work in 1996-8 to share knowledge of appointment process and evaluate outcomes between the GP directors and deaneries. With this transparency and evaluations, progress could be made. The three deaneries of North, South and Mid Trent began to use a regional assessment centre to select trainees in 2000. This evolved into a nationally-led recruitment process managed by the GP National Recruitment Office (GPNRO). The national GP recruitment process was based on a multi-source, multi-method assessment derived from a job analysis which enabled the competencies required for success as a GP to be identified (Patterson et al. 2000); this has subsequently been updated and used to revise the selection process (Patterson et al., 2013). (We describe the current selection process in detail in Chapter 2 of this report.) As part of initial development, work was undertaken to determine how the required competencies could be evaluated in written tests of clinical problem solving and situational judgement (also termed Professional Dilemmas), and in an assessment centre (Patterson et al. 2001, Patterson et al. 2009). A systematic approach to training and calibrating assessors was also developed and continues to be used.

A number of evaluations of the GP recruitment process have been published (see https://gprecruitment.hee.nhs.uk/Recruitment/Selection-System-Research-Evaluation), although these have been led by the team responsible for its development and thus may not be fully independent. We do not summarise all of these studies here, but refer to them in later chapters as appropriate.

Roger's view, in 2015, was that the GP selection system is not perfect, but there is very positive feedback from all involved, including the candidates. Ongoing internal and now external evaluation helps to ensure continuous improvement.

1.3 SELECTION IN OTHER POSTGRADUATE SPECIALTIES

Although the emphasis of the present report is on General Practice, it is perhaps worth emphasising that General Practice has been in the vanguard of developing formal, properly constructed, principled approaches to the selection of trainees. It is the largest of all the specialties which are selecting, and its methods are clearly documented and carried out consistently at national level. It is probably the case that that is not true of all other specialties. If in part some of GP selection is found wanting then that will partly reflect that GP selection has collected large amounts of systematic data and stakeholders are prepared to allow a proper assessment of it. We know of no comparable databases or analyses in other specialties.

1.4 COHORTS STUDIED

The data in this report are complicated as there are seven different selection cohorts included and not all cohorts have all measures. Figure 1.1 provides a synoptic overview of the data in a diagram similar to those developed by Charles Ibry in the 1840s for railway scheduling.

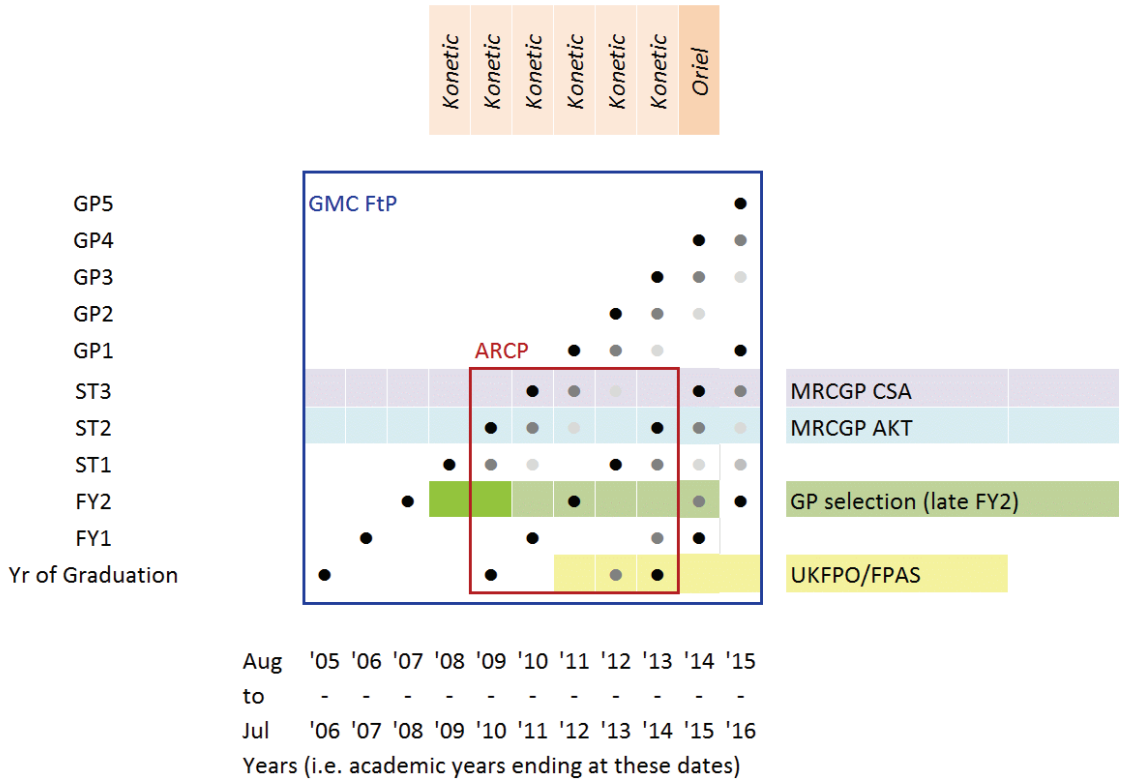
The horizontal axis shows time, expressed in academic years, and the vertical axis is the 'stations' through which a GP progresses in their training.

The doctors: The black dot in the lower left corner shows a final year medical student who is graduating in 2006. As each year passes the graduate moves up one step vertically, giving the diagonal line of black dots, so that this student/trainee/doctor begins GP training in Aug 2008, and enters the GP Register in August 2011, then progressing over the next years as a fully qualified GP. A problem for understanding the data is that many doctors in training take a year or more out, perhaps for personal or family reasons or due to examination failure or other professional problems. For the 2006 graduate that is shown by the dark grey dot at ST1, the doctor has taken a year out so that they move horizontally without progressing vertically. The paler grey dot shows a doctor who has taken two years out. After the year(s) out, progression continues as normal. Of course doctors can take time out at any stage during their training, but only a few examples are shown. Black dots have also been placed in the diagram for the 2010 and 2014 cohorts of graduates, with some grey dots to remind the reader that progression is not always linear.

Other aspects of the diagram show for which cohorts we have data.

GMC FtP data: The blue box indicates that GMC Fitness to Practice (FtP) data are available for all doctors in all cohorts.

» *Figure 1.1 Cohort diagram*



ARCP: The magenta box shows that ARCP data were only available for the academic years ending in 2010 to 2014. There is a delay in ARCP data being made available so that at the time of this report, December 2015, there was no ARCP data for the academic year August 2014-July 2015.

MRCGP examinations: The MRCGP AKT is typically taken in ST2 and the CSA in ST3, shown by the pale blue and pale purple boxes. Assessments have been made available to us for many years, although the format has changed over the years.

GP selection: National selection for General Practice began in 2005 for those entering training in August 2006, but the system was in flux for a year or two, and the present study has data only for doctors applying to enter training in ST1 in August 2009 (shown by the green boxes) until 2015. For doctors progressing directly from Foundation, selection itself takes place during FY2. The present study did not have data for applicants applying in November 2015 for training in August 2016. The two different colour greens indicate that Stage 3 had a different structure for the first two years than for the following five years.

UKFPO/FPAS: The UKFPO collected FPAS data for medical students graduating in 2012 and 2013, and these are shown in the pale yellow boxes.

Konetic/Oriel: Data on specialty training applications are collected nationally, and for the years 2008-9 to 2013-14 used the Konetic system, whereas for 2014-15 the Oriel system was introduced (the two systems being shown at the top of the diagram). Oriel in particular has ID numbers which allow linking across rounds and specialties, as well as collecting other data. We did not have access to Oriel data for 2015-16.

Of course the diagram does not include all of the other changes which have taken place in medical training since the millennium, all of which impact upon training, career choices, selection processes, etc. For example, MTAS was used just once, for training beginning in August 2007. Broad Based Training (BBT) was introduced in 2013 but has now ceased i.e. no applications for 2016. Within GP selection, the Stage 2 score thresholds and the design of Stage 3 have also changed. In the

relevant chapters, we have endeavoured to account for some of these major changes, but there may well be decisions and events that have impacted upon GP recruitment that we have not taken into account e.g. factors that have significantly altered the number of IMGs applying for GP training in the UK.

In reading Figure 1.1 it must be emphasised that it is **schematic**, rather than precise, and its intention is to give a sense of what can and has be done with the data. A more detailed diagram would subdivide academic years to make precise timing (e.g. of selection) clearer. Two examples of aspects of selection that can be seen are that a) not all cohorts of trainees can be followed up for the five (or more) calendar years which are necessary to assess whether trainees are on the GP Register within four years (our outcome for the economic evaluation in Chapter 7), particularly when taking Less Than Full Time and Out of Programme experience into account; and b) there are no doctors with FPAS data who have yet taken MRCGP AKT or CSA; AKT will not be taken until (at the earliest) the academic year 2015-16, and CSA the academic year 2016-17. Like all maps, the diagram can nevertheless help to clarify what is otherwise a complex and confusing terrain.

1.5 LIMITATIONS

For this evaluation of GP selection, it may seem that there should be simple answers to the simple questions, "Are the right people identified?" and "Can the selection system be improved?". This is not the case because of the number, size and messiness of the datasets used and the complexity of the issues. For example, reliability is a key concept for evaluating any selection system, but there are numerous ways that reliability can be calculated; all of these ways require an understanding to interpret them correctly, with the most appropriate way(s) depending on the data and context. Consequently, sophisticated analyses have been undertaken in this evaluation as it is important to understand this complex selection system and the even more complex social/medical/career system in which it is embedded.

To take an anecdotal example from Northern Ireland, we were told of a very high calibre intake, as a result of strong competition for GP training places. Those training in Northern Ireland also tend to have low exposure to GP at Medical School and during Foundation training. Does this mean that prior experience is unimportant? In Chapter 8, we suggest that such experience is important, but the point, here, is that with such a complex system, other factors are also at play and complex analyses have to be undertaken to investigate how different factors interact.

Doubtless, knowledgeable readers will be frustrated by many important issues that haven't been explored in detail in this report, if at all. For example, we have looked at differences between UK and international graduates, but not in sufficient detail to answer important questions of potential bias. Likewise, we have left many unanswered questions regarding differences between the four countries of the UK and between LETBs. Although it is relatively simple to find differences between countries or LETBs, it is much harder to understand 'why'. For example, if a particular LETB has a low rate of trainees completing GP training, would this be because of poorer trainees, poorer training, or other unknown factors?

In recent months, there have been many proposed changes to GP selection (see https://gprecruitment.hee.nhs.uk/Recruitment/Summary-of-Changes). We have included modelling of a bypass of Stage 3 for those with high Stage 2 scores, using the same cut-score as being used in 2016. However, many other proposals such as a pre-GP year did not seem to be amenable to analysis within this evaluation.

Chapter 9 is based on the survey for specialty selection applicants in Round 1 2015 distributed by HEE. We have analysed the questions that we designed to address certain questions, but we have not analysed questions on the selection system, particularly Oriel, that were of interest to HEE.

In Chapter 10, more detailed selection data were made available to us to compare GP selection with the BBT specialties: unfortunately time and space have precluded inclusion of these comparisons in the present report.

As indicated above, this report is structured round the Utility framework, but we say very little regarding acceptability and educational impact. WPG have undertaken several questionnaire surveys of GP candidates at Stages 2 and 3 so it did not seem appropriate or necessary to repeat this: their findings are that candidates generally felt tests were fair and relevant (e.g. Lopes, Ashworth and Tate 2013). One of us (CT) has previously run a conjoint analysis to try to ascertain stakeholders' relative

weightings on the different components of Utility for specialty selection in general; in that study, validity and reliability were certainly the most important to stakeholders (Thomas, Taylor, Davison et al. 2010). We did not consider it critical to repeat such work focusing on GP alone as we believed the results would be similar and that other analyses had a higher priority. Likewise, we have not investigated the educational impact of different methods of selection; but we do consider the issue in the final chapter.

1.6 THE PROFESSIONALISM AND COMMITMENT OF THOSE INVOLVED IN GP SELECTION

The three authors of this report have been involved in evaluations of undergraduate and postgraduate medical training over a number of years. Once again we have been struck by the professionalism and commitment of all those involved in the various aspects of selection and their willingness to critique processes that they have spent much time considering, developing and implementing, and to engage with potentially uncomfortable possibilities concerning future developments. Many people have gone out of their way to facilitate this evaluation, some of them we have acknowledged above and draw extensively upon their written and oral communications. We are confident that the changes to the GP selection system recommended here will be debated relatively dispassionately and whether or not changes are adopted, GP educators will continue to work hard to provide trainees with the best possible training; for it is by having the best trained practitioners that General Practice provides a fundamental corner-stone of British healthcare, and thereby provides high quality healthcare for the population.

~ This page is intentionally left blank ~

Appendix 1.1

Plain Language Summary of the Evaluation of GP Specialty Training

Appendix 1.1.

Plain Language Summary of the Evaluation of GP Specialty Training

This summary was agreed with the Advisory Group in March 2016. Although broadly undertaken, as described in Chapter 1, some elements have been expanded, other added, and some analyses have not been undertaken.

Health Education England (HEE) has commissioned this evaluation to see if the GP recruitment process is fit for purpose. There is national concern that too few GPs are being trained, but any lowering of selection standards may result in unacceptable increases in remediation costs (and threats to patient safety) if less competent trainees are appointed.

This evaluation investigates the reliability, validity, educational impact, acceptability of the method to the stakeholders, and cost. Reliability and validity are the focus of work package 1. Fairness is included here as our analyses will allow comparison between different groups of trainees. Cost – both in terms of the cost of selection itself and the costs associated with trainees who require additional training time or fail assessments – is also crucial as all NHS budgets are carefully scrutinised, so an economic evaluation is the focus of work package 2. Acceptability and educational impact are considered in work package 3, whilst work package 4 synthesises the findings.

WORK PACKAGE 1: SCORE ANALYSES

This work package investigates the reliability and validity of the GP selection process. *Reliability* (and its more detailed extension, known as 'generalizability') is concerned with whether the same candidates would be selected under different conditions, e.g. with different questions or assessors. Low reliability in any assessment is a concern, not only because of fairness issues, but also because good candidates may fail and weak candidates pass because of chance (random) factors, rather than differences in their true ability. *Validity*, in contrast, concerns how assessments at one stage relate to performance at a later stage, either as further assessments or in clinical practice. We are restricting our consideration of validity to how well success at selection predicts performance at summative assessment and whether those predictions differ between areas of the country or between various groups of trainees.

To undertake this work package, trainee data from different parts of the GP selection and assessment processes need to be linked using candidate codes. For selection, Stage 1 of the recruitment process (Application and Eligibility checking) is not envisaged to provide important information, so will not be included in the study. Stage 2 shortlists candidates using a Clinical Problem Solving Test (CPST) and a Professional Dilemmas test (PDT which is also called the Situational Judgement Test, SJT): the reliability of this stage has been previously investigated; however it would be useful to calculate the passmark reliability of the cut score. We will also consider these Stage 2 tests in terms of predictive validity. Stage 3 consists of LETB (Deanery) based Selection Assessment Centres (SAC), which comprise three simulated encounters (with a patient, a carer and a health professional) and a written exercise.

The reliability of Stage 3 of the selection process will be examined in three ways:

1. Cronbach's alpha will be calculated for the marks awarded within each simulation, the written assessment and also for Stage 3 overall. In addition a generalizability analysis will be carried out to take account of station selection, etc.

2. Pass mark reliability will be calculated for each area of the country: this is the proportion of candidates that we are confident would pass (or fail) again if the processes was repeated.

3. The main investigation into reliability of the selection system is a multi-facet Rasch model which investigates the extent to which differences between stations (the three simulations), questions, clinical and non-clinical attributes being assessed, examiners, role players and testing centres affect the outcome of the assessment.

We will estimate the predictive validity of the different elements of the selection system, separately and in combination. We will also look at differential validity for different groups of trainees; for example, do trainees who graduated abroad do better or worse in the summative assessments than expected given their selection scores?

WORK PACKAGE 2: COST AND ECONOMIC EVALUATION

The aim of this work package is to estimate the cost-effectiveness of the selection process. Such work is crucial to justify the extensive investment in the selection process. We will look at the potential impact of modifications to the current process e.g. altering the pass marks applied to selection at Stages 2 and 3; or using both Stage 2 scores as part of the final decision-making process.

We will consider the costs of: selection; unfilled posts; extensions to training; and trainees who do not qualify as GPs incurred over a five year period for a single cohort of doctors entering GP training. We propose a cost-effectiveness approach to the economic evaluation, using the number of trainees achieving CCT as the outcome measure. This analysis is based on the effectiveness of the selection process to determine training outcomes and will estimate the cost per trainee achieving CCT.

We will undertake sensitivity analyses to consider the implications of potential variations in different cost and effectiveness estimates. Such work will enable us to determine which inputs and/or outputs of the selection process are most influential in terms of cost-effectiveness.

In an ideal world, we would also undertake a cost-benefit analysis (CBA), which would enable us to estimate the rate of return to investments in selection. However this approach requires the application of a number of assumptions regarding the value of health care provided by trainees and the variation in this value amongst the cohort of doctors applying for GP training. We will consider the quality of the data available for these variables but do not envisage that the results of a CBA would be sufficiently robust to be useful to inform practice.

WORK PACKAGE 3: SURVEY DATA

This work package addresses educational impact and acceptability by adding questions to the surveys that are already being used in the selection system. We propose undertaking a brief literature search for similar questionnaire data from other specialties to compare their acceptability for candidates. The added questions will ask how candidates prepared for GP selection and their assessment of its wider educational value. We will also seek to understand why recruitment is less successful in some parts of the country. Similarly, it is important to know if candidates regard GP as their first choice and their reasons for this.

Now that applications for specialty training are all undertaken with Oriel, we intend to compare success at various stages of the selection process between GP and other specialties, for trainees who apply to more than one specialty. This may provide evidence regarding differences between specialties in terms of difficulty and popularity.

WORK PACKAGE 4: SYNTHESIS

Our synthesis will focus on illustrating how changes to the selection process could improve outcomes. The nature of these potential changes will depend on our initial findings, but here are some possibilities:

- If differential prediction analysis demonstrates differences in the relationship between selection and end of training assessment scores between parts of the country or groups of trainees (such as by gender), then modifications to the selection process may be recommended to improve fairness.

- Increasing the number of scenarios in Stage 3 of selection may be cost-effective in terms of reducing the number of trainees who have difficulties completing GP training. Alternatively, Stage 3 is a very expensive process and it may be that its incremental increase in predictive validity is too low to justify its inclusion, particularly if the predictive validity of Stage 2 is high.

- The cost-effectiveness of the selection system could be improved by including both Stage 2 scores in the final selection decision if this change increases predictive validity.

To our knowledge, we are using analytic techniques that are rarely or never used to this level of sophistication within medical education:

- Reliability: due to difficulties with data that are not fully crossed or nested, use of generalizability theory is usually restricted to very few factors. Consequently our proposed use of a multilevel framework increases the factors that can be included in the analysis. Likewise, multi-facet Rasch modelling allows examiner, simulator and case differences to be modelled in a sophisticated manner.

- Validity: differential validity and differential prediction analyses will allow us to explore issues related to fairness in greater depth and provide more accurate estimates for the economic analyses than could be otherwise undertaken.

- Economic analyses: our approach to evaluating cost-effectiveness provides a high-quality method of estimating the impact of potential changes to selection.

Appendix 1.2

Notes on the data provided

Appendix 1.2.

Notes on the data provided

Although not the main remit of this review. we cannot help but comment on the data available for carrying out the analyses. Much time and effort was spent making sense of the HEE data provided from Konetic and trying to merge together files from different years, and that has not been entirely satisfactory. When querying the coding of a few candidates we were informed that Konetic has been decommissioned, was inaccessible and potential errors could not be checked.

Since the intention of collecting selection data is presumably to allow assessment of predictive validity of the various measures, then to have a situation in 2015 whereby data even for the first two rounds of 2014 are now "inaccessible" seems pretty poor. If we had to make **one recommendation it would be that HEE should ensure that it has a comprehensive data-management system which goes back over a reasonable amount of time**[1]. To put it bluntly, the present data archiving system is barely fit for purpose. It may be that Oriel will be an improvement, but even assuming it is not decommissioned in half a decade or a decade, there should be a concerted effort to migrate data forwards onto any new system.

A consequence of some of the problems is that:

1. Whoever entered the data for 2010 did not record the round the candidates were taking part in. That information now seems to be lost in the mists of time. We were informed that in this case the member of staff with the password for the file had left and the data were not accessible.

2. Many of the codes used in different years (and even within years) were different in a host of ways. That is not satisfactory. Excel is always a bad way of storing data, not least as different people use different codes in different ways. The ARCP data had 61 different spellings of the University of Anglia, and variations on it.

3. Missing data are not updated. A result is that many candidates do not have GMC numbers. It would seem that this is because candidates can apply without GMC registration and the application form data are not updated when a candidate obtains full GMC registration.

   On the same point, we also notice that some GMC numbers are clearly wrong, having too few digits, etc.

4. Data are rarely 'cleaned backwards'. As an example, in ARCP there may be a typo in one year which says a trainee is in year 3 rather than year 2. However the next year, when they are in year 3, the erroneous data in Year 2 does not or cannot be corrected. That may not matter if there is only one ARCP record per year, but it can result in ambiguities when there are multiple records in a year, particularly when dates are not fully specified for the beginning and the end of an ARCP period.

Making sense of OOP and LTFT in ARCP was very problematic. Typically there was no information on the proportion of LTFT (so that calculating a likely end date was difficult), and codes for the reasons for both were not present.

[1] ICM is educational advisor to the MRCP(UK) and can access detailed, individual level data on candidates going back to 1997, and there is less complete data going back to the mid 1980s. Likewise, HESA has student data going back to the mid-1990s.

Part 1

The GP Selection System

~ This page is intentionally left blank ~

Chapter 2

Progression through GP selection

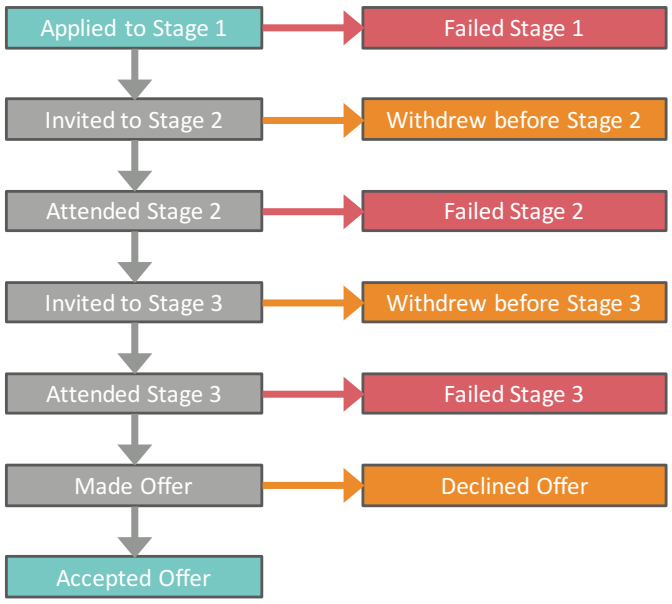
Chapter 2.

Progression through GP selection

2.1 INTRODUCTION

The GP selection program takes place in various stages, which are summarised in Figure 2.1[1]. Candidates pass through three Selection Stages, Stage 1 (Administrative Checks), Stage 2 (CPST and SJT), and Stage 3 (Selection Centre), and may fail (be rejected) at each stage. Candidates may also withdraw before taking Stage 2 or Stage 3. After Stage 3 the candidates may be made an offer, which they may accept (at bottom in blue) or decline (shown in orange). Leaving the selection process may therefore broadly be divided into rejection by the process (in red), withdrawal by the candidate (in orange) or accepting an offer (in blue).

» *Fig. 2.1: Progress through the GP selection process*



Progression of candidates through the selection process. The main part of the Table 2.1 below shows the number of **UK graduate candidates** leaving the process at each stage for the years 2009 to 2015[2] with various summary totals. The shaded part of the table shows numbers of **non-UK graduate candidates** but for simplicity does not also provide the details numbers at each stage of the process for each year and round.

[1] Candidates can also withdraw before Stage 1 and after Stage 3 before being made an offer; however, the data we have received do not go into this detail and so presumably such candidates are included in other categories.

[2] The figures are technically 'applications' and not candidates, since candidates may sometimes reapply in Round 2 or 3 within a year, or reapply across years. The analyses are based on all candidates in the data files supplied, and therefore include all known candidates, which includes some without GMC numbers who cannot be included in the more complex analyses linked across round or year and described elsewhere in the report, and therefore numbers may not always be compatible. A detailed point is that there are candidates who are flagged in the data files as having been rejected at Stage 1 but subsequently have Stage 2 or Stage 3 scores at the same round. We do not know the reasons for this but such candidates have been moved forward to Stage 2 or Stage 3 as appropriate since they appear to have taken place in selection.

The table shows many important features of selection in the last seven years:

1. Total numbers of Round 1 applications **of UK graduates have fluctuated somewhat,** from 3503 in 2009 to 4318 in 2013, and then down to 3696 in 2015. These are quite large fluctuations in a fairly short time period. 2015 is low but is not as low in fact as in 2009.

2. In contrast, **numbers of non-UK graduate candidates have fallen steadily from 2009 to 2015** (3012 to 1415 in Round 1).

3. Round 2 numbers for 2009 are low, but have been consistently higher since 2011, remaining stable for UK graduates, but falling for non-UK graduates.

4. Round 3 was only held in 2014[3]; numbers are small for UK graduates (51), higher for non-UK graduates (142), but overall are small compared with Rounds 1 and 2.

5. **Rates of offers being accepted overall are consistent for UK graduates, at 61%**, with little change by year. Acceptance rates for **non-UK graduates are much lower at about 24%**, with some suggestion of lower rates in more recent years.

6. Details of the stages of loss from the process are shown only for UK graduates, but there is a broadly similar pattern across the rounds and years.

7. Failure at Stage 1 and withdrawal before Stage 2 are only clearly differentiated in some of the years and rounds, but overall 8.2% of all UK candidates being lost for this reason, with probably about two-thirds due to failing Stage 1 and one-third due to withdrawing before Stage 2. See Chapter 8 for the reasons for the failures.

8. About 16% of candidates are lost due to rejection by the selection process, 8% at Stage 2 and 8% at Stage 3. In addition, 16 % of candidates withdraw, 3% before Stage 3 and 12% after receiving an offer at Stage 3.

9. Overall, amongst UK graduates who apply, **40% do not eventually end up accepting offers, comprised of about 22% due to failure at Stages 1, 2 and 3**, and **18% due to withdrawal**. See Chapter 10 for further investigation into these issues for the 2015 cohort.

10. Patterns amongst **non-UK graduates are rather different, as only 24% received offers**, with greater losses at Stage 1 (18%), Stage 2 (23%) and Stage 3 (30%), and much lower rates of withdrawal, totaling about 6% (made of 4% after Stage 2 and 2% of offers being declined). Losses amongst non-UK graduates therefore primarily reflect failures at Stages 1, 2 and 3, rather than withdrawals.

2.2 STAGE 2 SYSTEM

Stage 2 consists of two papers, a 100 item Clinical Problem Solving Test (CPST), and a 60 item Situational Judgement Test (SJT)[4], of which typically about 90 and 50 items are scored and the remaining items are being piloted. The CPST is a longer test because there are twelve clinical topics to be assessed across five competencies, whereas there are only three competencies for the SJT[5]. The CPST items have a standard 'best-of-five' format, whereas the SJT items have two separate formats, ranking of five items, or choosing three from eight items[5].

[3] 2015 round 3, pre-specialty GP and GP foundation program applications occurred too late for inclusion in this evaluation: see https://www.hee.nhs.uk/news-events/news/health-education-england-announces-big-jump-numbers-accepted-gp-training and http://generalpracticesurvival.com/2015/12/18/response-to-hees-announcement-of-a-big-jump-in-gp-recruitment-figures/ (accessed 17/01/2016).

[4] We are very grateful to Professor Moya Kelly and Dr Richard Jones for their detailed responses to our questions concerning the development, scoring and administration of the CPST and SJT. (Email communication dated 21st Sept 2015).

[5] It is also suggested that 50 items, "produce[s] a good reliability (over .70)". (Kelly and Jones, op cit.)

» *Table 2.1: GP selection progression for UK and non-UK graduates*

Year and Round	Detailed statistics for UK graduates with Total scores (in blue), Withdrawals(in green), and Rejections (in Red).								Statistics for non-UK graduates	
	Applied	Failed Stage 1	Withdrew before Stage 2	Failed Stage 2	Withdrew before Stage 3	Failed Stage 3	Offer Declined	Offer Accepted	Applied	Offer Accepted
2009 R1	3503	195 (5.6%)	112 (3.2%	284 (8.1%)	244 (7.0%)	206 (5.9%)	300 (8.6%)	2162 (61.7%)	3012	890 (29.5%)
2010 R1&2[6]	3699	265 (7.2%)	221 (6.0%)	340 (9.2%)	134 (3.6%)	114 (3.1%)	279 7.5(%)	2346 (63.4%)	2638	1013 (38.4%)
2011 R1	3706	214 (5.8%)	0[a]	251 (6.8%)	176 (4.7%)	425 (11.5%)	332 (9.0%)	2308 (62.3%)	1884	397 (21.1%)
2012 R1	4007	276 (6.9%)	0[a]	340 (8.5%)	127 (3.2%)	368 (9.2%)	424 (10.6%)	2472 (61.7%)	1908	397 (20.8%)
2013 R1	4318	266 (6.2%)	0[a]	322 (7.5%)	58 (1.3%)	346 (8.0%)	669 (15.5%)	2657 (61.5%)	1712	335 (19.6%)
2014 R1	3922	237 (6.0%)	0[a]	260 (6.6%)	48 (1.2%)	386 (9.8%)	584 (14.9%)	2407 (61.4%)	1553	331 (21.3%)
2015 R1	3696	242 (6.5%)	60 (1.6%)	105 (2.8%)	131 (3.5%)	272 (7.4%)	618 (16.7%)	2268 (61.4%)	1415	290 (20.5%)
Total R1	26851	1695 (6.3%)	393 na[a]	1902 (7.1%)	918 (3.4%)	2117 (7.9%)	3206 (11.9%)	16620 (61.9%)	14122	3653 (25.9%)
2009 R2	110	29 (26.4%)	5 (4.5%)	12 (10.9%)	5 (4.5%)	5 (4.5%)	20 (18.2%)	34 (30.9%)	441	114 (25.9%)
2011 R2	404	42 (10.4%)	0[a]	149 (36.9%)	6 (1.5%)	28 (6.9%)	17[b] (4.2%)	162[b] (40.1%)	1040	147[b] (14.1%)
2012 R2	389	43 (11.1%)	0[a]	103 (26.5%)	13 (3.3%)	44 (11.3%)	35 (9.0%)	151 (38.8%)	1035	105 (10.1%)
2013 R2	394	86 (21.8%)	0[a]	107 (27.2%)	5 (1.3%)	48 (12.2%)	25 (6.3%)	123 (31.2%)	760	110 (14.5%)
2014 R2	348	40 (11.5%)	0[a]	85 (24.4%)	6 (1.7%)	39 (11.2%)	24 (6.9%)	154 (44.3%)	735	139 (18.9%)
Total R2	1645	240 (14.6%)	5 na[a]	456 (27.7%)	35 (2.1%)	164 (10.0%)	121 (7.4%)	624 (37.9%)	4011	474 (11.8%)
2014 R3	51	11 (21.6%)	3 (5.9%)	2 (3.9%)	2 (3.9%)	6 (11.8%)	2 (3.9%)	25 (49.0%)	142	22 (15.5%)
UK grad Totals	28547	1946 (6.8%)	401 na[a]	2360 (8.3%)	955 (3.3%)	2287 (8.0%)	3329 (11.7%)	17269 (60.5%)	-	-
Non-UK Grad Totals	18275	2785 (15.2%)	423 na[a]	4190 (22.9%)	651 (3.6%)	5504 (30.1%)	432 (2.4%)	4290 (23.5%)	18275	4290 (23.5%)

[a] For these years it seems not possible to distinguish withdrawal from not invited

[b] Something in the raw data file is clearly wrong with these numbers and we have made the plausible assumptions that the declined and accepted values have been reversed, when they are compatible with previous years. We have no further explanation of the anomaly.

The CPST items have a standard 'best-of-five' format, whereas the SJT items have two separate formats, ranking of five items, or choosing three from eight items[7].

Scoring of both CPST and SJT items is by expert judgment (and in particular empirical scoring is not used in the SJT[8]).

Items for each test are banked after writing and then piloted within the test prior to contributing to candidates' scores within subsequent tests, and are subject to a range of quality control procedures involving both expert review and psychometric analysis. SJT questions go through a concordance panel where 'Subject Matter Experts' sit the SJT questions in test conditions; if the experts' level of agreement is good enough, the question is added to the bank. The CPST and SJT are administered on multiple occasions and therefore each test is split and recombined into eight versions which are administered to candidates at random, with the constraint that no candidate in Round 2 will get the same version as they had answered previously in Round 1. The different versions within a year are equated using common anchor items (up to a third of the questions may be the same in two versions to enable this equating).

Raw CPST and SJT scores within a year are converted to scales with a mean of 250 and SD of 40, meaning that the scales are norm-referenced, and cannot be compared across years[9]. For several purposes the raw scores are converted into 4 bands, based upon the standardized scores. Band 1 contains standardized scores up to 180[10]; Band 2 is up to 230; Band 3 is 231-290 and Band 4 is scores above 290. Therefore, there should be nearly the same proportion of candidates in each band each year (see 2013 to 2015 in Table 2.2, below).

Standard setting for the CPST bands was carried out by a Modified Angoff process until 2013 with experts judging the level to be expected of a minimally competent candidate. No standard setting was used for the SJT, "due to the difficulties in establishing what a 'minimally competent trainee' would score on the SJT given the way it is scored", and the bands would seem to be derived from the banding for CPST.

2.2.1 Calculation of the band scores in Stage 2:

As indicated above, the continuous CPST and SJT marks are converted to band scores in the range 1 to 4, and the band scores are used firstly in deciding which candidates go through to Stage 3, secondly the SJT band score is involved in decided whether candidates are appointable at Stage 3, and finally both band scores are combined with the Stage 3 scores to give a total outcome for Stage 3, which is used for ranking candidates (and we have used for our analyses). The setting of the thresholds for the bands is thus of practical importance.

We have found little about the reasoning behind the band thresholds, although a recent WPG summary document (Work Psychology Group, 2015a) describes a 2012 document entitled, **Review of SJT standard setting for GP Selection**, which we have not seen, which, "supports the cut score levels that have been used and suggests even higher cut scores are justifiable" (p.9)[11].

Table 2.2 shows, **for Round 1 only**, the percentages of candidates in the four bands, the mean and SD of the scores, and the thresholds for the bands.

[6] Data for 2010 did not distinguish rounds 1 and 2 and therefore all have been included in Round 1 here.

[7] The origins of the ranking format are somewhat obscure, being described in the document piloting the SJT for the UKFPO selection tests (Patterson et al., 2011) as due to Weekley in 2004 (para 3.1.8), although the reference does not seem to be available (Weekley, 2004) [and an email to Weekley received no reply]. There are concerns that the ranking scoring scheme for SJTs means that occasional missing marks are penalized in a draconian fashion (Harris et al., 2015). A recent, large-scale randomized comparison of different types of response format for SJTs (Arthur et al., 2014), suggests that the ranking format takes candidates longest, has the lowest reliability, and might also have greatest adverse impact.

[8] See Krokos et al., 2004 (REF at: http://www.researchgate.net/profile/Adam_Meade/publication/237302902_Empirical_Keying_of_Situational_Judgment_Tests_Rationale_and_Some_Examples/links/00b495317113da1049000000.pdf) for a discussion of empirical scoring.

[9] There is possibly a missed opportunity here to use statistical equating with Rasch modelling to allow direct comparisons of absolute scores across years.

[10] Up to 165 before 2013. We have been told that "this change made stage 3 more efficient and the numbers being made an offer were higher in 2013 that in 2012" Professor Moya Kelly (Email communication dated 23rd December 2015).

[11] The document also apparently discussed "higher cut scores for deaneries with different applicant ratios to available places" (p.9). We have however found no evidence that cut scores differ for different groups of candidates.

» *Table 2.2: Proportions of candidates in Bands 1 to 4 for CPST and SJT, together with band thresholds.*

Year	Band 1	Band 2	Band 3	Band 4	Mean	SD
			CPST			
2009	1.4% 0-52	31.4% 53-73	52.0% 74-88	15.1% 89+	77.60	10.51
2010	1.5% 0-44	30.1% 45-64	53.9% 65-78	14.6% 65-78	68.51	9.78
2011	3.5% 0-165	29.9% 166-230	54.1% 231-290	12.5% 291+	245.6	40.6
2012	3.2% 0-165	28.7% 166-230	55.2% 231-290	12.9% 291+	246.0	40.3
2013	5.6% 0-180	24.4% 181-230	56.8% 231-290	13.2% 291+	247.4	39.8
2014	7.4% 0-180	25.7% 181-230	54.8% 231-290	12.2% 291+	244.9	41.8
2015	5.8% 0-180	20.9% 181-230	59.1% 231-290	14.2% 291+	250.8	39.4
			SJT			
2009	2.2% 0-536	29.1% 537-626	56.4% 627-693	12.3% 694+	644.5	46.1
2010*	2.0% 0-552?	27.5% ?522-?	59.1% ?-?681	11.4% ?681+	639.2	40.2
2011	3.1% 0-165	31.1% 166-230	54.0% 231-290	11.8% 291+	244.7	40.8
2012	3.0% 0-165	31.1% 166-230	54.0% 231-290	11.9% 291+	244.6	40.8
2013	6.2% 0-180	26.1% 181-230	56.0% 231-290	11.7% 291+	246.2	40.4
2014	8.4% 0-180	25.6% 181-230	54.5% 231-290	11.5% 291+	243.9	42.9
2015	6.4% 0-180	19.8% 181-230	61.4% 231-290	12.3% 291+	249.7	40.0

* Some thresholds are very unclear for 2010 with overlapping of scores and bands.

The table of percentages, band thresholds and means and SDs shows a number of important features.

1. As described above, the marks are now standardized; therefore it is a norm-referenced process, and so cannot take account of differences in true candidate performance across years.

2. The thresholds for 2011-2015 are fixed between bands 2, 3 and 4. However the threshold from band 1 to 2 increased for 2013 onwards, from 165 to 180.

3. The increase in threshold for band 2 from 165 to180 was accompanied by an increase in candidates in band 1, rising from about 3% in 2011 and 2012, to about 7% in 2014-15.

4. To be invited to Stage 3, both CPST and SJT need to be in Band 2 or higher. The change in Band 2 threshold from 165 to 180 means that whereas in 2011, 284/5576 (5.1%) candidates were excluded from Stage 3, that figure had doubled in 2014 to 548/5438 (10.1%).

5. However, in 2014 Round 1, 5% of candidates (334/6688) met the criterion of both CPST and SJT Band 2 or higher but were not invited[12]. Also, this double cut-off means that someone with 183 for CPST and 181 for SJT was invited to Stage 3, whereas someone with 180 for CPS and 298 for SJT was not.

Interpreting the numbers is difficult, but the increase in the Band 2 threshold means that perhaps 250 candidates each year do not reach Stage 3 who might previously have done so. We are not clear of the justification for this change. Ultimately the acceptability of the altered threshold depends on the predictive validity of Stage 2 scores to both Stage 3 and subsequent professional attainment (see Chapter 6).

2.3 STAGE 3 SYSTEM

The Process:

Stage 3 consists of four stations, three (A, B and C) are simulations (scenarios) involving a simulator (actor) who role-plays a patient, a relative/carer, or a fellow professional[13]. The fourth station (W) consists of a writing task, in which candidates respond by producing a series of written texts similar to those required in day-to-day work in general practice (a prioritization scenario). All simulations are carefully piloted and documented, and assessors and simulators are well trained and calibrated (Anonymous, 2014a, b, Howorth, 2014). Simulators do not make formal judgements of candidates, although they are asked to indicate if they have serious concerns, but that would appear to be very rare and we have not analysed them further[14].

Each of the three researchers spent a day at a selection centre (two in the West Midlands, and one in London). The observational notes stem from these visits, in which staff were extremely helpful, allowing observation of role-play scenarios, sitting in on training and moderation, and discussion with organisers, moderators and assessors. We have no doubt about the professionalism of the Stage 3 assessments, the efficiency and the quality of their implementation, or the dedication and commitment of all of those who support them. Face validity is therefore extremely high, and there seems little doubt that the public would be reassured were it to see would-be GPs being assessed in this way.

Although as a process the Stage 3 assessments are good, that does not, unfortunately, mean that they are statistically reliable, that the various assessment domains are being effectively assessed independently of one another, or that the marks produced are valid predictors of outcome: see Chapter 3.

[12] This could be due to their preferred LETB being full, bit it is possible that these are coding errors, e.g. the candidates might have withdrawn.

[13] We are grateful to Dr Bob Kirk for providing us with examples of the various manuals used for the development and running of the Selection Centre in Stage 3.

[14] Of 3751 candidates seen at 11253 simulations in the 2015 Round 1, there were only 12 simulations in which simulators recorded concerns (0.11%), with no candidate receiving two or more concerns. Candidates with a concern recorded against them had much lower overall scores at Stage 3 (mean=38.7, SD=4.0) compared with other candidates (mean = 43.6, SD=5.1, t=3.39. 3749 df, p=.001). Where simulators had concerns those concerns were probably also being picked up in the judgements of assessors.

» *Table 2.3: Station by competency framework*

Station	Consultation with simulated...	Empathy and Sensitivity	Communication skills	Conceptual thinking and Problem solving	Professional Integrity
Simulation A	Patient	Assessed	Assessed	Assessed	Not assessed
Simulation B	Relative	Assessed	Not assessed	Assessed	Assessed
Simulation C	Colleague	Assessed	Assessed	Not assessed	Assessed
Written	-	Assessed	Assessed	Assessed	Assessed

* Some thresholds are very unclear for 2010 with overlapping of scores and bands.

There is one assessor at each station who awards marks for 3 or 4 competencies on a 1 to 4 scale (equating to something like: little, limited, satisfactory and strong evidence) using the grid shown in Table 2.3. The competencies are Empathy and sensitivity (ES), Communication skills (CS), Conceptual thinking and Problem solving (CT&PS), and Professional integrity (PI): see Appendix 2.1 for the positive and negative behavioural indicators that assessors are asked to look out for.

2.3.1 How offer decisions are made:

These scores and the SJT band are inputted into an Excel template. Formulae within the spreadsheet calculate an initial outcome: 1= demonstrated i.e. offered a post; 2= Review likely; 3= Review unclear; and 4=Not demonstrated i.e. rejected. Those with 2 or 3 are discussed at a moderation session, with 'Review likely' more likely to be offered a post; so the final outcome for all candidates is either 'Demonstrated' or 'Not demonstrated'.

The way the scores on the grid are converted to these initial outcomes is complex: 29 branches to this algorithm are given in the table below; no-one we spoke to at selection days could say exactly how it worked although they had a good intuitive feel for how the patterns of scores were likely to lead to different outcomes. This algorithm has been created from working though the GP National Recruitment Office Excel sheet that is used to convert the competency scores into an initial outcome.

For each of the 4 competencies, the mean score is calculated. The number of competencies with a mean score of 3 or greater is the major factor in determining the initial outcome. However also involved are the number of scores of 1 and any concerns raised by the assessors. This initial outcome usually moves up with an SJT band 4, stays the same with band 3, and moves down with SJT bands 1 or 2 (note that no-one with band 1 should have passed Stage 2). Sometimes whether there are 3 stations with 3 or 4 scores of 3 or 4 also affects the outcome.

- If all 4 competencies have a mean of 3 or more, and there are no scores of 1 or assessor concerns, then an SJT band of 3 or 4, means 'demonstrated', and a band of 2 means 'review likely'. If there are any scores of 1, then SJT band 4 is still 'demonstrated', but 3 means 'review likely' and band 2 means 'review unclear'. If there are any assessor concerns, then (irrespective of the number of scores of 1), SJT band 4 means 'review likely', band 3= 'review unclear' and band 2= 'not demonstrated'.

- With 3 competencies with a mean of 3 or more, no concerns, and not more than 1 score of 1, SJT band 4 means 'demonstrated', 3 = 'review likely' and 2= 'review unclear'. With one or more concerns or 2 or more scores of 1, the SJT band 4 = 'review likely', 3= 'review unclear' and 2= 'not demonstrated' (unless there are 3 stations with 3 or 4 scores of 3 or 4, in which case band 2 = 'review unclear'). If there are one or more concerns and 2 or more scores of 1, then SJT band 4 = 'review unclear' and everything else is 'not demonstrated'.

- With 2 competencies with a mean of 3 or more and no assessor concerns, SJT band 4 = 'review likely', 3= 'review unclear' and 2 = 'not demonstrated' (unless there are 3 stations with 3 or 4 scores of 3 or 4, in which case band 2 = 'review unclear').

» *Table 2.4: Stage 3 Algorithm.*

Row£	Competences scored 3 or 4	Scores of 1	Concerns	Pre-SJT outcome	SJT band$	3S*	Initial Outcome
21	4	0	0	1	3, 4		1
21	4	0	0	1	1, 2		2
22	4	1+	0	2	4		1
22	4	1+	0	2	3		2
22	4	1+	0	2	1, 2		3
25	3	0, 1	0	2	4		1
25	3	0, 1	0	2	3		2
25	3	0, 1	0	2	1, 2		3
23/24~	4	0 or 1+	1+	3	4		2
23/24~	4	0 or 1+	1+	3	3		3
23/24~	4	0 or 1+	1+	3	1, 2		4
26	3	0, 1	1+	3	4		2
26	3	0, 1	1+	3	3		3
26	3	0, 1	1	3	1, 2		4
26	3	0, 1	2+	3	1, 2	3	N/A#
27	3	2+	0	3	4		2
27	3	2+	0	3	3		3
27	3	2+	0	3	2	not3	4
27	3	2+	0	3	2	3	3
28	3	2+	1+	4	4		3
28	3	2+	1+	4	1, 2, 3		4
29	0, 1	-	-	4	-	not3	3
29	0, 1	-	-	4	-	3	4
30	2	-	0	3	4		2
30	2	-	0	3	3		3
30	2	-	0	3	2	not3	4
30	2	-	0	3	2	3	3
31	2	-	1+	4	4		3
31	2	-	1+	4	1, 2, 3		4

Competences: 4=strong evidence, 3= satisfactory, 2= limited, and 1=little evidence. Outcomes: 4= not demonstrated, 3=review unclear, 2=review likely, 1= demonstrated

£ Within the Excel spreadsheet.

*Number of stations with 3 or 4 scores of 3 or 4: only used on a few occasions.

~row 23 has no scores of 1, and row 24 has scores of 1+, but have same initial outcome so should be merged.

A42 should have F16>0 for concerns, and should be 'not demonstrated'.

$No-one with SJT band score of 1 should be at Stage 3, so we don't know why it is included.

- With 1 or 0 competencies with a mean of 3 or more, it is 'not demonstrated' (unless there are 3 stations with 3 or 4 scores of 3 or 4, in which case all SJT bands= 'review unclear').

The thrust of this scoring system seems to be that competencies and SJT scores are most important, but there are three further determinants of outcomes:

1. Scores of 1 are so poor they reduce the chance of being accepted

2. Assessor concerns also reduce the chance of being accepted

3. If a very poor station (or hawkish assessor) pulls down the competency scores in one station, a candidate's chance of being accepted is increased if their scores on all other stations are high (i.e. at least 3/4 on all competencies) with high scores.

Candidates who are assessed as 'review likely' or 'review unclear' are discussed at a moderation session when all assessors who have worked together meet at the end of the half-day session. The session is led by a moderator; all of the marks of the assessors on the candidate are displayed. Moderators had slightly different approaches, but always asked the assessors to say in turn what behaviours they have observed and/ or why they had made their judgements; the moderator then makes an overall judgement. There was no discussion of the final judgement. Assessors are not aware of other assessors' judgements except for the candidates who are being moderated. The focus is on competencies with means of less than three. One moderator always asked the assessor who gave the lowest marks to explain their reason first. Moderators differed in how much agreement they sought before deciding. The candidate has to be deemed to pass all competencies, but seemed to be given the benefit of the doubt if they had high scores elsewhere. Also, when asked to justify marks, some assessors gave general comments i.e. they were not focussed on competency judgements.

The impact of the moderation is unclear, and it would benefit from both statistical and sociolinguistic analyses. Superficially it allows the words and the behaviours of candidates to be discussed properly, rather than just numerical ratings, and that is admirable. A null hypothesis would be that the process contributes little beyond what was already contributed by the candidate's overall mark; a negative hypothesis would be that the process of moderation exaggerates the importance of one or two utterances of a candidate, which are viewed as positive or negative indicators, and so makes the system less reliable. Whether a moderated decision has greater predictive validity than a decision based on the mean mark is the key question; ideally this would require the relationship of the various marks to a later outcome (and MRCGP marks would be the obvious ones). However, those who fail do not go onto GP training and we have not explored this further.

2.3.2 How the final score is calculated

Table 2.3 shows the competencies that are assessed at each station. Each individual mark is in the range 1 to 4, and there are 13 marks (Y in Table 2.3), which give a total in the range 13 to 52. In 2015, this total score was used to rank candidates.

ES has four marks and its total is in the range 4 to 16, whereas the raw totals for CS, CT&PS and PI are in the range 3 to 12. Prior to 2015, they were therefore rescaled to also be in the range 4 to 16. As a result, scores of 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12, become rescaled marks of 4, 5, 7, 8, 9, 11, 12, 13, 15 and 16. Note that the latter is non-linear with steps of one, except between 5 and 6, 7 and 8, and 10 and 11, which have a step size of 2 (from 5 to 7, 9 to 11 and 13 to 15). It is not clear what the rationale was for this non-linear rescaling . The grand total for the Stage 3 stations was then in the range 16 to 64.

Prior to 2015, the final score at the end of Stage 3 included the band scores for CPST and SJT and was calculated as the grand total for the Stage 3 stations (range 16 to 64) + 4 times the SJT Band score (range 4 to 16) + 4 times the CPST Band score (range 4 to 16), giving a final score at the end of Stage 3 in the range 24 to 96. This final score was not used to decide whether a candidate was offered a place, although it is used to determine the subsequent allocation of posts to those who are. Note that prior to 2015 Stage 2 was worth 32/96 = 33%, but in 2015 it was not included in the final score, except as a tie-break to create a unique rank in Oriel for offering purposes[16].

[16] Kelly Chambers, personal communication, 5/01/2016.

In 2014, the lowest Stage 3 score that led to being offered a training place was 60/96 (the written test competency scores were 1, 1, 1 and 2 respectively: clearly at moderation, it was decided this should not override the scores of 3 or 4 for the scenarios), the highest score of a rejected candidate was 79.

7.3 SUMMARY

Candidates pass through three Selection Stages, Stage 1 (Administrative Checks), Stage 2 (CPST and SJT), and Stage 3 (Selection Centre), and may fail (be rejected) at each stage. Between 2009 and 2015, Round 1 applications for UK graduates rose from 3,503 in 2009 to 4,318 in 2013, and then down to 3,696 in 2015; non-UK graduate applications have halved in that time. Rates of offers being accepted has remained around 61% for UK graduates, with about 22% failing at Stages 1, 2 and 3, and 18% withdrawing. For non-UK graduates, the acceptance rate is much lower at about 24%, and perhaps falling; about 71% fail at Stages 1, 2 and 3 and just 6% withdraw.

The process of developing and testing Stage 2 questions has been outlined. CPST and SJT scores are put into 4 Bands; a candidate in either Band 1 is not invited to Stage 3. The Band 2 threshold was raised in 2011 meaning about 10% of candidates were excluded from Stage 3, instead of the previous 5%. However, in 2014 Round 1, five percent of candidates (334/6,688) met the criterion of both CPST and SJT Band 2 or higher but were not invited to Stage 3.

Stage 3 face validity was thought to be very high. There are four stations in Stage 3: three simulations/ scenarios that we view as OSCE-style role plays and a written exercise. One assessor at each station awards marks for 3 or 4 competencies on a 1 to 4 scale (equating to something like: little, limited, satisfactory and strong evidence). These scores and the SJT band are converted to outcomes in a complex way: we have traced 29 branches of this algorithm. For each of the four competencies, the mean score is calculated. The number of competencies with a mean score of 3 or greater is the major factor in determining the initial outcome. However, the number of scores of 1 and any concerns raised by the assessors are also involved in the decision making process. This initial outcome usually moves up with an SJT Band 4, stays the same with Band 3, and moves down with SJT Band 2. Sometimes whether there are 3 stations with 3 or 4 scores of 3 or 4 also affects the outcome.

This algorithm leads to one of the following outcomes: 1= demonstrated i.e. offered a post; 2= Review likely; 3= Review unclear; and 4=Not demonstrated i.e. rejected. Those with 'Review likely' or 'Review unclear' are discussed at a moderation session, with 'Review likely' more likely to be offered a post; so the final outcome is either 'Demonstrated' or 'Not demonstrated'. In 2014 Round 1, 874/1,173 (75%) candidates reviewed at moderation were offered places. It may be better to use a total score and no moderation instead of the current complex algorithm followed by moderation: some possibilities are modelled in Chapter 7. Chapter 3 considers the distinctiveness of the four competencies and four stations which has important implications for this issue. We discuss possible changes in the discussion (Chapter 11).

~ This page is intentionally left blank ~

Appendix 2.2

Examples of Behavioural Indicators for Target Competencies

Appendix 2.1

Examples of Behavioural Indicators for Target Competencies

Downloaded from: http://www.gp-elearning.com/downloads/Competencies_2012.pdf (08/12/2015).

EMPATHY & SENSITIVITY

Capacity and motivation to take in patient/colleague perspective, and sense associated feelings. Generates safe/understanding atmosphere. The scoring scheme for decision-making is complex.

Negative Indicators:

- showed very little visible interest/understanding
- was quick to judge, make assumptions
- appeared isolated or authoritarian
- lacked warmth in voice/manner; failed to encourage
- created uncomfortable atmosphere

Positive Indicators:

- responded to needs/concerns with interest/understanding
- acted in open, non-judgemental manner
- was co-operative/inclusive in approach
- spoke and behaved with warmth and encouragement
- generated safe / trusting atmospheres

COMMUNICATION SKILLS

Capacity to adjust behaviour & language (written/spoken) as appropriate to needs of differing situations. Actively and clearly engages patient and colleague in equal/open dialogue.

Negative Indicators:

- restricted dialogue by overuse of closed questions
- was unable to adapt language/behaviour as needed
- was often unclear when contributing ideas/questions
- failed to engage patient/colleague at non-verbal level
- use of language too functional/narrow/inflexible

Positive Indicators:

- where possible used open, patient-centred questions
- adjusted style of questioning/response as appropriate
- was able to express ideas clearly (written/spoken)
- used effective non-verbal behaviour (voice, posture etc)
- used inventive language (humour/analogy etc)

CONCEPTUAL THINKING & PROBLEM SOLVING

Capacity to think/see beyond the obvious, with analytical but flexible mind. Maximises information and time efficiently and creatively.

Negative Indicators:

- made immediate assumptions about problems
- dealt with issues narrowly or dogmatically
- was unable to suggest "workable" outcomes
- was disorganised/unsystematic
- focused on non-important/peripheral issues

Positive Indicators:

- attempted to think "around" issues
- was open to new ideas/possibilities
- generated functional solutions
- prioritised information/time well
- was able to identify key points

PROFESSIONAL INTEGRITY

Capacity and motivation to take responsibility for own actions (and thus mistakes). Respects/defends contribution & needs of all.

Negative Indicators:

- lacked sufficient respect for others
- treated issues as problems rather than challenges
- avoided taking responsibility for poor decisions/ideas
- showed more concern for some than others
- was tentative when explaining decisions/actions

Positive Indicators:

- demonstrated respect for patient/colleague
- was positive/enthusiastic when dealing with problems
- was able to admit mistakes/learn from them
- •was committed to equality of care for all
- backed own judgement appropriately

RATING SCALE

For ALL exercises:

4. Strong display of specified positive behavioural indicators [and possibly others].
   Few negative indicators displayed, and these considered minor in impact.

3. Satisfactory display of specified positive behavioural indicators.
   Some negative indicators displayed, but none causing concern.

2. Limited number of specified positive behavioural indicators displayed.
   Many negative indicators displayed, one or more causing concern.

1. Little evidence of specified positive behavioural indicators.
   Mostly negative indicators displayed, one or more decisively.

~ This page is intentionally left blank ~

Chapter 3
Reliability and generalizability

Chapter 3.

Reliability and Generalizability

3.1 INTRODUCTION

Measures used in high-stakes assessments should be reliable. Educationalists require that measures are reliable (or more accurately, generalizable) so that when it is decided that a candidate has passed or failed then that is a decision in which examiners can be confident. No measures are however perfectly reliable, so that there are always false positives and false negatives; a false positive means that an incompetent candidate is passed, and a false negative that a competent candidate is failed. False positives and negatives have consequences for individual candidates but also for the system of which they are a part. An incompetent candidate who falsely passes may go on to cause damage to patients, whereas a competent candidate who fails may be lost to the specialty or even the profession, and cannot contribute their skills to the health system.

Although usually couched in terms of the consequences for individuals, low reliability also has economic consequences. Incompetent candidates who pass impose a range of costs on a health system, such as requiring extra training, dropping out of training, and providing inappropriate treatments to patients, perhaps involving further treatment costs, legal costs, and so on. Competent candidates who fail are lost to the specialty, and prior training costs until that point are lost. As a selection measure becomes less reliable so the average competence level of those passing will fall, producing a less competent cohort of trainees and doctors, which presumably has overall consequences for the costs of a health service, and a reduction in positive health-related outcomes for patients. An ideal economic analysis of selection would therefore take the reliability of selection instruments into account. In the extreme, a costly but completely unreliable selection method costs money but contributes nothing beyond a lottery. These issues will be considered in Chapter 6.

This chapter will concentrate on assessing the reliability and generalizability of the assessments in Stages 2 and 3 of GP selection using a variety of techniques. In advance, it is anticipated that Stage 2, which consists of machine-marked multiple-choice assessments should have relatively high reliabilities since most tests of reasonable length do have acceptable reliability, and Stage 2 of the GP tests are of reasonable length. Stage 3 is more likely to have low reliability since it is a relatively brief OSCE-type assessment, with a candidate seeing three simulators and carrying out one open-ended written assessment. Although we have not investigated these, it is likely that all (or almost all) other specialty selection systems will face this same challenge. Stage 3 is also more expensive to run than the machine-marked Stage 2 (see Chapter 5), and therefore its cost-effectiveness needs careful scrutiny.

We really want to know if those accepting offers at selection will succeed in training: this is a question of predictive validity, which will be addressed in Chapter 4. As predictive validity depends upon reliability, we must first consider reliability in detail.

3.2 WHAT IS ALREADY KNOWN ABOUT THE RELIABILITY OF STAGES 2 AND 3?

3.2.1 Stage 2

<u>The reliability of Stage 2 CPST and SJT:</u>

The reliabilities of the Stage 2 SJT and CPST are reported in several places and are summarized in Table 3.1. In all cases reliability seems to have been calculated using Cronbach's alpha.

» *Table 3.1: Estimated reliabilities from published and other sources for Stage 2 CPS and SJT.*

Year and source	CPST			SJT			CPST-SJT correlation
	Test details	Alpha	Mean (SD)	Test details	Alpha	Mean (SD)	
2006 (Patterson et al., 2009a)	"Around 100 items"; 90 minutes	**0.89**		Pilot. 50 items	**0.80 to 0.83**		0.39
2007 (Lievens and Patterson, 2011, Patterson et al., 2013) SJT also reported elsewhere (Carr et al., 2009)	98 items, 90 minutes	**0.88**		50 items; 90 minutes	**0.87/ 0.88**		0.50
2008 (Patterson et al., 2009b). n=8195. SJT also reported elsewhere (Carr et al., 2009)	100 items	**0.91**		50 items	**0.83**		0.51
2009 (Patterson et al., 2009b) (Carr et al., 2009) [*] n=8195	100 items	**0.86**	[77.60*] [(10.51*]	50 items	**0.85**	[644.5*] [46.1*)]	0.53 [0.537*]
2010 (National Recruitment, 2010) [*] (Work Psychology group, 2011)	105 items (86 scored); 90 minutes	**0.84**	69.2 (9.5)	60 items (50 scored); 120 minutes	**0.80**	643.9 (35.7)	0.56 [0.557*]
2011 [*](Work Psychology group , 2012)	-	**0.88**	69.1 (11.0)	-	**0.80**	641.0 (31.9)	0.58 to 0.66 [0.603*]
2012 [*](Work Psychology group , 2013)	-	**0.88**	66.9 (10.4)	-	**0.80**	627.1 (31.9)	0.54 to 0.64 [0.577*]
2013 [*](Work Psychology group, 2014)	-	**0.88**	67.7 (10.2)	-	**0.81**	659.9 (34.6)	0.52 to 0.62 [0.569*]
2014 [*](Work Psychology group, 2015b)	-	**0.88**	68.3 (10.1)		**0.80**	663.9 (34.0)	0.52 to 0.58 [0.547*]
2015 (Work Psychology group , 2015a)	-	**0.87**	67.3 (9.80)		**0.80**	664.7 (33.48)	0.52 to 0.65

[*] based on data from the present study (see below).

Taken overall, the median quoted reliability for CPST is 0.88 (mean = 0.88) compared with a median quoted reliability for the SJT of 0.81 (mean = 0.82).

These findings suggest that the CPST is more reliable than the SJT, but it is longer (about 90 scored items vs 50). For a 90 item test with median reliability of 0.88, the Spearman-Brown formula would predict that a 50 item test would have a reliability of 0.80, which is a little less than the SJT. Item-for-item the SJT may therefore be more reliable.

The correlation of CPST and SJT had a median value of 0.56 (mean=0.55) from 2006 to 2009. Disattenuating the correlation in the individual years (i.e. looking at the true underlying correlation after taking account of the lack of perfect reliability of each of the measures) gave a median value of 0.67 (mean=0.64). This means that about 45% of the true latent variance in one measure is being assessed by the other. The remaining 55% of true variance suggests that the CPST and SJT are indeed measuring substantively different things, and therefore might have different predictive value.

In describing the reliabilities of the Stage 2 tests it should be emphasized that we have not seen raw data for these assessments, and are analyses are mostly based on summaries of various documents, and in particular a useful 2015 summary.

Later in this document we assess alternate-forms reliability of the CPST and SJT using a different methodology and compare the values with those for Cronbach's Alpha reported in Table 3.1.

3.2.2 Stage 3

For this part of our analysis, we are entirely concerned with the performance of the Selection Centre component of Stage 3, and when we refer to the **Stage 3 total score**, we are referring to the total score from the Selection Centre alone. The outcome of the whole of Stage 3 will be referred to in terms of the **Stage 3 final score**, and the **Stage 3 decision**.

Stage 3 for the selection rounds of 2009 and 2010 had a different structure with only three stations, one of which was a group exercise. Since 2011 the structure has been stable, and it is that format which we will be analyzing here.

Reliability of Stage 3:

The reliability of Stage 3 is not clearly stated in the various sets of documentation, either in terms of the methods used for calculating it, or the values found. We have been provided with raw data for Stage 3 from 2009 to 2015, and will therefore calculate reliabilities and generalizabilities using various approaches. The reliability of Stage 3 is of particular importance because Stage 3 is necessarily an expensive process involving many candidates, assessors and simulators at special centres. We should emphasize that when we speak generically of 'reliability' we are including the various types of reliability coefficients, as well as the different types of generalizability coefficient.

Information about Stage 3 of selection for the 2010, 2011 and 2012 selection rounds is available in executive summaries to the NRO (Patterson and Empey, 2012a, b, c), although no information is provided on overall reliability[1]. The Stage 3 selection process has however been described recently (Patterson et al., 2013), albeit for the 2007 round of GP recruitment, when the format of the selection centre was somewhat different to the current selection process. Further information on the 2007 round is also available in a separate 2011 paper (Lievens and Patterson, 2011). For its sample 1 (2007 recruitment round)

[1] There are comments about internal reliability of the various SC exercises. Thus for the 2010 exercise it is said that, "Scores for all three SC exercises (total and case-by-case) demonstrated good to high internal consistency (Cronbach's alpha scores of .88 to .92), which suggests the underlying constructs within each exercise are correlated with [one] another" (p.2). A similar claim is made for 2011 for the four exercises ("alpha scores of .84 to .91"), and 2012 ("alpha scores of .76 to .92"). These appear to be reliabilities between the three or four separate scales within each exercise (ES, CS, CT&PS and PI) and since these are non-independent, all being made by the same judge, it is hardly surprising that they are correlated and hence alpha is high. The judgements are highly correlated, and it is unclear whether the constructs are in fact independent to mean that there are actually four separable scores. What is meant by "total" in the phrase "total and case-by-case" is very unclear indeed. It may be alpha based on the totals of all 3 or 4 simulation exercises, but if so it should have been reported separately. Assuming that it is indeed that, and the lowest values are the most likely to be that alpha, then alpha values of .88, .84 and .76 may be for the independent scores at each simulation, giving an average alpha of .83. It would however then be strange that the alphas based on four simulations are lower than the one based on only three simulations.

the 2013 paper says that **Cronbach's alpha for the selection centre was 0.87** (p. e737). No estimate of reliability is provided in the 2011 paper.

Information on the reliability of an earlier selection centre is available in a 2009 paper on the 2006 selection round (Patterson et al., 2009a), which says that there was a reliability of **alpha = 0.89** (p.52), presumably for "the mean scores across three simulation exercises" (p.52). The selection centre is not described further but is said to be that described in 2005 (Patterson et al., 2005), which had three components (simulated consultation, group exercise, and a written exercise), but for which no formal reliability statistic was provided, it merely being said that, "We found good internal reliability for the new selection system for recruiting general practice registrars". Neither paper gives further information of how reliability was calculated except to say that it is Cronbach's alpha.

Taken overall the situation concerning the published reliability of the Stage 3 selection centre is rather unsatisfactory, as only two clear statements for reliability estimates are available (0.87 and 0.89), and both are for earlier versions of the selection centre.

Wakeford (2014) has been skeptical about the recently claimed reliability of 0.87, saying,

> "those of us who are responsible for devising OSCE assessments would be grateful to learn how one with three stations, single marked, can be devised such that its reliability (Cronbach's alpha) is 0.87, when considerably longer similar assessments give far lower reliability assessments" (p.71).

As examples of other exams he cites MRCGP CSA with a reliability of 0.77 based on 13 stations, iMRCS Part B with a reliability of 0.68-0.72 and 0.76-0.78 based on 8 stations and 10 stations respectively. Wakeford, via an FoI request, obtained raw data from one Deanery for the 2011 selection round, and for the four components of the Stage 3 selection centre calculated **alpha = 0.62** (Wakeford, 2014).

Taken overall it is clear that there is considerable uncertainty about the reliability of Stage 3, with the Wakeford estimates seeming to be incompatible with those of Patterson et al. A further problem is that all of these reliability estimates use Cronbach's alpha, which is generally not regarded as suitable for an OSCE-type assessment, where generalizability should be calculated. The difference between **reliability** and **generalizability** will be considered in more detail in the next section, where **alternate forms reliability (coefficient of stability and equivalence)** will also be introduced.

Finally, we will mention in passing that Jolly and Wakeford (personal communication to ICM, 21st June 2012) have submitted the Wakeford FoI data to a generalizability analysis, and based on a G study of the three simulator stations, in a D study estimated the generalizability coefficient (G^2) to be **0.63** for three stations or **0.64** were there to be four stations. Later we show that that G estimate is compatible with some of our analyses of the 2011 to 2015 data.

There is only one systematic review of reliabilities and generalizabilities in OSCEs (Brannick et al., 2011), although probably not all of the figures reported can completely be trusted[3]. For 18 OSCEs with 2 to 5 stations for which alpha was reported, the median was **0.60** (range 0.19 to 0.85; mean=0.55, SD=0.19), and for ten OSCEs analysed using Generalizability Theory, **the median generalizability coefficient was 0.19** (range 0 to 0.63, mean=0.26, SD=0.21). The median alpha is compatible with that of Wakeford, and even the highest alpha is less than that reported by Patterson et al. The Jolly and Wakeford G is somewhat

[2] Strictly the generalizability coefficient, shown here as G, is symbolised as Ep2 (the expectation of rho squared), but that is sufficiently off-putting to mean that G is probably more straightforward.

[3] For instance an assessment with only two stations seems unlikely to have a Cronbach's alpha of about .84 as reported in figure 9.2 of the paper. Most likely is that repeated measures by judges have erroneously been treated as independent, or there was some other form of non-independence. Cronbach's alpha is often misused, but that is much rarer with Generalizability since it is harder to run, not being available in SPSS, and those doing it tend to know what assumptions they are making. The generalizabilities are therefore more trustworthy.

3.3 RELIABILITY, GENERALIZABILITY, PRECISION AND THE STANDARD ERROR OF MEASUREMENT

3.3.1 Reliability

Reliability (Meyer, 2010, Webb et al., 2007) is a fundamental concept in psychometrics, and in measurement in general. A measure is reliable if it is accurate and if measured again will give the same result or outcome. With a ruler or a stopwatch, that is straightforward, and a physicist may say that the length of an object is 483 mm ± 1 mm, or that a time interval is 40.3 seconds, ± 0.2 seconds. In so far as exactly the same answer is not obtained on each occasion, the measurement procedure is not perfectly reliable, and an estimate of the amount of unreliability or uncertainty is contained within the expressions ± 1 mm and ± .2 seconds.

An equivalent process occurs for psychometric measures. Brennan (Brennan, 2001c) said that,

> "Reliability, broadly conceived, involves quantifying the consistencies and/or inconsistencies in examinee scores. [As a result] …replications in some sense are necessary to estimate reliability". (p.295)[4].

The question of what counts as a replication is not straightforward, and Brennan quoted the AERA/APA/NCME Standards for Educational and Psychological Testing (American Educational Research et al., 1999) which said that,

> "The ideal approach to the study of reliability entails **independent replication of the entire measurement process**" (our emphasis).

The 2014 AERA/APA/NCME Standards (American Educational Research et al., 2014) are much more explicit, and are worth quoting more extensively:

> "For most testing programs, scores are expected to generalize over alternate forms of the test, occasions (within some period), testing contexts, and raters (if judgement is required in scoring). To the extent that the impact of any of these sources is expected to be substantial, the variability should be estimated in some way. It is not necessary that the different sources of variance be estimated separately. The overall reliability/ precision, given error variance due to the sampling of forms, occasions, and raters, can be estimated through **a test-retest study involving different forms administered on different occasions and scored by different raters.**" (p.37; emphasis added)

Despite the growing recognition of the importance of different forms on different occasions with different raters, Brennan (2001c) also said that,

> "… it is still relatively rare for reliability to be estimated using two full-length operational forms" (p.309).

The current analysis, as well as using conventional methods of assessing reliability, will also describe a method for estimating reliability from candidates taking independent, full-length versions of the Stage 2 and Stage 3 assessments.

3.3.2 Types of reliability coefficient

There are three different types of reliability coefficient that have been proposed in the psychometric literature.

Test-retest reliability (coefficient of stability) is the simplest way to approach reliability. A test is administered on one occasion and then the same test is administered again on a second occasion, after some time interval, to the same participants. If a

[4] As a result, Brennan says that,

> "… in order to understand reliability and meaningfully interpret any estimate of it, an investigator must have a clear answer to the following question: (1) What are the intended (possibly) idealized replications of the measurement procedure? [To answer this question] … a second question must be considered also: (2) What is the data collection design used to estimate reliability? … The first question is concerned with defining the parameter(s), and the second question relates to estimation." (p.296, emphases added).

measurement procedure is reliable then it should give broadly similar results when used today as when used as a measure on the same individuals or objects on a previous occasion.

Alternate-form reliability (coefficient of stability and equivalence). Often different versions of a measuring instrument are available (different rulers, different clocks, different tests, or whatever), and these different versions should give similar results. Two variant forms of a 100-item SJT may have **different items** but be designed to measure the same thing. If SJT-A is used on one occasion, and then, after a time interval, SJT-B is used, then **alternate forms** are being administered. If the same subjects are tested on the two versions on different occasions, then both the stability and the equivalence of the two tests are being assessed. The correlation between the scores on SJT-A and SJT-B is known as the **coefficient of stability and equivalence**, or the **alternate-form reliability**. Alternate forms are usually administered after a delay, although one could be given almost immediately after the other (when a coefficient of equivalence would be calculated). The longer the delay then the lower the likely correlation as other factors may also have changed during the interval[5].

Alternate forms matter because they contain different items. A candidate might then do better or less well because they were lucky or unlucky in the particular choice of items in their test. The set of items in the test can be seen as a **random** selection from the universe of possible items that could have been included in the test. In so far as a coefficient of equivalence is different from a coefficient of stability with the same time interval, then candidates are indeed being lucky or unlucky. The role of luck in an assessment is relatively less as the test gets longer. If a person is asked a single question and does or doesn't know the answer then they may just have been lucky or unlucky, but with ten questions the role of luck is far less, and with a hundred questions it is less still. Alternate-forms with 100 or more items should therefore be fairly similar in the outcomes if the tests are well constructed[6]. As a result most knowledge assessments using multiple-choice questions tend to have a hundred or more items, and luck plays relatively little role (although it can never be eliminated). In contrast OSCE-type assessments usually have relatively small numbers of stations, cases or scenarios, typically twenty or fewer, and in the case of Stage 3, either three or four. Luck is therefore expected to play a greater role in OSCE-type assessments, and that process has to be taken into account in statistical analysis of OSCE-type assessments, the most powerful method being with generalizability theory.

Internal reliability is in some ways the most difficult of the types of reliability to understand, despite it being the most frequently reported reliability coefficient. To estimate test-retest reliability and alternate-forms reliability requires candidates to be assessed on two occasions. **Internal reliability attempts to calculate reliability merely from the items in a single administration of a single test.** That of course looks very attractive in practical terms. Conceptually one can consider a 100-item SJT as being comprised of two 50-item SJTs being administered immediately one after the other, or even as two interleaved 50-item SJTs (say, the odd numbered items being SJT-A and the even-numbered items being SJT-B). Scores can be calculated for the two subtests, and the correlation between them found. The correlation between those two 50-item tests can then be scaled up to calculate the likely correlation for a 100-item test[7]. It may be that odd- versus even-numbered items is not the best way of dividing the 100 items into two sets of 50. There is in practice a near infinite number of ways of dividing 100 into two sets of 50 and the much used coefficient of internal reliability known as **Cronbach's alpha** has the nice feature that it is the average reliability which would be obtained for all possible ways of comparing half of the items with the other half.

Although it is very tempting, internal reliability, test-retest reliability and alternate-forms reliability cannot be treated as equivalent[8]. Internal reliability is almost always higher than test-retest reliability, which in turn is higher than alternate forms reliability (Brennan, 2001d, p.129). Internal reliability may be easier to calculate, but it is really alternate-forms reliability which is needed in order to be able to generalize the results of an assessment.

[5] Brennan (2001c) Manual for mGENOVA, version 2.1. Iowa City, IA: Iowa Testing pPrograms (available at http://www.uiowa.edu/~casma/computer_programs.htm) stresses that the reliability across different time intervals is very informative (p.313).

[6] Not to know the answer to one question might be construed as unfortunate, but not to know the answers to one hundred questions would have to be construed as carelessness or ignorance.

[7] This is done using what is known as the Spearman-Brown formula, which for a test of length m allows the calculation of the reliability of a similar test of length n.

[8] The AERA/APA/NCME Standards are clear that, "Internal consistency, alternate-form, and test-retest coefficients should not be considered as equivalent, as each incorporates a unique definition of measurement error." (AERA/APA/NCME Standards, 2014 p.44)

3.3.3 Judgments made by raters

A key feature of all forms of reliability, particularly internal reliability measures such as Cronbach's alpha, is that the **items have to be statistically independent.** If ten assessors separately rate a person's ability then those ten judgements are statistically independent. However if one person rates a person on ten different attributes then those ten judgements are not independent and cannot be treated as such. That problem occurs with most OSCE-type assessments, such as Stage 3, where a single assessor may rate a candidate on three or four scales, but those judgements cannot be treated as independent. It is possible, but unlikely, that the judgements are independent, but that needs testing, and cannot be assumed. An important consequence is that measures such as Cronbach's alpha cannot (and should not) be used when raters are making multiple judgements of the same candidate. In the specific case of Stage 3, Cronbach's alpha could in principle be used (but with certain caveats, and see below) for the total marks obtained at each station (since these are rated by different assessors) but Cronbach's alpha could not be used for the different judgements made at a particular station, across the set of thirteen judgments made at four stations by four assessors. It should also be said that Cronbach's alpha cannot be used as a measure of the unidimensionality or otherwise of multiple judgements made by single assessors and that question instead needs answering using factor analysis and related techniques.

3.3.4 Fixed and random effects

In a single sitting of a OSCE-type examination, some candidates may encounter a set of four particular scenarios, and receive a single mark for each scenario. It is tempting then to calculate the reliability of the total mark calculated from the separate scores for each of the four encounters using Cronbach's alpha. However an important consideration is that those four scenarios are a random set from a potentially infinite number of sets of four possible scenarios. However Cronbach's alpha only provides an estimate of the reliability for those particular four scenarios. The assumption therefore is that the test items (the scenarios) are fixed. Cronbach's alpha is therefore only a description of how well exactly the same measure with exactly the same set of items would work if repeated immediately. In practice that is not what is wanted in medicine, where one wants to use test results to predict how candidates will perform with a much wider range of items chosen from the domain of clinical practice, and which candidates will meet when encountering patients in the future[9]. Brennan (Brennan, 2001c) emphasizes that,

> "if generalization is intended over occasions, then such coefficients [as Cronbach's alpha] will overestimate reliability. … [I]f data are collected on a single occasion, an estimate of reliability based on such data will almost certainly overestimate reliability when interest is in generalizing over occasions" (p.151; emphases in original).

In all the contexts of interest to postgraduate medical examinations, it is therefore very probable that Cronbach's alpha is overestimating reliability.

Medical assessments, be they MCQ or OSCE-type, wish to generalize to different scenarios or test items administered at a later date, so that both stability and equivalence need to be taken into account; and Cronbach's alpha does neither of these, whereas the alternate-forms reliability (the coefficient of stability and equivalence), does so. It is not always necessary to separate out which of a host of possible factors are contributing to variation in stability and equivalence (as the 2014 Standards(American Educational Research et al., 2014, p.37) makes clear), but it can be useful to separate them out, and instead of reliability one then needs to consider **generalizability**, and in particular a **generalizability analysis**.

3.3.5 The limitations of coefficient alpha

Cronbach, at the end of his life, made very clear that alpha is a coefficient that is much misused and misunderstood within psychology (Cronbach and Shavelson, 2004). Two specific problems arose.

Firstly, alpha is not a good description of a test, relying on too many factors of which the variability of the persons taking the test is the most important. As a result Cronbach said that,

[9] This can be seen in synthetic data set 3 of Brennan (2001d) Manual for urGENOVA: Version 2.1. Iowa City, IA: Iowa Testing Programs (available at http://www.uiowa.edu/~casma/computer_programs.htm). for either there is the same set of four random items crossed with occasion (p x O x I) or there are different sets of four random items nested within occasions (p x I:O).

> "I am convinced that the standard error of measurement ... is the most important single piece of information to report concerning an instrument, and not a [reliability] coefficient." (p.413).

The use of the standard error of measurement (SEM) rather than reliability, particularly in sequential assessments, has been stressed elsewhere by one of us (Tighe et al., 2010), with standard errors often being of more practical importance. Although we restrict ourselves mainly in this report to reliability and generalizability, invariably SEMs can be calculated straightforwardly[10].

Secondly, although, as Cronbach says,

> "The alpha coefficient ... is appropriate enough for objectively scored tests where items can be considered a sample from the domain [as in machine-scored MCQ tests. However,] for instruments that make use of judgments of scorers or raters [as in OSCE-type assessments], a simple p x i design [as assumed by alpha] is inadequate" (p.414).

Generalizability is the appropriate way to approach the sort of data obtained in OSCE-type assessments in which raters make judgements, and alpha should not be used, despite the fact that it frequently is.

Cronbach's alpha in particular provides a reassuringly over-optimistic answer to the question which is central to high-stakes OSCE-type assessments, "On a different occasion and with a different set of questions, or with different raters, would the candidate gain a similar score?".

3.3.6 The problem of restriction of range

Restriction of range frequently occurs in studies of selection, and the problem has been known about for a century or so (see Burt, 1943), and affects both questions of reliability and validity (McManus et al., 2013a). The essential problem for validation is that selection takes place on a much larger population than that of validation. Validation only occurs in those who have been selected, and they usually score much higher than those who have been rejected, and they also have a lower range (variance). Correlations in the selected individuals are therefore much lower than in the population of candidates, but the validation of selection properly requires knowledge of the correlation that would have occurred had candidates of all ability levels entered. A similar problem can occur with resit examinations, where the candidates retaking an assessment have failed the first time and are therefore less good and have a narrower range of marks (McManus, 2012). The problem arises in assessing reliability in the present analyses where candidates taking Stage 2 or Stage 3 in successive years (i.e. they retake because they were not offered a place first time) are typically less good than average candidates.

Restriction of range can be handled by many techniques, but a very powerful technique introduced in recent years treats the issue as a problem of missing data (and it can be argued that the outcomes for rejected candidates are indeed nothing but missing data, with missingness conditional on previous marks obtained). That problem can be handled using the EM algorithm (Wiberg and Sundström, 2009). An important advantage of using the EM algorithm, as opposed to earlier methods such as those of Thorndike, is that multivariate problems can be handled simultaneously, so that all information about candidate performance across a range of measures can be handled at the same time. Multiple imputation can also be used for more complicated analyses.

3.3.7 Generalizability

Generalizability theory (Bloch and Norman, 2012, Brennan, 2001d) is a complex set of techniques developed by Cronbach and his colleagues (Cronbach et al., 1972) which considers measurement error in a wide range of situations of which conventional test-retest reliability and so on are special cases. A fundamental feature of generalizability theory is that it is a general theory of random effects upon performance, using **random** in the statistical sense of being the opposite of **fixed**.

[10] For a reliability coefficient, SEM=SD.√(1-reliability), where SD is the standard deviation of the candidates' marks, and reliability is the coefficient which is being used. Calculation of SEMs in generalizability theory is more complex, but the standard program, GENOVA, gives both lower-case δ (SEM for relative measures) and upper-case Δ (SEM for absolute measures).

OSCE-type assessments differ from multiple-choice exams in that it is not only the items in the test which are random. MCQ exams are machine-marked and there should always be precisely the same actual mark after different markings. That is not the case when a human is making judgements about written material (as in the written test of Stage 3), or when an assessor is observing a candidate interacting with another person and making judgements about the quality of their performance. Examiners and assessors may differ in their stringency, colloquially being referred to as hawks and doves (McManus, 2006), and a candidate may be unlucky if they happen to have hawks marking their performance. Examiners are generally not fixed, in the statistical sense, but instead are random, being drawn from a large set of possible examiners, which in effect can be regarded as infinite[11].

OSCE-type assessments have other random factors as well. An actor may simulate the role of a patient or a professional colleague, and actors can also be hawkish or doveish in the way they play a patient. Actors are also, in effect drawn from a large pool, and candidates may get lucky or unlucky in whom they happen to see. In addition, examination centres may differ, with examiners perhaps having been briefed differently, or conditions being different in other ways. All are random factors which can affect a candidate's final mark.

Those involved in GP selection work hard to reduce these random factors by training assessors and simulators, the design of the scenarios etc: this chapter assesses the effectiveness of these processes. A separate claim is that moderation ameliorates the impact of this unwanted variability. We have not examined this claim statistically, but return to this point, later.

Generalizability theory partitions the total variance in an assessment into different sources (the candidate, the test item, the assessor, the actor, the occasion, etc) and then estimates how reliable the assessment is. Generalizability also distinguishes between two types of decision, those which are **relative** (as in ranking candidates from best to worst on a single test), and those which are **absolute** (comparing candidates across occasions on different tests against a fixed standard). Postgraduate examinations typically make absolute judgements, where each candidate is compared to a fixed standard which should be unvarying across years. In contrast, in competitive assessments there is a particular number of places available to be filled, which should go to the best candidates, making the assessment relative. Whether GP selection is absolute or relative is a moot point and will be returned to. If all places on a program are always filled, albeit sometimes by weaker candidates than on other occasions, then judgements are relative, but if places are left unfilled because too few candidates of the required calibre are available, then the judgements are absolute.

Generalizability produces two different types of coefficient:

Relative measures:

- The generalizability coefficient, strictly symbolized as $\boldsymbol{E}\rho^2$ (the estimate of rho-squared), but more simply referred to as G, is an equivalent of a reliability coefficient for generalizability studies with relative measures.
- The relative error variance, symbolized as $\sigma^2(\delta)$ (sigma squared lower case delta), is the variance of an individual relative measure, and its square root is the equivalent of a standard error of measurement, and will here be called SEM_G .

Absolute measures:

- The **index of dependability**, symbolized as Φ (Phi), is an equivalent of a reliability coefficient for generalizability studies with absolute measures.
- The absolute error variance, symbolized as $\sigma^2(\Delta)$ (sigma squared upper case Delta), is the variance of an individual absolute measure, and its square root is the equivalent of a standard error of measurement, and will here be called SEM_Φ.

[11] Generalizability Theory actually allows finite sizes for the pool of examiners or other characteristics, but in practice unless there is only a small pool then it makes little difference to the calculations.

- Absolute measures also differ from relative measures in that while SEM_G can be regarded as equivalent at all levels of the scale of measurement, from highest to lowest, the absolute standard error of measurement, SEM_G, varies at different levels of attainment. Typically it should be measured at the cut-point which differentiates pass from fail. If a level is not specified then conventionally it is measured at the mean.

- Just as SEM can vary at different levels of performance, so Φ can also be calculated for a particular performance level, λ (lambda), to give a generalizability coefficient, known as Φ_λ. Φ_λ is always lowest at the mean, and is higher as λ gets further from the mean. That may sound counter-intuitive, but for an assessment with an average mark of, say, 50, and a spread of marks from 25 to 75, it easier to say that a mark is reliably different from a mark of 20 than it is to say that a mark is different from a mark of 50[12].

- In general, absolute measures produce lower reliabilities and higher standard errors of measurement than do relative measures.

Whether G, Φ or $\Phi\lambda$ is the appropriate measure of generalizability always needs careful consideration.

3.3.8 G studies and D studies

A final distinction in generalizability theory is between G (Generalizability) studies and D (Decision) studies. In order to understand the different types of variability in assessments it is ideal to have as many sources of variance crossed as is possible (so that, for instance it is better to have several assessors at an encounter rather than one, so that the effects of different assessors on different candidates can be estimated). G studies therefore need not always be representative of the final form of the examination, but are used to estimate **variance components**. Once a G study has been carried out then the variance components can be used in a D study to make decisions about the likely generalizability of an assessment with different numbers of stations, examiners, or whatever. In the analyses below, a G study will be based on those candidates taking Stage 3 in both rounds 1 and 2, and will be used in a D study to estimate the generalizability of the more typical situation in which a candidate only takes the assessment in one round.

It should be clear that generalizability is not straightforward, but it does provide a general coherent method for approaching an analysis of the reliability of the forms of complex assessment found in typical OSCE-type examinations. **It also cannot be emphasized sufficiently that simpler approaches, such as that of Cronbach's alpha, are not appropriate for OSCE-type assessments with multiple assessors and small numbers of randomly chosen situations.**

3.4 ABSOLUTE AND RELATIVE STANDARDS IN GP SELECTION

A difficult aspect of both selection and high-stakes assessments is the extent to which measures are making absolute or relative decisions. In general, postgraduate assessments should set **absolute standards**, where a decision is taken about a particular level of performance or competence, and **each individual candidate** is assessed against that standard, typically using such method as **criterion referencing**. A feature of absolute standards is that in principle all candidates could pass or all candidates could fail the assessment. In contrast, in competitive systems, the intention is not to ask whether individual candidates meet a specific level of competence, but rather to ask "who are the best candidates?", in order that they should be the ones who are appointed. Competitive systems are an extreme case of **relative standards** which use processes such as **norm referencing**, where each candidate is compared to all other candidates. Many systems, although not strictly competitive, also use norm referencing more for administrative convenience and simplicity than because competition is necessary[13].

[12] An interesting variant on this is that it is possible using Φ_λ, as in the example above, to calculate the reliability of an assessment which all the candidates pass.

[13] A difficult example of norm referencing is the current use of deciles in the Educational Performance Measure (EPM) of the Foundation Programme Application System (FPAS), which are calculated within medical schools, and yet are used to compare candidates across medical schools despite there being strong reasons to believe that deciles across medical schools are not equivalent.

The utility of absolute and relative standards differs according to the type of a selection system. If there are more well-qualified candidates than places (i.e. the specialty is **selecting**) then a selection system is relative, looking for the best candidates. However if there are fewer well qualified candidates than places (i.e. the specialty is **recruiting**) then a prime concern is that all candidates accepted reach a minimum level of competence which is necessary for appropriate practice and patient safety, and then absolute criteria are necessary.

Whether the GP selection is making absolute or relative judgements is **not entirely clear** to us. The judgments in Stage 3 have the appearance of being absolute (or at least trying to be absolute), with a clear attempt at examiners being trained and calibrated against absolute standards (Anonymous, 2014a), and candidates described in terms of having **demonstrated or not demonstrated** competence and hence suitability for training. That clearly seems to be criterion referencing.

The status of Stage 2 is though far less clear. The nature of standard setting for Stage 2 is important because the threshold for going through to Stage 3 has been raised, with twice as many candidates being excluded from Stage 3 in 2014 as in 2011. **The nature of the standard setting for Stage 2 therefore needs further exploration and justification.** In Chapter 2 we outlined how CPST and SJT scores are placed into bands, which decides whether or not a candidate goes on to Stage 3, are used in the final Stage 3 score, and the SJT band influences selection at Stage 3. The average marks in CPST and SJT are currently normalized to a mean of 250 and a standard deviation of 40, without there being statistical equating across years. Therefore, Stage 2 is norm referenced as it is using relative standards. The signature of absolute standards using judgmental or statistical equating across years is the presence of variation across years (Anonymous, 1990), and the relatively small variation across years (since 2013) in mean scores and success rates indicates that norm referencing and relative standards are being used.

In summary, Stage 2 is making relative judgements and the selection centre in Stage 3 is attempting to make absolute judgements; consequentially the final decision based on Stage 3 scores and the moderation process uses a hybrid of relative and absolute judgements as the marks from Stage 2 and Stage 3 are combined. There is an argument that the Stage 2 scores should be made into a more absolute form of assessment, most straightforwardly by using statistical equating of the CPST and SJT across years by using Rasch modelling. That would be particularly important as the GP CPST and SJT are transformed into the Specialty Recruitment Assessment (SRA) which will be used for all specialties which wish to use it (and currently it is being used by GP, Psychiatry and Ophthalmology).

An important statistical implication of absolute and relative standards is that, within generalizability theory, the generalizability statistic G is the appropriate measure of reliability for relative standards, whereas for absolute standards the appropriate measure of reliability generically is Φ (Phi), and for a specific absolute standard is Φ_{λ}. The reliability of absolute standards is always less than for relative standards, since variation occurs not only between candidates but also between assessments, but the reliability of a particular absolute standard can be higher, particularly if it is far from the mean.

3.5 RELIABILITY AND GENERALIZABILITY IN THE 2009 TO 2015 DATA

In this section we will calculate the reliability and generalizability of the Stage 2 and Stage 3 scores using various approaches, including:

1. Cronbach's alpha for Stage 3. Although this approach is not recommended we carry it out for consistency with previous studies.

2. Alternate forms reliability for Stage 2 and Stage 3 by considering candidates taking the assessments on two or more occasions.

3. Generalizability analyses for Stage 3. This is the gold standard for OSCE-type assessments but is complex and the conclusions depend in part on the assumptions that are made.

4. Multi-facet Rasch modelling (MFRM). Although not strictly an analysis of reliability, MFRM, does provide its own estimate of reliability, which needs careful interpretation, but allows comparison between different sets of data analysed using MFRM.

We will also compare our reliability estimates with those provided elsewhere.

3.5.1 The data

We have been provided only with summary data on individual candidates taking Stage 2 during the selection rounds from 2009 to 2015. Many statistical analyses can be carried out with these data, but analyses at the level of individual item responses cannot. For Stage 3 we have been provided with raw, item-level, data for selection from 2009 to 2015 which allows a wider range of re-analyses.

3.5.2 Cronbach's alpha

Stage 2: Cronbach's alpha is available for the CPST and SJT, and is reported in the table earlier. No reliability estimate is available for the total Stage 2 score. The median alphas are about **0.88** for CPST and about **0.81** for the SJT. The alpha reliabilities seem fairly constant from 2009 to 2015.

Stage 3: Despite the many problems with it, analyses of OSCE-type examinations often provide values of Cronbach's alpha, and several values are available for Stage 3 (see above). We calculated Cronbach's alphas for Round 1 of Stage 3 in 2011, 2012, 2013, 2014 and 2015 using the four mean scores of candidates at the three simulations and the written station, all of which are statistically independent. Values of alpha were **0.64, 0.65, 0.62, 0.62** and **0.55** respectively, with an average value of **0.62**. The value of 0.64 for 2011 is very similar to the value of 0.62 reported by Wakeford for a subset of data from one Deanery (Wakeford, 2014) and the average value of 0.62 is reasonably close to the average of 0.55 (median=.60) for the 18 small OSCEs in the systematic review reported above (Brannick et al., 2011). There is no obvious explanation of why the alpha for 2015 seems to be quite a lot lower than those for previous years, but it may be part of a trend which will also be apparent later in the generalizability analyses.

The average alpha value for 2011 to 2015 of 0.62 is substantially below the values of 0.87 and 0.89 described in reports on the GP selection centre (Lievens and Patterson, 2011, Patterson et al., 2013). Without going into details, we note that an alpha calculated for the thirteen individual Stage 3 measures for Round 1 of 2014 (rather than the four independent mean scores) gives a value of 0.85, which is close to .87 and .89, and might explain the origin of those values. It is however an inappropriate way to calculate alpha as the scores are not independent of each other.

Overall there seems little doubt that the alpha levels reported in documents describing the Stage 3 assessments are not an accurate representation of the reliability. The typical phraseology says, "All cases … demonstrated **high internal reliability**" (Work Psychology, 2015a, p.3; emphasis in original). This seems to be describing the association between the sub-scores within a case (e.g. between the scores for ES, CS and CT&PS in station A), but since they are all made by the same assessor it is not surprising that they are highly correlated. **If a single Cronbach's alpha statistic is required for Stage 3, with all the caveats on the inappropriateness of Cronbach's alpha, then it is the average value of 0.62 reported earlier.**

3.5.3 Alternate forms reliability (coefficient of stability and equivalence) in Stages 2 and 3

Internal coefficients of reliability such as alpha cannot be calculated without raw data. However, as suggested earlier, alpha can often provide an overly inflated estimate of reliability, not least because one often wishes to generalize not only over the particular version of the test which was administered (items or stations) but to alternate versions with different but equivalent items or stations administered at a different time.

The **alternative forms reliability (coefficient of stability and equivalence, r^{se})**[14] can be calculated for Stages 2 and 3 because there are candidates who apply for the selection programme in successive years. These candidates take different versions of the CPST and SJT, and different versions of the Stage 3 selection centres at an interval of a year or so. If the traits being

[14] Although Brennan and others refer to the 'coefficient of stability and equivalence', symbolised as rse , the AERA/APA/NCME Standards refers to 'alternate forms reliability'. We will follow the latter in referring to 'alternate forms reliability', but shall follow Brennan in symbolising it as rse since that makes the mathematical derivation clearer.

assessed by the measures are stable and equivalent then candidate scores across the occasions should be strongly correlated. Since the intention of the selection tests is not merely to determine candidate abilities on the day of the test, but to assess candidate abilities into the future, certainly across the three years of a training programme, but potentially ahead into their future working lives, rse is a robust and practical measure of the effectiveness of Stages 2 and 3.

A problem in calculating r^{se} is that candidates retaking stage 2 and particularly stage 3 are only subsets of those taking the assessments on a single occasion, and they will differ in average scores and the range of their scores. The problem of restriction of range will be handled using the EM algorithm (Wiberg and Sundström, 2009), as described earlier. Only a summary of the key results will be presented here. Elsewhere as a separate document, we can provide a fully worked example of the calculations, along with an illustration of the method using synthetic data (McManus et al., 2015).

Data were available for the selection rounds from 2009 to 2015. Although Stage 3 changed in 2011, the total scores can still be used in the analyses across the years[15]. The analysis calculated the correlations between the scores of candidates taking the assessment on two or more separate occasions, and therefore data across years are available for 2009-10, 2010-11, 2011-12, 2012-13, 2013-14 and 2014-15. In each case we will consider all candidates who applied in year N and any candidates who also applied in year N+1, etc., analyzing the total score in Stage 2, as well as the CPST and SJT sub-scores, and the Stage 3 total score i.e. the total marks attained in the Stage 3 selection centre, and not the final Stage 3 score which also incorporated the Stage 2 band scores[16].

Stage 2 assessments are usually only taken once by an individual candidate in a year[17]. Stage 3 assessments are though sometimes taken twice by candidates in both rounds 1 and 2. We have restricted the analyses here to attempts in round 1 since that is when the main selection process occurs.

3.5.4 Alternate forms analysis of the reliability of the Stage 2 total score, the CPS and SJT, the Stage 3 total score, the final decision, and candidate acceptance

The calculation of alternate forms reliability, r_{se}, requires, as Brennan emphasizes, "at least two independent instances of the measurement procedure" (Brennan, 2001c, p.305). The GP selection system has been run in broadly similar form from 2009 to 2015, so that there have been seven selection cycles.

Table 3.2, below, shows the number of candidates in Round 1 of each year taking Stage 2 (in bold), and Stage 3 (in italics). Altogether 34,515 took Stage 2, but in addition 3,911 candidates took the exam in two successive years, 1356 two years apart, 601 three years apart, 243 four years apart, 134 five years apart and 52 six years apart. The equivalent figures for Stage 3 are 30,002, 2,742, 876, 365, 124, 60 and 29. Overall therefore 6,297 candidates had taken Stage 2 on at least two occasions, and 4,196 Stage 3 on at least two occasions. Across the seven years there were 25,122 candidates taking Stage 2 on just one occasion, 3530 on two occasions, 557 on three occasions, 125 on four occasions, 25 on five occasions, 5 on 6 occasions, and one who had taken Stage 2 on all seven occasions[18].

Clearly there are many candidates who have taken the assessments on two or more separate occasions. An obvious problem is that they are probably not representative of the population of candidates as a whole, as has been shown elsewhere for those taking MRCP(UK) on repeated occasions (McManus, 2012), being noteworthy particularly in that they had done less well at their first attempt[19]. Simple raw correlations between the marks at the various attempts are therefore likely to under-estimate

[15] Since undertaking these analyses, we understand that the 2015 Stage 3 total score does not include weighting of competencies; although we have not allowed for this, we doubt it will have an important impact.

[16] In the various spreadsheets, the mark "Stage3score" includes weighted components of both Stage 2 and Stage 3. The mark exists for candidates who do not take Stage 3 but is not useful as they are awarded a mark of 0 for stage 3. Stage3score is therefore only used when Stage 3 has been taken.

[17] An exception was in 2015 when low-scoring candidates could re-take Stage 2. That is potentially problematic since candidates gaining low but passing scores at Stage 2 were not allowed to retake, even though re-taking to get higher scores could have advantaged them. The result was that candidates who failed at Round 1 had higher eventual scores than those who passed at Round 1. Whether that is fair is open to debate.

[18] Those making repeated attempts tended to be non-UK grads and to be non-white, with only a very small interaction between the two effects. There was no relationship to sex.

[19] The mean score at the first attempt at Stage 2 for those taking the exam only once was 514.7, whereas those taking it two, three or four or more attempts had mean first attempt scores of 477.0, 457.1 and 429.6.

» *Table 3.2: Numbers of candidates taking Stage 2 (bold) and Stage 3 (italics) in Round 1 for each combination of years.**

	2009	2010	2011	2012	2013	2014	2015
2009	**5434** 4526						
2010	**814** 567	**5118** 4270					
2011	**298** 209	**546** 337	**4677** 4259				
2012	**185** 127	**282** 141	**583** 455	**5045** 4509			
2013	**90** 53	**138** 66	**281** 196	**660** 543	**5001** 4620		
2014	**76** 40	**86** 33	**143** 84	**292** 196	**634** 466	**4957** 4160	
2015	**52** 29	**58** 20	**67** 38	**135** 88	**203** 134	**674** 374	**4283** 3658

*The values on the diagonal are the total numbers of candidates taking the assessment that year.

the true correlations due to **range restriction**. Range restriction is a recurrent problem in research on selection due to those selected inevitably having higher scores with less variability than those who are not selected. Various solutions have been proposed to the problem, but a powerful one is the method described by Wiberg and Sundström (2009) which treats it as a problem of missing data; in effect the outcome measures for the non-selected group being 'missing'[20]. Wiberg and Sundström use the EM algorithm (Baghaei et al., 2009), which is a standard method for estimating correlations in the presence of missing data (Graham, 2009, Pigott, 2001), and which is described elsewhere in more detail (McManus et al., 2015)[21].

The result of using missing data imputation is that an estimate of the correlation matrix and the means and SDs of the measures can be obtained as if all of the candidates had taken all of the assessments, taking account of restriction of range, differences in mean scores, and so on. The separate document (McManus et al., 2015) provides a simple example using synthetic data where some values are deleted and it is shown that the method can reconstruct the original values. Worked examples are also provided using subsets of the present data.

The full analysis of the GP selection data can estimate the correlations between the various measures for all possible combinations of years, and for a range of variables. In practice the more variables that are included, the better since the algorithm uses all possible information to infer the correlation matrix.

[20] In an ideal world all candidates would take the entire selection test, but in practice that is rarely possible. Wiberg and Sundström in fact described an unusual situation with the Swedish driving test there all of the candidates taking the theory test went on to take the practical assessment. They could only find one previous example of such data in the literature.

[21] Note that the EM algorithm is not strictly a method for imputing (i.e. estimating) actual missing data points (and that is what is done by procedures such as mean substitution or multiple imputation), but instead is a method for estimating an unbiased mean, standard deviations and a correlation matrix in the presence of missing data which are biased in the sense of not being missing completely at random (MCAR).

3.5.5 Alternate forms reliability for Stage 2, Stage 3 and the outcomes of selection

An alternate forms reliability coefficient can be calculated for each of the various pairs of years when candidates took the assessments. Thus there is a coefficient for those taking the exam in 2009 and 2010, another for those taking it in 2009 and 2011, and so on, for all of the possible 21 combinations. For simplicity those coefficients in Table 3.3, below, are averaged over how many years apart are the assessments, and also producing a grand average.

It might be expected that alternate form reliability coefficients would decline somewhat with time, primarily because there is change in the thing that is being measured. Some candidates retaking an examination might therefore acquire more knowledge before taking the assessment on a second occasion and hence do better, and thereby reduce the coefficient. Table 3.3 shows that doesn't actually happen. Consider the Stage 2 total score. The coefficient is 0.75 for those taking the assessment one year apart, and is broadly similar for those taking it up to six years apart. The overall average coefficient of 0.73 is therefore a good estimate of the reliability of the Stage 2 total mark[23]. A similar pattern across time occurs for most of the measures in the table. Two important conclusions can be drawn: i) the measures which are being assessed are probably

» Table 3.3: Alternate forms reliability coefficients estimated across time intervals from 1 to 6 years, along with the overall average.

	Stage 2 Total	Stage 2 CPS	Stage 2 SJT	Stage 3 Total	Stage 3 Score	Stage 3 Offer made	Stage 3 Offer accepted
EM Mean	504.7	252.8[a]	253.3[a]	50.35a	75.4a	79.1%	66.1%
(SD n=5 or 7)[a]	(3.16)	(1.50)	(1.04)	(3.91)	(.907)	(5.28)	(5.55)
EM SD	68.7	38.0	37.91	6.91	10.17	40.8%	47.3%
(SD n=5 or 7)[a]	(4.19)	(.466)	(.754)	(.916)	(.337)	(3.48)	(2.28)
Alternate forms reliability coefficient, r_{se}, across different numbers of years [mean (SD)]							
1 year delay (n=6)	.745	.743	.606	.432	.574	.318	.100
	(.036)	(.052)	(.022)	(.048)	(.065)	(.062)	(.072)
2 years delay (n=5)	.717	.721	.558	.434	.578	.343	.148
	(.073)	(.059)	(.071)	(.056)	(.075)	(.056)	(.135)
3 years delay (n=4)	.730	.732	.549	.446	.578	.358	.171
	(.058)	(.038)	(.068)	(.085)	(.056)	(.095)	(.077)
4 years delay (n=3)	.729	.763	.576	.388	.551	.410	.171
	(.105)	(.124)	(.035)	(.050)	(.112)	(.046)	(.105)
5 years delay (n=2)	.720	.684	.577	.42	.519	.329	.136
	(.032)	(.081)	(.029)	(.039)	(.104)	(.072)	(.017)
6 years delay (n=1)	.755	.732	.560	.469	.512	.380	.041
	(-)	(-)	(-)	(-)	(-)	(-)	(-)
Average[22]	.733	.729	.571	.43	.552	.356	.128
	(.051)	(.059)	(.038)	(.046)	(.069)	(.023)	(.068)

[22] i.e. the average of the averages for the six delays of 1, 2, 3, 4, 5 and 6 years.

[23] Brennan points out that the correlation is in fact the reliability of the test if used on a single occasion, which is what is needed to be found.

stable across time; and ii) the **correlations can be used as acceptable measures of the reliability of each assessment when it is carried out on a single occasion**[24].

In interpreting these values it must be emphasized that these coefficients are for all candidates taking all of the assessments in round 1 of two different years. Thus the coefficient for Stage 3 is not the conventional one which is merely for those taking the assessment but is as if all candidates had taken Stage 3, with the missing data being estimated. In practice such coefficients are more useful as they allow one to model the entire selection process.

A comparison of the average coefficients of stability and equivalence in Table 3.3 shows important results. The highest reliability is for the Stage 2 total score, with a value of 0.733. That value is very similar to the 0.729 for the CPST and higher than the 0.57 for the SJT. Since the total score is the sum of the two measures on standardized scales, it is somewhat surprising that the total score is only a little better than the CPST alone. The SJT is less reliable, and also correlates relatively weakly with CPST (and the value corrected for restriction of range of 0.54 is shown in the lower part of Table 3.4). CPST and SJT when added together are mixing two separate types of variance, and hence the total is not much more reliable than that for CPST.

The reliability of the Stage 3 total (i.e. the marks on just the four stations) has a value of 0.43, which is not very high (and is certainly much, much lower than some of the reliabilities which have been claimed for it – see the earlier discussion – although later it will be compared with various generalizability theory analyses). The low reliability of Stage 3 total means that when it is combined with Stage 2 marks to give Stage 3 score[25], there is still only a lowish reliability of 0.55. That value of 0.55 is a measure of the reliability of the numerical assessment produced at the end of selection. The final stages of selection are an offer being made and that offer being accepted. The reliability of an offer being made or not being made is lower still at 0.36, and the value is lower than that for the Stage 3 total, as the Stage 3 total score is a continuous variable, whereas an offer being made is binary, and binary marks usually have lower reliability than continuous scores[26]. The final stage of selection is that an offer is accepted or not. The reliability of that is very low (0.13), and is barely different from chance. The most likely explanation is that other events happen in candidates' lives, which may affect whether or not they accept an offer, and those events may well be relatively random[27].

Table 3.4 in its upper part shows EM-estimated correlations, within a single year for simplicity, between the various selection variables. These are corrected for restriction of range, and give a good idea of the correlations that are likely within a single selection season. They should however be treated with care as they do not take into account lessened reliability due to alternate forms of the assessments being used across years.

The lower part of Table 3.4 shows the equivalent EM-estimated correlations across years. These correlations are all lower than those in the upper part of the table because as well as taking into account unreliability within an assessment they also take into account the unreliability between alternate, different versions of the same assessment which are used across the years.

[24] There seems to be a paradox in that despite the candidates potentially having revised and therefore improved across occasions there is nevertheless a high correlation. The paradox is in part because not only will failed candidates potentially improve, so also it is probably the case that passing candidates are continuing to improve, acquiring more knowledge and skills. The assessment of reliability is ultimately concerned only with correlations, and hence the ordering of candidates, and not their absolute level, and hence improvement across all those reapplying is not involved in the calculation of the reliability. For a more detailed analysis of how candidates behave when they resit assessments they have previously failed, see McManus and Ludka (2012).

[25] The raw Stage 2 marks are not combined with the Stage 3 total, but instead it is the Stage 2 Band scores which are combined with the Stage 3 marks. The band scores, having only 3 applicable categories, are inevitably less reliable than the raw marks which are continuous, and that is a second reason why the Stage 3 score is less reliable than several of its components. It should also be remembered that the Stage 2 Band scores make up less than half of the Stage 3 total score, so that the more reliable components are weighted less.

[26] The decision to offer a training place depends on the algorithm and moderation described in Chapter 7; that process is much more difficult to model and has not been attempted, here.

[27] Candidates may be made offers in other specialties, they may have to move to other parts of the country because of a partner's career, they may take maternity leave, and so on.

» *Table 3.4:EM-estimated correlations between selection measures, within years (top) and across years (bottom).**

	Stage 2 Total	Stage 2 CPS	Stage 2 SJT	Stage 3 Total	Stage 3 Score	Stage 3 Offer made	Stage 3 Offer accepted
Correlations between selection variables within year, averaged across all of the seven years							
Stage 2 Total	1. (-)	.878 (.010)	.876 (.010)	.505 (.057)	.729 (.077)	.448 (.048)	.339 (.059)
Stage 2 CPST	.878 (.010)	1. (-)	.540 (.034)	.404 (.060)	.620 (.061)	.339 (.048)	.257 (.056)
SJT	.876 (.010)	.540 (.034)	1. (-)	.487 (.050)	.662 (.069)	.452 (.038)	.343 (.048)
Stage 3 Total	.505 (.057)	.404 (.060)	.487 (.050)	1. (-)	.909 (.026)	.734 (.031)	.587 (.052)
Stage 3 Score	.729 (.077)	.620 (.061)	.662 (.069)	.909 (.034)	1. (-)	.744 (.111)	.592 (.113)
Stage 3 Offer made	.448 (.048)	.339 (.048)	.452 (.038)	.734 (.031)	.744 (.111)	1. (-)	.737 (.069)
Stage 3 Offer accepted	.339 (.059)	.257 (.056)	.343 (.048)	.587 (.052)	.592 (.113)	.737 (.069)	1. (-)
Correlations between selection variables across years, averaged across intervals of one to six years. The values in bold across the diagonal are the estimates of alternate forms reliability and are very similar to those in Table 3.3[28]							
Stage 2 Total	**.731 (.057)**	.687 (.058)	.597 (.052)	.447 (.062)	.593 (.054)	.386 (.051)	.228 (.053)
Stage 2 CPST	.687 (.058)	**.732 (.063)**	.475 (.057)	.352 (.057)	.508 (.048)	.289 (.053)	.156 (.049)
SJT	.597 (.052)	.475 (.057)	**.574 (.050)**	.438 (.064)	.534 (.055)	.389 (.053)	.244 (.059)
Stage 3 Total	.447 (.062)	.352 (.057)	.434 (.064)	**.430 (.055)**	.477 (.052)	.372 (.049)	.245 (.067)
Stage 3 Score	.593 (.054)	.508 (.048)	.534 (.055)	.477 (.052)	**.564 (.070)**	.416 (.054)	.258 (.060)
Stage 3 Offer made	.386 (.050)	.289 (.053)	.389 (.053)	.372 (.049)	.416 (.054)	**.349 (.066)**	.222 (.069)
Stage 3 Offer accepted	.228 (.053)	.156 (.049)	.244 (.059)	.245 (.067)	.258 (.060)	.222 (.069)	**.136 (.091)**

*Note that all values apply to the entire population of candidates, and not only those who happened to take Stage 3, etc. The values in brackets approximate a standard error.

[28] Note that the values on the diagonal are slightly different from those presented as the average of all the coefficients of stability and equivalence since here they are averages of all 21 correlations, whereas in the previous table they were the averages of the six averages for the varying delays. The differences are minor.

3.5.6 Generalizability of Stage 3 scores

In carrying out generalizability analyses of the marks from Stage 3, two separate issues need to be considered:

1. Analysis of the scores in terms of relative marks (i.e. using the coefficient G), or in term of absolute marks (using the coefficient Φ) or absolute marks at the particular pass mark (Φ_λ). The latter is what is really required, although many generalizability analyses restrict themselves to G.

2. The problem of 'hidden facets' across occasions. Brennan (2001b, p.149) emphasizes that analyses often have hidden facets, where generalizability should be across occasions, but data are only available for a single occasion. We have therefore carried out analyses firstly for a single occasion (in the conventional way), and then for those candidates taking Stage 3 twice within a year, when there are two occasions. The latter probably gives a better estimate of the true generalizability of the assessment, although the calculations are more complex.

3.5.7 The 'pass mark' for Stage 3

As described in Chapter 2, Stage 3 does not have a pass mark; at the decision-making stage, the Stage 3 scores are combined with SJT scores using a complex algorithm to arrive at an initial decision, which is then converted into a final decision (demonstrated/not demonstrated) during the moderation process. However, calculating Φ_λ, the absolute measure of generalizability at a particular cut-score, requires that a pass mark or cut score is defined. There is also a difficulty that the four Stage 3 sub-scales have different numbers of items contributing to them. The generalizability analysis works on the averaged marks for the four stations (since these are independent), and hence on the average of those scores, although the individual averages are based on three (different) sub-scales for A, B and C and four for W. To get round these problems, an average of the mean A, B, C and W marks was calculated for those passing and those failing Stage 3 overall. That analysis suggests that a mark of about **3.1**[29] is a good estimate of the point above which candidates pass and below which they fail. It is below a typical average for the Stage 3 marks of about 3.3 and that is compatible with about 62% of candidates passing Stage 3.

3.5.8 Generalizability within the first selection round of each year.

Table 3.5 summarises the generalizability analyses for Round 1 of Stage 3 in each year. Rows 1, 2 and 3 show the numbers of candidates, the mean score for each year, and the SD[30] of all scores within a year. There is a clear trend for the number taking the assessment to decrease across years, and there is a suggestion that the mean score is increasing across the years and the SD is decreasing across years.

Variance components:

In generalizability analyses the various facets (e.g. the stations that candidates encounter) can either be **crossed** or **nested** in relation to other facets. If all candidates saw exactly the same four stations (including the same examiners, simulators and scenarios) then Persons[31] (P: the candidates) would be fully crossed with Stations (S), symbolized as SP (i.e. SxP). For some selection centres, Persons are fully crossed with Stations, but that is not the case across the assessment as a whole, as different centres have different examiners and simulators; therefore we have carried out the analysis on the basis that Stations are nested within Persons (S:P)[32].

[29] Generalizability analysis traditionally works on the scale of individual marks and therefore this represents a likely 'pass mark' based on the four-point scale used by assessors from 1 to 4.

[30] This is the simple SD of all of the judgements made, so that if there are 1,000 candidates assessed on four scales it is the SD of all 4,000 judgements that were made.

[31] We follow Brennan in using P for persons, although C for candidates, or D for doctors, is also used in other analyses elsewhere.

[32] We have repeated the analysis using SP rather than S:P and it makes relatively little difference, although the G and Φ coefficients are different.

G studies. Rows 5 and 6 show the estimated variance components calculated in the G study, for P and S:P[33], as well as the degrees of freedom (df). There is a clear tendency for the variances due to P and S:P to decline across time (but as has been shown earlier, the overall SD decreases with years as well).

D studies. The remaining rows of the table show the results of the D studies, which assess the likely generalizability which would be achieved for different designs which vary in the number of stations (symbolized as n_s'). The actual Stage 3 has four stations, and in rows 8, 13 and 18 the values for $n_s' = 4$ are put in bold. We have also calculated generalizabilities for 'what-if' scenarios, considering the likely generalizability with 3 stations (equivalent to that carried out in 2009 and 2010), or with a larger Stage 3 with 6,10 or 20 stations, to allow possible design options to be considered. Mean values across the five years are shown in the final column.

The three separate parts of the D study consider the generalizability for relative judgements (G: rows 7 to 11), absolute judgements (Φ: rows 12 to 16), and absolute judgements to a cut score of 3.1 (Φ_λ: rows 17 to 21).

Generalizability for four stations. Consider firstly the mean values in the final column for the actual situation where there are four stations (ns' = 4; rows 8, 13 and 18). With a simple nested design such as the present one, G and Φ are the same, and give an average value of 0.60. In contrast Φ_λ is a bit higher at 0.67, reflecting the fact that the cut-score of 3.1 is somewhat lower than the mean scores in the assessment.

Trends from 2011 to 2015. Comparing across the five years it is also apparent that G, Φ and Φ_λ are all lower in 2015 than in 2011. That is not unexpected given the decreasing SDs shown in row 4 and the decreasing variance due to both P and S:P in rows 5 and 6. However P is declining at a greater rate than S:P, and hence the generalizabilities are lower in 2015 than in 2011. No immediate explanation of the trend is apparent, and it requires further exploration, but the implication is that Stage 3 is becoming less reliable (and that conclusion is compatible with the decreasing Cronbach alpha coefficients reported earlier).

Six, ten or twenty stations. The other rows of the D study show the effect on generalizability of increasing the number of stations to example values of 6, 10 or 20. If the selection centre in Stage 3 is felt to be important, and it is also felt important that it has an acceptable generalizability, then a 10 (or 20) station Stage 3 would give G, Φ and Φ_λ of 0.79, 0.79 and 0.84 (or 0.88, 0.88 and 0.91).

3.5.9 Generalizability for candidates taking Stage 3 in both Rounds 1 and 2 within a year

Table 3.6 provides a summary of the generalizability analyses for those candidates who took Stage 3 on two occasions, once in Round 1 and then again in Round 2 of the same year. As before, rows 1, 2 and 3 show the numbers of candidates, the mean score for each year, and the SD[34] of all scores within a year. Again there is a clear trend for the numbers taking the assessment to decrease across years, and also there is a suggestion that the mean score is increasing across the years, although unlike in Table 3.5 the SD is stable across the years.

Variance components. The generalizability analysis is more complicated than when a candidate only takes the assessment once. The candidates (P) each take Stage 3 twice, and hence Occasion (O) is crossed with P, giving PO (i.e. P x O). Candidates will almost always see different stations on the second occasion to those they saw on the first occasion, and stations also differ between candidates within each occasion, so that Stations are nested within Person x Occasion (i.e. S:PO).

G studies. There are more variance components than in the previous analysis, with estimated components shown in rows 5, 6, 7 and 8, along with their degrees of freedom. The variance for Persons (.057) is less than in the analysis of Table 3.5 as those retaking are a restricted subset of all of the candidates, in particular having failed on the first occasion. There is variance due to Occasion (.049), probably reflecting higher scores on the

[33] There is only observation per cell and hence P and S:P together account for all of the variance.

[34] This is the simple SD of all of the judgements made, so that if there are 1,000 candidates assessed on four scales it is the SD of all 4,000 judgements that were made.

» *Table 3.5:Summary of Generalizability analyses for Stage 3 total for the years 2011 to 2015. The Generalizability coefficients are in grey to remind the reader that they are based only on Round 1 and hence are probably overly high (see discussion below).*

(1)	Year		2011	2012	2013	2014	2015	Mean
(2)	N		4259	4509	4619	4159	3751	4259
(3)	Mean		3.228	3.227	3.332	3.304	3.371	3.292
(4)	SD		.666	.693	.637	.635	.602	.647
	G Study							
	Effect	df	Variance Components					
(5)	P	N-1	.1320	.1502	.1153	.1095	.0762	.1166
(6)	S:P	3.N	.3114	.3303	.2908	.2938	.2864	.3025
	D study	ns′						
(7)	G	3	.560	.577	.543	.528	.444	.530
(8)		4	.629	.645	.613	.598	.516	.600
(9)		6	.718	.732	.704	.691	.615	.692
(10)		10	.809	.820	.799	.788	.727	.789
(11)		20	.895	.901	.888	.882	.842	.882
(12)	Φ+	3	.560	.577	.543	.528	.444	.530
(13)		4	.629	.645	.613	.598	.516	.600
(14)		6	.718	.732	.704	.691	.615	.692
(15)		10	.809	.820	.799	.788	.727	.789
(16)		20	.895	.901	.888	.882	.842	.882
(17)	Φ_λ	3	.588	.601	.636	.607	.610	.608
(18)		4	.655	.668	.699	.673	.676	.674
(19)		6	.741	.751	.777	.755	.758	.756
(20)		10	.827	.834	.844	.853	.839	.838
(21)		20	.905	.910	.921	.911	.913	.911

+ although Φ is usually regarded as an absolute assessment of performance it is here actually a relative assessment since there is only a single assessment and therefore no way of distinguishing relative standards from absolute standards, and therefore G and Φ give identical results.

second occasion than the first, but there is little Person x Occasion variance (PO) suggesting that the increase across occasions is similar in the various candidates. Stations are nested within PO and have the largest variance component (.407), which, as ever, reflects case-specificity, different candidates doing differently on different stations. Although PO and S:PO seem stable across the years, there is a tendency, as before, for P to decline across the years, and there is also an increase in the O variance across the years.

D studies. The D studies in Table 3.6 are shown in grey to indicate that they are only included for completeness, so that the table is analogous to Table 3.5. The D study estimates are not however readily interpretable, referring to the expected generalizability for candidates who resit Stage 3, which is not a measure of much practical utility. Rows 9 to 23 show the D study results for different designs. Note that all of these D studies have the number of occasions, no', set to 1, since we wish to know about a single administration of the Stage 3 assessment. As previously, since the actual Stage 3 has four stations the values in rows 10, 15 and 20 with values for n_s' =4 have been put in bold. The table also has results for n_s' =3, 6, 10 and 20, with mean values across the five years shown in the final column.

Four stations. Consider firstly the mean values in the final column for the actual situation where there are four stations (n_s' = 4; rows 10, 15 and 20). The present analysis has two occasions, and hence there is a difference in relative and absolute generalizability, the average for G being 0.34 and that for Φ being .266.

Trends from 2011 to 2015. Comparing across the five years it is clear, as before, that all of the generalizability coefficients are lower in 2015 than in 2011. That is not unexpected given the variance due to both P in row 5 and the increasing variance in S:PO in row 7. Once more there is a need for further exploration of why Stage 3 seems to be becoming less reliable.

Six, ten or twenty stations. The other rows of the D study show the effect on generalizability of increasing the number of stations to example values of 6, 10 or 20. These values however apply only to the situation for resit candidates, and not for candidates as a whole, and are included only for completeness in relation to Table 3.5.

Restriction of range. The candidates in Table 3.6 are clearly less able than those in Table 3.5 having lower mean scores (2.81 vs 3.29), and they are resitting Stage 3 because they failed on the first occasion. That needs taking into account in assessing the generalizability scores, and that has a number of technical aspects. The approach has not to our knowledge been fully described previously in the literature, although the method is implicit in the approach of Brennan (2001b), and as a result we have put a detailed description into Appendix 3.1 so that others can rework the calculations if necessary.

Generalizability coefficients corrected for restriction of range in those taking both Round 1 and Round 2. Table 3.12 in Appendix 3.1 shows the general form of the calculations for the corrected generalizability coefficients (called G**, Φ** and Φ_λ** to distinguish them from the values of G, Φ and Φ_λ in Table 3.6 which are not corrected for range restriction), and in particular shows the specific values of 0.57, 0.48 and 0.51 for the 2011 candidates. Values for the 2012 to 2015 candidates are shown below in Table 3.7.

3.6 GENERALIZABILITY AND RELIABILITY ANALYSES: SUMMARY

3.6.1 Stage 3

Calculation of the reliability/generalizability of OSCE-type assessments has mostly used Cronbach's alpha, which has well-known flaws, although some studies have used generalizability theory (Brannick et al., 2011)[35]. As described earlier, the reliability of Stage 3 of GP selection has not so far been well described in the literature, although there is an important and relatively recent claim that Cronbach's alpha for the selection centre was **0.87** (Patterson et al., 2013). Wakeford (2014) was skeptical

[35] Generalizability analysis can also have problems if the model does not allow estimation of particular effects, with the probems of 'hidden facets' being particularly problematic, especially if the facet is 'Occasion'.

» Table 3.6: Summary of Generalizability analyses for Stage 3 total for candidates taking Stage 3 in both Round 1 and Round 2 in years 2011 to 2015. Note that the D study values are shown in grey to indicate that they are present only for completeness, and are not readily interpretable (see text for further details).

(1)	Year		2011	2012	2013	2014	2015	Mean
(2)	N		491	499	379	372	267	402
(3)	Mean		2.738	2.738	2.767	2.833	2.949	2.805
(4)	SD		.694	.715	.693	.693	.697	.698
	G Study							
	Effect	df	Variance Components					
(5)	P	N-1	.0667	.0754	.0536	.0396	.0476	.057
(6)	O	1	.0422	.0328	.0293	.0579	.0830	.049
	S:PO	6.N	.3929	.4186	.4083	.4037	.4095	.407
(7)	PO	N-1	.0010	.0012	.0040	.0091	0.**	.003
(8)	D study*	ns′						
(9)	G	3	.336	.349	.277	.216	.232	.282
(10)		4	.402	.416	.336	.264	.287	.341
(11)		6	.501	.515	.427	.341	.376	.432
		10	.623	.637	.545	.444	.501	.550
(12)		20	.763	.773	.687	.574	.668	.693
(13)								
(14)	Φ	3	.277	.303	.240	.164	.158	.228
(15)		4	.320	.352	.284	.191	.182	.266
(16)		6	.380	.421	.346	.228	.214	.318
		10	.447	.499	.420	.269	.249	.377
(17)		20	.515	.579	.500	.312	.285	.438
(18)								
(19)	Φ_λ	3	.188	.252	.314	.102	.124	.196
(20)		4	.222	.296	.365	.120	.144	.229
(21)		6	.272	.360	.434	.146	.171	.277
(22)		10	.323	.436	.512	.177	.202	.330
(23)		20	.392	.516	.592	.210	.304	.403

* Dstudy is for a single occasion

** negative variance estimate set to zero

about that claim, and his own analyses of data obtained via an FoI request to one UK deanery suggested a Cronbach's alpha of **0.6**2, and an unpublished generalizability analysis with Jolly gave an estimate of G of **0.64** for a four station assessment.

Cronbach's alpha is essentially a measure of **internal reliability,** and it does not take account of sources of variance other than items within a single administration of a test. Estimates of reliability will therefore be overly liberal. Table 3.7 summarises our calculations of Cronbach's alpha for the five years from 2011 to 2015, and finds an average value of **0.62.** A simple generalizability analysis of the same data (see Table 3.5) finds broadly similar coefficients with a mean G of **0.59**. Since these generalizability estimates are based on a single set of data they cannot take into account variation across assessments in scenarios, assessors and simulators, and hence they are, like Cronbach's alpha, only a measure of internal reliability. That G and Phi in rows 2 and 3 of Table 3.7 are identical, there being only a single assessment, also shows that the analyses are of necessity concerned only with relative standards. However an assessment such as Stage 3 should also take absolute standards into account, the intention being that candidates in different years are assessed against the same absolute level of performance. Neither Cronbach's alpha nor any analysis of only one Round 1 can do that[36].

The AERA/APA/NBME Standards (American Educational Research et al., 2014) are clear, as also is Brennan (2001a), that complete replications are the ideal method of assessing reliability, **the alternate forms reliability** being the ideal, but requiring a second complete, parallel assessment at a later date, which in a OSCE-type assessment would mean different scenarios, different assessors and different simulators at a different selection centre. Practicality prohibits most candidates receiving such an

» *Table 3.7: Summary of estimates of reliability of Stage 3 using different methods across years.*

	Year	Type	2011	2012	2013	2014	2015	Mean
	Internal consistency analyses							
(1)	Cronbach's Alpha from raw data	Relative	.635	.652	.621	.662	.552	.616
	Round 1 analysis (Table 3.5)							
(2)	G	Relative	.629	.645	.613	.528	.516	.586
(3)	Φ	Relative+	.629	.645	.613	.528	.516	.586
(4)	Φ_λ	Relative+	.656	.668	.699	.607	.676	.661
	Alternate forms analyses							
	Round 1 and Round 2 analysis with correction for range restriction (see Table 3.12 and Appendix 3.1)							
(5)	G**	Relative	.571	.587	.521	.499	.427	.521
(6)	Φ**	Absolute	.483	.520	.460	.393	.291	.430
(7)	Φ_λ**	Absolute	.512	.545	.555	.474	.447	.507
(8)	EM analysis of Stage 3 correlations across years (see Table 3.3)	-	-	-	-	-	-	.432
(9)	EM analysis of Stage 3 correlation within year	Relative	.614	.599	.483	.499	.380	.515

+ although Φ is usually regarded as an absolute assessment of performance it is here actually a relative assessment since there is only a single assessment and therefore no way of distinguishing relative standards from absolute standards, and therefore G and Φ give identical results.

[36] Consistency of standards across different selection centres is also important, and in principle can be analysed in terms of generalisability.

entire second assessment, as those passing on the first attempt will not wish subsequently to take another assessment. Earlier in this chapter we have described two different ways of calculating an alternate forms reliability which takes account of the inevitable restriction of range of resit candidates. One method looks at candidates retaking the Stage 3 (and Stage 2) assessments across one or more years, so that the minimal interval is one year, and uses the EM algorithm to handle what can be regarded as missing data. The other method considers candidates within a selection year who take Round 1, fail it, and then re-take it in Round 2, and we modify a conventional generalizability analysis to take into account the restriction of range. Table 3.7 in rows 5 to 7 shows the corrected estimates of alternate forms generalizability coefficients within a year, and row 8 shows the EM estimates of alternate forms reliability across one or more years. Without going into the details, row 9 of Table 3.7 also contains a second EM estimation of alternate forms reliability, within each year assessing the correlation between the Round 1 and Round 2 total marks[37].

G** in row 5 has a mean value of 0.52, which is rather lower than the internal consistency G in row 2 which had a mean value of 0.59. The EM analysis in row 9 of the same data, but carried out in an entirely separate analysis, gave a very similar set of results, the average reliability being 0.52, which is very similar to the average G** of 0.52. That provides strong support for the EM method being a valid way of assessing alternate forms reliability. The EM estimate of alternate forms reliability in row 8, which is across years, rather than within years, has a lower mean value of 0.43, and this may reflect the fact that in row 8 there is a year between assessments, whereas in row 9 there is only a few months (and as Brennan emphasizes, the effects of different time intervals are always of interest and importance). All of the values in rows 5, 8 and 9 are assessing relative performance. Taking the results together, the alternate forms reliability of Stage 3 for relative judgements is likely to be between about 0.43 and 0.52, with a middle estimate of about **0.48**. Stage 3 is, though, intended to be an assessment of absolute performance, in which case either Φ** or Φ_λ** is the appropriate coefficient. The latter is set at about the level of a notional pass mark for the selection centre in Stage 3, and in row (7) it can be seen to have an average value of **0.51**.

Taking all of the various calculations together, a best estimate of the alternate forms reliability is probably of the order of **0.50** [38] .

Changes in Stage 3 reliability across years. Earlier we have commented that there seem to be suggestions that the reliability of Stage 3 in more recent years may be lower than in earlier years. Looking at all of the assessments summarized in Table 3.7 there seems to be little doubt that across all of the methods of calculation the reliability in 2014 and 2015 is lower than that in 2011 and 2012. The reasons for this are not clear at present, but it might well mean that the overall estimate of 0.50 given in the previous paragraph is overly optimistic, at least for the assessment in its present form.

D study analyses of Stage 3. It would seem that Stage 3 is not very reliable. An important practical question is whether the reliability could be increased by increasing the length of the selection centre. Without giving detailed results, particularly for Φ_λ**, we will simply draw attention to the greyed out parts of Table 3.5, particularly for rows 7 to 11, which show how G would change with greater numbers of stations (and broadly similar proportional effects could be expected for Φ_λ**). Increasing to 10 or even 20 stations does inevitably improve reliability, but that of course would come at a substantial cost in workforce requirements and candidate time.

Stage 2 reliability. We have not had access to raw data to recalculate reliability estimates, but earlier in Table 3.1 we summarized the published estimates for CPST and SJT, which have mean Cronbach's alphas of 0.88 and 0.81, the lower reliability of the SJT probably due to it having fewer items. The CPST and SJT correlated with one another about 0.55 (correlation disattenuated for unreliability 0.64), suggesting that they share about 40% of their variance). As emphasized earlier, Cronbach's alpha is a measure of internal consistency. Although we did not have access to raw Stage 2 data, we used the EM algorithm, as described for Stage 3 (see Table 3.3) to estimate alternate forms reliability, which was about 0.73 for CPST, 0.57 for SJT and 0.73 for the total Stage 2 score. The values for CPST and SJT are rather lower than might be expected from Cronbach's alpha,

[37] We also carried out this analysis using multiple imputation in SPSS and, as expected, obtained very similar results.

[38] This is derived from the relative judgements having a range of from .43 to .52, and Φ_λ having a value of .507, so that .50 seems a reasonable estimate. Other readers may choose different values to summarize the results in Table 3.7. What is clear is that the reliability is certainly not the 0.87 which has been claimed, and even the values of .62 and .641 by Wakeford and Jolly are probably over-estimates, being only internal consistency measures.

but the alternate form reliability estimates can probably be trusted since a) the results from Table 3.7 suggest that the EM approach gives broadly similar results to more conventional measures such as generalizability analyses, and b) the SJT, which shows the largest change in reliability from 0.81 to 0.57 is based on only 50 items, which is a fairly short multiple-choice test (and the scoring is not straightforward for candidates). Overall there might be an argument that the reliability of both the CPST and the SJT needs increasing, perhaps by adding more items.

Predictive validity. Although reliability is important, of greater importance is validity – the extent to which selection tests correlate with important later outcomes, such as exit examinations, professional behaviour, and, of course, patient care in all its manifold senses. We consider these issues in Chapter 4. None of these though are possible without reliability, and that is why the alternate forms reliability after a delay is so much more important than internal reliability. Internal reliability measures something now, but it may be gone by tomorrow. Alternate forms confirm that a competency is stable across time and ultimately a selection programme is not asking only about a few weeks or even a few years of a training programme, but decades of professional behaviour. In that sense the longer the delay before the alternate form then the more useful that measure is of reliability.

3.7 FACETS MODELLING (MFRM) OF STAGE 3 IN DIFFERENT DEANERIES/LETBS

The analyses so far of reliability/generalizability have assessed reliability and generalizability in the sense of the extent to which candidates attain similar marks on different occasions. However those analyses have not been able to work out the various influences which mean that candidates might get a different mark on different occasions. Generalisability analysis can, in principle, partition variance into different components, such as scenarios, assessors and simulators, but it requires data in a very structured format to be able to do that. Multi-facet Rasch modelling is a rather better technique for exploring the influences on candidate scores, not least as it not only estimates variance but also the effect for each individual station, assessor, or simulator.

In order to separate the effects of station, assessor and simulator they have to be crossed with one another. If each candidate sees the same combination of a particular station with a particular assessor and a particular simulator then it is impossible to tease apart the various components as they are all confounded with one another. Fortunately Deaneries/LETBs have adopted different approaches to organizing Stage 3, and in some cases there are sufficient data to analyse. This section begins therefore by looking at the different ways in which Stage 3 has been organized.

Organisation of Scenarios, Assessors and Simulators within Deaneries. Making sense of how Deaneries organize their Stage 3 assessments is not straightforward. Detailed information on the identity of the Assessor, the Simulator and the Case (Scenario) are only available for 2015 Round 1. Various codes are entered into the spreadsheets, and these vary from simple numerical codes ('3'), through apparent sets of initials ('EPL', 'AKB', etc), and sometimes very lengthy complex codes (e.g. 'HR/SIMA/RM07', where although it is the simulator code which should be provided, there seems also to be information on the station and perhaps also the case or the room). Table 3.8 summarizes information about the method of Stage 3 selection at each of the Deaneries. Deaneries can be broadly grouped into two types. Some Deaneries are relatively small, and the analyses described below are tedious, and therefore restricted to the two largest deaneries in each of the two groups, London (n=701) and North West (n=408) for Type 1, and South West (n=324) and East of England for Type 2 (n=304). These four deaneries should give a reasonable idea of the extent to which candidate scores are dependent on the particular combinations of assessors, simulators and cases that the candidates encounter.

[44] We also carried out this analysis using multiple imputation in SPSS and, as expected, got very similar results.

[45] This is derived from the relative judgements having a range of from .43 to .52, and Phi-Lambda having a value of .507, so that .50 seems a reasonable estimate. Other readers may choose different values to summarize the results in Table 7. What is clear is that the reliability is certainly not the 0.87 which has been claimed, and even the values of .62 and .641 by Wakeford and Jolly are probably over-estimates, being only internal consistency measures.

Type 1 deaneries, at the top of the table, are those in which, by and large, each simulator always works with the same assessor so that one can consider the analysis in terms of a 'Simulator-Assessor Team'. For such deaneries it is not possible to disentangle the effects of Assessors from effects of Simulators. The North West and London deaneries both show similar patterns. For North West, for instance, there are codes for 44 assessors and 39 simulators, but a small proportion have only 1 or 2 instances, and may be the result of typos or someone else stepping in for just one assessment (we have no way of checking the data). Because of the nature of the 'carousels' (circuits) around which the candidates rotate, there are only 11 separate combinations of assessors across the four stations, each candidate getting one of them. Likewise, although there are two different cases at each of stations A, B, C and W, which in principle would lead to 2x2x2x2=16 combinations, in fact there are only 5 combinations to which candidates are exposed.

The Type 2 deaneries, at the bottom of the table appear to have Assessors nested within Simulators (i.e. each simulator will work with several assessors, but each assessor works with only a single simulator). Consider the East of England[39] Deanery. There are 304 candidates and hence 1216 encounters, each candidate being assessed at four stations. Assessor codes are mostly three letters and might be initials. There are 118 different codes, mostly ranging from 18 cases down to 1 case, with a single code occurring 36 times, perhaps due to two assessors having the same initials. The median number of assessments per assessor is therefore 9, perhaps representing a half day of assessment. There are 32 Simulator codes for this Deanery for the 912 OSCE-type stations, with a median of 30 per simulator. Each simulator seems only to have a single case at a single station, and similarly each assessor seems to work only with a single case at a single station. On average each simulator will work with about 3 assessors. Finally there are two different cases at each of the four stations, but there are only four different combinations which actually occur, candidates seeing either cases 3, 1 and 1 or 4, 2 and 3 at stations A, B and C, and then then seeing either case 2 or 3 at the written station.

The detailed data within Deaneries can be assessed in several ways. We will approach the data here using the FACETS program (Linacre, 2004), which carries out Multi-Facet Rasch modelling, which is a multivariate extension of Item-Response Theory (IRT). Introductions to FACETS can be found elsewhere (Eckes, 2011), as also can introductions to IRT (DeMars, 2015, Hambleton et al., 1991). We have used FACETS in other studies for analyzing medical assessment data, and in particular have used it to assess the extent to which examiners are 'hawks' or 'doves' (McManus, 2006). In complex assessments, it is not only examiners who may differ in 'stringency', but some simulators may be more difficult to obtain information from than others, and so being more hawkish, and some cases may be harder than others (as we have demonstrated in the PACES Part 2 assessment (McManus et al., 2013b)).

The key feature of FACETS is that it puts candidates, assessors, simulators, stations and cases onto a single common metric, which allows them to be directly compared, along with an assessment of the variation due to each. The scores on each scale are shown on a common **yardstick**, which will be explained below.

3.7.1 FACETS analysis of the Type 1 deaneries.

London selection centre yardstick:

For the Type 1 deaneries the individual marks are a function of a candidate's ability, the particular Simulator-Assessor team that they saw at a case, and the particular case. The yardstick is shown in Figure 3.1, and has five columns. The measure scale at the extreme left is the common unit of measurement, which is on a logistic scale. Individual candidate marks are shown in the next column, with an asterisk indicating 8 candidates, and a "." indicating 7 or fewer candidates. The distribution of candidate ability is mostly normally distributed, with perhaps a small number of candidates getting very high scores, which means they have performed very well. Apart from the candidate scale, the other measurement scales are centred at zero. The middle column shows the 'stringency' for the seven simulator-assessor teams, and to a large extent the differences are small, as can be seen by comparing their variability (range) with the range of candidate variability. Likewise the eight combinations

[39] Dates and times of assessments are not available and so that cannot be checked.

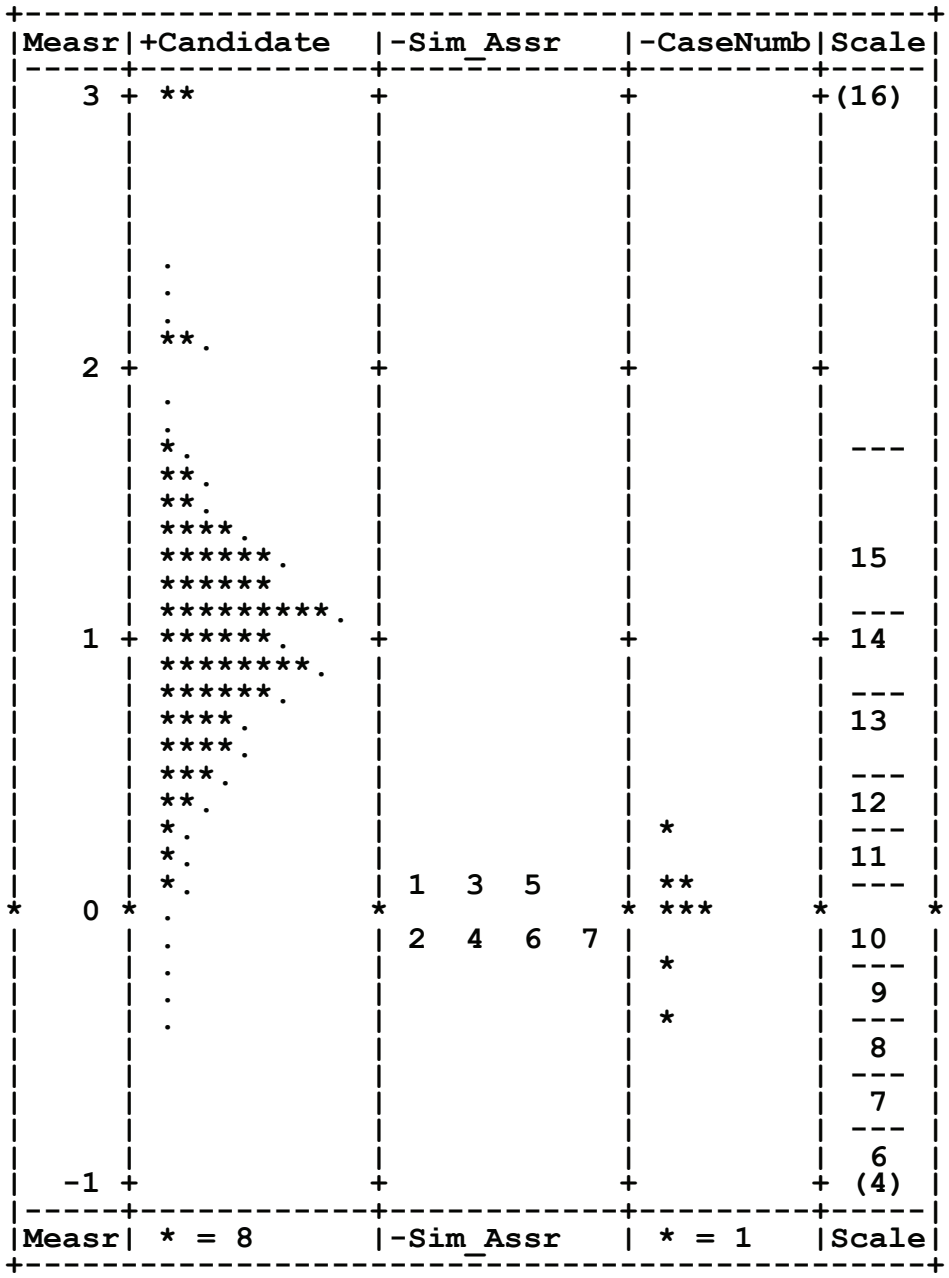
» *Table 3.8: Summary of Assessors, Simulators and Cases at each of the Deaneries Deaneries/LETBs are sorted into two broad types, with detailed information provided only for the larger deaneries in each group since the modelling described is only practical with a reasonably large sample size. .*

Selection Deanery	N total	Assessors (A)	Simulators (S)	Cases (C)	Assessors x Cases (excluding small, i.e. <5)	Design summary (P=Person; T= Assessor x Simulator Team)
Type 1: Assessor and Simulator confounded (each assessor always works with the same simulator at a station)						
London	701 (662)	21 assessors each assessing 89 to 202 candidates. 7 combinations of assessors overall (89-100 candidates). 13 combinations of only 1 or 2 removed.	21 simulators each working with the same assessor	Two cases at A, two at B, two at C and two at W, with 13 combinations from 21 to 98 candidates. 2 combinations of only 1 or 4 candidates removed.	The 7 Assessor combinations saw 2,3,4 or 5 case combinations. The 13 case combinations were mostly assessed by 2 assessor combinations, with 1 assessed by 1, 2 by 3 and 1 by four assessor combinations.	P x (T:C) Not balanced.
Northern Ireland	123					
North West	408 (292)	44 assessors each assessing 13 to 27 candidates. 11 combinations of assessors overall (52-100 candidates). 2 assessors seeing <5 candidates, and 2 combinations of assessors seeing only 4 candidates were removed.	39 simulators almost but not entirely working with the same assessor. Assessors and simulators treated as teams within assessors.	Two cases each at A, B, C and W, with 5 combinations from 23 to 62 candidates.	The 11 Assessor combinations mostly saw all 7 case combinations, with 4 and 1 seeing 6 or 5 combinations. T x C was therefore well distributed.	P x (T:C) Nearly balanced.
Scotland	297					
West Midlands	284					
Yorks and Humberside	284					

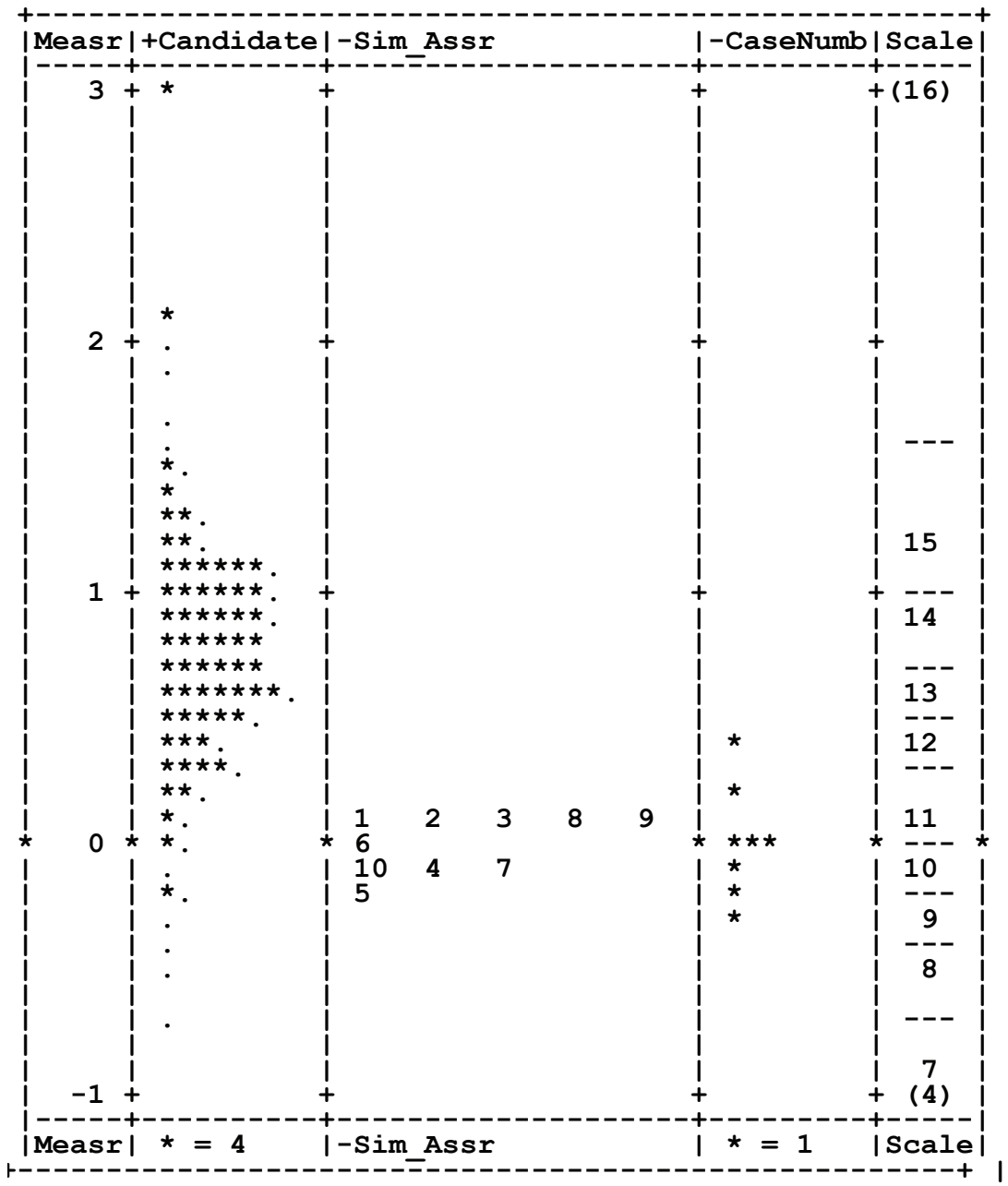
» *Table 3.8: Continued.*

Selection Deanery	N total	Assessors (A)	Simulators (S)	Cases (C)	Assessors x Cases (excluding small, i.e. <5)	Design summary (P=Person; T= Assessor x Simulator Team)
Type 2: Assessors within Simulators (i.e. several assessors work with each simulator)						
East of England	304	118 different assessor codes, with one code of frequency 36, and the remainder with frequencies distributed evenly from 18 to 1 (median = 9).	32 different simulator codes for 912 clinical stuations, with a median of 30 encounters per simulator. Each simulator therefore works with about three assessors	Two different cases at each of stations A, B, C and W, but only four of the possible 16 combinations occurred, with either 311 or 423 for ABC, and 2 or 3 for W.	Many combinations of assessor, simulator and case.	P x (A:S) x C
East Midlands	153					
Northern	103					
South West	324	138 different assessor codes, with frequencies from 30 to 1 (median=10), and 50 having frequencies of 4 or less (median = 1). Median = 12 for codes with frequency >4.	46 different simulator codes from 43 to 1 (median = 28), and 18 having frequencies of <=1 (median = 1). For simulators with frequency > 4, median = 34.	Two different cases at each of stations A, B, C and W, but of the 16 possible combinations, 70% consisted of 3112 or 4233 (i.e. for ABCW), with a further 20% for 3113 and 4232.	Many combinations of assessor, simulator and case.	
Wales	122					
Wessex	135					
Other (too small or missing information)						
Defence	30	Too few candidates				
Thames Valley	163	No assessor codes				

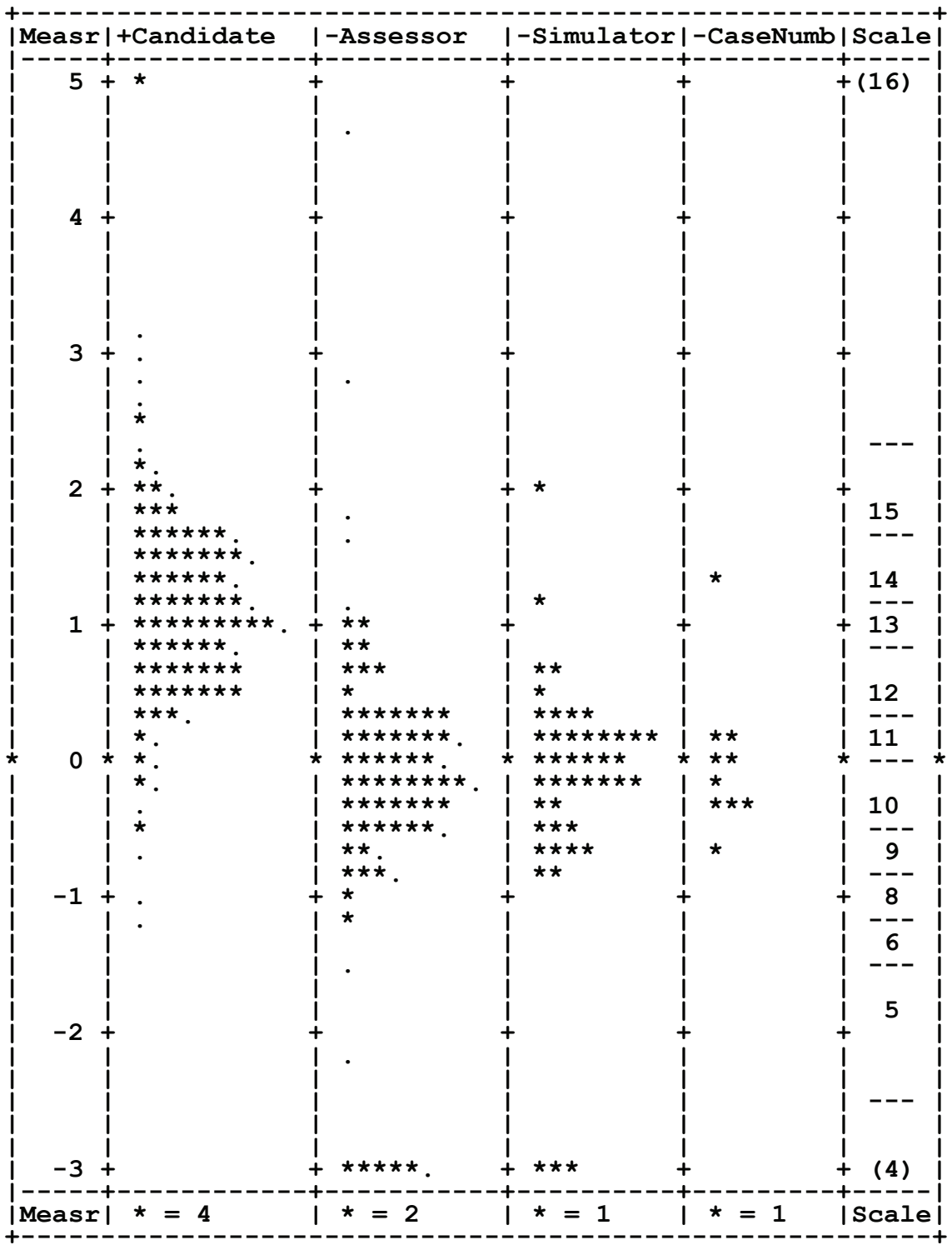
» *Figure 3.1: Yardstick for London Selection Centre (n=662 candidates)*



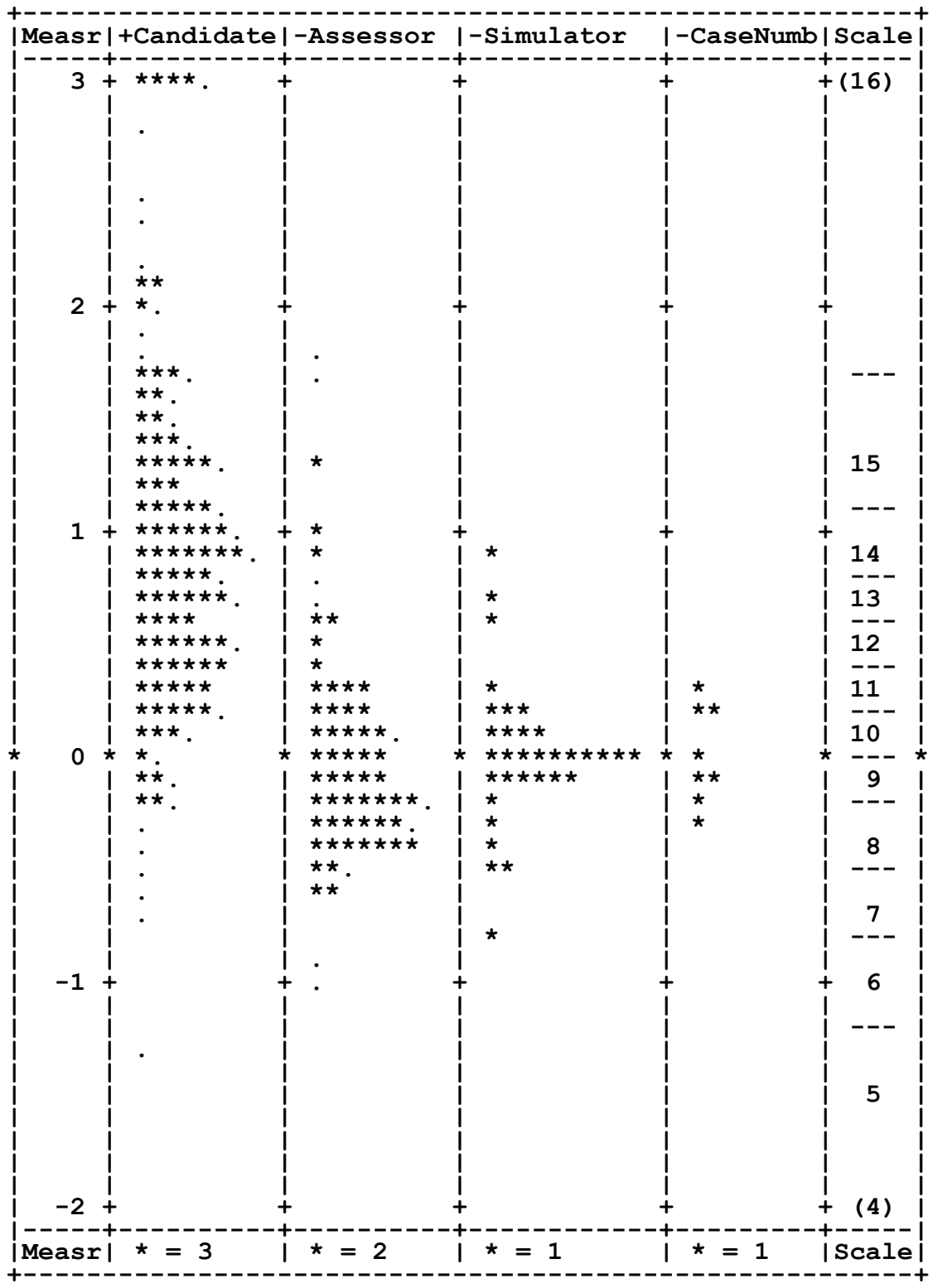
» *Figure 3.2: Yardstick for North West Deanery (n=292 candidates)*



» *Figure 3.3: Yardstick for the South West Deanery (n=324 candidates)*



» *Figure 3.4: Yardstick for the East of England Deanery (n=304 candidates)*



of cases, which are shown in the column second from the right are mostly very similar in difficulty. Finally, the right-most column shows scores on the raw scale (which has been converted so that marks are in the range 4 to 16).

The FACETS program also provides various statistics describing the overall performance of the scale, including a variant of a reliability coefficient. This reliability should be treated with care as it is, in many ways, similar to Cronbach's alpha, not taking stability across time, or variation between forms into account, and hence it tends to produce an inflated estimate of reliability (Brennan, 2001c, d, Webb et al., 2007). Having given all those provisos, for the London Deanery, the IRT reliability is **0.38**, which is low.

North West Deanery yardstick:

The yardstick for the North West Deanery, which like the London Deanery is of type 1, is shown in Figure 3.2. The picture is very similar to that of Figure 3.1, with relatively small amounts of variance for the Assessor-Simulator team, and between cases. The IRT reliability is slightly higher at 0.49.

3.7.2 FACETS analysis of the Type 2 deaneries.

The yardstick for the South West Deanery:

The type 2 deaneries, of which South West and East of England are the largest, are slightly more complex as Simulator and Assessor are separable to some extent. Figure 3.3 shows the yardstick and the single column in the middle has now been replaced by two separate columns, for Assessor and Simulator, and it can be seen that there is more variance, and the variance is present both in the Assessors and the Simulators, both groups of whom therefore vary in their hawkishness (hawks have higher values on the yardstick, it taking a higher ability level to satisfy the requirement of a hawk). One of the Cases is also rather harder than the others. The IRT reliability of the candidate facet is nevertheless higher at 0.65 and that probably reflects the fact that some of the variance in raw scores has been accounted for in assessor and simulator variance, that variance having been removed from the 'fair scores'.

The yardstick for the East of England Deanery:

East of England, like South West, is a type 2 deanery, with variance for assessors and simulators being separable. The yardstick is shown in Figure 3.4, and it is very similar in its general pattern to that seen in Figure 3.3 for the South West Deanery. Once again the IRT reliability can be calculated and it is **0.55**. For these 2 deaneries, **very roughly**, Assessors, Simulators and Cases have similar variances and so contribute **approximately** equally to unreliability.

3.7.3 Summary of the FACETS analyses of Assessors, Simulators and Cases in four large deaneries

The four FACETS analyses give broadly similar results, with IRT reliabilities of 0.38, 0.49, 0.65 and 0.55, the overall average being **0.53**, with somewhat higher values when assessor and simulator variance are taken separately into account (mean=0.62), than when they are combined (mean=0.44)[40]. The latter two analyses, in which simulators and assessors are separable, indicate that there is undoubtedly variance due to differences in stringency (hawkishness) between assessors, with it also being the case that some simulators are more hawkish (i.e. it is harder to get information or whatever from them), and some cases are also harder than others. If those factors are taken into account then a higher IRT reliability can be obtained. The implication is also that in the analyses of alternate-form reliability reported earlier, some of the variance in candidate performance is due to differences in difficulty of assessors, simulators and cases, and if that variability were to be taken into account then the reliability of the assessment as whole might be increased. In principle it is possible to use MFRM to 'correct' marks for a candidate having been assessed on a more difficult case, by a more hawkish assessor or a more hawkish simulator, although

[40] All of the FACETS yardstick analyses should be treated with care as in each case, due to the designs not being fully randomised, there is evidence of disjoint subsets, there being 7, 10, 69 and 88 subsets in the four analyses. Such problems could be avoided if the collection of data were part of a properly designed study in which, in effect, assessors, simulators, cases and candidates are allocated more at random.

[41] See Till et al. (2013) for an example in medical student selection (REF: http://journals.lww.com/academicmedicine/Abstract/2013/02000/Improving_Student_Selection_Using_Multiple.26.aspx)

that is still at an experimental stage and is probably not suitable for routine use[41]. The MFRM analyses are however a clear reminder that assessors and simulators, despite extensive training, are not identical in their treatment of candidates.

The MFRM analysis also emphasizes that it is only possible to assess whether, say, assessors are more or less hawkish if those assessors assess on a range of simulated cases and with a range of simulators. If case, assessor and simulator always work together then there is no possibility of teasing apart variation due to each of them. In designing any assessment system it therefore makes sense to try and make the system so that different combinations of case, assessor and simulator occur and then in principle the effect of each can be disentangled.

3.8 THE RELIABILITY OF THE ES, CS, CT&PS AND PI SUB-SCORES AND THE SEPARATE SCORES ON THE FOUR STATIONS IN STAGE 3.

The analyses so far have considered only the psychometrics of the total scores at stations. However the scoring of Stage 3 involves the separate assessors at each of the three simulation stations (A: patient, B: carer; C: co-professional) and the Writing Station to rate each candidate on three or four separate scales, ES (Empathy and Sensitivity), CS (Communication Skills), CT&PS (Conceptual Thinking and Problem Solving) and PI (Professional Integrity). The simulator stations are rated on various combinations of three of the four scales, and the writing station is rated on all four scales (see Chapter2). A key question for such a marking scheme is whether there is separable, reliable variance associated with each of these four sub-scales or with each of the four station types. The remaining analysis in this chapter therefore considers the scores on the sub-scales in Round 1 over the five years from 2011 to 2015 when the structure of Stage 3 was stable.

Correlation between sub-scales. The correlations between the scores on the sub-scales across 23,817 assessments over the five years are shown in Table 3.9 below. Correlations are high, varying from 0.59 to 0.74 (mean=0.68). In interpreting these correlations it must be remembered that they are mostly made by the same assessors and hence are not statistically independent.

The eigenvalues for this correlation matrix are 3.03, 0.45, 0.27 and 0.25 which strongly suggests one major factor, with visual inspection of the scree-slope suggesting that there may be a smaller second factor. Principal component analysis finds a very strong first (common) factor, with a second factor contrasting ES with PI and CT&PS.

» *Table 3.9: Simple correlations of the ES, CS, CT&PS and PI subscores across all stations in Stage 3.*

	ES total	CS total	CT&PS total	PI total
ES total	1.	.669	.601	.591
CS total	.669	1.	.738	.729
CT&PS	.601	.738	1.	.732
PI	.591	.729	.732	1.

» *Table 3.10: Simple correlations of scores on stations A, B, C and W in Stage 3*

	A station total	B station total	C station total	W station total
A station total	1.	.387	.358	.270
B station total	.387	1.	.355	.272
C station total	.358	.355	1.	.259
W station total	.270	.272	.259	1.

Correlations between station scores. Correlations between the marks on the four stations are shown in Table 3.10 below, and are much lower than those between scales, with a range of 0.27 to 0.39 (mean=0.32). The correlations between the three simulations (all over 0.35) are clearly higher than those between each simulation and the written exercise (all under 0.28). The between-station scores are each made by separate assessors and therefore are properly independent statistically.

The eigenvalues for the correlation matrix are 1.96, 0.78, 0.65 and 0.61, once again suggesting a large first component, with perhaps a smaller second component. Principal component analysis showed that all four factors loaded on the first component, and that the second component contrasted the writing station with the three simulation stations.

Reliability of sub-scales and station scores. Analyzing the reliability of sub-scales scores and station scores began by extracting **unstandardised residuals** from separate multiple regressions, in which each sub-score (or station score) was regressed on the other three sub-scores (or station scores). The residual therefore describes the extent to which a sub-score or station score was different from that predicted by the other three measures. The residuals were rounded to whole numbers, and can be interpreted as the difference in marks (on a scale of 4 to 16) between the actual score and the score predicted from

» *Table 3.11: Alternate forms reliability for ES, CS, CT&PS and PI residuals, and residuals for Stations A, B, C and W (top), and correlations with Stage 2 total score and CPS and SJT scores (bottom).**

Delay	ES residual	CS residual	CT&PS residual	PI residual	Station A residual	Station B residual	Station C residual	Writing station residual
Alternate forms reliability, r_{se}, across different numbers of years [mean (SD)]								
1 year	.187	.049	.009	.013	.048	.089	.059	.080
(n=4)	(.023)	(.016)	(.036)	(.044)	(.015)	(.022)	(.015)	(.062)
2 years	.141	-.015	.024	-.003	-.023	.080	.101	.035
(n=3)	(.061)	(.049)	(.093)	(.047)	(.105)	(.077)	(.005)	(.073)
3 years	.148	.013	-.013	-.047	.050	.010	.024	.033
(n=2)	(.023)	(.067)	(.008)	(.023)	(.010)	(.092)	(.067)	(.012)
4 years	.126	.040	.062	-.019	.209	.032	.144	.005
(n=1)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
Average[42]	**.150**	**.022**	**.021**	**-.014**	**.071**	**.053**	**.082**	**.037**
	(.022)	**(.033)**	**(.034)**	**(.029)**	**(.048)**	**(.023)**	**(.022)**	**(.023)**
Correlations with main selection variables, averaged across each of the five years								
Stage 2 score	.260	.083	.043	-.001	.152	.162	.144	.129
(n=5)	(.027)	(.035)	(.035)	(.023)	(.026)	(.063)	(.024)	(.027)
CPS	.208	.064	.037	-.010	.119	.118	.109	.113
(n=5)	(.029)	(.020)	(.027)	(.014)	(.011)	(.053)	(.017)	(.027)
SJT	.242	.081	.037	-.002	.145	.164	.141	.110
(n=5)	(.027)	(.044)	(.038)	(.025)	(.036)	(.063)	(.028)	(.023)

*Note that "(n=4)" etc refer to the number of years across which the correlations were averaged. The number of candidates on which the correlations are based can be found in Table 3.2.

[42] i.e. the average of the averages for the four delays of 1, 2, 3 and 4 years.

the other three measures. For ES, the SD was 1.49, and a range from -6 to +5, some candidates having relatively high or relatively low ES scores given their other three scores. It is clear therefore that there is variance between candidates, but the key question is whether that variance is systematic and reliable. Similar patterns were found for CS, CT&PS and PI, as well as for the station sub-scores.

The reliability of the residuals is a good test of whether there is independent variance associated with each of them, and hence whether each may independently be contributing something unique to the total score at Stage 3. Sub-score reliability can be assessed, as previously, by computing the alternate forms reliability of the sub-scores across different year intervals (from one year to five years). The assessments in different years will have different assessors, different simulators, and different scenarios, so that if there are reliable sub-scores which depend on candidates then the residuals should be correlated across the years. The EM algorithm was used to calculate expected correlations across years in a model which took into account range restriction, and included for each year the total Stage 2 score and the CPST and SJT scores, the total Stage 3 mark (i.e. sum of the ES, CS, CT&PS and PI totals), the four ES, CS, CT&PS and PI residuals, and the four station residuals.

Reliability of sub-scales and station scores across years. Table 3.11 shows the alternative forms reliability of the sub-score residuals, averaged across the four time-intervals (one year to four years). Of the four sub-scale residuals, only ES shows any evidence of being reliable, the remaining coefficients of stability and equivalence being close to zero. ES probably does have a small amount of reliable variance but with a coefficient of 0.15, it is unlikely to be of much practical significance.

The four stations show little evidence of having variance that is separable, the alternate forms reliabilities varying from 0.04 to 0.08, all of which are very small and barely different from zero.

The table also shows the correlations, within year (and averaged across the five years) for each residual with the Stage 2 scores. Of the sub-scales, only ES shows evidence of being correlated separately with Stage 2 scores, correlating 0.26 with the total score, and with a slightly higher correlation with the SJT (0.24) than with the CPS (0.21). The four stations all correlate weakly with Stage 2 scores to much the same extent, although the writing station shows the lowest correlations.

Taken overall these data suggest that **there is no evidence that the four separate stations are contributing specific variance to the Stage 3 total, and amongst the four sub-scales, it is only ES that shows any evidence of specific variance, and it is small and not particularly reliable.** Those conclusions have implications for any re-design of the Stage 3 assessments, since global scores at stations are probably as good in general as sub-scales, and the different station types perhaps do not need separating out in terms of increasing reliability. Of course including different station types and skills to ensure assessment against a blueprint is a different matter, and can be justified independently of the psychometric reliability of the components.

The low reliability of the separate skill and station scores has potential implications for the use of the algorithm and moderation process (see Chapter 2), which should not put undue emphasis upon individual scores or different profiles of scores, as most of the key information is probably in the total score.

9.9 SUMMARY

Claims have been made that the reliability of the GP selection system assessments is high. Previous estimates of reliability have relied on using Cronbach's alpha, which is inappropriate, particularly for OSCE-type assessments; this is because alpha is a relative, fixed-effects measure of internal reliability, so cannot be generalized to other questions or test occasions. Guidelines recommend the use of alternate-form reliability, which is an absolute, random-effects measure than is generalizable to other questions, test occasions etc. Our re-analyses of GP selection data from 2009 to 2015, particularly looking at candidates who take different forms of the test in different years, suggests that that **the reliability of the assessments may be lower than is desirable** for what is a high-stakes assessment. That has consequences both for **fairness and equity**, but also economically in terms of **cost-effectiveness**, both in additional training costs, and in increased mortality and morbidity.

The Stage 2 assessments are fairly reliable, and currently are not weighted very highly in the overall decision. Alternate forms reliability is of the order of 0.73 for CPST and about 0.57 for the rather shorter SJT. The numbers of items could probably be

increased and reliability improved. An important question concerns the extent to which CPST and SJT have different predictive validities, which will be considered in Chapter 4.

The Stage 3 (simulation) assessments have low reliability, with alternate forms reliability/generalizability of the order of 0.5 or less. Any improvement will not be easy without increasing the number of stations from the current four, which is low for OSCE-type assessments but may be typical for specialty selection; any substantial increase would be expensive. The alternate forms reliabilities are remarkably stable over time for both Stage 2 and Stage 3, as the coefficients are pretty consistent with delays of between one and six years.

Multifacet Rasch modelling using FACETS shows that both assessors and simulators (role-players/actors) differ in their level of stringency (hawkishness), and cases different in levels of difficulty; so all three factors contribute measurement error to Stage 3.

Analysis of the four subscales (ES, CS, CT&PS and PI) suggests that they are highly correlated, with only ES having some suggestion of being independent to a small extent. Likewise, the four stations (A, B, C and W) mostly behave in a very similar way.

We return to these important issues in the discussion (Chapter 11) to consider possible changes to the GP selection system.

APPENDIX 3.1
Correcting generalizability coefficients for resit candidates having range restriction

Appendix 3.1

Correcting generalizability coefficients for resit candidates having range restriction

The generalizability coefficients in Table 3.6 are based only on resit candidates, who are less able than typical candidates, meaning that their range is restricted, and the calculated coefficients are therefore not useful in practical terms as one wishes to be able to generalize to all candidates taking Round 1. A correction for range restriction is therefore required. Table 3.12 shows the theoretical basis for the calculations and provides a single example calculation for the 2011 data.

Generalizability calculations are usually carried out using the program GENOVA (Crick and Brennan, 1983), using the methods described by Brennan (2001b). Table 3.12 shows some of the key measures calculated by GENOVA when running an analysis. Rows 1 to 12 show the calculations for the data described in Table 3.5 for all candidates taking Round 1, with GENOVA names in the second column in upper case, and the symbolic notation used by Brennan (2001b) in the third column. The fourth column shows the formulae for calculating derived variables, in particular the generalizability coefficients, and the final column shows a numerical example for 2011, with the D Study based on 4 stations and a single occasion (i.e. the standard situation for Round 1).

The Generalizability Coefficient (G) is calculated from the Universe Score variance in row 3 (a measure of the variability between the persons, p) and the relative error variance (in row 4), the variance due to measurement error. G is the ratio of the Universe Score variance to the total variance, the formula and the result being shown in row 11. The universe score variance of 0.13201 (row 3) is larger than the relative error variance due to measurement error (row 4; 0.07785), so that G is 0.629 (row 10).

The square root of the relative error variance, in row 7, is the standard error of measurement corresponding to G (i.e. it is a relative measure), and it has a value of 0.27902, so that an individual candidate's score has a 1 SEM confidence interval of ± 0.279.

Row 5 shows the absolute error variance, and since there is only a single occasion in this model, it is the same as the relative error variance, so that Phi, the Dependability, in row 11 is the same as G, and the absolute SEM in row 8 is the same as the relative SEM.

Row 12, at the end of the calculations for the model calculated from only Round 1, shows Phi-Lambda, which is the generalizability for a particular cut-score, lambda (row 2), in relation to a particular mean (in row 1). Although the mean is typically the actual mean of the data it need not be and DBAR and LAMBDA in rows 1 and 2 can actually be any values, and need not correspond to the actual data, as will be useful below.

Calculating Phi-lambda is more complex than calculating Phi (see pp 48-49 of Brennan). Phi-Lambda takes into account the difference between the cut score and the mean, and from the formula in row 12, which requires a knowledge of the values in row 9, itself calculated from the values in row 6, the error variance for the mean (see Brennan pp. 44 and 48) it can be seen that as the cut score gets further from the mean so the generalizability increases[43]. For the overall mean of 3.228, and a pass mark (cut score) of 3.1, the dependability (Phi-lambda) is 0.69, a higher value than the phi of 0.63.

Similar calculations can be repeated for the generalizability analyses shown in Table 3.6. To make clear that all of the estimates in rows 13 to 24 are for the range restricted group who took Round 1 and Round 2, the symbols all have a superscript asterisk attached to them. The formulae are exactly equivalent except that the asterisked variables have replaced the non-asterisked variables. A notable difference from the earlier analysis in the results is that the relative and absolute error variances in rows 15 and 16 are now different, since data are collected on two occasions, and hence G and Phi in rows 22 and 23 are also different. Likewise the relative and absolute SEM differ, with absolute judgements in row 20 being less accurate than relative judgements. The generalizability coefficients are much lower than the values in rows 10 and 11, but they are of little interest, only applying to resit candidates.

A key thing to realise about the two analyses is that the variances due to measurement error in rows 15 and 16 (and hence also the SEMs) are higher when there are two occasions than when there is only a single occasion (in rows 4 and 5). Candidates are varying across occasions, probably because they are seeing different scenarios, different assessors and different simulators, and candidates will have changed over the intervening time; the marks in the two stations are therefore less accurate. Although the G and Phi coefficients in rows 22 and 23 cannot be compared with the coefficients in rows 10 and 11, the standard errors in rows 19 and 20 can be compared with the standard errors in rows 7 and 8. **The standard errors in rows 19 and 20 therefore correspond to the standard errors expected from an alternate forms reliability, whereas the standard errors in rows 7 and 8 correspond to a measure of internal consistency.** The only remaining challenge is to calculate G, Phi and Phi-Lambda from the analysis of Rounds 1 and 2 for the entire set of candidates taking Round 1.

The error variances in rows 16 and 17 are valid across any set of values, including the situation where candidates take only Round 1. Range-restriction corrected generalizability coefficients, shown as G**, Phi** and Lambda-Phi** in rows 25 to 27 can therefore be calculated by using the estimates of the universe variance from the Round 1 candidates (in row 3) with the estimates of measurement error taking occasions into account from rows 16 and 17 in the analysis of Round 1 and Round 2 resit candidates. The formulae for rows 25 and 26 therefore have asterisked values for error variances, but non-asterisked values for other values (so that, for instance, G** consists of .13201/(.13201+.09928) [from rows 3 and 16) giving G**=0.57. G** and Phi** are lower than G and Phi as they take the additional unreliability in the alternate forms design into account. An analogous calculation can be used for Phi-Lambda**, and it takes a value of 0.51, some what lower than the value of 0.69 in row 12 (which is an internal consistency generalizability) but a little higher than the Phi** value of 0.48.

Of course it is also possible to calculate values of as G**, Phi** and Lambda-Phi** for other years and other values of N_s' and N_o'.

[43] On p.49 Brennan shows in a graph that Phi-Lambda is at a minimum at the mean, and as the cut-score gets further from the mean so Phi-lambda increases. Less obvious is that although Phi-lambda is at a minimum at the mean, it is actually lower than Phi, the overall dependability, and it only becomes higher than Phi when lambda is sufficiently far from the mean. Phi and Phi-Lambda will be identical when the unbiased estimate of $(\mu-\lambda)2$ in row 9 is zero, which occurs when $(\mu-\lambda)2$ is the same as as the error variance for the mean in row 5. for the right-hand graph of table 2.12 in Brennan (p.49) a more accurate picture can be obtained if the value of Phi of 0.77 is also drawn horizontally across the graph. Phi-lambda is less than Phi in the range 0.44 to 0.67 and higher than Phi outside of that range. Phi can then be seen as a weighted average of possible values of Phi-lambda for differing values of lambda.

» Table 3.12: Generalizability coefficients for candidates sitting Round 1 and Round 2, corrected for restriction of range. Values in UPPER CASE are the names used in GENOVA output. Generalizability estimates are in bold.

	Estimates based on all candidates taking Round 1	Symbol	Formula/Notes	Example values (2011) N_S '=4 N_o'=1
(1)	Mean for Lambda(DBAR)	μ	Need not be actual mean	3.228
(2)	LAMBDA	λ	Can be any value	3.1
(3)	UNIVERSE SCORE VARIANCE	$\hat{\sigma}^2(p)$	Variance of persons	.13201
(4)	LOWER CASE DELTA VARIANCE	$\hat{\sigma}^2(\delta)$	Relative error variance	.07785
(5)	UPPER CASE DELTA VARIANCE	$\hat{\sigma}^2(\Delta)$	Absolute error variance	.07785
(6)	MEAN VARIANCE	$\hat{\sigma}^2(\overline{X})$	Error variance for the mean	.00005
(7)	LOWER CASE DELTA STANDARD DEVIATION	$\hat{\sigma}(\delta)$	Relative SEM	.27902
(8)	UPPER CASE DELTA STANDARD DEVIATION	$\hat{\sigma}(\Delta)$	Absolute SEM	.27902
(9)	Estimate of $(\mu\text{-}\lambda)^2$ (Brennan, 2001, Eq'n 2.55)	$(\widehat{\mu-\lambda})^2$	$(\mu-\lambda)^2 - \hat{\sigma}^2(\overline{X})$	.010633
(10)	GENERALIZABILITY COEFFICIENT [G]	$E\rho^2$	$\frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p)+\hat{\sigma}^2(\delta)}$	**.629**
(11)	PHI [Dependability]	$\widehat{\Phi}$	$\frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p)+\hat{\sigma}^2(\Delta)}$	**.629**
(12)	PHI (LAMBDA) [Based on DBAR and LAMBDA in rows1 and 12]	$\widehat{\Phi}(\lambda)$	$\frac{\hat{\sigma}^2(p)+(\widehat{\mu-\lambda})^2}{\hat{\sigma}^2(p)+\hat{\sigma}^2(\Delta)+(\widehat{\mu-\lambda})^2}$	**.691**
	Estimates based on all candidates taking both Round 1 and Round 2	Symbol	Formula/Notes	Example values (2011) N_S '=4 N_o'=1
(13)	Mean for Lambda (DBAR)	μ^*	Need not be actual mean	3.371
(14)	LAMBDA	λ^*	Can be any value	3.1
(15)	UNIVERSE SCORE VARIANCE	$\hat{\sigma}^2(p)^*$	Variance of persons	.06670
(16)	LOWER CASE DELTA VARIANCE	$\hat{\sigma}^2(\delta)^*$	Relative error variance	.09928
(17)	UPPER CASE DELTA VARIANCE	$\hat{\sigma}^2(\Delta)^*$	Absolute error variance	.14153
(18)	MEAN VARIANCE	$\hat{\sigma}^2(\overline{X})^*$	Error variance for the mean	.0426
(19)	LOWER CASE DELTA STANDARD DEVIATION	$\hat{\sigma}(\delta)^*$	Relative SEM	.31508
(20)	UPPER CASE DELTA STANDARD DEVIATION	$\hat{\sigma}(\Delta)^*$	Absolute SEM	.37620
(21)	Estimate of $(\mu\text{-}\lambda)^2$ (Brennan, 2001, Eq'n 2.55)	$[(\widehat{\mu-\lambda})^2]^*$	$(\mu^*-\lambda^*)^2 - \hat{\sigma}^2(\overline{X})^*$	-.02621
(22)	GENERALIZABILITY COEFFICIENT [G*]	$E\rho^{2*}$	$\frac{\hat{\sigma}^2(p)^*}{\hat{\sigma}^2(p)^*+\hat{\sigma}^2(\delta)^*}$	**.402**
(23)	PHI*	$\widehat{\Phi}^*$	$\frac{\hat{\sigma}^2(p)^*}{\hat{\sigma}^2(p)^*+\hat{\sigma}^2(\Delta)^*}$	**.320**
(24)	PHI (LAMBDA)* [Based on DBAR and LAMBDA in rows13 and 14]	$\widehat{\Phi}(\lambda)^*$	$\frac{\hat{\sigma}^2(p)^*+((\widehat{\mu-\lambda})^2)^*}{\hat{\sigma}^2(p)^*+\hat{\sigma}^2(\Delta)^*+((\widehat{\mu-\lambda})^2)^*}$	**.222**
	Generalizability coefficients calculated from Round 1 and Round 2 corrected for range restriction	Symbol	Formula/Notes	Example values (2011) N_S '=4 N_o'=1
(25)	GENERALIZABILITY COEFFICIENT [G**]	$E\rho^{2**}$	$\frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p)+\hat{\sigma}^2(\delta)^*}$	**.571**
(26)	PHI**	$\widehat{\Phi}^{**}$	$\frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p)+\hat{\sigma}^2(\Delta)^*}$	**.483**
(27)	PHI (LAMBDA)** [Based on DBAR and LAMBDA in rows1 and 2]	$\widehat{\Phi}(\lambda)^{**}$	$\frac{\hat{\sigma}^2(p)+(\widehat{\mu-\lambda})^2}{\hat{\sigma}^2(p)+\hat{\sigma}^2(\Delta)^*+(\widehat{\mu-\lambda})^2}$	**.512**

Chapter 4

Outcome measures and their relationship to selection measures

Chapter 4.

Outcome measures and their relationship to selection measures

4.1 INTRODUCTION

The purpose of selection is to predict outcomes, with the intention that those who were selected will have better outcomes than those who were not selected, the outcomes preferably being several years after selection has finished.

The validity of selection measures comes from demonstrating the relationship between selection scores and training outcome measures. The subtle but important point that training outcomes are only known in those who are selected means that demonstrating validity depends on making an inference about the probable performance of those who were not selected. In the first instance that is usually done by showing that selection measures and training outcomes are related in those who are selected, although more sophisticated methods allow an inference across all candidates.

This chapter will firstly describe each of the various outcome measures, and their inter-relations with one another, and will also relate each of the outcome measures to a set of selection measures, mainly within the selected group itself.

4.1.1 Outcome measures for training

Outcome measures for training are not all equivalent for many reasons. The measures may be assessing different abilities or skills, and may have different practical consequences. From the point of view of providing GPs for the NHS there is an argument that the most important **primary outcome measure** is the binary variable of **being on the GP Register**, as a trainee who does not get on the Register is not a GP, but money has been spent on training that doctor. For workforce planning, and an economically effective training system which takes into account the costs of training, **the time to get onto the GP Register** is also of practical importance. Doctors who are expected to enter the Register within three years[1] but do not do so, a) incur additional costs during training, and b) impose costs as a result of not providing service to the NHS as a GP. **The primary outcome measures have, therefore, to be entry to the GP Register and its timing.** Other **secondary outcome measures** of importance are ARCP and MRCGP results. Passing the **MRCGP CSA and AKT** are both necessary prior conditions for getting onto the GP Register[2]; they assess different skills and have good reliability[3], and although the ultimate outcome is binary (pass/fail), marks in the examination are continuous measures with greater statistical power for detailed analyses and prediction. Candidates are (currently) allowed up to four attempts at each of CSA and AKT, and **time to pass CSA and AKT** are therefore also important outcomes, as delay in passing either incurs costs for both training and service delivery. **Progression during training is assessed by ARCP**, which is carried out at least yearly during training. The various ARCP outcomes not only indicate the level of progression but also record factors such as being less than full time (LTFT) or out of programme (OOP). The professional behaviour of trainees is important both before and after qualifying as a GP, and can be indexed by **fitness to practice (FtP) issues as identified by the GMC and recorded on the LRMP**. Although a **tertiary outcome measure**, FtP issues are potentially particularly important as they have implications for effective and safe delivery of health care, and need therefore to be considered in relation to selection, particularly as FtP issues can be very expensive for the NHS.

[1] In general in our analyses we consider the target of getting onto the GP Register after a certain amount of training expressed as 'full-time equivalents' (FTEs). Doctors who are part-time or out of programme do not cost the system in any direct way, and to treat them as an expense would be to discriminate unfairly against those who are less than full-time or out of programme.

[2] The third component is based on Work Place Based Assessments (WPBAs) for which we were provided with no direct information, although they are implicit in ARCP.

[3] see: http://www.rcgp.org.uk/training-exams/mrcgp-exams-overview.aspx

4.1.2 Selection measures and selection endpoints

The selection measures used will consist of the **Total Stage 2 score** (and its separate component scores of **CPST** and **SJT**[4]), and the **Total Stage 3 score**[5] (i.e. the total score from the selection centre, not the final score which includes both the Stage 3 and Stage 2 scores). For the 2011 and subsequent cohorts we also report the scores on the four subscales of ES, CS, CT&PS and PI (for more details see Chapter 3). Although not strictly a selection method at present (although it could in principle be one for UK graduates), we also have information on the attainment of UK-trained doctors[6] during their undergraduate training, provided by UKFPO, and having separate assessments of **educational progress (EPM) during medical school, with separate components of decile, further degrees, and publications, and the FPAS Situational Judgement test** (which was taken in the final year at medical school). If the hypothesis of an academic backbone is correct (McManus, Woolf, Dacre et al., 2013), then better medical school performance should lead to better performance in GP selection tests, better performance in GP training, and hence better performance at MRCGP and ARCP[7]. We also describe a series of outcomes from selection itself which we call '**selection endpoints**', with the final endpoint being selected into training.

4.1.3 Demographic and other factors

There is plentiful evidence that performance during selection and in training are related to demographic factors such place of Primary Medical Qualification (PMQ), sex, and ethnicity, although the reasons for those differences are likely to be complex. We do not consider them in the present study since the outcomes of selection should be blind to such factors, in order that selection can be fair and at the level of the person.

4.2 SAMPLING FRAME

The present chapter is mainly concerned with the eventual outcomes in training of those selected for GP training, in terms of ARCP, MRCGP and being on the GP Register, and it relates those outcomes to the measures taken during selection at Stage 2 and Stage 3, as well as those available from FPAS. The chapter then asks how candidates who were **not** selected **might** have performed had they been selected in a different selection process, and that can form the basis for an economic analysis of the costs of various selection processes. The analysis has therefore to consider two main groups, those selected and those applying, and it has to consider data from several separate sources, including the selection process, ARCP, MRCGP and the GP Register, as well as FPAS. Those five processes collect data separately and there inevitably is relatively little in the way of internal consistency checking, and there is almost no consistency checking across the separate data sources. Some data are also collected 'live', so that not all variables in selection are necessarily updated as a result of later changes in a candidate's status. Likewise the separate ARCP assessments may be made by different people in different years and sometimes in different deaneries, and measures and outcomes may be inconsistent as a result. Some candidates and trainees also 'slip through the net' for various reasons. Applicants may not always have a GMC number at the time of application, and GMC numbers may not always be entered accurately into the system. The GMC number is however the sole item of data which allows linkage across the various databases, and therefore anomalies will inevitably occur. A more sophisticated system might be able to avoid them, but in a research project such as this which is working within a tight timescale it is inevitable that some inconsistencies will occur. Table 4.1 shows estimates of the numbers of applicants at each stage during the selection process in the years 2009 to 2015. The numbers have been discussed in Chapter 2, and the various problems discussed. For present purposes the most important column is that of 'Offer made and accepted' with 21,979 offers apparently having been

[4] For the 2009 and 2010 selection cohorts the SJT and CPST scores have been rescaled so that they are equivalent to the mean and SD of the scores from 2011 onwards.

[5] The 2009 and 2010 selection cohorts had a rather different selection centre from that used for 2011 onwards. The 2009 and 2010 total scores have been rescaled to be similar to those from 2011 onwards.

[6] FPAS is not of course available for most non-UK trained doctors (IMGs). Most of those IMGs applying for GP training will have taken the GMC's PLAB assessments (and we have discussed their performance on MRCGP elsewhere (McManus & Wakeford, 2014), and will also have a score on the language test, IELTS. Unfortunately the GMC was not able to provide us with data on PLAB and IELTS for the present study, despite a number of requests.

[7] Later, in the next chapter, we will discuss the limitations of FPAS data due to them only being available for the 2012 and 2013 cohorts of UK graduates as the doctors in these cohorts who entered GP training are likely to have taken AKT but they will not be taking CSA until 2016 or later.

accepted. 182 of the 21,979 accepted offers represent duplicates, having the same GMC number in different years/rounds, and removing them leaves 21,797 accepted offers. 131 accepted offers came from doctors in the selection database who did not have GMC numbers at the time of application[8], and removal of them leaves 21,666. Finally, 6 applications had GMC numbers which were not possible (e.g. five digits in length). Overall that means that there were **21,660 uniquely identifiable candidates who were made offers** (see the last column of Table 4.1), and these 21,660 candidates provide **the main sample** for the following analyses. It should also be noted that Year of Application and Year of Acceptance are often quite different with multiple applications, and for instance, of those applying first in 2009, 3,173 were accepted that year, with a further 628, 186, 102, 48, 40 and 0 being accepted from 2010 to 2015[9]. A result is that the total numbers here may appear quite different from those in other chapters which looked only at selection within individual years.

4.2.1 Applications, applicants and the handling of repeated applications

A problem in any linked dataset such as the present one is that applicants can apply on multiple occasions, and hence take selection tests on two or more occasions, but necessarily individuals undergo training only once. There are two possible ways of approaching this problem, which essentially are to analyse data at the level of the **applicant** or at the level of the application. Both have their advantages and disadvantages. **The present chapter will carry out the majority of its analyses at the level of the applicant** (i.e. the individual doctor) since the interest is mainly in how individuals perform at different stages of selection and training. In the next chapter, though, where the interest is mainly in selection and how it takes place, it is more appropriate that **analysis will be at the level of the application.** When there are only data for a single application cycle, as in the study of Patterson et al (Patterson, Kerrin, & Ashworth, 2015), analyses at the level of the applicant and the application are equivalent, but may give somewhat different results to the findings presented here.

Considering the issue in more detail, although the seven cohorts from 2009 to 2015 have 48,200 applications, those applications do not come from 48,200 separate doctors. Candidates, for multiple reasons not always related to failure, reapply in separate rounds or separate years[10]. That makes a statistical problem as most statistical analyses assume statistical independence of the individuals in the analysis. In the main part of this chapter[11] the compromise we have adopted is as follows: when a candidate applies on multiple occasions then for analyses of scores on the various selection measures the score is chosen at which the candidate reached the highest outcome level on the seven categories shown on Table 4.1, the most recent year and most recent round being chosen in the rare cases of there being a tie. There potentially is a **problem that candidates taking tests repeatedly may eventually do better because of chance fluctuations in their performance**, perhaps because of unreliability of measures, but despite that the approach is straightforward[12]. An inevitable consequence of using the highest attaining application is that better attaining candidates are somewhat over-represented, so that although 46% (21,979/48,200) of applications result in accepted offers, for 33,920 unique candidates the highest selection endpoint attained was that an offer was accepted by 64% (21,797/33,920). A further complication is that not all candidates are properly identifiable, mostly due

[8] The selection database is not updated when erroneous or missing data are present and more accurate information is available.

[9] The data are doubly truncated, which means that some candidates accepted in 2009 may have applied several or many years previously, and those rejected by 2015 may be accepted at some time in the future.

[10] Some of these re-applications are complex, and, for instance, there are candidates who applied and were accepted for training in 2009, undertook training, failed to get the MRCGP or had serious ARCP problems, and as a result dropped out of training, but then reapplied in 2015. Such candidates are entered twice in Table 4.1, once as accepted in 2009 and again as rejected at Stage 1 in 2015.

[11] Later, in carrying out the multiple imputations, we will carry out analysis at the level of the application rather than the applicant, since from the point of view of selectors it is the applications that matter within a year.

[12] A more complex analysis may be possible using multilevel modelling with repeated testing nested within individual applicants, but that has not been attempted within the short time window available.

[13] As discussed in another chapter, the 2011 Round 2 is problematic as the data file suggests that of 333 who had "demonstrated" after Stage 3, the Application Outcome was 'Offer Accepted' for 23, "Offer declined" for 1, and "Withdrawn" for 2. For the remainder the largest group was "Final Assessment Demonstrated" for 282 candidates, with another 25 "Offered". In other second rounds the vast majority offers are accepted. Linkage to other data suggests that many of those who not accepted an offer seem subsequently to have taken MRCGP or to have ARCP GP records. The "Offer Deanery" was Defence Medical Services in 10 cases, "Yorkshire and Humber" in 39 cases, and "Null" in the remaining 284 cases. Clearly there has been some form of error in entering the final destination of candidates into the raw data file. For present purposes the only feasible solution is to treat all of the cases who were made an offer as having accepted unless there is explicit evidence of the offer being declined or the candidate having withdrawn. The 2011 figures should therefore be treated with caution (and indeed that should perhaps be the case for the detailed numbers in any of the files…).

» *Table 4.1: Summary of numbers of applicants and destinations in each year, and the numbers of unique identifiable accepted candidates who can be followed through into training.*

Year	Selection endpoint Numbers of applicants at various stages during each selection year								Unique identifiable accepted candidates
	1. Applied and rejected Stage 1	2. Invited Stage 2 but withdrew	3. Failed Stage 2	4. Invited Stage 3 but withdrew	5. Failed Stage 3	6. Offer made but withdrew	7. Offer made and accepted	8. Total	
2009	727	328	601	563	1,254	393	**3,200**	7,066	**3,142**
2010	690	353	782	236	555	362	**3,359**	6,337	**3,320**
2011[13]	699	4	1,101	244	1,541	411	**3,034**	7,034	**2,984**
2012	693	0	1,238	213	1,563	507	**3,125**	7,339	**3,098**
2013	665	2	1,206	118	1,205	763	**3,225**	7,184	**3,193**
2014	663	12	1,130	83	1,135	650	**3,078**	6,751	**3,046**
2015	773	175	806	255	799	723	**2,958**	6,489	**2,877**
Total	4,910	874	6,864	1,712	8,052	3,809	**21,979**	48,200	**21,660**
Unique identifiable candidates	**2,015**	**453**	**2,647**	**852**	**3,068**	**2,825**	**21,660**	**33,520**	**-**

to not having GMC numbers at the time they applied for GP training. As a result they cannot be linked into other databases, since linkage is by GMC number. **Unique identifiable candidates** are candidates who are applying once only in Table 4.1[14], and have an acceptable GMC number. They are shown in bold, across the bottom of the table and in the right-most column. Most of the analyses which follow will be restricted to these individuals. To summarise, 48,200 applications were made for GP training of which 33,520 were from individually identifiable doctors, and of which 21,660 resulted in an offer of a training post being made and accepted. Some of those 21,660 doctors may on some other occasion have advanced less far through the system.

4.2.2 Outcome Categories in relation to selection measures.

As mentioned earlier, the main selection measures are the scores from Stage 2 and Stage 3, although FPAS can be treated as a selection score since it is prior to GP training.

Stage 2 and Stage 3 scores:

Stage 2 has a total score, which is the weighted sum of the scores on the CPST and SJT. The total score has had a single standardised score since 2009 and can be used unaltered. The scores on the CPST and SJT were scaled differently for 2009 and 2010 from the scaling in later selection rounds, and the 2009 and 2010 scores have therefore been rescaled to make them equivalent to those from 2011 onwards. A very few candidates are missing one of the total score, CPST or SJT, for various reasons, including scores which are clearly out of range. A single imputation has been used to replace these values to make the summary tables simpler[15]. The **Stage 3** scores are more complicated as the selection centre changed from 2011 onwards. The total Stage3 score was used for all years, with the 2009 and 2010 values being rescaled to make them compatible with 2011 onwards. Full sets of subscores for ES, CS, CT&PS and PI were only available for 2011 onwards, and for simplicity analyses for these measures are restricted to 2011 onwards. All candidates from 2011 onwards with Stage 3 measures had a full set of sub-scores, and therefore imputation was not necessary.

The means (SDs) of Stage 2 and Stage 3 scores for the various outcome categories are shown in Table 4.2 at the level of individual candidates, not applicants. The Ryan-Einot-Gabriel-Welsch Q statistic was used to assess homogenous subsets (i.e. pairs or groups of outcomes which are not significantly different), using $p>.001$ in view of the large sample sizes, so that pairs which are not homogenous can be regarded as different with $p<.001$.

Many of the results in Table 4.2 are not surprising. Selection at Stage 2 is on the Stage 2 scores and it is therefore unsurprising that outcome group 3, who failed Stage 2, have lower scores on the Stage 2 selection measures (and similarly for the performance on the Stage 3 measures for those failing Stage 3). Of more interest is that those invited to Stage 3 but withdrew had lower Stage 2 scores than those who subsequently were made offers. In contrast amongst those made offers, those who withdrew after receiving offers did not score differently at stage 2 ("6=7") but did score significantly lower at Stage 3, albeit only be about one point overall. **Withdrawal after receiving offers is not therefore losing particularly weak or strong candidates in terms of Stage 2 scores, and therefore cannot be predicted from Stage 2 selection scores.**

CPST and SJT show broadly similar patterns of scores. The residuals however show an interesting finding, in that those failing Stage 3 performed as relatively poorly on the SJT as did those failing Stage 2, but there was no such effect for the CPST residual. The implication is that **the SJT is a better predictor of Stage 3 performance than is the CPST.** This point will be returned to later.

FPAS scores:

The UKFPO FPAS scores have two components, an **Educational Performance Measure (EPM)**, which, since 2013, has consisted of three separate scores, **Decile of performance** within medical school, **Degree score** (based on intercalated and other degrees), and **Publication score** (based on peer-reviewed publications), and a separate **Situational Judgement Test score**, based on a

[14] A few candidates apply twice within a single year, but they are ignored for present purposes, their highest outcome being used for simplicity.

[15] Missing values varied from 10 to 18 out of 42,328. Imputation used a single iteration of the SPSS multiple imputation procedure with a fixed seed so that results were replicable.

» *Table 4.2 Mean (SD) score on the various selection measures of candidates in relation to the maximum level outcome level attained across all selection rounds. Note that whereas Table 4.1 is for applications, this table is for individual applicants.*

Selection scores	Selection endpoint: Maximum outcome level attained by individual candidates across selection rounds								Homogenous subsets (p>.001)
	1. Applied and rejected Stage 1	2. Invited Stage 2 but withdrew	3. Failed Stage 2	4. Invited Stage 3 but withdrew	5. Failed Stage 3	6. Offer made but withdrew	7. Offer made and accepted	8. Total	
N stage 2	*na*	*na*	*2647*	*852*	*2992*	*2824*	*21664*	*30979*	
Stage 2 total	na	na	424.1 (99.3)	506.9 (57.6)	461.8 (55.2)	522.4 (49.4)	522.1 (53.8)	507.5 (66.6)	6=7
Stage 2 CPST	na	na	212.7 (55.9)	252.4 (33.8)	235.1 (33.4)	260.6 (33.4)	260.4 (32.7)	253.7 (38.2)	6=7
Stage2 SJT	na	na	211.9 (56.5)	254.4 (33.0)	226.7 (31.9)	261.8 (28.3)	261.8 (30.4)	253.9 (37.3)	6=7
N Stage 3 total					*3068*	*2824*	*21664*	*27556*	
Stage 3 total	na	na	na	na	41.8 (5.83)	53.9 (4.83)	54.9 (4.99)	53.3 (6.52)	-
N stage 3 subscores (2011 onwards)					2345	2303	15200	19848	
Stage 3 ES	na	na	na	na	10.5 (1.77)	13.5 (1.35)	13.9 (1.36)	13.4 (1.78)	-
Stage 3 CS	na	na	na	na	10.4 (1.95)	13.5 (1.55)	13.9 (1.58)	13.4 (1.97)	-
Stage 3 CT&PS	na	na	na	na	10.4 (1.84)	13.3 (1.54)	13.6 (1.59)	13.2 (1.91)	-
Stage 3 PI	na	na	na	na	10.9 (1.87)	13.7 (1.56)	13.9 (1.56)	13.5 (1.86)	-

paper which currently has 70 questions (60 live and 10 pilot items) answered in 2 hours and 20 minutes (Patterson, Ashworth, & Good, 2015). Scoring changed somewhat in 2013 and 2012 scores have been equated to those of 2013[16].

FPAS data were available for UK doctors graduating from 2012 to 2015, and were collected as part of the selection programme for Foundation. Applications to GP selection are during F2, two years later, and hence only candidates applying for GP selection for 2014 or 2015 could have FPAS scores. For those applying for selection in 2014 there were 2,008 candidates with FPAS scores from 2012, and for selection in 2015 there were 2,663 doctors with FPAS scores, 1,774 from FPAS 2013, and 889 who had delayed application to selection or had been rejected in the previous year, with FPAS from 2012. Table 4.3 shows mean scores for the various selection groups. Note that the alpha level for this table has been set at a more liberal level of 0.05 since there are rather fewer data, and the table is more exploratory.

Table 4.3 shows that those made offers (columns 6 and 7) scored more highly than those who were rejected at Stage 1, Stage 2 or Stage 3 (columns 1, 3 and 5) on all of the EPM and SJT measures ($p<.001$ for a priori contrast), suggesting that GP selection is picking out more highly qualified applicants. Comparing the candidates who withdrew before Stage 3 (column 4) or after receiving an offer (column 6), there was no significant difference on any of the measures, suggesting that those withdrawing are equally able as those accepting offers.

The FPAS measures are important as candidates applying for GP selection can be compared with values for all FPAS candidates in 2012 and 2013 (including those who applied for GP). Table 4.3, in its final column, shows mean (SD) scores for all of the 14,478 FPAS applicants in 2012 and 2013 who had EPM scores, and the 7536 FPAS applicants in 2013 who took the SJT. Visually it is clear that that GP applicants overall have lower mean scores (column 8) than do UK graduates in general. The GP applicants are about 0.28 standard deviations below non-GP applicants for the EPM total score, and about 0.09 standard deviations below non-GP applicants for the FPAS SJT score[17]. These effects need replicating in further data.

Correlation between FPAS and Stage 2 and Stage 3 selection scores:

The measures assessed by FPAS, essentially academic attainment in medical school and an SJT score, collected about two or sometimes three, years before Stage 2 and Stage 3 performance are measured during GP selection. Table 4.4 shows the correlations between the measures in FPAS, Stage 2 and Stage 3[18]. The correlation, significance level and N in the top half of the table are raw correlations based on the actual data. Many of the data are, of course, missing, both for structural reasons (only those who pass Stage 2 go on to take Stage 3), and also because FPAS data only became available for 2012 and 2013 graduates, who would have applied to GP selection in 2014 and 2015. The correlations in parentheses at the bottom of each cell are therefore calculated using the MVA function in SPSS, using the EM algorithm to take missingness and bias into account, and find better estimates of the correlations taking the entire set of data into account.

The discussion here mainly concentrates on the FPAS EPM decile score and SJT scores (shown in blue in Table 4.4), the Stage 2 CPS and SJT scores (shown in green) and the Stage 3 total score (shown in purple). FPAS decile and FPAS SJT correlate only at a level of 0.40 (0.28 in the raw data), which may reflect the fact that the decile is calculated within medical schools, whereas SJT is calculated on a similar basis for candidates at all medical schools. The correlation of the Stage 2 CPST and Stage 2 SJT

[16] FPAS in 2012 used quartiles scored as 34,36, 38 and 40, whereas FPAS in 2013 onwards used deciles scored in integer steps from 34 to 43. FPAS 2012 quartiles were therefore given scores of 34.8, 37.2, 39.8 and 42.2 so that the scores for 2012 and 2013 are equivalent. The publication score was also changed so that it maximum of 2 rather than 5, but here we have left the publication scores for 2012 unchanged as then the mean and SD of the total score are equivalent across 2012 and 2013. It should also be noted that the SJT was only introduced in 2013, and so results are not available for 2012.

[17] It is not straightforward to test these differences using the data in Table 4.3, which are a guide only. However the Oriel data file in 2015 for all applicants to specialty training can be linked to the FPAS file. Of 14,110 applicants, 6,530 were for General Practice. 6,725 applicants had FPAS EPM scores and 4,564 had FPAS SJT scores. Applicants applying for GP had lower scores on EPM total, EPM deciles, EPM degree scores and EPM publication scores (all $p<.001$, effect sizes (Cohen's d)= -.284,- .208, -.217 and -.207), and lower scores on the FPAS SJT ($p=.002$, effect size = -.093) than those applying to specialties other than General Practice.

[18] Although we looked at the four subscores of Stage 3 (ES, CS, CT&PS and PI), their patterns of correlation with all of the measures were very similar to that shown by the total Stage 3 score, and for simplicity we have therefore omitted them here. In Chapter 3 we looked at the four sub-scores in more detail and concluded that, with the possible exception of ES, there was little evidence that they were psychometrically separable.

» *Table 4.3 Mean (SD) score on the EPM and SJT scores of FPAS in relation to the maximum level outcome level attained across all selection rounds. Note that whereas Table 4.1 is for applications, this table is for individual applicants.*

Selection scores	Selection endpoint: Maximum outcome level attained by individual candidates across selection rounds								Homogenous subsets (p>.05)	Mean (SD) for all FPAS candidates in 2012 & 2013 to all specialties
	1. Applied and rejected Stage 1	2. Invited Stage 2 but withdrew	3. Failed Stage 2	4. Invited Stage 3 but withdrew	5. Failed Stage 3	6. Offer made but withdrew	7. Offer made and accepted	8. Total		
N	*221*	*30*	*135*	*91*	*214*	*655*	*2898*	4244		*14478 (SJT 7536)*
FPAS EPM total score	40.06 (3.74)	39.14 (3.53)	39.44 (3.30)	40.47 (3.68)	40.41 (3.68)	41.02 (3.75)	40.50 (3.68)	40.45 (3.70)	1=2=3=4=5; 1=3=4=7; 1=4=6=7	40.77 (3.80)
FPAS EPM decile score	37.88 (2.79)	37.14 (2.55)	37.70 (2.55)	38.42 (2.74)	37.45 (2.62)	38.68 (2.68)	38.47 (2.81)	38.39 (2.79)	1=2=3=4=5; 1=3=4=7; 4=6=7	38.50 (2.81)
FPAS EPM degree score	1.86 (1.67)	1.77 (1.68)	1.51 (1.69)	1.67 (1.67)	1.66 (1.70)	1.95 (1.70)	1.76 (1.67)	1.78 (1.68)	1=2=3=4 =5=6=7	1.92 (1.68)
FPAS EPM publication score	1.15 (1.46)	.83 (1.21)	1.19 (1.57)	.86 (1.27)	.96 (1.43)	1.19 (1.48)	1.21 (1.52)	1.18 (1.50)	1=2=3=4 =5=6=7	2.00 (.63)
N	*100*	*17*	*42*	*53*	*102*	*296*	*1098*	*1708*		
FPS SJT score	39.96 (3.35)	40.98 (1.86)	38.99 (3.42)	39.72 (3.88)	39.20 (3.19)	40.84 (3.17)	40.66 (3.24)	40.50 (3.28)	1=3=4=5; 1=2=4=6=7	40.59 (.63)

» *Table 4.4. Pearson correlations between FPAS scores and the scores of Stage 2 and Stage 3 in the Main Sample. Correlations in parentheses indicate correlations corrected for missingness using the MVA algorithm. Key: ***p<.001; **p<.01; *p<.05; NS NS.*

Correlation N [EM correlation]	FPAS EPM total score	FPAS EPM decile score	FPAS EPM degree score	FPAS EPM publication score	FPAS SJT	Stage 2 total	Stage 2 CPS	Stage 2 SJT	Stage 3 total
FPAS EPM total score	1.	.862*** N=4244 (.886)	.646*** N=4244 (.640)	.464*** N=4244 (.455)	.258*** N=1708 (.373)	.451*** N=3993 (.580)	.468*** N=3993 (.586)	.242*** N=3993 (.435)	.148*** N=3767 (.317)
FPAS EPM decile score	.862*** N=4244 (.886)	1.	.199 *** N=4244 (.237)	.142 *** N=4244 (.161)	.281 *** N=1708 (.400)	.505 *** N=3993 (.633)	.532 *** N=3993 (.647)	.263 *** N=3993 (.466)	.145 *** N=3767 (.326)
FPAS EPM degree score	.646*** N=4244 (.640)	.199 *** N=4244 (.237)	1.	.690 *** N=4244 (.692)	.101 *** N=1708 (.142)	.132 *** N=3993 (.190)	.127 *** N=3993 (.181)	.082 *** N=3993 (.154)	.068 *** N=3767 (.131)
FPAS EPM publication score	.464*** N=4244 (.455)	.142 *** N=4244 (.161)	.690 *** N=4244 (.692)	1.	.032 NS N=1708 (.098)	.078 *** N=3993 (.111)	.086 *** N=3993 (.116)	.035 * N=3993 (.079)	.043 ** N=3767 (.076)
FPAS SJT	*.258**** *N=1708* *(.373)*	*.281 **** *N=1708* *(.400)*	*.101 **** *N=1708* *(.142)*	*.032 NS* *N=1708* *(.098)*	*1.*	*.390 **** *N=1591* *(.532)*	*.316 **** *N=1591* *(.458)*	*.324 **** *N=1591* *(.480)*	*.158 **** *N=1496* *(.339)*
Stage 2 total	.451*** N=3993 (.580)	.505 *** N=3993 (.633)	.127 *** N=3993 (.181)	.078 *** N=3993 (.111)	.390 *** N=1591 (.532)	1.	.883 *** N=31277 (.883)	.879 *** N=31270 (.879)	.428 *** N=27701 (.487)
Stage 2 CPS	.468*** N=3993 (.586)	.532 *** N=3993 (.647)	.082 *** N=3993 (.154)	.086 *** N=3993 (.116)	.316 *** N=1591 (.458)	.883 *** N=31277 (.883)	1.	.553 *** N=31270 (.552)	.334 *** N=27701 (.395)
Stage 2 SJT	.242*** N=3993 (.435)	.263 *** N=3993 (.466)	.055 ** N=3008 (.105)	.035 * N=3993 (.079)	.324 *** N=1591 (.480)	.879 *** N=31270 (.879)	.553 *** N=31270 (.552)	1.	.404 *** N=27700 (.465)
Stage 3 total	.148*** N=3767 (.317)	.145 *** N=3767 (.326)	.068 *** N=3767 (.131)	.043 ** N=3767 (.076)	.158 *** N=1496 (.339)	.428 *** N=27701 (.487)	.334 *** N=27701 (.395)	.404 *** N=27700 (.465)	1.

(and the measures can be regarded as broadly similar to the EPM decile and the SJT in FPAS) is 0.55, which is a little higher than the 0.40 for the FPAS measures. The FPAS decile and the Stage 2 CPST correlate 0.65, which suggests they are measuring related things, presumably medical knowledge. The FPAS SJT and Stage 2 SJT correlate somewhat less with a value of 0.48, a correlation similar to that of the 'cross-correlations' of FPAS decile and Stage 2 SJT of 0.47, and of FPAS SJT with Stage 2 CPST of 0.46. The two academic measures are likely to be stable, both measuring medical knowledge, and being based on a number of medical school assessments for the decile, and 100 or so items for the CPST. The SJTs are somewhat shorter (and hence less reliable) and there may also be less trait stability.

Finally, it should be noted that the Stage 3 total score correlates 0.40 and 0.47 with the Stage 2 CPST and SJT scores, and slightly less with the FPAS decile and SJT scores (0.33 and 0.34). In each case the SJT measure correlates a little more strongly with the Stage 3 score than does the knowledge measure. Considering that the Stage 2 scores are measured two to three years later than FPAS, the similar predictive validity of FPAS and Stage 2 scores suggests a moderate degree of stability in the skills being measured by Stage 3. It also raises the possibility of replacing Stage 2 with FPAS or similar scores.

The relationship between FPAS scores and Stage 2 and Stage 3 scores is important, partly for reasons of validation, and partly because FPAS scores may provide a more cost-effective method for specialty selection, at least for UK applicants (and perhaps for all applicants if a UK Medical Licensing Examination is introduced). That FPAS decile and Stage 2 CPST correlate quite highly (0.65) cross-validates each measure against the other. The measures were collected several years apart and use different methods, thereby suggesting they are measuring something robust, probably the attainment of clinical knowledge. The correlation of the SJTs in FPAS and Stage 2 (0.48) also helps to cross-validate the measures, although they are assessed using similar questions. SJTs are relatively new in postgraduate selection in the UK, and the correlation across time suggests that something consistent is being assessed, and that it is relatively independent of clinical knowledge. The relatively low but similar correlations of Stage 3 with FPAS decile and Stage 2 CPST (0.29 and 0.32) and the somewhat higher correlations of Stage 3 with the FPAS and Stage 2 SJTs (0.31 and 0.35) suggests that Stage 3 may be measuring something more to do with inter-personal communication and social judgement than with clinical knowledge, although the lowish correlations overall may simply be due to the low reliability of Stage 3 (see Chapter 3). The predictive validity of the FPAS, Stage 2 and Stage 3 measures is therefore of especial interest and will be discussed further.

4.3 OUTCOMES FROM SELECTION

4.3.1 Dropouts, delays and missed trainees

Not all candidates accepting offers will have entered training in the August after selection, and some of those apparently not accepting offers will end up in GP training. Estimating the proportions of missed trainees and dropouts is not easy and requires combining ARCP, MRCGP and GP Register data. ARCP is the most fallible of the data (particularly in 2010, the first year), and therefore when it is missing it will be triangulated against having taken MRCGP exams or being on the GP Register, as a doctor who has taken MRCGP or is on the Register must have received training, even if there are no ARCP records (unless they have come through the CEGPR route, about which we have no data).

There are various outcomes of training in relation to selection that need to be considered. All potentially have economic implications for health care. A broad grouping is as follows:

1. Applicants accepted for training and have accepted the offer
    - Acceptance only after two or more applications
    - Training delayed after acceptance
    - Dropout after acceptance but before training begins
2. Applicants who were never accepted but nevertheless underwent training (Lost trainees)

3. Applicants who began training but did not complete it

4.3.2 Applicants accepted for training only after two or more applications.

Not all doctors accepted to GP training are accepted in the first year that they apply. Table 4.5 shows the first year in which a doctor applied for GP training and the year in which they were eventually accepted. Note that there is double censorship in the table as some of those applying in, say, 2009, may have applied previously but there is no record of that as records only begin in 2009. Likewise, some of those applying in 2015 and being rejected will probably reapply in later years. The top row of Table 4.5 probably gives the best picture of repeated application and eventual acceptance. Of doctors who eventually enter GP training, only about three-quarters will have been accepted in the year of their first application, a further 10 to 15% will be accepted the year after, about 5% will be accepted two years later, and the remaining 5% or so accepted from three to six years later, with some perhaps being accepted even later.

Table 4.6 explores the scores at Stage 2, Stage 3 and FPAS of doctors who enter GP training after a delay from their first application. The overall pattern is relatively straightforward, doctors with no delay in entering training have higher scores on all of the Stage 2 and Stage 3 measures[19]. However those delaying 1, 2 or 3 years mostly did not differ in their Stage 2 and Stage 3 scores, which were all lower than those with no delay. The pattern is similar to that which occurs when candidates are allowed to resit an assessment as in many postgraduate examinations (McManus & Ludka, 2012). Candidates who pass the exam on their first attempt have a wide range of marks, some at or just above the pass mark, and others high above the pass mark. Candidates who fail the first attempt though tend on their second or later attempt only just to be above the pass mark, giving them a lower average score at the passing attempt. That is probably the picture seen here.

Table 4.6 also shows FPAS EPM scores for candidates who enter GP training with or without a delay. Once again, candidates who are delayed in entering GP training have lower scores on the total score and the decile score, suggesting that candidates who only enter GP training on a later attempt are academically less able. There was no difference between the groups on the EPM degree score. Somewhat surprisingly the EPM publication score was higher in candidates who were delayed in entering training. No immediate explanation for that is available, but it may indicate other mechanism are also involved[20].

Summarising, **doctors who enter GP training in a later year than their first application have lower selection and FPAS scores, and it is possible therefore that they will encounter additional problems in further professional assessments.**

4.3.3 Applicants who accept offers but who dropout before training begins

Any doctor entering GP training should have ARCP records for their general practice training, although only the 2009 to 2013 trainee cohorts have ARCP results. The 2009 cohort has five years of ARCP records, the 2010 cohort has four years, and the 2011, 2012 and 2013 cohorts have three, two and one year of records. The data for the later cohorts are therefore right-censored. Values are shown for all cohorts, but a '-' is entered where data cannot be calculated or are unreliable.

Table 4.7 shows the numbers of selected trainees for whom ARCP GP records were found. Trainees in the Defence Medical Services mostly did not have ARCP records (with a handful of exceptions). Data for the 2014 and 2015 cohorts are, as expected, almost completely missing, and the 2013 cohort also has quite a few missing cases, perhaps because of either dropouts or delays in starting training. For the 12488 non-military trainees accepting places in the 2009, 10, 11 and 12 cohorts, there were 345 (2.8%) doctors without ARCP records. These doctors are probably genuine dropouts in most cases as only 8 (2.3%) of these 345 trainees showed any other evidence of GP training in the form having taken either part of MRCGP[21]. We can therefore assume that **about 2.5% of accepted trainees do not continue into GP training.**

[19] In Table 4.6, as in other tables, groups are shown which are statistically homogenous. If two groups are not homogenous then they are significantly different with p<.001. For instance, the Stage 2 total scores are homogenous for groups 1, 2 and 3+, indicating that groups 1 and 2, groups 2 and 3+, and groups 1 and 3+ are not significantly different with p<.001. However group 0 (no delay) is significantly different from group 1, group 2 and group 3.

[20] One possibility is that candidates with peer-reviewed publications were more likely to turn down a training place in order that they could have a second attempt at entering hospital medicine or some other specialty which is perceived as more academic.

[21] The power of that analysis can be confirmed for the 2013 cohort, where 393 cases in year 2013 in Table 4.7 appear to have no ARCP GP records, but 148 of them in fact have taken MRCGP AKT, suggesting that the problem is with the ARCP records.

» *Table 4.5: Year of acceptance into GP training in relation to Year of first known application to GP training. Percentages are of all those first applying in a particular year who are accepted in the various years.*

Year first applied to GP training	Year Selected to GP training							Total
	2009	2010	2011	2012	2013	2014	2015	
2009	*3181*	*632*	*187*	*103*	*48*	*42*	*29*	*4222*
	75.3%	*15.0%*	*4.4%*	*2.4%*	*1.1%*	*1.0%*	*0.7%*	
2010	-	2702	283	153	70	42	44	3294
		82.0%	8.6%	4.6%	2.1%	1.3%	1.3%	
2011	-	-	2529	267	169	72	43	3080
			82.1%	8.7%	5.5%	2.3%	1.4%	
2012	-	-	-	2565	353	143	67	3128
				82.0%	11.3%	4.6%	2.1%	
2013	-	-	-	-	*2533*	*396*	*149*	*3078*
					82.3%	*12.9%*	*4.8%*	
2014	-	-	-	-	-	2314	408	2722
						85.0%	15.0%	
2015	-	-	-	-	-	-	2136	2136
							100.0%	
Total	3181	3334	2999	3088	3173	3009	2876	21660

» *Table 4.6 Stage 2, Stage 3 and FPAS scores of GP trainees who were accepted into training after a delay of 0,1,2 or 3+ years from first application to GP selection. Note that trainees who had taken FPAS could have been delayed a maximum of one year because of the relatively recent introduction of FPAS in 2012, and that no delayed entrants could have taken FPAS SJT which was introduced in 2013.*

Year first applied to GP training	Delay between first application and acceptance on GP training				Total	Homo-genous subsets (p>.001)
	0 years	1 year	2 years	3+years		
N stage 2	*17960*	*2339*	*801*	*560*	*21660*	
Stage 2 total	526.4 (51.80)	500.8 (56.8)	506.8 (61.3)	498.2 (59.2)	522.1 (53.8)	1=2=3+
Stage 2 CPST	262.2 (32.0)	250.6 (34.3)	254.5 (35.6)	251.2 (33.6)	260.4 (32.7)	1=2=3+
Stage2 SJT	264.2 (29.2)	250.1 (32.3)	252.3 (34.1)	246.4 (35.8)	261.8 (30.8)	1=2=3+
N Stage 3	*17960*	*2339*	*801*	*560*	*21660*	
Stage 3 total	55.1 (4.92)	53.5 (5.16)	53.94 (5.01)	53.05 (5.01)	54.85 (4.99)	1=2=3+
N Stage 3 subscores (2011 onwards)	12123	1714	801	560	15198	
Stage 3 ES	13.94 (1.34)	13.71 (1.34)	13.64 (1.38)	13.29 (1.37)	13.88 (1.35)	1=2=3+
Stage 3 CS	13.94 (1.66)	13.62 (1.58)	13.43 (1.65)	13.18 (1.65)	13.85 (1.58)	1=2
Stage 3 CT&PS	13.65 (1.58)	13.40 (1.59)	13.28 (1.62)	13.16 (1.57)	13.58 (1.59)	1=2; 2=3+
Stage 3 PI	13.96 (1.55)	13.68 (1.55)	13.59 (1.62)	13.41 (1.52)	13.89 (1.56)	1=2=3+
N	2715	183	-	-	2898	
FPAS EPM total score	40.57 (3.69)	39.44 (3.46)	n/a	n/a	40.50 (3.68)	-
FPAS EPM decile score	38.53 (2.81)	37.62 (2.68)	n/a	n/a	38.47 (2.81)	-
FPAS EPM degree score	1.77 (1.67)	1.57 (1.60)	n/a	n/a	1.76 (1.67)	0=1
FPAS EPM publication score	1.18 (1.51)	1.57 (1.60)	n/a	n/a	1.21 (1.52)	-
N	-	-	-	-	-	
FPS SJT score	n/a	n/a	n/a	n/a	n/a	

[22] We are aware that there are very minor anomalies in this table, so that, for instance, for 2009 GP trainees, there are 3050 with ARCP records found, but the total on the right-hand side is 3046. Of the chasing of such inconsistencies there is no ending, not least as the databases are not entirely consistent.

» *Table 4.7: Estimation of dropout and delay rates in the various cohorts.*

Cohort	ARCP GP records found				Year of first ARCP GP record found (percent of "Any Found")[22]				
	Not Found (Defence)	Not Found (other)	Any Found	Total	2009	2010	2011	2012	2013
2009	28 (0.9%)	103 (3.2%)	3050 (95.9%)	3181	2264 (74.3%)	728 (23.9%)	49 (1.6%)	3 (0.1%)	2 (0.1%)
2010	27 (0.8%)	59 (1.8%)	3248 (97.4%)	3353	-	3033 (94.2%)	157 (4.9%)	26 (0.8%)	4 (0.1%)
2011	36 (1.2%)	119 (4.0%)	2844 (94.8%)	2999	-	-	2697 (94.0%)	118 (4.1%)	13 (0.5%)
2012	23 (0.7%)	64 (2.1%)	3001 (97.2%)	3088	-	-	-	2859 (95.7%)	126 (4.2%)
2013	29 (0.9%)	383 (12.1%)	2761 (87.0%)	3173	-	-	-	-	2718 (100%)
2014	18 (0.6%)	2981 (99.1%)	10 (0.3%)	3009	-	-	-	-	-
2015	0 (0%)	2868 (99.7%)	8 (0.3%)	2876	-	-	-	-	-
Total	161	6577	14922	21660					

4.3.4 Applicants accepted for training but training delayed after acceptance

There are many potential reasons why doctors may not enter GP training in the year that they first apply. Table 4.7 gives the training year in which trainees with ARCP records are first found with a GP record. The 2009 cohort records are different from others, and may well be inaccurate as that was the first ARCP year. However the 2010, 2011 and 2012 cohorts shows a consistent picture with 4.9%, 4.1% and 4.2% delaying one year (mean=4.4%), and 0.8% and 0.5% (mean=0.7%) delaying two years, with a very tiny proportion delaying longer than that. We can therefore assume that **about 5% of accepted trainees begin their GP training a year after the year in which they have been selected.**

4.3.5 Lost trainees: applicants who were never accepted but nevertheless underwent training

Given the state of the various databases, it is possible that there are some doctors who have not apparently been selected for training or been through it and yet who have actually been trained. The numbers of such individuals can be assessed by considering the group of applicants who have not been accepted into training. Of 33,445 individual doctors who had applied for GP training, 21643 were accepted, leaving 11802 who were not accepted according to the records. However 295 of these individuals had either taken a part of the MRCGP at some time (AKT: 239; CSA 206), had ARCP records from GP (258), or were on the GP Register (177). These doctors seemed to be concentrated in GP selection during 2011 (n=178) which suggests that there may have been administrative errors in that year. In practical terms, a conclusion is probably that there are somewhat more GPs being trained than the conventional records from GP selection might indicate.

In summary, **about 2.5% of doctors accepted as trainees do not appear to enter training programmes, about 5% enter programmes a year later than expected, and there is an additional 2% of doctors who despite apparently not having been selected have progressed through training programmes.**

4.4 OUTCOMES FROM TRAINING

GP training and assessment in principle are straightforward, but in practice are much more complex. A trainee who is accepted onto a GP training programme by a Deanery/LETB will, all things going well, be a trainee for three years, during each year of which their progress will be monitored by the **Annual Review of Competence Progression (ARCP)** process. After one year or so trainees will take the **MRCGP Applied Knowledge Test (AKT)**, a multiple-choice assessment of clinical knowledge, and after two or so years they will take the **MRCGP Clinical Skills Assessment (CSA)**, which is an OSCE-like assessment with thirteen stations in the form of a simulated surgery, candidates being observed by an examiner while interacting with a trained role-player. Currently, AKT and CSA can each be taken a maximum of four times. When both parts are passed, and assuming that **work-placed based assessments (WPBAs)** are also completed satisfactorily (as indicated by a satisfactory ARCP outcome) then the doctor can be placed on the **GP Register of the General Medical Council's LRMP (List of Registered Medical Practitioners).** Being on the GP Register is one of the requirements for a doctor being on the **National Performers List**, which are maintained by the Area Teams of NHS England, the Health Board in Wales, Local Medical Committees in Scotland and the HSC Business Services Organisation in Northern Ireland[23].

Although the process is straightforward for a typical trainee, there are a number of complications for some trainees. Firstly, trainees may spend some of their training working **Less Than Full-Time (LTFT)**, sometimes but not always for childcare[24] or other personal considerations, and mostly but not always as 60% of full-time (i.e. three working days). Secondly, trainees

[23] We do not consider the Performers Lists further in this document.

[24] Maternity/paternity leave is strictly not considered as a part of being OOP (which is mostly restricted to training purposes, although exceptional family and childcare responsibilities can be included under OOPC, with C for Childcare), and while presumably it can result in a trainee being LTFT, when maternity leave results in the trainee not being present at all it presumably is 'statutory leave', during which time ARCP outcomes should not be issued (The Gold Guide, 2014, p.70 para 7.74). A consequence for the present study is that it is not clear that there is any definitive evidence for time not spent in training due to full-time maternity/paternity leave, the ARCP records seeming to be blind to statutory leave. More problematic still is that time out of training (TOOT) can only be inferred from a difference between the start and end dates of the first and last ARCP records available and the cumulative time in training accounted for by those records. ARCP records are however often imprecise in their timings. An exploratory analysis in late December 2015 found substantial differences in estimated TOOT between Deaneries/LETBs, with some having 10% of trainees with 24+ weeks of TOOT, whereas others had almost none. Although the larger,

may be given permission to be **Out-of-Programme (OOP)**, which may occur for a host of reasons (see footnote 24). Neither process is well recorded within the computerised records, and often there is no indication of the percentage of time LTFT, or of the reasons for LTFT or OOP.

The major outcomes from General Practice training for present purposes are therefore:

- ARCP outcomes
- MRCGP outcomes
- Entering the GMC's GP Register
- Fitness to Practice issues

We will consider each in turn and their inter-relationships, and later we will look at how they relate to the selection measures described earlier. ARCP will be considered first, since it occurs from the beginning of training onwards, then it is convenient to look at entry to the GP Register, we will then consider MRCGP and then briefly look at Fitness to Practice issues. For each outcome we will also assess the relationship to the selection measures in Stage 2 and Stage 3.

4.4.1 ARCP outcomes, including LTFT and OOP

ARCP data are complicated[25]. We were provided with two files one for England and one for the other countries of the UK. The English file contained 240,957 records for the ARCP years 2010 (37,697), 2011 (40,386), 2012 (44,463), 2013 (57,807) and 2014 (60,244) and had data on all trainees in all specialties. Although candidates typically have one record per year, some have multiple records in a year for different time periods, so that a few candidates had as many as 18 records. The 2010 trainees included those who were far advanced in their training, and hence there were no records for their earlier training careers (i.e. the data are left-truncated). For candidates in 2013 and 2014 there were also ARCP records for Foundation training. Each record was classified according to the Programme Specialty, and many post-foundation candidates had records from two or more specialties. The non-English data contained 9119 records (2010: 1,572; 2011: 2,027; 2012: 1,760; 2013: 1,807; 2014: 1,953), consisting only of records for which the training programme was designated as General Practice. Analyses of movements between specialties are therefore only possible for the English data. All records had a GMC number, and therefore the multiple records for individual trainees could be combined. Overall data were available for 82,354 doctors. For most purposes the data could be reduced to consider only those on GP training programmes, but in the chapter on choosing a career in GP we also used the English data to examine movement of trainees between different training programmes.

Information available in the ARCP data as supplied included for each record:

- The ARCP year, the submitting Deanery, the Programme Specialty, the Training Level, and flags to indicate Academic Trainees and Military Trainees
- The date of the ARCP review, and the beginning and the end date to which the review applied
- The outcome of the ARCP (or its equivalent in RITA) expressed as 1, 2, 3, 4, 5, 6, 7.1, 7.2, 7.3, 7.4, 8, 9, C, D, E, F or G. A code was also available for Out of Programme (OOP)

more recent Deaneries/LETBs appear to have fairly consistent and high rates of TOOT, earlier Deaneries/LETBs are much more variable, and they are the ones who are providing the majority of data for the 2009 and 2010 cohorts of entrants to GP training. Ultimately this problem is one that cannot be addressed in the present study given the quality of the data in ARCP and the report deadline. We have therefore continued on the basis that OOP and LTFT provide evidence of reduced numbers of FTEs when compared with calendar years, and, as ever, can only counsel caution when interpreting numbers of trainees not reaching the GP Register within specific numbers of FTE years.

[25] The data we received had been pre-processed by Daniel Smith at the GMC, and he had already prepared a number of derived variables which were of use, and he had also cleaned up a number of inconsistencies in the data. We are grateful to him for his help on understanding and interpreting the ARCP data, and for his extensive work on the data.

- A derived score looked at unsatisfactory outcomes, and separate codes were also available for unsatisfactory outcomes which were exam-based or non-exam-based

- A flag was also available to indicate that a trainee was less than full time (LTFT), although no indication was available of the proportion of hours worked

The major interest for the present analyses is in the time taken by a GP trainee at each stage of their training (ST1, ST2, ST3) and these might be covered by multiple ARCP records (although in the simplest case there will be a single year-long record for each training stage). Records were therefore aggregated by training stage, and the total time in, say, ST1 or ST2, was then calculated across all records for that training level. Similarly a flag was created if any record at a level was OOP or LTFT. The most important outcome at each level is the overall ARCP outcome and we expressed that using a modification of the scale developed by Tiffin and his colleagues (Tiffin, Illing, Kasim et al., 2014), which is ordinal and indicates the relative severity of ARCP problems:

The 'Modified Tiffin score'

The Modified Tiffin (mTiffin) score is calculated separately at each training level as the maximum for any of the records at that level. The Tiffin scheme does not separate an outcome of particular importance for GP training which is "ARCP Outcome 4" ("Released from Training Programme"). Our modification therefore splits Tiffin et al's stage 4 into Modified Tiffin 4 (Extended training time required) and Modified Tiffin 5 (ARCP Outcome 4). The scale is therefore:

- Modified Tiffin score 1: Satisfactory progression

- Modified Tiffin score 2: Insufficient evidence presented

- Modified Tiffin score 3: Targeted training required (but training time not extended)

- Modified Tiffin score 4: Extended training time required

- Modified Tiffin score 5: Left programme (i.e. ARCP outcome 4)

4.4.2 Data integrity for ARCP

The ARCP data are far from 'clean', having numerous errors of various sorts[26]. As an example, some checking was carried out to assess whether candidates in the 2009 and 2010 cohorts had ARCP outcomes for ST1, ST2 and ST3, which most should have had. Of 6,466 trainees in those cohorts, 4,678 (72.3%) had data from all three training levels, 1,292 (20.0%) from two training levels, 269 (4.2%) from one training level, and 227 (3.5%) from no training levels[27]. Clearly the 167 trainees with an ST1 and ST3 record but no ST2 record should have an ST2 record (and as a specific single example, candidate 2420 in the sequential datafile had ST1 and ST3 outcomes but no ST2 outcome, with detailed checking of the raw data suggesting that ST2 had been mistyped as ST3). Given the time window of the present study, and the lack of access to detailed data from Deaneries/LETBs, it is probably impossible to resolve most such issues, and no attempt has been made. Instead where data are missing an attempt will be made to impute them at a later stage (see below). Unless such errors are systematic they are unlikely seriously to distort the present findings.

4.4.3 Academic and military trainees

The ARCP records include flags for being an academic or military trainee, and although military flags are present for 3.9%, 3.9% and 3.3% of the 2009 to 2011 trainees it seems probable that most of these are recorded in error and they will not be considered further here. Academic flags are present for 1.4%, 0.7%, 0.8%, 0.7% and 1.0% of the 2009 to 2013 trainees[28].

[26] We are told that there is little data checking for ARCP, even within records entered within a single year, so that end dates can be before start dates, or years can be erroneous. Data checking across years inevitably would be much harder.

[27] Of course a few of the missing cases will represent trainees who left training in ST1 or ST2 but these are probably not the vast majority of missing cases.

4.4.4 Modified Tiffin Scores

Table 4.8 shows the number of trainees at each mTiffin score in relation to training level and cohort. The table needs interpreting with some care.

» *Table 4.8: ARCP outcome at ST1, ST2 and ST3 for 2009 to 2013 training cohorts.*

Level Training cohort	Modified Tiffin score					Total
	1: Satisfactory progression	2: Insufficient evidence presented	3: Targeted training required (not extended)	4: Extended training time required	5: Left programme (i.e. ARCP outcome 4)	
ST1						
2009	2207 (88.8%)	179 (7.2%)	73 (2.9%)	14 (0.6%)	17 (0.7%)	2,490
2010	2671 (86.9%)	233 (7.6%)	117 (3.8%)	28 (0.9%)	26 (0.8%)	3,075
2011	2457 (89.2%)	178 (6.5%)	97 (3.5%)	19 (0.7%)	4 (0.1%)	2,755
2012	2433 (85.5%)	292 (10.3%)	95 (3.3%)	20 (0.7%)	6 (0.2%)	2,846
(2013)	2304 (85.4%)	277 (10.3%)	95 (3.5%)	20 (0.7%)	1 (0.03%)	2,697
Average pct (exc 2013)	87.1%	8.4%	3.4%	0.7%	0.4%	
ST2						
2009	2422 (84.1%)	255 (8.9%)	150 (5.2%)	39 (1.4%)	15 (0.5%)	2,881
2010	2589 (85.6%)	236 (7.8%)	143 (4.7%)	51 (1.7%)	7 (0.2%)	3,026
2011	2130 (84.7%)	249 (9.9%)	119 (4.7%)	14 (0.6%)	3 (0.1%)	2,515
(2012)	1927 (83.5%)	247 (10.7%)	113 (4.9%)	18 (0.8%)	3 (0.1%)	2,308
Average pct (exc 2012)	84.5%	9.2%	4.9%	1.1%	0.3%	
ST3						
2009	1905 (69.6%)	284 (10.4%)	75 (2.7%)	330 (12.0%)	145 (5.3%)	2,739
2010	1905 (69.0%)	361 (13.1%)	86 (3.1%)	299 (10.8%)	111 (4.0%)	2,761
(2011)	1385 (72.0%)	328 (17.1%)	59 (3.1%)	146 (7.6%)	5 (0.3%)	1,923

Not all training levels are available for all cohorts in Table 4.8. Consider ST1 which is the first year of specialist training. A candidate in the 2009 cohort who is progressing satisfactorily will have a satisfactory result in the 2010 ARCP. However if that candidate requires extra time then the full details of the eventual outcome, which may include leaving the programme, will only be available in the 2011 ARCP[29]. ARCP results are available for 2010, 2011, 2012, 2013 and 2014[30]. For the 2009, 2010, 2011 and 2012 cohorts there will therefore be almost complete ST1 records available (i.e. ST1 in the first and a possible

[28] The reasons for the missing flags in the later cohorts of trainees are unclear.

[29] Trainees can have up to six months or in exceptional circumstances a year of additional training. That year does not include time spent OOP or LTFT, and that has not been taken into account in the table shown. However all available ST1 records, for the example given, will have been included.

[30] ARCP for 2015 are not likely to be available before November or December 2015.

second year). However the 2013 cohort will have its first records for ST1 in the 2014 ARCP, but there is no information on those having an extension, and therefore the 2013 cohort is in parentheses in the table to indicate that results are provisional and are probably under-estimates of the most serious outcomes. For ST2 and ST3 the consequence of having only five years of ARCP results is that there are only three and two cohorts who have relatively complete data. The problem can be seen in ST3, where there are many more trainees receiving a Modified Tiffin score of 4 or 5 in the 2009 and 2010 cohorts, having reached the end of their training programme, than there are in the 2011 cohort, some of the latter presumably having extensions, and the final outcome not yet being clear.

Despite the problems for some records, the ARCP provides a good first estimate of the proportions of trainees who progress satisfactorily at each training level, or who have problems at different levels, as shown in Table 4.8.

4.4.5 Modified Tiffin outcomes in relation to selection measures

Table 4.9 shows performance of trainees with the various Year 3 Tiffin outcome categories in relation to performance at Stage 2 and Stage 3 selection measures. All predictors were highly significant ($p<.001$). The subscores for Stage 3 have been omitted for simplicity. Overall it can be seen that mTiffin outcome 5 (ARCP outcome 4) have the lowest overall on Stage 2 and Stage 3 measures. mTiffin 1 and 2 (Satisfactory progression vs Insufficient evidence) show no evidence of differences on any of the scores. mTiffin 3 perform less well than 1 and 2 on all measures, and mTiffin 4 are generally worse than mTiffin 3, but better than mTiffin 5. Overall it is clear that **lower scores on the various selection measures are related to worse ARCP outcomes.**

An important question is whether Stage 3 predicts mTiffin Scores after taking the Stage 2 scores into account. A one-way ANCOVA with the Stage 2 scores as covariates, found a significant predictive effect of Stage 3 scores after taking Stage 2 scores into account ($p<.001$; $\eta^2 = .018$) although the variance accounted for was far less than when the covariates were not present ($\eta^2 = .078$). In contrast, Stage 2 total on its own accounted for substantially more variance ($\eta^2 = .229$), which was reduced only somewhat by including Stage 3 total as a covariate ($\eta^2 = .179$). The implication is that **most of the prediction of ARCP outcomes is by Stage 2, although Stage 3 contributes a little additional variance.**

» *Table 4.9 Mean (SD) score on the various selection measures of trainees in relation to the mTiffin Outcome Score at ST3*

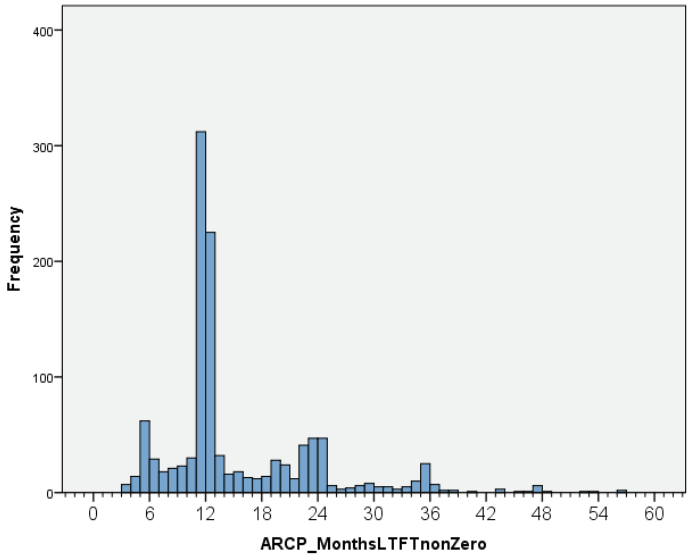
	1: Satisfactory progression	2: Insufficient evidence presented	3: Targeted training required (not extended)	4: Extended training time required	5: Left programme i.e. ARCP outcome 4	Homo-genous subsets (p>.001)
N (total=7446)	5212	977	221	776	260	
Stage 2 total	532.7 (49.2)	530.8 (46.3)	482.7 (57.1)	462.9 (48.8)	443.7 (45.2)	1=2
Stage 2 CPST	266.1 (31.1)	265.0 (29.6)	237.5 (37.0)	227.3 (31.9)	216.7 (31.6)	1=2, 3=4
Stage2 SJT	266.6 (27.2)	265.8 (26.5)	245.3 (28.9)	235.6 (29.6)	227.0 (28.4)	1=2
Stage 3 total	54.9 (5.01)	54.41 (5.11)	52.3 (5.00)	51.2 (4.72)	49.9 (4.54)	1=2, 3=4, 4=5

4.4.6 Less than full-time (LTFT) and Out-of-Programme (OOP) trainees

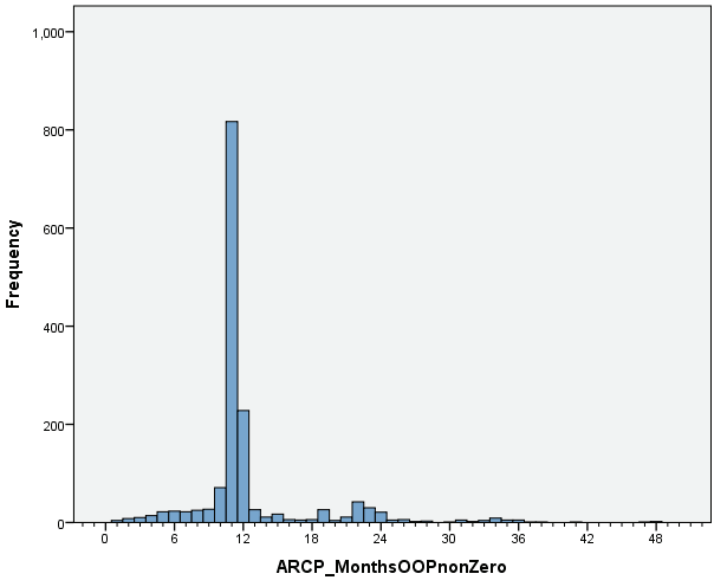
Trainees for many reasons may spend some of their time less than full-time, or may take time out-of-programme. Time LTFT or OOP can be roughly estimated by summing the time for those ARCP records in which LTFT or OOP

was indicated, although it is accepted that this may sometimes be very inaccurate[31]. Figure 4.1 shows a histogram of estimated number of calendar months spent LTFT, and Figure 4.2 shows a histogram of estimated number of months spent OOP. Note that there are strong modes at 6, 12, 18, 24, 30 and 36 months. Not surprisingly the proportion taking LTFT or OOP varies with training year, as is seen in Table 4.10, where probably only the 2009 cohort had sufficient ARCP records for the typical amounts of LTFT and OOP. Even some of the 2009 cohort of trainees may have extended beyond 2014.

» Figure 4.1 Histogram of estimated months of LTFT for all trainees. Note that longer values are underestimated due to later cohorts having less time to accrue months LTFT (see Table 4.10).



» Figure 4.2 Histogram of estimated months OOP for all trainees. Note that longer values are underestimated due to later cohorts having less time to accrue months OOP (see Table 4.10).



[31] We are told that when a trainee is OOP then typically there is an ARCP at the beginning and end of the period, and so estimates of OOP duration are probably reasonably accurate. Trainees may however be LTFT for less than the full time to which the ARCP period applies. Given that there are no other centrally held data which look at LTFT and OOP, the present method is probably reasonable for estimating the extent of LTFT and OOP.

» *Table 4.10: Percentage of each cohort that has been LTFT or have taken time OOP*

Cohort	Any LTFT	Months LTFT* mean (SD)	Any OOP	Months OOP mean (SD)	Any LTFT or OOP	Any LTFT and OOP
2009	18.2% 561/3090	15.4 (9.4)	13.4% 414/3090	13.5 (6.9)	25.4% 784/3090	6.2% 191/3090
2010	11.6% 378/3270	15.1 (8.0)	15.9% 520/3270	12.7 (5.4)	22.5% 749/3270	4.6% 149/3270
2011	5.1% 137/2777	13.8 (5.1)	12.0% 348/2905	11.7 (4.5)	15.5% 449/2905	1.7% 48/2905
2012	2.0% 63/3028	11.5 (1.4)	6.2% 187/3028	11.4 (2.8)	7.8% 236/3028	0.4% 11/2947
2013	0.1% 2/3194	**	1.9% 57/3014	10.4 (2.8)	1.9% 57/3014	0% 0/3014

* i.e. months at 60% LTFT, so that 15.4 calendar months LTFT is equivalent to 9.2 full-time equivalent months;

** N too small to be useful

Using the most complete data, which is from the 2009 cohort, it seems that during their training about 25% of trainees will take time either as LTFT or OOP. The mean total time taken is 31 calendar months, with means of 15 months LTFT and 14 months OOP. For the 2009 cohort the modal time OOP is 12 months and for LTFT it is also 12 months.

4.4.7 Modified Tiffin outcomes in relation to selection measures

Trainees taking time LTFT or OOP are not necessarily a random subset of trainees. Table 4.11 shows the mean Stage 2 and Stage 3 scores for trainees who had taken time LTFT or OOP (or both).

Simple t-tests showed that trainees who had taken time LTFT scored **lower** on all four selection measures, whereas those taking time OOP scored **higher** on all four selection measures (p<.001). A more detailed analysis of the four groups of LTFT by OOP showed that the principle difference was between candidates taking only LTFT and all the

» *Table 4.11 Mean (SD) score on the various selection measures of trainees in relation to having taken time LTFT, OOP or*

	0: Neither LTFT nor OOP	1: LTFT only	2: OOP only	3: LTFT and OOP	Homogenous subsets (p>.001)
N (total=7446)	*6316*	*516*	*430*	*184*	
Stage 2 total	522.0 (54.3)	497.3 (60.8)	523.9 (59.2)	529.1 (63.9)	0=2=3
Stage 2 CPST	260.1 (33.8)	245.9 (37.1)	263.4 (36.2)	263.6 (39.0)	0=2=3
Stage2 SJT	262.0 (29.3)	251.3 (33.5)	260.5 (30.8)	265.5 (30.9)	0=2=3
Stage 3 total	54.2 (5.20)	53.2 (4.89)	54.6 (5.23)	55.2 (4.96)	0=1, 0=2=3

other three groups. A specific comparison of the effects of taking time OOP in those who were never LTFT found no significant differences on any of the four selection measures. The main conclusion therefore has to be that **candidates taking LTFT have lower scores on the Stage 2 and Stage 3 selection measures**[32]. That has implications for the likelihood of those trainees getting on to the GP Register and the time it takes them to do so.

4.5 THE GP REGISTER OF THE LRMP

Only doctors who are on the GP Register of the GMC's LRMP can practice independently as general practitioners. The List has been established since 1996, and provides a good indication of the numbers of qualified GPs, and is a key outcome indicator for studies of selection. The GMC Register was downloaded on 27th August 2015 for the present analyses, and therefore includes doctors entering the Register in August 2015[33].

Table 4.12 shows the time from starting training to entering the GP Register, for the various cohorts of trainees. A 'standard' GP trainee should enter the GP Register three years after beginning training. It would also be accepted that a trainee entering the Register within four years would be an acceptable outcome. In this and further tables, cells marked in green indicate trainees who would be entering the GP Register three years after starting training and those entering between three and four years are in light green (both being acceptable outcomes), while those in red indicate those who would take longer than four years[34]. For obvious reasons the follow-up times are less for more recent cohorts. Nevertheless there is a broad consistency between the cohorts, with almost exactly a half of trainees getting onto the Register three years after beginning training (which would be the expected, normal trajectory), and 71% within an acceptable four years. The estimated cumulative percentages show that only 80% of trainees are on the Register within five years, and a further 5% enter the Register from 5 to 6 years after beginning training. About 15% of trainees do not seem to enter the GP Register within 6 years of starting training. As noted earlier, a substantial minority of trainees have either LTFT working or are OOP for some of their training, and these doctors need considering separately.

4.5.1 Expected time to GP Register for those who are OOP or LTFT

Table 4.12 shows the expected time on the GP Register as within four calendar years of entering training. That is correct for a typical trainee, but for those who are OOP or LTFT the expected time has to be extended[35].

Table 4.13 compares trainees who have never been LTFT or OOP, with those who have. Unsurprisingly those without LTFT or OOP progress more quickly, with 61% on the GP Register within three years and 83% within an acceptable four years, with 9% are still not on the Register 6 years after beginning training. Trainees who have been LTFT or OOP are less likely to be on the Register within four years, and indeed only about 74% are on the Register six years after starting training (compared with 91% of those who are not LTFT or OOP). Although it is straightforward to state that the expected time on the Register for a non-LTFT/OOP trainee is three years and an acceptable limit is within four years of beginning training, it is harder to calculate the expected time on the GP Register for those who have been LTFT/OOP.

[32] It should be emphasised here, as in most of the other analyses in this chapter, that no account is taken of the possible interactive, moderating and mediating effects of background variables such as sex, ethnicity or place of PMQ. That is a separate set of analyses, but since those measures are not used during selection (and arguably cannot and should not be used in selection) they are not relevant to the present analyses of the selection process.

[33] The majority of GPs enter the Register during the first week of August, with a subsidiary peak in February, and a few others entering sporadically during the rest of the year.

[34] In practice we have set the limits to three years and two months and four years and two months to allow for the fact that for administrative reasons most doctors enter the Register a week or more after the beginning of August, and hence it makes more sense to use a year centred at the end of September.

[35] In this section it should be noted that although trainees who are part-time due to maternity/paternity leave can be taken into account due to being registered as LTFT, full-time statutory leave does not seem to appear in ARCP records or other data to which we have access and therefore might result in further delays in time to the Register which cannot here be accounted for, Being LTFT might though be a flag for statutory leave in some cases, and may account for some of the extra delays for entering the GP Register for those who have been LFTT which are described here.

» *Table 4.12: Time from starting GP training to entering the GP Register: All trainees. Years after starting training are calendar years and take no account of time LTFT or OOP, which are discussed later.*

Cohort	Years after starting training to GP Register								Total
	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5	5-5.5	5.5-6	Not on GP register	
2009	Aug 2012 49.1% 1549/3156	Feb 2013 11.1% 350/3156	Aug2013 12.7% 401/3142	Feb 2014 5.0% 157/3156	Aug 2014 3.8% 121/3156	Feb2015 2.4% 77/3156	Aug2015 1.0% 33/3156	14.8% 468/3156	3156
2010	Aug 2013 48.5% 1613/3324	Feb 2014 10.0% 332/3324	Aug2014 12.9% 430/3324	Feb 2015 4.9% 163/3324	Aug 2015 2.3% 77/3324	-	-	21.3% 709/3324	3324
2011	Aug 2014 53.3% 1591/2986	Feb 2015 7.5% 224/2986	Aug2015 8.5% 253/2986	-	-	-	-	30.7% 918/2986	2986
2012	Aug 2015 48.9% 1508/3081	-	-	-	-	-	-	51.5% 1573/3081	3081
Average percent	**50.0%**	**9.5%**	**11.4%**	**5.0%**	**3.1%**	**2.4%**	**1.0%**	-	-
Cumulative percent	**50.0%**	**59.8%**	**71.2%**	**77.1%**	**80.2%**	**84.1%**	**85.1%**	**[14.9%]**	-

Consider a trainee who has one year OOP and one year LTFT, with the latter being assumed to at 60% FTE (full-time equivalent). Instead of taking the expected three years, the OOP means that the trainee should take an additional calendar year to get onto the Register. The LTFT is a little more complicated as the trainee will have worked LTFT for a whole calendar year, but since the trainee is 60% FTE, only about 7 months of those 12 calendar months will have been working in GP and hence gaining appropriate experience. In terms of experience the trainee will therefore be five months short, so that the expected date of entry to the GP Register will be delayed by 5 months. The expected time on the Register is therefore after 4 years and 5 months (i.e. 3 years + 1 year OOP + 5 months extra training because of being LTFT).

Similar calculations can be carried out on an individual basis for various periods OOP and LTFT. The bottom two rows of Table 4.13 show the expected proportions of LTFT/OOP trainees who should be on the Register within different time windows, with the cumulative percentage the most useful values. Notice that only 12.5% of these trainees should be on the Register within three years (compared with 100% of non-LTFT/OOP trainees). By 3.5 years after starting training, 46% of trainees should be on the Register, 80% within 4 years, 89% within 4.5 years, 85% within 5 years, 97% within 5.5 years and 99 within 6 years. Notice that the colour scheme in the lower half of Table 4.13 is different to flag up in the green areas that up to about four and a half years is acceptable for most LTFT/OOP trainees, that more can occur due to differences in OOP and LTFT (grey), but that beyond about 5.5 years (pink) is probably beyond acceptable expectations, even taking LTFT and OOP into account.

Using the 2009 cohort, for whom the most complete data are available, and calculating an individual expected time on Register for each trainee, 65% (1,491/2,306) of non-LTFT/OOP trainees were on the Register within the expected three years and **85% (1,963/2,306) were on the Register within the acceptable limit of four years.** For LTFT/OOP trainees, only 31% (244/784) trainees were on the Register after three years of training (taking account of LTFT and OOP), and **63% (496/784) were on the Register within the acceptable limit of four years of training.** The conclusion has to be that, even taking into account a lengthening of the expected time to be on the Register, the majority of LTFT/OOP trainee, 37% of LTFT trainees are not on the Register after four years of full-time equivalent training, compared with 15% of non-LTFT/OOP trainees after four years of full-time training . The reasons for that need exploring further. Clearly, being LTFT or OOP has a substantial effect not only on the time taken to get onto the Register, but also on whether the trainees ever get on to the Register, with 26% still not being on the Register after 6 years, compared with 9% of non-LTFT/OOP trainees, and an expected proportion of 1%. It is possible that some of the LTFT/OOP trainees will get on to the Register after six years, but that seems unlikely given the ever declining rates of additional entrants after five and then six years.

4.5.2 Selection measures and time to GP Register

Table 4.14 shows the mean selection scores in relation to the time taken to enter the GP Register for trainees who entered training in 2009 and 2010 and who had taken no time LTFT or OOP.

In simple terms there is a clear picture, with those getting onto the GP Register within three years have the highest Stage 2 and Stage 3 scores, and those not on the Register after 5 years having the lowest selection scores. However the pattern is slightly more complex as 3 and 4 are higher than 3.5, and 4 and 5 are higher than 4.5 years, suggesting that a general downwards trend is complicated by those entering the Register on half-years having lower selection scores than those entering after integer number of years. No obvious explanation of that pattern is apparent, although it may reflect 6 month extensions due to failing CSA or be due to LTFT or OOP which has not been registered within the ARCP process. The broad pattern nevertheless is clear, that **higher Stage 2 and Stage 3 scores are associated with entering the Register more quickly, particularly within four years.**

4.6 ARCP OUTCOMES AND ENTERING THE GP REGISTER

ARCP outcomes, as their name implies, are a record of the progression of competence of trainees. Those without problems in ARCP should therefore be competent and should be fit to enter the GP Register. That can be looked at, as in Table 4.15, by

[36] The cumulative percentages are averaged across all cohorts for which there is data. As a result the number not on the register, despite being based mainly on the 2009 cohort is not precisely the same as that for 2009.

[37] Note the important caveat that this result might in part be due to full-time statutory leave for maternity/paternity not being included within ARCP or other records.

» *Table 4.13: Time from starting GP training to entering the GP Register: Trainees with and without LTFT/OOP time analysed separately.*

Cohort	Years after starting training to GP Register								Total
	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5	5-5.5	5.5-6	Not on GP register	
All trainees without any LTFT or OOP									
2009	Aug 2012 65.0% 1491/2294	Feb 2013 10.4% 238/2294	Aug2013 10.2% 234/2294	Feb 2014 2.5% 58/2294	Aug 2014 1.7% 40/2294	Feb2015 0.7% 16/2294	Aug2015 0.4% 9/2294	9.1% 208/2294	2294
2010	Aug 2013 62.0% 1561/2519	Feb 2014 10.2% 258/2519	Aug2014 10.8% 273/2519	Feb 2015 3.1% 79/2519	Aug 2015 1.2% 30/2519	-	-	12.6% 318/2519	2519
2011	Aug 2014 63.8% 1561/2451	Feb 2015 7.3% 180/2451	Aug2015 8.4% 207/2451			-	-	20.4% 501/2451	2451
2012	Aug 2013 53.6% 1494/2786					-	-	46.4% 1292/2786	2786
Average percent	**61.1%**	**9.3%**	**9.8%**	**2.8%**	**1.5%**	**0.7%**	**0.4%**	-	-
Cumulative percent	**61.1%**	**72.9%**	**82.7%**	**87.1%**	**88.6%**	**90.5%**	**90.9%**	**[9.1%]**	-

Continued on next page.

» *Table 4.13 continued.*

Cohort	Years after starting training to GP Register								Total
	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5	5-5.5	5.5-6	Not on GP register	
All trainees with either LTFT or OOP									
2009	Aug 2012 5.1% 40/784	Feb 2013 14.0% 110/784	Aug2013 20.9% 164/784	Feb 2014 12.4% 97/784	Aug 2014 10.3% 80/784	Feb2015 7.8% 61/784	Aug2015 3.1% 24/784	26.4% 207/784	784
2010	Aug 2013 4.8% 36/746	Feb 2014 9.4% 70/746	Aug2014 20.4% 152/746	Feb 2015 11.3% 84/746	Aug 2015 6.3% 47/746	-	-	47.9% 357/746	746
2011	Aug 2014 2.9% 13/448	Feb 2015 7.8% 35/448	Aug2015 8.9% 40/448			-	-	80.4% 360/448	448
2012	Aug 2013 0% 0/236					-	-	100.0% 236/236	236
Average percent	**3.2%**	**10.4%**	**16.7%**	**11.9%**	**8.3%**	**7.8%**	**3.1%**	-	-
Cumulative percent	**3.2%**	**14.7%**	**31.4%**	**49.2%**	**57.5%**	**70.5%**	**73.6%**	**[26.4%]**	-
Expected percent on GP Register	12.5%	33.0%	34.2%	9.3%	5.9%	2.4%	1.5%	[1.2%]	
Expected	12.5%	45.5%	79.7%	89.0%	94.9%	97.3%	98.8%		

» *Table 4.14 Mean (SD) score on time to entering GP Register in relation to the various selection measures for 2009 and 2010 trainees who had no record of time LTFT or OOP.*

Cohort	Years after starting training to GP Register						Homogenous subsets (p>.001)
	3: (2.5-3)	3.5: (3-3.5)	4: (3.5-4)	4.5: (4-4.5)	5: (4.5-5)	Not on Register after 5 years	
N (total=4813)	3052	496	507	137	70	551	-
Stage 2 total	533.0 (47.4)	487.7 (56.1)	525.9 (53.5)	492.7 (61.5)	502.6 (69.7)	480.0 (64.4)	3=4=5, 3.5=4.5=5=Not
Stage 2 CPST	266.2 (30.8)	241.2 (35.9)	261.2 (32.3)	244.1 (36.5)	247.2 (42.3)	236.2 (40.4)	3=4=5, 3.5=4.5=5=Not
Stage2 SJT	266.8 (26.2)	246.5 (30.8)	264.8 (29.4)	248.6 (33.1)	255.4 (35.4)	242.8 (34.2)	3=4=5, 3.5=4.5=5=Not
Stage 3 total	54.7 (5.12)	52.2 (5.43)	54.4 (5.13)	51.9 (5.61)	53.7 (4.96)	52.0 (5.40)	3=4=5, 3.5=4.5=5=Not

» *Table 4.15: Time to entering the GP Register in relation to modified Tiffin score in ST3 for all 2009, 2010 and 2011 trainees without any time LTFT or OOP. Note that not all 2010 and 2011 trainees have had sufficient time to enter the Register 4+ or 5+ years after training (see Table 4.13) and hence values in grey should be treated with great care.*

Modified Tiffin score in ST3	Years after starting training to GP Register								Total
	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5	5-5.5	5.5-6	Not on GP register	
Tiffin 1 Satisfactory progression	81.8% 3661/4477	5.1% 229/4477	8.9% 398/4477	1.0% 45/4477	0.8% 35/4477	0.2% 7/4477	0.1% 4/4477	2.2% 98/4477	4477
Tiffin 2 Insufficient evidence presented	83.6% 704/842	3.3% 28/842	8.4% 71/842	1.3% 11/842	1.2% 10/842	0.1% 1/842	0.1% 1/842	1.9% 16/842	842
Tiffin 3 Targeted training required (not extended)	46.2% 72/156	16.7% 26/156	7.7% 12/156	2.6% 4/156	2.6% 4/156	0% 0/156	0% 0/156	24.4% 38/156	156
Tiffin 4 Extended training time required	5.0% 31/625	52.3% 327/625	11.0% 69/625	9.6% 60/625	2.1% 13/625	1.0% 6/625	0.2% 1/625	18.9% 118/625	625
Tiffin 5 Left programme (i.e. ARCP outcome 4)	0.5% 1/195	4.1% 8/195	9.2% 18/195	4.6% 9/195	1.0% 2/195	0.5% 1/195	0.5% 1/195	79.5% 155/195	195

comparing the time until entering the GP Register with the modified Tiffin scores for ARCP in ST3. The analysis is restricted to those in the 2009, 2010 and 2011 cohorts as these have been at least four years since starting training (see Table 4.12) and for simplicity is also restricted to those who had neither been LTFT nor OOP. It should be remembered that the 2009, 2010 and 2011 cohorts are followed up for six, five and four years respectively, and that although it is known if trainees entered the GP Register in 2015 there are no ARCP records as yet for the 2014-2015 training year. Several features are clear in Table 4.15.

Firstly there is **no discernible difference in outcome for those with mTiffin scores 1 and 2**, both performing equally well, and both groups do well, 83% entering the GP Register in year 3 and 96% within four years. In contrast Tiffin groups 3 and 4 perform much less well, only 71% and 68% not being on the Register within four years, a large difference from the 96% of mTiffin groups 1 and 2. The mTiffin group 3 is relatively small (n=151) and they do better than mTiffin group 4 at getting onto the Register in Group 3 (seemingly supporting the view that an extension of training was not required). Whether the extra training for Group 4 really benefited them is difficult to say. Finally, it is hardly surprising that mTiffin Group 5 (with their ARCP outcome 4) did badly, and perhaps the really surprising thing is that 14% of them still ended up on the GP Register within four years (and a few after that), although the mechanism and process is far from not clear if they had indeed "left programme". The lack of difference between mTiffin Score 1 (Satisfactory), and Score 2 (Insufficient evidence) is consistent with recent analyses of ARCP in Core Medical Training, where the majority of Outcome 5 (mTiffin Score 2) results are due to "minor infringements [because] …the assessment focuses on process, not practice" (Dormandy & Laycock, 2015), and it is unlikely that such problems are predictive of major outcomes.

4.6.1 The impact of poor ARCP outcomes in ST1 and ST2

Table 4.15 suggests that poor ARCP outcomes in ST3 are associated with a lengthened time to entering the GP Register, and a lower rate of eventually reaching the Register. An important question is therefore the extent to which ARCP in ST1 and ST2 predicts entry to the GP Register, taking into account ARCP in ST3, and LTFT and OOP. In Table 4.16 are shown only 2009, 2010 and 2011 trainees who attained a mTiffin 1 or 2 in ST3, and had been neither LTFT nor OOP. mTiffin scores of 3, 4 or 5 are relatively rare in ST1 and ST2 and therefore the data are tabulated according to the worse mTiffin score in either ST1 or ST2. As in Table 4.15 there is little evidence of a difference between mTiffin 1 and mTiffin 2, 96% of both groups being on the Register within 4 years. Neither is there any obvious difference between mTiffin 1&2 and mTiffin 3, 94% of the latter being on the Register within four years. However the relatively small group with mTiffin 4 outcomes, with only 31 trainees, does show a lower rate of GP Register entry at year 3 (but then they had required extended training), and also at year 4, with only 65% being on the Register[38]. Only 1 trainee in the group has a mTiffin 5 score (ARCP outcome 4).

4.6.2 ARCP outcomes and entering the GP Register: Summary

There is much variation in the time of entering the GP Register, only about 50% of trainees entering after an expected time of three years. About 25% of trainees are either LTFT or OOP, and that clearly affects progression on to the Register. However although 65% of non-LFTT/OOP trainees were on the Register within an expected time of three years, only 31% of LFTT/OOP trainees were on the Register within an extended expected time which takes account of time out of programme and reduced experience when LTFT. For those without LTFT or OOP a mTiffin Score of 3, 4 or 5 in ST3 predicts both delay and a lower final rate of going on to the Register. mTiffin scores of 1 or 2 in ST3 do, though, seem to have similar outcomes. Higher mTiffin scores in ST1 and ST2 have little impact on time to entering the Register, with no obvious differences between mTiffin scores of 1, 2 or 3, and only the very small group with a mTiffin score of 4 in ST1 or ST2 having lower rates of getting onto the Register.

4.7 MRCGP OUTCOMES

Obtaining the Membership of the Royal College of General Practitioners (MRCGP) is necessary before a trainee can enter the GP Register. The examination is in two parts, the AKT (Applied Knowledge Test) and the CSA (Clinical Skills Assessment), with the two parts being taken in any order, although in practice most trainees take the AKT first since it can be taken in ST2

[38] A comparison of 20/31 in mTiffin 4 not being on the Register within 4 years with 167/4360 in mTiffin 1 not being on the Register within four years is highly significant, chi-square = 79.3, 1 df, p<.001.

» *Table 4.16: Time to entering the GP Register in relation to the higher modified Tiffin score in ST1 and ST2 for all 2009,2010 and 2011 trainees without any LTFT or OOP time who achieved a Tiffin score 1 or 2 for ST3. Note that not all 2010 and 2011 trainees have had sufficient time to enter the Register 4+ or 5+ years after training (see Table 13) and hence values in grey should be treated with great care.*

Modified Tiffin score in ST3	Years after starting training to GP Register								Total
	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5	5-5.5	5.5-6	Not on GP register	
Tiffin 1 Satisfactory progression	83.7% 3651/4360	4.5% 198/4360	7.9% 344/4360	1.0% 45/4360	0.7% 30/4360	0.2% 7/4360	0.1% 5/4360	1.8% 80/4360	4360
Tiffin 2 Insufficient evidence presented	76.7% 522/675	4.7% 32/675	14.8% 100/675	1.0% 7/675	1.2% 8/675	0.1% 1/675	0% 0/675	2.5% 17/675	675
Tiffin 3 Targeted training required (not extended)	79.0% 188/238	6.3% 13/238	8.8% 21/238	0.4% 1/238	1.7% 4/238	0% 0/238	0% 0/238	3.8% 9/238	238
Tiffin 4 Extended training time required	29.0% 9/31	22.6% 7/31	12.9% 4/31	9.7% 3/31	9.7% 3/31	0% 0/31	0% 0/31	16.1% 5/31	31
Tiffin 5 Left programme (i.e. ARCP outcome 4)	-	1/1	-	-	-	-	-	-	1

» *Table 4.17: Time from starting GP training to first taking AKT: All trainees (top) and trainees without LTFT or OOP (bottom).*

Training cohort	Years after starting training to first taking CSA											Total
	1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-.5	4.5-5	5-5.5	5.5-6	AKT not taken	
2009	18.2% 561	32.6% 1,008	29.4% 909	4.6% 142	4.8% 148	1.4% 44	1.2% 37	0.2% 7	0.7% 7	0.3% 10	6.5% 202	3,090
2010	25.0% 819	27.8% 908	28.6% 935	6.0% 195	3.7% 120	1.3% 41	1.6% 53	0.4% 13	-	-	5.7% 186	3,270
2011	29.0% 843	31.8% 924	22.6% 656	3.6% 104	3.7% 107	1.1% 33	-	-	-	-	8.2% 238	2,905
2012	35.7% 1,081	29.4% 891	18.6% 563	4.6% 138	-	-	-	-	-	-	11.7% 355	3,028
2013	34.7% 1,047	29.5% 888	-	-	-	-	-	-	-	-	35.8% 1,079	3,014
Average percent	28.5%	30.2%	24.8%	4,7%	4.1%	1.3%	1.4%	0.3%	0.7%	0.3%	-	-
Trainees with no time LTFT or OOP												
2009	20.7% 477	35.9% 827	30.5% 704	3.5% 80	2.0% 47	0.4% 9	0.3% 7	0% 0	0.1% 2	0.1% 2	6.5% 151	2,306
2010	25.0% 702	30.8% 777	30.7% 775	4.5% 114	1.2% 31	0.3% 8	0.4% 9	0.04% 1	-	-	4.1% 104	2,52
2011	31.5% 773	34.4% 846	23.0% 566	2.4% 59	2.0% 48	0.6% 15	-	-	-	-	6.1% 149	2,456
2012	37.6% 1,050	31.1% 869	18.9% 529	3.9% 110	-	-	-	-	-	-	8.4% 234	2,792
2013	35.3% 1,045	30.0% 888	-	-	-	-	-	-	-	-	34.6% 1,024	2,957
Average percent	30.6%	32.4%	25.8%	3.6%	1.7%	0.4%	0.4%	0.02%	0.1%	0.1%	-	-

» *Table 4.18: Time from starting GP training to first taking CSA: All trainees (top) and trainees without LTFT or OOP (bottom)*

Training cohort	Years after starting training to first taking CSA									N
	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5	5-.5	5.5-6	CSA not taken	
2009	11.0% 339	60.5% 1870	5.0% 155	9.5% 294	2.4% 75	3.0% 94	1.2% 36	1.3% 40	6.1% 187	3,090
2010	9.8% 320	56.8% 1858	5.0% 163	12.4% 404	3.3% 109	2.8% 91	-	-	9.9% 325	3,270
2011	11.9% 345	56.6% 1645	6.5% 190	8.4% 245	-	-	-	-	16.5% 480	2,905
2012	21.9% 662	45.1% 1366	-	-	-	-	-	-	33.0% 1000	3,028
Average percent	13.7%	54.8%	5.5%	10.1%	2.9%	2.9%	1.2%	1.3%	-	-
Trainees with no time LTFT or OOP										
2009	12.7% 294	71.0% 1637	3.4% 79	5.9% 135	0.8% 18	0.9% 20	0.2% 5	0.2% 4	4.9% 114	2,306
2010	11.4% 288	68.4% 1725	3.8% 95	8.4% 213	1.5% 38	0.8% 21	-	-	5.6% 141	2,521
2011	13.1% 322	64.9% 1593	5.7% 141	5.9% 144	-	-	-	-	10.4% 256	2,456
2012	23.6% 659	48.4% 1352	-	-	-	-	-	-	28.0% 781	2,792
Average percent	15.2%	63.2%	4.3%	6.7%	1.2%	0.9%	0.2%	0.2%	-	-

in ST1 and ST2 have little impact on time to entering the Register, with no obvious differences between mTiffin scores of 1, 2 or 3, and only the very small group with a mTiffin score of 4 in ST1 or ST2 having lower rates of getting onto the Register.

4.7 MRCGP OUTCOMES

Obtaining the Membership of the Royal College of General Practitioners (MRCGP) is necessary before a trainee can enter the GP Register. The examination is in two parts, the AKT (Applied Knowledge Test) and the CSA (Clinical Skills Assessment), with the two parts being taken in any order, although in practice most trainees take the AKT first since it can be taken in ST2 whereas CSA can only be taken in ST3. AKT and CSA currently have a limit of four attempts[39], although that has only been introduced relatively recently.

4.7.1 Time of taking AKT and CSA

AKT can currently be taken during or after ST2[40], and can be sat in January, April or October. For candidates entering training in, say, August 2012, the first opportunity would therefore be in October 2013, 15 months later. CSA can currently only be taken during ST3, and can currently be sat in November, December, January, February, March, April and May[41], and therefore a candidate entering training in August 2012 could first take CSA in November 2014, 28 months later.

Table 4.17 shows the time from starting training until the first attempt at AKT, for all trainees (top) and trainees without either LTFT or OOP (bottom). As in earlier tables, cells marked in green indicate trainees who are on target for entering the GP Register three years after starting training, while those in red indicate those who are behind target. Most trainees are taking AKT within three years of starting training[42], and those who fail it still have several more opportunities to pass and be on target. Noticeable in Table 4.17 is **a tendency in more recent years for candidates to take AKT at the first time possible, the proportion rising from 18% to 35%.**

Table 4.18 shows the time of first taking CSA, and in comparison with AKT it is clear that most trainees are not taking CSA at the first possible opportunity, but are waiting until the final half-year of training (January to May), which leaves them fewer opportunities to retake should they fail. There is however a **recent trend towards trainees first taking CSA at the earliest opportunity, the proportion rising from 11% to 22% over the four years.**

4.8 MRCGP MARKS

4.8.1 AKT

MRCGP AKT, which has 200 questions overall, currently has three subscores, clinical medicine (CM), Evidence Interpretation (EI) and Organisational (OQ), with 80%, 10% and 10% of questions in each category. Since the numbers of questions in the EI and OQ sections are likely to be too small to give reliable sub-scores, and since also the structure of the exam has not been constant since 2009, we will for the analyses reported below only consider the total AKT score. The pass mark for each diet of AKT varies according to the standard setting, and since 2009 it has been set between 132 and 143, with a mean of 136.47 and a median of 136. For convenience, therefore, all marks have been adjusted so that they are percentage points relative to the pass mark expressed as a percentage difference score, a candidate with a score of zero just passing and a candidate with a score of 1 just failing the exam, having a percentage mark which is one percentage point below the pass mark. As is conventional in studies of academic attainment (McManus & Ludka, 2012) we will use the mark gained by candidates at their first attempt as the major outcome variable[43]. The distribution of marks is shown in Figure 4.3.

[38] A comparison of 20/31 in mTiffin 4 not being on the Register within 4 years with 167/4360 in mTiffin 1 not being on the Register within four years is highly significant, chi-square = 79.3, 1 df, p<.001.

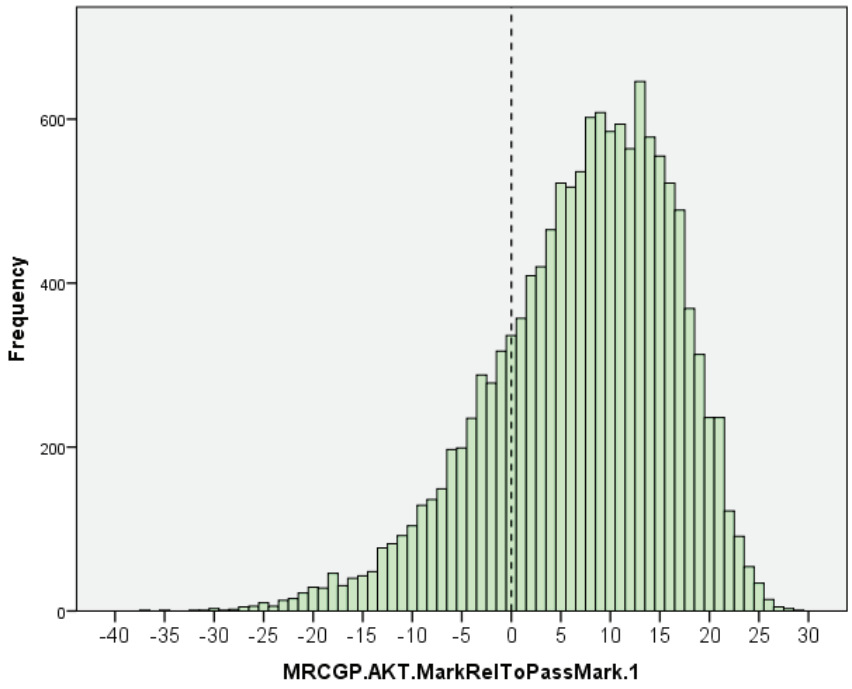
[39] These regulations were introduced in August 2012. Some candidates can, for various reasons, make a fifth attempt.

[40] The current regulations were for candidates entering training programmes from 1st August 2010.

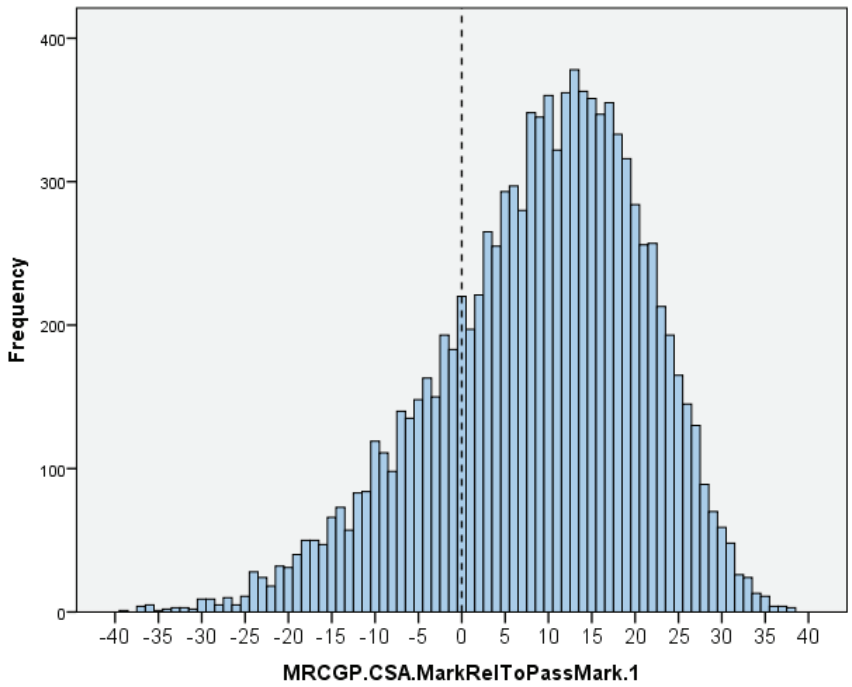
[41] http://www.rcgp.org.uk/training-exams/mrcgp-exams-overview/mrcgp-clinical-skills-assessment-csa.aspx

[42] The data are calculated in terms of calendar half-years, so that 1.5 years means within the third six-month period from commencing training. ST1 is 0.5 and 1 year, ST2 is 1.5 and 2 years, and ST3 is 2.5 and 3 year. For a trainee starting in August 2012 that would mean taking AKT between July and December 2013, at the beginning of ST2. Those within three years of training are therefore in the 1.5, 2 or 2.5 columns. Normal exit from training would be in the period 2.5-3.

» *Figure 4.3: Distribution of AKT marks at the first attempt relative to the pass mark. Abscissa is in percentage points.*



» *Figure 4.4: Distribution of CSA marks at first attempt relative to the pass mark. Abscissa is in raw examination scale marks.*



4.8.2 CSA

MRCGP CSA changed its structure in September 2010, and currently has 13 cases, each of which is marked on three domains of Data Gathering Skills, Clinical Management Skills and Interpersonal Skills, each marked on a four-point scale from 3 (Clear Pass), through 2 (Pass), 1 (Fail) and 0 (Clear Fail). The mark on each station is therefore from 0 to 9, and the summed station marks have a range of 0 to 117 marks[44]. The overall pass mark is set using a borderline groups method which takes account of day-to-day variation in the difficulty of cases, and in addition an adjustment is made which takes into account the standard error of measurement of the marks. Because the pass marks vary from day to day, we will here, as with AKT, consider a candidate's performance at their first attempt at CSA in terms of the difference of their mark from the pass mark for the day that they took the examination. Note that unlike with AKT, where results are expressed as percentage points, for CSA we express differences in terms of actual marks on the marking scale itself, as these are more easily understood in relation to the marking scheme[45]; see Figure 4.4.

4.8.3 Association of MRCGP AKT and CSA marks with selection measures

The association of first attempt AKT and CSA marks with Stage 2 and Stage 3 selection marks is best described in terms of Pearson correlations. The correlations can be calculated for all trainees who have taken either part of the MRCGP assessment. Table 4.19 shows the correlations between the various measures. The key correlations are in the final two columns, and are shown in bold, but the correlations between the Stage 2 and Stage 3 measures are also included. Performance at the AKT and CSA assessments correlates fairly highly at 0.56, suggesting that candidates who are better at the knowledge assessment are also better at the clinical assessment. The pattern of correlations is interesting with the knowledge assessment of Stage 2 (CPST) correlating mostly highly with AKT (0.74) and the SJT correlating rather less at 0.47, and Stage 3 correlating much less well with the AKT at 0.26. The MRCGP clinical assessment, the CSA, correlates most highly with the total Stage 2 score (0.60), and to almost equal extents with Stage 2 CPST and SJT (0.512 and 0.508). The Stage 3 correlates somewhat more with the CSA (0.37) than it did with the AKT (0.26), but it is still a rather lower correlation than the written assessments of Stage 2. Overall it seems clear that **the Stage 2 selection assessments are better predictors of AKT and CSA than is the Stage 3 selection centre assessments. CPST is a much better predictor of AKT than is the SJT, whereas the SJT and the CPST are equally good predictors of CSA.**

4.9 GMC FITNESS TO PRACTICE ISSUES

The LRMP of the GMC was examined to find which of the 21,643 trainees on the Register at 27th August 2015 had evidence of Fitness to Practice (FtP) issues. Overall there were 10 doctors who, at some time, had been erased, 73 who had been suspended, 122 who had conditions on their practice, 56 who had undertakings and 100 who had warnings. Some of the FtP categories are overlapping, and overall there were 273 doctors (1.3%) with a history of erasure, suspension, conditions, undertakings or warnings (ESCUW). Given the small numbers in some categories, analysis was restricted to comparing those with any ESCUW to those without. Table 4.20 shows the mean scores of those with or without FtP issues. Note that the analysis considers all GP trainees from all years.

Differences between the groups are highly significant ($p<.001$) for all four outcome measures. Overall it is clear **that doctors who have FtP issues have lower scores at both Stage 2 and Stage 3 selection measures**, with medium to large effect sizes. The question of causality has not been taken into account and some doctors may have had FtP issues at the time of taking the selection tests, which may have adversely affected their performance on those tests.

[43] The bivariate distribution of mark at second attempt against mark at first attempt shows the statistical problem of using later marks, since the distribution of marks at the first attempt is necessarily truncated at zero (as those candidates gaining zero or above passed the exam and therefore did not need to take it a second time).

[44] http://www.rcgp.org.uk/training-exams/mrcgp-exams-overview/~/media/F0E9EF64C6224E279090C5E769213B14.ashx

[45] For instance a candidate with a score of -1 was one point below the pass mark, which is equivalent to a single station on which they scored 2 rather than 3 or 0 rather than 1.

» *Table 4.19: Correlations between MRCGP AKT and CSA assessments at first attempt and selection measures in all trainees who have taken the AKT and CSA assessments. All correlations are highly significant with p<.001.*

	Stage 2 total	Stage 2 CPST	Stage 2 SJT	Stage 3 total	MRCGP AKT 1st attempt	MRCGP CSA 1st attempt
Stage 2 total	1	.883 (n=30979)	.878 (n=30979)	.426 (n=27480)	.718 (n=13582)	.595 (n=10536)
Stage 2 CPST	.883 (n=30979)	1	.552 (n=30979)	.333 (n=27480)	.743 (n=13582)	.512 (n=10536)
Stage2 SJT	.878 (n=30979)	.552 (n=30979)	1	.403 (n=27480)	.470 (n=13582)	.508 (n=10536)
Stage 3 total	.426 (n=27480)	.333 (n=27480)	.403 (n=27480)	1	.264 (n=13582)	.372 (n=10536)

» *Table 4.20: Mean (SD) score in all trainees on the various selection measures of trainees in relation to Fitness to Practice issues.*

	No FtP issues	FtP (ESCUW) issues	Effect size (Glass's delta)
N (total=21643)	21370	273	-
Stage 2 total	522.6 (53.5)	485.0 (58.3)	-0.70
Stage 2 CPST	260.6 (32.6)	240.0 (35.5)	-0.61
Stage2 SJT	262.0 (30.3)	244.9 (32.3)	-0.56
Stage 3 total	54.88 (4.98)	52.84 (4.99)	-0.41

4.10 SUMMARY

This chapter has primarily looked at the various outcome measures, both for GP selection itself and for the outcome of GP training. Analysis has primarily been at the level of the **applicant** which is important for looking at individual doctors.

The outcomes of selection are varied, with not all applicants eventually being selected into GP training, and, not surprisingly, those selected differing quite substantially on the selection tests from those not selected.

The most important question concerns the predictive validity of the selection tests, and the various analyses show that both the Stage 2 and Stage 3 selection tests showed predictive validity for better outcomes in the AKT and CSA tests of MRCGP. In addition, trainees who had poor ARCP outcomes, or had FtP issues, scored lower on the Stage 2 and Stage 3 selection tests suggesting further validity. However the predictive validity of Stage 3, particularly as measured by the correlation with AKT and CSA performance, was substantially lower than that for the multiple-choice tests of Stage 2, with the CPST particularly predicting AKT, and the CPST and SJT predicting CSA equally well. An important implication is that the CPST and SJT are probably measuring different underlying abilities which are assessed separately by AKT and CSA, presumably clinical knowledge and some form of clinical communication or situational judgement.

Finally we should reiterate that since our emphasis is upon selection, it is not appropriate to carry out analyses taking account of sex, ethnicity or country of PMQ, since these are not and should not be a part of the selection process, which should as far as possible be blind to applicant characteristics[46]. Neither have we considered issues to do with the fairness of selection in relation to various demographic groups since that was not a part of our remit, and would have extended the analyses.

[46] Of course demographic measures can and should be used in a differential prediction analysis to identify possible issues related to fairness of the selection process worthy of further investigation, but that is beyond the scope or the remit of the current study.

~ This page is intentionally left blank ~

Chapter 5

Assessing the predictive validity of selection and the consequences of various selection methods using multiple imputation

Chapter 5.

Assessing the predictive validity of selection and the consequences of various selection methods using multiple imputation

5.1 INTRODUCTION

Selection studies of necessity always have some missing data, because outcome measures are always absent in those who have not been selected. The result is that correlations between selection measures and outcome measures can only be calculated **in those individuals who have been selected.** However individuals who have been selected invariably have higher mean selection scores and lower variances than those who have been rejected, since selection chooses which individuals to select on the basis of the selection scores. The problem in general is called **range restriction**, and it has been known about for many years (Burt, 1943) , with various techniques having been developed to overcome it (Pearson, 1903; Thorndike, 1949; Schmidt & Hunter, 1977; Schmidt & Oh, 2006), although additional problems can arise when the reliability of measures needs also to be taken into account, as in meta-analysis.

The present study of GP selection also has problems of range restriction, as the GP selection process inevitably rejects many candidates, be it at Stage 2 or Stage 3, **and evaluation of the effectiveness of GP selection requires an assessment of how rejected candidates would have performed had they been selected.** The problems of restriction of range become particularly acute when economic analyses require answers as to the cost-effectiveness of different selection methods from those which actually have been used. Whereas psychometrics asks questions about validity, of how much better selected candidates would have performed compared with rejected candidates, economic analyses instead ask about the overall systemic costs of appointing selected candidates compared with the potential costs of appointing candidates who in fact had been rejected. In this chapter we describe how a solution to such problems can be provided using modern methods of handling missing data. Earlier, in Chapter 3 we have described the approach of Wiberg and Sundström to estimating reliability using the EM algorithm (Wiberg & Sundström, 2009), and here we develop that further using the techniques of multiple imputation (Graham, Taylor, Olchowski, & Cumsille, 2006).

5.1.1 Background on missing values and imputation

Missing data are a problem which is universal in complex statistical analyses for a host of reasons (Enders, 2010), with many possible reasons for data being missing. Handling missing data was put on a formal, systematic basis by Little and Rubin with their clear classification of different reasons for missingness, and the introduction of modern computational methods for handling missingness and assessing its consequences (Little & Rubin, 1987; Little & Rubin, 2002). Little and Rubin, in the preface to both editions of their work have stressed that "we continue to find that many statistical problems can be usefully viewed as missing value problems even when the data set is fully recorded" ((Little & Rubin, 1987, p viii). Restriction of range in selection is one such problem, the problem arising because of structural missingness in the data themselves.

5.1.2 MCAR, MAR and MNAR

Missing values are a recurrent problem in all large-scale social science research (Allison, 2002; Enders, 2010), and over the years there have been many techniques developed to deal with them (Graham, 2009). A host of ad hoc approaches have been developed including list-wise deletion, pair-wise deletion, mean substitution, and various forms of 'hot-decking', and all of them have quite serious statistical problems, which needn't be discussed here. The seminal 1987 study by Little and

Rubin (Little & Rubin, 1987) put approaches to missing data onto a principled statistical basis, by distinguishing between data which are Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). MCAR data are the most straightforward case and includes, but is not restricted to, the situation where some data values are missing entirely at random (and such data can be described as 'moth-eaten', little holes in the full data matrix being missing in a way which is completely unstructured and unrelated to the values of the data). Purely MCAR data are rare in practice, there almost always being other mechanisms as well. MAR data are systematically missing, but missingness is a function only of measures which are already in the analysis, which means that they are MCAR when conditioned on what is already known about the data. MCAR data can be handled by a host of methods (and although in technical terms they are 'ignorable', they cannot actually be ignored in the everyday sense of the term, but need dealing with). MNAR data are much more problematic, and essentially data are missing because of some event or process of which there is no measure in the dataset. MNAR data are very hard to handle, and technically are 'non-ignorable', but there is no statistical test to decide whether data are MAR or MNAR. Fortunately, in most cases treating MNAR data as MAR does not produce seriously biased results (Enders, 2010), and the inclusion in the dataset of a range of variables known to predict outcomes helps to mitigate against bias arising (Morris, White, & Royston, 2014).

Modern methods for handling missing data take two approaches (Enders, 2010; Raghunathan, 2016), the EM algorithm and multiple imputation, both of which generally give similar results in well-structured data.

5.1.3 EM algorithm

One method, already described in Chapter 3, uses the EM algorithm to estimate an unbiased covariance/correlation and mean matrix. This method is known to have good statistical properties, and in Chapter 3 is used for estimating reliability. A problem with the EM algorithm is that although it gives unbiased estimates of statistics such as means and covariance/correlation variances, it cannot be used as such for replacing missing values in the data, which limits its utility, although for simulation one can randomly generate data from the estimated means and covariance matrix.

5.1.4 Imputation, single and multiple

An ideal approach to handling missing data would be to impute (or 'fill in') numbers for those data cells where values are missing, so that all data are present rather than some being missing. Once data are complete then standard software methods can be applied. Imputation methods vary, and as mentioned earlier, there are many ad hoc approaches, none of which are satisfactory. A dataset can be modelled using techniques such as regression, which allow a prediction of what a missing value should be, given the other measures which are known for all of the other cases. In effect this fits a regression line to the data, and calculates what an expected value for the missing data. However that expectation must be on the regression line itself, and hence it will be less variable than real data, which are randomly scattered around the line. Single imputation therefore takes an estimate of the predicted variable and adds an appropriate amount of random variation to that estimate so that it is no longer exactly on the regression line. That added variation is random, and a problem with single imputation is that on a different occasion different random values would be added. Multiple imputation confronts that problem head on, making a virtue out of the variability. The single imputation is carried out M times, so that different imputations have different values imputed into the missing values. Each of the M imputations therefore gives different outcomes when any conventional statistical analysis is carried out on it. Those different outcomes might seem at first to be a problem but are in fact a strength, as the variability of the resulting statistics in the conventional statistical analyses provide an estimate of how robust are the results and how sensitive are the conclusions to the method of imputation. If the M imputations give very similar results (i.e. they have a low variance), then the multiple imputation process is probably robust. Rules for the combination of results from multiple imputations are relatively straightforward (Little & Rubin, 2002).

5.1.5 Different approaches to multiple imputation

Multiple imputation algorithms can be found in both SPSS and R. SPSS is simpler to use and it provides a number of different approaches. Several technical problems arise with any imputation, particularly when some variables are ordinal, but SPSS allows for this. SPSS also allows imputed values to be forced to be integer and it also allows imputed values to be categorical. The 'linear' approach to imputation SPSS means that imputed values can be extrapolated outside of the range of the values

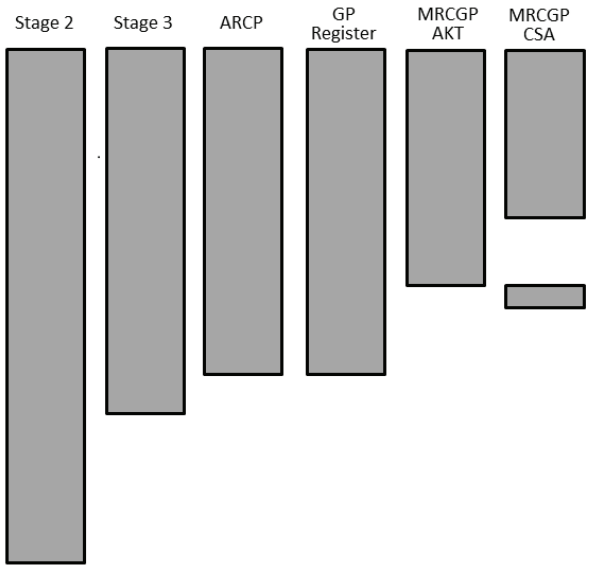
attained by actual candidates and that can be necessary when selection data are being analysed, weak candidates potentially having true values below those of any candidates who are actually selected[1].

5.2 MULTIPLE IMPUTATION FOR THE GP SELECTION DATA

5.2.1 The pattern of missingness

Multiple imputation works particularly well when data show a monotone pattern of missingness. Figure 5.1 shows a schematic diagram of the pattern of missingness in the GP selection data, shown in a similar way to that of Little and Rubin (Little & Rubin, 2002). Selection in general leads to monotonicity, so that in the present case, all participants have Stage 2 data, but only some of those with Stage 2 data have Stage 3 data, and only some of those with Stage 3 data are selected into training. All of those selected into training have a record of whether or not they eventually end up on the GP Register[2], and most of those selected for training have ARCP data. Not all of those selected for training take MRCGP, but of those who do take MRCGP, most take AKT, there are a few who take AKT but not CSA, and there is a small group who take CSA but not AKT (and the latter group mean that the data are not entirely monotone). Monotonicity is not essential for modern software, which typically breaks the dataset down into multiple subsets which are monotone, but it is desirable where possible. Multi-chain Monte Carlo (MCMC) methods can handle arbitrary patterns of missingness.

» *Figure 5.1 Schematic diagram of the pattern of missingness in the GP selection data.*



[1] SPSS also allows Predictive Mean Matching (PMM), a semi-parametric approach to imputation, and we explored it for the GP selection data but did not use it for several reasons. Perhaps the most important reason is that it often did not give sensible results in SPSS (and we note that the implementation in SPSS has been criticised for using only k=1, where k is a specific parameter defining the extent to which nearby existing values can be used for the imputation (Morris et al., 2014)). We also looked at the R program **aregImpute** in the library Hmisc (which has a number of desirable qualities (Morris et al., 2014)), but it recommends not using PMM when "mechanistic missingness requires the use of extrapolation during imputation" (p.14). There is a clear mechanism for the missingness of much of the data in the present study, which arises from low scoring candidates not being selected for later selection stages or for training, and that inevitably may require extrapolation outside of the range of the data.

[2] Note that although it is known for every candidate taking Stage 2 whether they ended up on the GP Register, it is only of interest for those who were selected into training, and those who were not selected for training cannot end up on the Register and therefore need to be regarded as structurally missing. Imputation can then decide whether those candidates who were not selected might have ended up on the Register were they to have been selected.

5.2.2 The imputations

We used the multiple imputation programs in SPSS, not least as it was more straightforward to implement, all of the databases already being within SPSS, and it provides a number of different approaches. SPSS allows variables to be ordinal or categorical, and continuous variables can be set to have integer values which is useful with, say, the Stage 3 scores. Number of FTE training years before entering the GP Register in particular was set as ordinal. The algorithm used was 'Fully conditional' which is an iterative Markov Chain Monte Carlo (MCMC) method[3]. The scale model was linear, which means that imputed values can be outside of the range of the values attained by actual candidates, and that is clearly a possibility with, say, imputed values for Stage 3 scores, and MRCGP results.

Overall we used 10 imputations[4] as that gives an adequate sense of the variability in the imputations without imposing major computational constraints. The multiple imputation datasets were stored in SPSS for processing, and where necessary exported to Excel and Stata for further processing. Estimates of the various statistics from the ten multiple imputations were combined as the arithmetic mean of the ten estimates, and a measure of variability calculated as the standard deviation of the estimates. In a larger number of imputations it would be expected that 95% of imputations would be within two standard deviations of the mean. On occasion we also plot data separately for all of the ten imputations to allow a visual sense of the variability resulting from the imputation process.

5.2.3 The variables in the imputation

The variables in the imputation were:

- The Stage 2 scores, total, CPST and SJT
- Invited to Stage 3 but withdrew, as a binary variable
- The Stage 3 total score (i.e. the score from just the selection centre) and the four point score for the initial outcome, which was set as ordinal. The four values in order were 'Not demonstrated', 'Review unclear' (i.e. for review but unclear if the candidate would succeed'), 'Review likely' (i.e. for review and likely that the candidate would succeed) and 'Demonstrated'.
- Made an offer of a training place but withdrew, as a binary variable
- MRCGP marks at AKT and CSA, the AKT marks being in percentage points relative to the pass mark for the diet expressed as a percentage, and the CSA marks as marks on the raw mark scale relative to the pass mark for that day of examining.
- The ARCP data were coded as three variables:
    - Ever LTFT, as a binary variable
    - Ever OOP, as a binary variable
    - ARCP outcome 4 (i.e. modified Tiffin Group 5), as a binary variable

[3] The algorithm takes each variable in turn as the dependent variable, using all other variables as predictors, and then uses those predictions to impute new imputed values, the progress iterating until the maximum number of iterations has been reached, the SPSS default being 10.

[4] Many sources, following Little and Rubin, say that five imputations is adequate, but we chose to use ten. As the areglmpute manual in R says, "i.impute=5 is frequently recommended but 10 or more doesn't hurt".

- Time in FTE years of training for entry to the GP Register. This variable was ordinal, and took the values <= 3 years, 3-3.5 years, 3.5 – 4 years, 4 – 4.5 years, 4.5 to 5 years, 5 to 5.5 years, 5.5 – 6 years, and 6 years + (including not yet on the Register)[5].

- Demographic variables of UK or non-UK medical school graduate, Sex (male vs female), and ethnicity (BME vs White), all of which are binary variables. The demographic variables were present in most cases, but in the tiny proportion where they were missing they were not imputed but instead used only as background variables in the imputation of other variables[6].

- Round of application. For the 2010 dataset only, as described elsewhere, the raw data did not specify whether applications were in Round 1 or Round 2, and that information seems to have been irremediably lost. For 2010 only we therefore imputed a selection round for use in some of the models, using information available from the other years for which round was available[7].

5.2.4 The level of the analyses

The analyses in Chapter 4 were carried out at the level of the candidate, rather than the application, since it was the performance of individual doctors that was of interest. The present chapter, which also sets up the key data sets for Chapter 7, wishes however to compare different selection processes, and is therefore **at the level of the application**. Selection runs on annual cycles, as does training, and changes to selection processes take place in annual cycles. The primary interest is therefore in those **applications** which are accepted within a particular year, and results in doctors who do or do not enter GP training, so that the application is the natural level of analysis.

5.2.5 Years analysed

Data are available for seven selection years, from 2009 to 2015. However data for many of these years are structurally incomplete as insufficient time has passed to allow data to be collected. The problem is particularly acute with the key outcome measure of time in FTE years to enter the GP Register, for which, as Chapter 4 shows, it is really only the 2009 and 2010 cohorts which have sufficient data over a long enough time period to track through to five FTE years after starting training. The 2009 and 2010 cohorts did have slightly different criteria for the CPST and SJT in Stage 2, and Stage 3 had a number of differences for the two earliest years, but nevertheless it is assumed that the properties of the assessments are broadly similar. Time to GP Register was set as missing for all cohorts from 2011 onwards as follow-up was too short to be reliably complete. For similar reasons, ARCP outcomes were only set as available for the 2009 and 2010 cohorts. MRCGP outcomes were not available for all cohorts, insufficient time having passed, and CSA was only set as available for cohorts up to and including 2013, and AKT for cohorts up to and including 2014. It should be noted that we did not attempt to include the FPAS measures in the analysis since they are only available for the latest cohorts, and those cohorts have no data available for MRCGP AKT or CSA. This pattern is very similar to that described by Little and Rubin as "file matching" (p.5), and they describe the inevitable limitations that arise when data for calculating certain correlations are missing for all participants (p.158), although approximations can sometimes be made if some assumptions are acceptable.

5.2.6 A brief note on how SPSS handles imputations

Without going into technical details it may help the reader to know how a program such as SPSS handles imputations. The raw data file contains 42,017 records, one for each application in the seven years for which complete Stage 2 data were available.

[5] As noted elsewhere, we have allowed a two month grace period when calculating time to GP Registration. A trainee obtaining GP Registration in exactly 4 years and 2 months would be grouped in the 3.5 – 4 years group.

[6] General advice is that where there is a possibility that data are NMAR then the effects can be mitigated by the inclusion of known correlates of the outcome variables. In many studies it has been shown that non-UK graduates, males, and non-white applicants and trainees perform less well at selection and training measures, and hence these demographic variables were included. As noted earlier, there is no definitive way of ensuring that data are MAR rather than NMAR.

[7] Not all years used Round 1 and Round 2, and some recent years have used other schemes. Here we designate Round 1 as the first selection round of the year and all other applications as Round 2.

SPSS creates a variable called Imputation_ and for the raw data Imputation_ has the value 0 (i.e. not an imputation). SPSS then creates ten imputed sets of data, each 42,017 records long, with values of Imputation_ from 1 through to 10. At the end the single file has 42,017 + 42,017*10 = 462,187 records, the first 42,017 from the raw data and the remaining 420,170 for the ten sets of imputed data. For any variable which has been imputed all of the values are complete, any missing values having been 'filled in'. All non-imputed variables in the raw data file remain in the raw data file and also are copied across into the imputed data sets, missing or not, and are available for statistical analysis. SPSS uses its 'split file' command to allow the independent imputations to be analysed separately. In addition the split file command recognises when the split file variable is Imputation_ and the program can then, for a range of commands, automatically calculate pooled estimates of the various parameters which are calculated, and can calculate a standard error[8].

5.3 CHECKING THE ACCEPTABILITY OF IMPUTED ESTIMATES

Multiple imputation is a complex process and can go wrong unless careful checks are made on the acceptability of imputed estimates for missing data. In this section we describe a graphical approach to checking acceptability which should also provide a visual demonstration to the reader of how the imputation process is working on a range of measures. We will begin by looking in detail at how actual and imputed Stage 3 total marks[9] (i.e. the pure selection centre score) are related to the total Stage 2 score. It should be remembered that Stage 3 total marks are only available in those who have been invited to the Stage 3 selection centre and therefore many values are missing.

Actual Stage 3 total marks are available for 32,266 candidates, with a mean score of 51.98 and a standard deviation (SD) of 7.28. The ten imputations had a rather lower mean score of 50.86, with a range from 50.83 to 50.88, with all of the imputations having a lower mean than the actual scores. The standard deviation of the imputed means (which can be treated as an approximate standard error in this case) was 0.013. **The imputed means are therefore significantly lower than the raw mean, which is to be expected with selection.**

The standard deviation of Stage 3 total marks in the ten imputations had a range of 7.58 to 7.62, all of which are higher than SD in the actual marks which was 7.28. The approximate standard error of the imputed standard deviations was 0.013[10]. **The imputed standard deviations are therefore significantly higher than the raw standard deviation, which also is to be expected with selection due to range restriction.**

A graph of the relationship between Stage 3 total marks and total Stage 2 scores is a useful way of checking the acceptability of the imputed values and seeing how imputation is working. Figure 5.2 shows a plot of Stage 3 total mark (vertical) against Stage 2 score (horizontal). The solid blue points are actual, raw data for candidates who took both Stage 2 and Stage 3. The blue points for the lowest Stage 2 scores have wide confidence intervals as there are few candidates with these scores who got through to Stage 3[11]. The open, coloured circles show the mean Stage 3 total marks in the imputed samples. For high Stage 2 scores the plotted points are a mixture of actual and imputed values (and it should be remembered that not all candidates with high Stage 2 scores accept the invitation to take Stage 3, and for these candidates the imputations will vary slightly in the values predicted, and may have a different mean to those of candidates who actually took Stage 3). Candidates with the lowest Stage 2 scores were not invited to take Stage 3 and therefore have no actual measured scores (although in principle they could have been asked to attend the selection centre). However imputation has estimated Stage 3 scores for all of those individuals using all of the information that it has available on them. The mean Stage 3 scores continue to

[8] The relationship between the standard error and the standard deviation can be confusing. The standard deviation is a description of the variability of a set of numbers. The typical standard deviation is a measure of the estimate of the variability of raw data. However when there are separate estimates of a parameter then the standard deviation of those estimates is what is usually called the standard error, as in the present case, where the standard deviation of the M imputed estimates is actually an estimate of the standard error of the estimate.

[9] The term 'Stage 3 total mark' was used on the raw data files with which we were provide and is the integer score between 16 and 64 for performance on the stations at the selection centre. It should not be confused with the Stage 3 final score which includes Stage 2 band scores (see Chapter 2).

[10] This is approximate as it would probably be better, although with only minimal difference, to have calculated the standard errors of the variances and not the standard deviations.

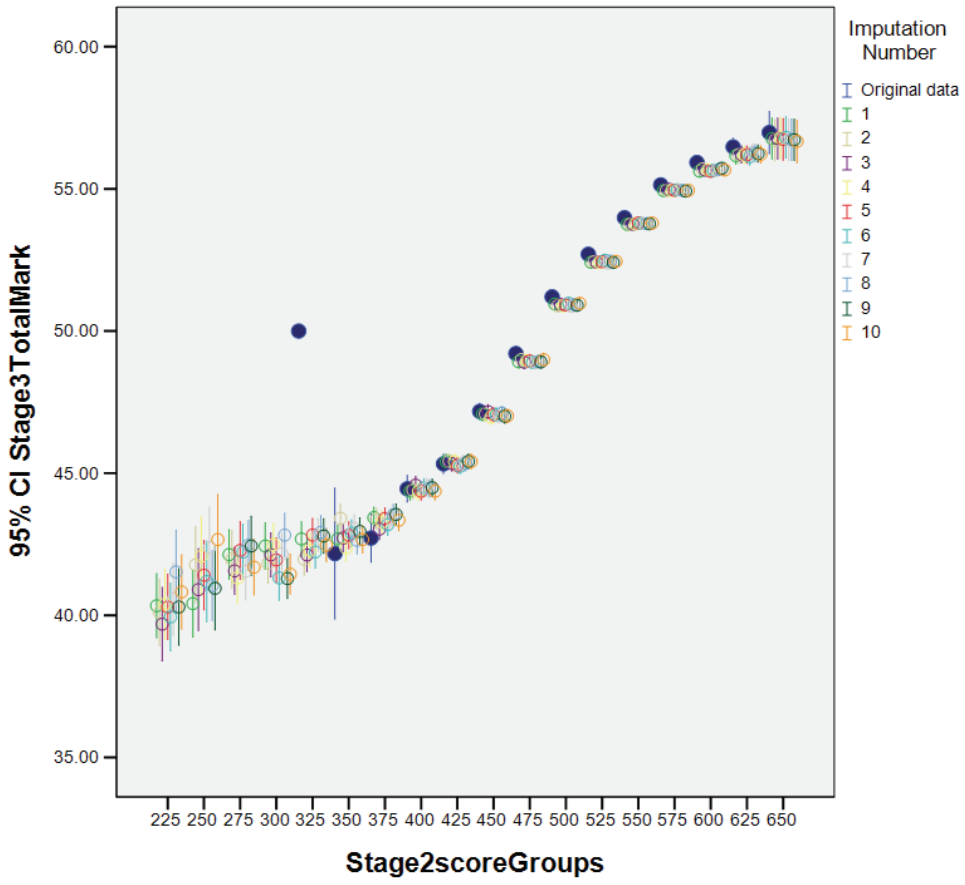
[11] The process by which these very few individuals took Stage 3 has not been explored further here.

decline as Stage 2 scores get lower (reflecting a known positive correlation between Stage 2 and Stage 3 performance), but the confidence intervals become wider, with more variation between the imputations, reflecting a growing uncertainty as to the precise levels of performance.

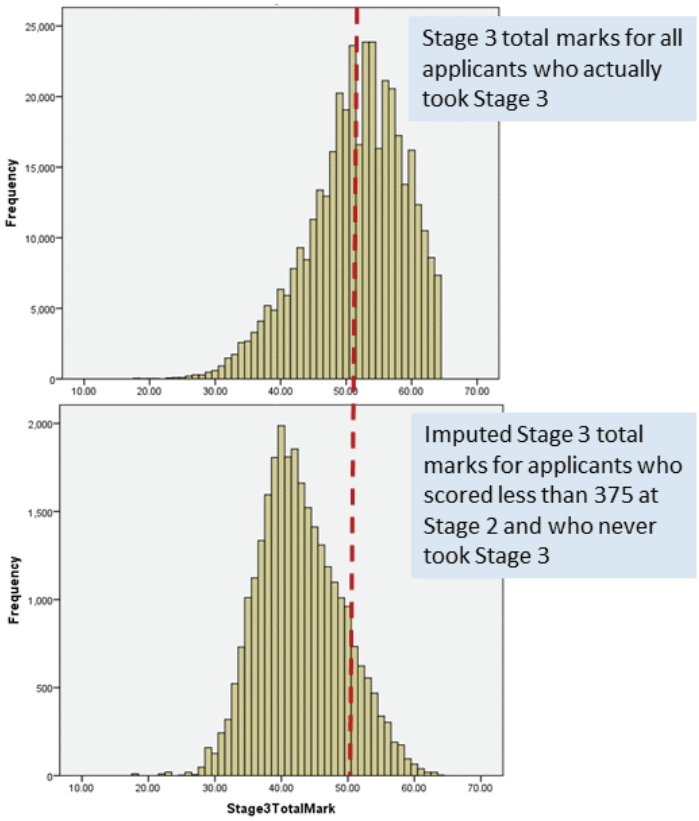
The confidence intervals in Figure 5.2 are confidence intervals for the mean, and do not directly show the variability of imputed values in candidates. The top part of Figure 5.3 shows the distribution of actual total marks attained by candidates taking Stage 3. The lower part shows the distribution of imputed scores for candidates who scored less than 375 at Stage 2 and did not take Stage 3. Although those who did not take Stage 3 have visibly lower imputed scores than the actual scores of those who did take Stage 3, the distributions nevertheless overlap substantially. Many of those taking Stage 3 had lower scores than the imputed scores for those who did not take it. The median score for those actually taking Stage 3 is 52, and 11% of the imputed scores for those not taking Stage 3 are above the median for those actually taking it. Figure 5.3 emphasises that there is a wide variability in imputed Stage 3 scores, just as there is a wide variability in those actually taking Stage 3. The imputed scores for Stage 3 are lower than the actual scores as Stage 2 scores are a valid predictor of Stage 3, although the correlation is not very high, primarily because of unreliability in Stage 3. The result is that many low scoring Stage 2 candidates might actually have achieved high marks at Stage 3.

Figure 5.2 and Figure 5.3 show how imputation can generate plausible marks for candidates who have not actually taken Stage 3[12].

» *Figure 5.2: Solid blue circles show raw (actual) Stage 3 total marks, with 95% confidence limits in the original data in relation to total Stage 2 score grouped with a bin width of 25. Note that there are very few data points for the lowest actual Stage 2 scores (and only one for the lowest, so that there is no confidence interval). Other colours with open symbols indicate mean and confidence intervals for imputed values.*



» Figure 5.3: Top: Actual Stage 3 total scores of all candidates taking Stage 3. Bottom: Imputed Stage 3 total scores for candidates scoring less than 375 at Stage 2 and who did not take Stage 3. The red vertical line shows those scores of 52 or above on Stage 3



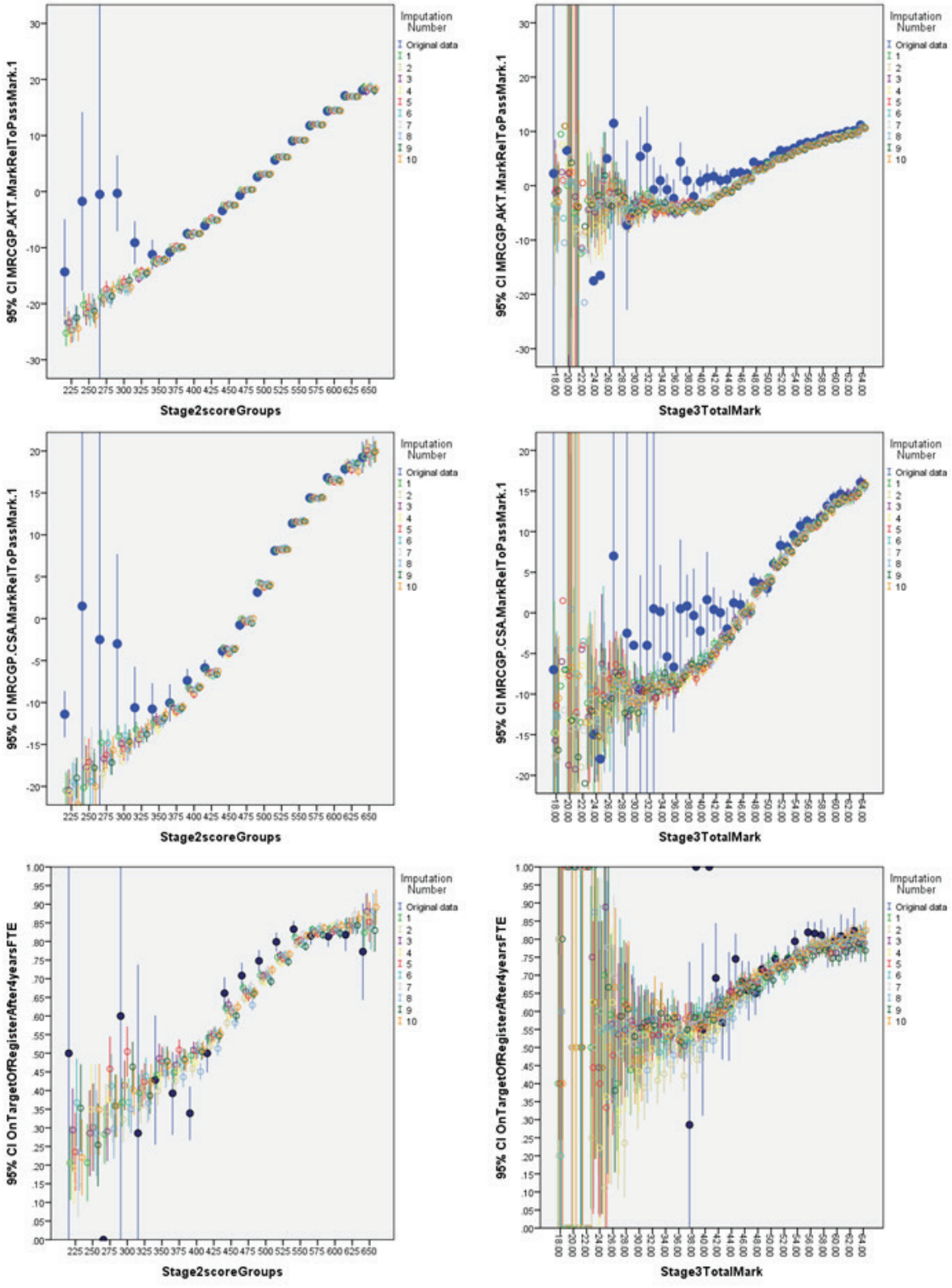
Imputed values should show plausible estimates when plotted against Stage 2 and Stage 3 scores, which are the key selection variables, and a range of imputed values are shown in Figure 5.4 and Figure 5.5.

Figure 5.4 shows the relationship between Stage 2 and Stage 3 scores and the major outcome variables of MRCGP AKT, MRCGP CSA and being on the GP Register within 4 years of FTE training. Note that while Stage 2 scores are present for all GP candidates, Stage 3 scores are only present for a proportion of candidates, and hence some values have been imputed. In all cases there is a strong relationship between Stage 2 scores and MRCGP AKT, CSA and being on the GP Register. As previously there are few candidates taking AKT and CSA who have low Stage 2 scores, and therefore those solid blue points have very wide confidence intervals (and hence are visually dominant but have little actual data). Imputed values for all three outcome measures are strongly linear with low Stage 2 scores predicting poorer outcomes. Stage 3 scores show a broadly similar picture although the relationship tends to be less strong, reflecting the lower reliability of Stage 3 scores, and while there is much variability estimated for imputed values for very low Stage 3 scores a monotonic trend is still visible.

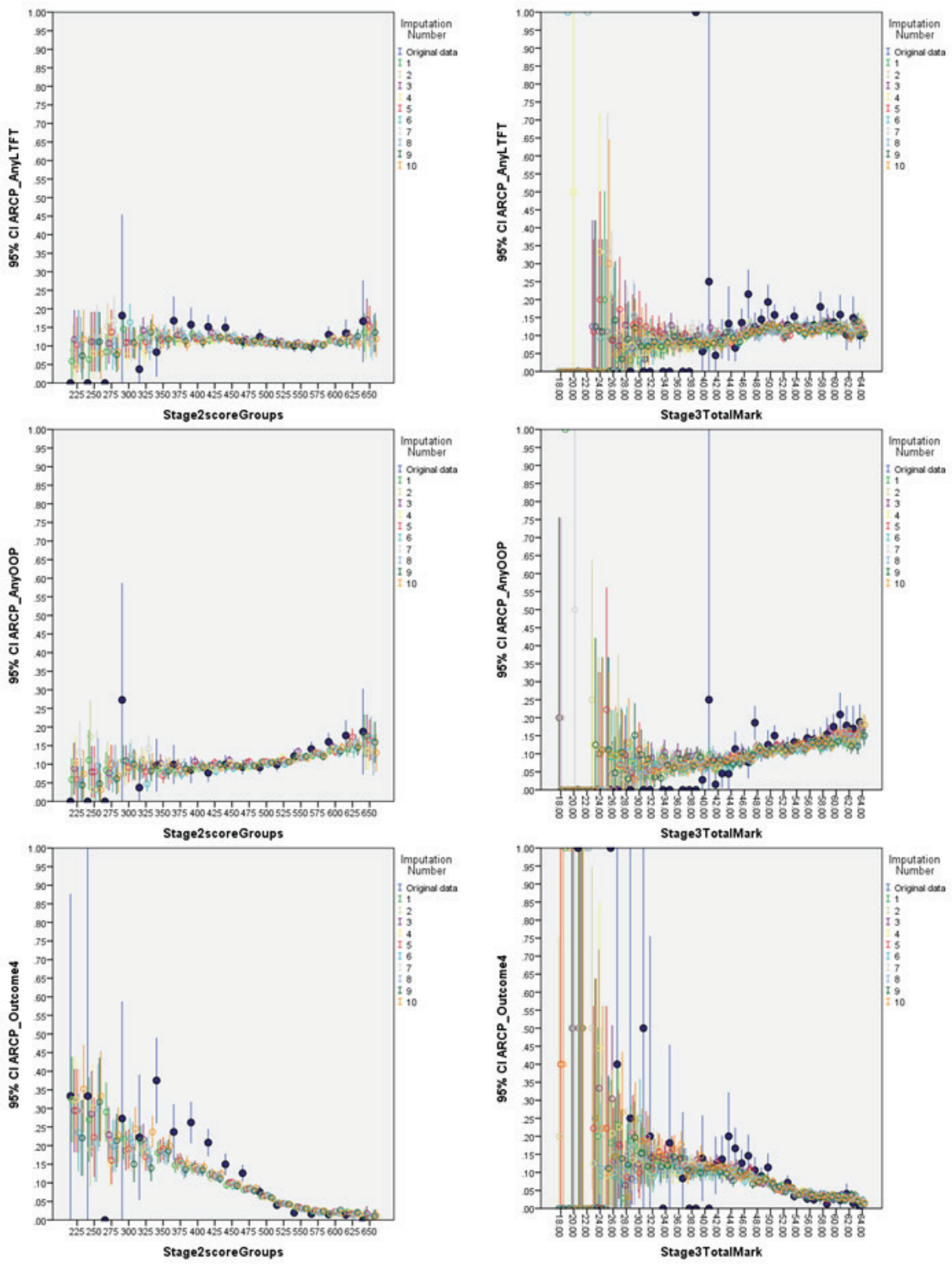
Figure 5.5 shows the relationship of Stage 2 and Stage 3 scores to three ARCP outcomes, at the top, the probability of a trainee being LTFT, in the middle of a trainee being OOP, and at the bottom of a trainee receiving an ARCP outcome 4 (mTiffin group 5). LTFT shows no relationship to with Stage 2 or Stage 3, and hence is not predictable from selection scores. In contrast, trainees

[12] A technique known as 'over-imputation' allows the plausibility of imputed marks to be checked more directly. A proportion, say 10%, of actual marks are artificially set as missing, values for those marks are then imputed, and the imputed marks can be compared with the actual marks as a quality control check. Time constraints however meant that it was not possible here. It should also be remembered that over-imputation will become less and less good as a) the reliability of Stage 3 gets lower and lower, and b) the correlation between Stage 2 and Stage 3 becomes lower.

» *Figure 5.4: Relationship of actual and imputed values to Stage 2 scores (left) and Stage 3 scores (right); note that Stage 3 scores are a mixture of true and imputed values. Outcome variables are MRCGP AKT (top row), MRCGP CSA (middle row), and proportion of trainees on GP Register after 4 four years (bottom row). The format is the same as that for Figure 5.2 (and see main text for further descriptions).*



» Figure 5.5: Relationship of actual and imputed values to Stage 2 scores (left) and Stage 3 scores (right); note that Stage 3 scores are a mixture of true and imputed values. Outcome variables are Any LTFT in ARCP (top row), Any OOP in ARCP (middle row), and ARCP Outcome 4 (mTiffin group 5) (bottom row). The format is the same as that for Figure 5.2 (and see main text for further descriptions). Note that in the bottom row it is the rate of poor outcomes which is being plotted.



who are OOP have higher scores on both Stage 2 and Stage 3. Finally it can be seen that there is a strong relationship between ARCP outcome 4s (leaving the training programme) and lower scores on Stage 2 and, with rather more variability, on Stage 3.

All of the imputed values seem plausible (as do others not shown here), and therefore can reasonably be used to model the consequences of different approaches to selection, as well as calculating predictive validity across the entire range of candidates, rather than only on those who have been selected, thereby providing a partial solution to the recurrent problem of range restriction in selection studies.

5.4 EVALUATING THE OUTCOMES OF DIFFERENT APPROACHES TO SELECTION

The imputed data mean that all candidates for GP training, irrespective of their endpoint in the selection process, have estimated values for all of the selection measures as well as all of the outcome measures. Such a dataset means that it is possible, virtually, to recreate the probable outcomes for various methods of selection. There are also ten imputed datasets which means that it is possible to estimate the likely variability or uncertainty in the various outcomes as a result of the random component of selection. Figure 5.6 shows a range of different selection models and the likely outcome of those selection models. All of the models are included which are later used in Chapter 7 to assess the economic consequences of various selection methods, but in addition there are some other models which are conceptually simpler and therefore help to illustrate important points about the nature of the selection process. Figure 5.6 is complicated, and therefore the reader will be walked through it step-by-step.

5.4.1 Modelling

Although the imputation process has used all possible data for the selection years 2009 to 2015, the modelled outcomes in Figure 5.6 are based on selection for the years 2011 to 2014 when the selection processes were relatively stable.

5.4.2 Explaining Figure 5.6

The columns of Figure 5.6, labelled A to AB, correspond to the various outcome measures, and the rows, labelled 1 to 35, to different selection methods.

Model 2 (Baseline), in bright green in rows 6 to 8, shows what we call the 'Baseline model', which corresponds to the present selection system. A strong contrast with the Baseline model is Model 1 (Random selection, n=3,250), shown in orange in rows 3 to 5, where trainees are selected entirely at random from candidates until the available places are filled. Although the random selection model is not realistic, it does provide a conceptually important comparison against which all other models, including baseline, can be compared. The comparisons make a key point that in general, if there is a reasonably high selection ratio, then, for those selected, any form of selection is better than no selection (i.e. filling posts at random)[13].

Columns C to AB show various outcome measures, grouped to make it easier to see how the data are organised. Column C shows the total number of trainees selected. For the Baseline model it averages 3,014/year, which is less than the target we have used of 3,250 per year. Such a shortfall has economic consequences which will be discussed later. In contrast the random selection model in rows 3 to 5 has its quota of 3,250 trainees, since places are filled until the target is reached.

The outcome measures take various forms, with the most important being the time taken for trainees successfully to enter the GP register. Columns D to H (in black font) show the number of candidates who entered the Register in 3 years, 3 to 3.5 years, 3.5 to 4 years, 4 to 5 years, and 'Other' (either 5+ years or never), the numbers totalling those in training. Columns I to M (in blue font) show the cumulative percentages of trainees on the Register within 3, 3.5, 4, 4.5 or 5 years. Although the percentage of doctors on the GP Register within 4 years is higher with baseline than random selection (76.8% vs 73.9%), the absolute numbers at four years are higher for random selection (1,692+352+358=2,402 vs 1,699+296+321=2,316) as random selection takes in more trainees and hence more get through to the Register, even given that on average they do less well. Since there are costs to not having doctors in post then both percentage success rates and absolute numbers need to be compared for different selection methods.

» Figure 5.6 Modelled consequences of different selection methods for a range of outcomes. See text for further details.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1			Total	Number on GP Register in [illegible] year					% on GP Register in N years				
2	Model	Imp#	Trainees	<=3	3-3.5	3.5-4	4-5	Other	3yrs	3.5yrs	4yrs	4.5yrs	5yrs
3	Model 1: Random selection n=3,250												
4	Mean	10	3,250	1,692	352	358	158	691	52.1	62.9	73.9	76.9	78.7
5	SD			[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
6	Model 2: Baseline												
7	Mean	10	3,014	1,699	296	321	144	554	56.4	66.2	76.8	79.7	81.6
8	SD		0	21	14	25	10	14	0.68	0.71	0.83	0.84	0.46
9	Model 3: Best with Stage 2 score >=575, no withdrawals												
10	Mean	10	1,461	975	94	150	56	186	66.7	73.2	83.4	86.0	87.2
11	SD		0	21	9	14	9	14	1.42	1.55	0.95	1.00	0.94
12	Model 4: Lower Stage 2 cut score to pre-2013 value of 166 on each test												
13	Mean	10	3,108	1,733	308	330	91	61	55.8	65.7	76.3	79.2	81.2
14	SD		0	21	15	26	7	7	0.68	0.74	0.63	0.63	0.48
15	Model 5: 3250 with best Stage 2 - no withdrawals												
16	Mean	10	3,250	2,009	245	342	135	519	61.8	69.4	79.9	82.6	84.0
17	SD		0	46	22	34	18	30	1.43	1.36	0.91	0.96	0.93
18	Model 6: No Stage 3												
19	Mean	10	3,250	1,892	307	354	143	554	58.2	67.7	78.5	81.3	83.0
20	SD		0	24	15	24	8	18	0.75	0.87	0.69	0.62	0.54
21	Model 7: Stage 3 bypass (at 575)												
22	Mean	10	3,061	1,720	2021	2348	2436	2495	56.2	66.0	76.7	79.6	81.5
23	SD		1	21	22	20	14	14	0.69	0.72	0.65	0.61	0.46
24	Model 8: Equal weight to all (3,250 posts filled in Round 1)												
25	Mean	10	3,250	1,849	325	352	93	56	56.9	66.9	77.7	80.6	82.3
26	SD		0	22	15	25	7	7	0.68	0.78	0.65	0.50	0.40
27	Model 9: Stage 3 only, 48+ - No withdrawals												
28	Mean	10	4,304	2,334	421	455	220	874	54.2	64.0	74.6	77.5	79.7
29	SD		6	34	50	50	24	40	1.28	1.22	0.97	1.11	0.95
30	Model 10: Best with Stage 3 score, no withdrawals												
31	Mean	10	3,250	1,817	299	349	163	622	55.9	65.1	75.9	78.6	80.9
32	SD		0	35	18	56	20	28	1.08	1.05	0.94	1.10	0.85
33	Model 11: Stage 3 only (3,250 posts filled in Round 1)												
34	Mean	10	3,250	1,805	343	353	92	63	55.5	66.1	77.0	79.8	81.7
35	SD		0	24	14	25	7	10	0.75	0.77	0.73	0.72	0.47

	A	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1		MRCGP AKT		MRCGP CSA		LTFT, OOP and ARCP outcome 4						All trainees			UK grads	
2	Model	mean	SD	mean	SD	nLTFT	nOOP	nARCP4	%LTFT	%OOP	%ARCP4	%UK	%BME	%Female	%BME	%Female
3	Model 1: Random selection n=3,250															
4	Mean	4.90	[illegible]	6.74	[illegible]	361	362	200	11.1	11.1	6.2	71.4	50.0	60.4	35.4	64.8
5	SD	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
6	Model 2: Baseline															
7	Mean	7.46	[illegible]	10.23	[illegible]	373	397	118	12.4	13.2	3.9	85.6	40.3	65.5	32.9	66.8
8	SD	0.02	[illegible]	0.00	[illegible]	17	15	6	0.57	0.50	0.20	0.00	0.00	0.00	0.00	0.00
9	Model 3: Best with Stage 2 score >=575, no withdrawals															
10	Mean	13.44	[illegible]	15.48	[illegible]	160	190	34	10.9	13.0	2.3	98.1	21.6	72.9	21.2	73.0
11	SD	0.06	[illegible]	0.06	[illegible]	8	8	4	0.56	0.55	0.25	0.00	0.00	0.00	0.00	0.00
12	Model 4: Lower Stage 2 cut score to pre-2013 value of 166 on each test															
13	Mean	7.27	[illegible]	9.94	[illegible]	386	409	124	12.4	13.2	4.0	84.2	40.6	65.2	27.5	55.8
14	SD	0.02	[illegible]	0.00	[illegible]	13	15	7	0.56	0.48	0.23	0.00	0.00	0.00	0.00	0.00
15	Model 5: 3250 with best Stage 2 - no withdrawals															
16	Mean	10.40	[illegible]	12.60	[illegible]	349	397	106	10.7	12.2	3.3	92.8	32.1	66.3	29.0	67.3
17	SD	0.06	[illegible]	0.10	[illegible]	13	16	6	0.53	0.48	0.20	0.00	0.00	0.00	0.00	0.00
18	Model 6: No Stage 3															
19	Mean	7.99	[illegible]	10.27	[illegible]	383	411	128	11.8	12.6	3.9	85.7	39.6	64.6	32.6	66.3
20	SD	0.08	[illegible]	0.13	[illegible]	13	17	8	0.56	0.45	0.27	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
21	Model 7: Stage 3 bypass (at 575)															
22	Mean	7.38	[illegible]	10.11	[illegible]	379	402	122	12.4	13.1	4.0	85.0	39.8	65.2	27.5	56.5
23	SD	0.03	[illegible]	0.09	[illegible]	17	16	7	0.56	0.51	0.23	0.01	0.02	0.02	0.02	0.02
24	Model 8: Equal weight to all (3,250 posts filled in Round 1)															
25	Mean	7.34	[illegible]	9.59	[illegible]	387	409	138	11.9	12.6	4.2	82.5	42.0	63.9	27.5	54.3
26	SD	0.08	[illegible]	0.16	[illegible]	19	16	8	0.57	0.40	0.26	0.4	0.2	0.1	0.2	0.2
27	Model 9: Stage 3 only, 48+ - No withdrawals															
28	Mean	6.64	[illegible]	9.34	[illegible]	504	528	194	11.7	12.3	4.5	81.7	43.7	63.9	34.4	65.8
29	SD	0.05	[illegible]	0.10	[illegible]	19	20	7	0.46	0.48	0.16	0.2	0.1	0.1	0.1	0.1
30	Model 10: Best with Stage 3 score, no withdrawals															
31	Mean	7.45	[illegible]	10.80	[illegible]	387	415	128	11.9	12.8	3.9	86.5	39.2	66.4	32.4	67.8
32	SD	0.05	[illegible]	0.09	[illegible]	15	18	5	0.48	0.56	0.17	0.2	0.2	0.1	0.1	0.1
33	Model 11: Stage 3 only (3,250 posts filled in Round 1)															
34	Mean	6.65	[illegible]	9.54	[illegible]	406	419	145	12.5	12.9	4.5	83.2	40.8	64.9	26.9	55.4
35	SD	0.04	[illegible]	0.10	[illegible]	18	16	6	0.54	0.50	0.19	0.1	0.1	0.1	0.1	0.1

Although we will not emphasise the point in the remainder of this section, it is worth mentioning that the difference between the success rates of 73.9% and 76.8 for the two selection methods shown in column K is statistically meaningful, the multiple imputations allowing approximate standard errors to be calculated, which are shown as the SDs in rows 5 and 8, with values of 0.63% for baseline and 1.10% for random selection. Random selection also tends to be more variable than baseline selection, in part because the numbers of very weak trainees who do poorly in training depends on sampling variation. The table provides standard deviations across imputations for all of the measures and such comparisons are therefore always possible.

Columns N to Q show outcomes at the MRCGP exams, for AKT (columns N and O) and CSA (columns P and Q), with the mean scores in red font in columns N and P and the standard deviations in columns O and Q. Baseline selection results in candidates who perform better at their first attempt at both AKT and CSA (mean scores of 7.46 and 10.23) than does random selection (4.90 and 6.74) reflecting the use of the Stage 2 and Stage 3 scores in baseline selection.

Columns R to AB show the extent to which different selection methods impact upon other outcomes which are potentially of interest and importance. Columns R to W show various ARCP outcomes, firstly in columns R to T as absolute numbers of trainees and then in columns U to W as percentages of trainees. Percentages are easier to interpret but absolute numbers inevitably relate better to service provision implications. Columns U to W show the percentages of trainees who at some time are LTFT or OOP, and the percentage who have an ARCP Outcome 4 (released from training). The most striking difference between baseline and random selection is the much higher proportion of trainees with Outcome 4, 6.2% vs 3.9%. ARCP outcome 4 is expensive for the training process, and a rate nearly 60% higher inevitably will have cost implications.

Columns X to AB show the demographic characteristics of selected trainees, in terms of place of primary medical qualification divided into UK and non-UK graduates, ethnicity (white vs BME), and sex (male vs female). Such measures are important in determining whether different selection methods may have adverse impact upon particular demographic group. Columns X to Z show the proportion of UK, BME and female trainees for all trainees. Since non-UK trainees tend to be BME and are more likely to be male, the numbers in columns X to Z tend to co-vary. Columns AA and AB therefore also show the proportion of trainees who are UK graduates who are BME or female. Compared with Model 1, random selection, Model 2, baseline selection, results in trainees being more likely to be UK graduates and to be female, and to be less likely to be BME, with a similar pattern amongst UK graduates where baseline selection results in more female and fewer BME trainees.

5.4.3 Comparing different models of selection

Figure 5.6 compares a number of different models of selection. The previous section has already compared baseline selection, the system currently used, with random selection, which can be used to assess the extent to which selection is useful at all. Random selection results in lower marks at MRCGP AKT and CSA, a lower proportion of doctors on the GP Register after 3, 4 or 5 years, and more trainees with the problematic ARCP 4 outcomes. On the other hand it is more inclusive, with more non-UK graduates as trainees, more BME doctors as trainees, and somewhat more male trainees. Baseline selection consists of Stage 2 selection in the form of the CPST and SJT tests, followed by the Selection Centre in Stage 3. An immediate question asks the extent to which Stage 2 or Stage 3 is the more effective in selection, or whether it is a combination of the two which works particularly well. Such questions can be asked in two ways, either by selecting only candidates who do particularly well at the various stages, when numbers of trainees may not be sufficient to satisfy needs, or using the selection measures in a form which guarantees that at least 3,250 trainees are selected, so that there are sufficient trainees entering training (although of course not all may exit onto the GP Register).

Stage 2 selection variants:

Variations on models which primarily use Stage 2 are shown in Models 3 to 6 in the figure, all of which are shown in light blue. The impact of Stage 2 selection in a pure form can be seen by looking at **Model 3 (Best with Stage 2 score >= 575, no withdrawals)** in rows 9 to 11, in which the only form of selection is that candidates with a (high) Stage 2 score of 575 or more are selected. This is a purely illustrative model, and not intended to be practical, not least as it is assumed that all candidates scoring >= 575 will accept offers and go into training, but it is a useful 'thought experiment'. These high-scoring Stage 2

trainees are very successful, with a very high score at AKT (13.44) and CSA (15.49), and a very low rate of ARCP outcome 4 (2.3%), and 83.4% are on the Register after 4 years of FTE training (column F). Inevitably there has to be a catch, and it is shown in column C, where only 1461 trainees are accepted under such a criterion, far below the numbers that are actually needed. Strong selection on Stage 2 is therefore effective at increasing pass rates and MRCGP performance, but it is at a cost of not having sufficient trainees, the numbers being far short of what is required. It is also worth noticing the impact on the demographic mix of trainees, 98% being UK graduates, only 21% being BME and 73% being female, the proportions being very similar in just the UK graduates since almost all of the trainees are UK graduates.

A conceptually similar approach to **Model 3**, is in **Model 5 (3,250 with best Stage 2 – no withdrawals)**, where selection is still entirely on Stage 2 total score, but the 3,250 candidates with the highest scores on Stage 2 are offered training places (and in this simple model, all accept them). Model 5 is shown in rows 15 to 17. There are now 3,250 trainees (column C), so that all places are filled, and the pass rate at 4 years is higher than for the baseline model at 79.9% (column K), and that in part is explained by high AKT and CSA scores of 10.4 and 12.6, with an ARCP outcome 4 rate of 3.3% (column W) which is a little but significantly lower than for baseline. Once again there is though an impact on demographics, with 92.8% of these trainees being UK graduates, and 32% being BME, although there is little impact on the proportion of female trainees. Model 5 shows therefore that a selection method can be better than Baseline selection, both in getting a higher success rate at outcome, and also filling all training places.

Model 5 is somewhat unrealistic, in assuming that all candidates accept places, etc, but **Model 6 (No Stage 3)** in rows 18 to 20, which is used in the economic modelling in Chapter 7, does model dropout, etc, and behaves in a similar way to Model 5, filling all training places. Compared with Model 5, there is a slightly lower four year success rate at 78.5%, which higher than the 76.8% in the baseline model. MRCGP marks and ARCP outcome 4 are very similar in Model 6 to those at Baseline, and the demographic mix is similar to that at Baseline. The major advantage of Model 6 is therefore that the 3,250 training places are filled with little obvious detriment.

The remaining model in this section is not a pure Stage 2 model but is a variant on Baseline, altering the way that Stage 2 is used. The cut scores for the bands at the Stage 2 CPST and SJT were increased in 2011, and **Model 4 (Lower Stage 2 cut score to pre-2013 value of 166 on each test)**, in rows 12 to 14, explores the effects of reducing the cut scores to their earlier values. Compared with Baseline there is a slightly lower four year success rate and slightly slower AKT and CSA scores, but a similar ARCP Outcome 4 rate. There are slightly more trainees (3,108 compared with 3014), the total being closer to the notional target of 3,250, but the number is still not sufficient.

Stage 3 selection variants:

The three models in purple at the bottom of the figure, in rows 27 to 35, show the effects of selection using only Stage 3, so that all candidates would attend the selection centre. **Model 9 (Stage3 only, 48+ No withdrawals)** in rows 27 to 29 selects only candidates with a total score on the selection centre of 48 or more. The value of 48 was chosen as a candidate scoring 3 at each criterion on each of the stations would probably be deemed to have demonstrated their competency, and they would have a total score of 48. The first thing to notice is that there would be 4304 candidates who were selected, more than is required. Of those trainees, 74.6% would be on the Register within four FTE years, which is lower than the Baseline model value of 76.8%). Average marks on AKT and on CSA are also a little lower than for Baseline. The latter is particularly interesting as it might have been expected that pure Stage 3 selection should result in higher attainment on CSA since it is also a selection-centre/OSCE type of assessment of similar competencies. The rate of ARCP outcome 4s is also higher at 4.5%. In terms of demography, Model 9 results in somewhat fewer non-UK graduates being selected. **Model 10 (Best with Stage 3 score, no withdrawals)**, in rows 33 to 35, is similar in structure to Model 5 for Stage 2 except that it takes the 3,250 best candidates in terms of Stage 3 scores (but does not account for withdrawals etc). In this model all training posts are filled, and 75.9% of trainees are on the Register within 4 FTE years, which is probably not different from the Baseline model. MRCGP AKT scores are similar to Baseline, but MRCGP CSA scores are slightly higher, ARCP outcome 4 rates are very similar to Baseline, as also is the demographic mixture. The final model in this section, **Model 11 (Stage 3 only (3,250 posts filled in Round 1)**, in rows 33 to 35, is one of those used in Chapter 7 for assessing the economic consequences of selection, and it is similar to Model 10 in that all 3250 posts are filled, except that withdrawals etc are taken into account. The proportion of those on the Register after four FTE years is similar to baseline, but performance is somewhat lower on both MRCGP AKT and CSA, and ARCP outcome 4 rates are higher at 4.5%, with demographic impacts being broadly similar to Baseline, but with somewhat more non-UK

graduates. **Taken overall**, it is clear that selection on just Stage 3, with all candidates attending a selection centre, is broadly similar to Baseline in its consequences, although there is a tendency for MRCGP AKT and CSA scores to be a little lower, and since these predict ARCP outcome 4, for Outcome 4 rates to be somewhat higher.

Hybrid selection models:

The figure also shows, in brown in rows 21 to 26, two hybrid selection models (in addition to Baseline, which of course is also a hybrid model), using both Stage 2 and Stage 3 scores. **Model 7 (Stage 3 bypass at 575)** is a model in which candidates scoring 575 or more on Stage 2 by-pass the Stage 3 selection centre and are accepted immediately into GP training. Model 7 results in only 3061 trainees being selected, which is a little higher than the Baseline rate of 3,014, but is not the desired 3,250. The proportion of trainees on the Register after 4 FTE years is almost identical to Baseline, the MRCGP AKT and CSA marks are very slightly lower than Baseline, the ARCP outcome 4 rate is the same as at Baseline, and the demography is almost identical to Baseline except that amongst UK graduates there are fewer BME and female trainees. Overall a Stage 3 bypass has very little effect on the pattern of outcomes. The other hybrid model, **Model 8 (Equal weight to all, 3250 posts filled in Round 1)**, weights the CPS, SJT and Stage 3 equally, and after accounting for dropouts takes the 3,250 best qualified candidates. All training posts are now filled, the four year completion rate is a little higher than for Baseline, MRCGP AKT and CSA marks are slightly lower than for Baseline, ARCP outcome 4 rates are perhaps a little higher at 4.2%, there is a little higher rate of non-UK graduates, but fewer BME trainees amongst UK graduates. Taken overall, perhaps the main conclusion is that hybrid selection models tend to be midway between pure Stage 2 and pure Stage 3 models in their consequences.

5.4.4 Summarising the different selection models

Figure 5.6 is complicated, and the different models show different processes in action, some of which have effects that pull in different directions, and make any simple interpretations hard to make. Having said that, a number of principles seem to be visible in the models of Figure 5.6, and the results of Figure 5.6 can also be put into a broader context.

- **Any form of selection is better than not selecting (random selection).** Random selection (Model 1) is always worse than other selection methods. This fits the general view of economists that selection always pays off compared with no selection. The question therefore is not whether selection is effective at all, but how selection itself can be the most cost-effective.

- **Strong selection has excellent outcomes, but it is at the price of not filling all training posts.** This is clearest for Model 3, which takes only those scoring very highly at Stage 2. 83% of trainees would be on the GP Register within four years, but there would be only 1,461 trainees selected, leaving a serious short fall in trained GPs.

- **Candidates vary and less good ones tend to do less well.** Not all candidates are equally good, for a host of reasons, and less good ones will have less good outcomes. The further one has to go down the list of candidates then the less good will be outcomes.

- ***Lower* success rates can sometimes result in *more* trainees on the GP Register.** 'More is sometimes more' as a slightly lower success rate can be compensated for by the percentage being applied to more trainees. This has been shown even when comparing Baseline selection with Random selection. If 'people in posts' is all that is required then random selection can be effective, but probably only at much higher later costs.

- **Selection is more powerful when selection tests are more reliable.** Comparing selection based only on Stage 2 and only on Stage 3 shows that Stage 2 is far more effective. That is probably because Stage 2 is a more reliable assessment than Stage 3, and hence high Stage 2 scores are more likely to deliver better later performance than high Stage 3 scores. A totally unreliable selection test is, of course, a lottery, and it would not be different from random selection, except in so far as it incurs all of the costs of administering the 'test'. Higher reliability always makes assessments more effective.

- **Hybrid selection methods give hybrid outcomes.** Averaging across multiple methods gives results which tend to be the average of the methods taken individually. If one method is doing particularly well then adding in other

methods will tend only to dilute that excellence. Although it is tempting to say that, "a mixture of measures covering a broad range of attributes is likely to provide the best way of selecting our future doctors" (Morrison, 2016), the reality may be that if one tries to select on everything then in effect one selects on nothing, each selection component being diluted by all of the others (McManus & Vincent, 1997).

- **Selection scores co-vary.** Although Stage 2 and Stage 3 are conceptually distinct (as are CPST and SJT within Stage 2, and the ES, CS, CT&PS and PI scores within Stage 3), assessing different skills or competencies, the reality is that candidates who score highly on one tend to score highly on all of the others (although the results for that are not shown in the present chapter). That means that a) it is difficult to fine-tune selection for particular attributes, and b) that selecting on the most reliable and the most easily administered tends to result in candidates who are better on all of the attributes.

- **Some things cannot be predicted by selection tests.** The 'pure' models, such as Models 5 and 10, in which the best candidates on Stage 2 or Stage 3 are selected, perform better than those such as Models 6 and 11 in which realistic expectations about dropout from the selection process or not accepting offers are included in the models. Dropout tends not to be well correlated with selection scores and hence is not well predicted by selection scores.

- **Stage 2 and Stage 3 scores are not measures of motivation.** None of the models result in more than 87% of trainees being on the GP Register after 5 years. The reasons for such failures are unlikely to be lack of academic ability (or the selection tests would predict them), and are more likely to be a composite of factors due to motivation, lack of interest, competing career attractions, or simply the 'stuff' of everyday life, perhaps involving illness, personal relationships, family problems, domestic responsibilities, or a host of other factors that would never be predicted by Stage 2 and Stage 3 scores. Whether they are predictable at all is another matter.

- **There are other outcomes than are important but cannot necessarily be measured.** Although Figure 5.6 considers a number of outcomes, there are many other which may be important but cannot be considered. Perhaps the most important are patient-care outcomes. Doctors are trained to help patients have better healthcare, and lower morbidity and mortality, and those should be the ultimate outcome measures. Although it is tempting to suggest that 'mere knowledge' tests, such as CPST in Stage 2, are not relevant to such outcomes, one of the very rare studies that looks at academic assessments in trainee doctors in relation to mortality in their patients a decade or more later, finds a strong relationship, doctors doing better in assessments having patients who do better (Norcini, Boulet, Opalek, & Dauphinee, 2014). That result suggests that, say, selecting only on Stage 2 might be effective in wider terms.

- **Assessment has its costs, as also does appointing the wrong trainees or not appointing sufficient trainees.** The comparisons of the models at present take no account of the costs of the various assessments and the costs incurred by the different outcomes. That is a complex issue and will be left until Chapters 6 and 7.

5.5 PREDICTIVE VALIDITY OF STAGE 2 AND STAGE 3

A key question for evaluating the Stage 2 and Stage 3 selection tests is their predictive validity for the outcome variables of MRCGP AKT and CSA, ARCP outcome 4, and being on the GP Register within four FTE years of training. A problem in any such analysis of selection, as described at the beginning of this chapter, is that not all candidates take Stage 3, and not all of those taking Stage 3 progress into training, so that there is restriction of range, which can reduce the validity coefficients which are calculated.

In a study carried out at the same time as the present study, Patterson et al were commissioned by the GMC to look at the relationship between GP selection scores and MRCGP outcomes (Patterson, Kerrin, Baron, & Lopes, 2015). The study looked at trainees entering training between 2008 and 2012 for whom Stage 2 and Stage 3 scores could be matched to MRCGP AKT and CSA scores. In addition IELTS and PLAB scores were available to the researchers[14]. Restriction of range corrections were

[14] Despite a number of requests, we were not able to obtain these data.

applied (p.15), although separately for UK and IMG trainees, and no values are given for the entire set of trainees. Although a slightly different dataset was used to the present one, and the methods are different at various places, comparisons are possible, and should give broadly similar results. Correlations, which are validity coefficients, are shown in Table 5.1. The Patterson et al study does not report total Stage 2 scores, but only gives separate scores for CPST and SJT, and it reports correlations to only two decimal places. Patterson et al. divide trainees into UK and IMG groups, and while they report simple correlations for all trainees and UK and IMG groups separately, they only provide correlations corrected for range restriction for the separate UK and IMG groups. Despite the various differences between the datasets, it is clear that the raw correlations are similar in the two studies[15], the mean differences of the six correlations being -.001, with a standard deviation of the differences being .029. For the present study the eight correlations corrected for range restriction are about 17% higher than the raw correlations. Range restriction does therefore have an effect in reducing the apparent predictive validity of the Stage 2 and Stage 3 selection tests, the effect being larger for Stage 3 (a 38% increase) compared with Stage 2 (increases of 8% and 15% for CPST and SJT respectively), the benefit to Stage 3 being because range restriction is greater for it. However Stage 3 still has a rather lower predictive validity (.401 and .516) than CPST (.790 and .582) and SJT (.564 and .604).

5.5.1 Incremental validity

Patterson et al report estimates of the incremental validity of Stage 3 over Stage 2 for predicting MRCGP AKT and CSA. Here we report similar analyses, although not all are identical[16].

AKT:

In the current raw data, CPST and SJT together explain 57.1% of the variance in AKT, and the inclusion of Stage 3 increases that proportion to 57.2%, an increment of 0.1%. Those values are similar to the values of 55.3% and 0.4% reported by Patterson et al. Correcting for range restriction results in about 64.4% of variance in AKT being accounted for by the Stage 2 measures, but with only an additional 0.1% accounted for by the Stage 3 scores.

CSA:

For the present raw data, CPST and SJT together account for 35.7% of variance in CSA, with the Stage 3 scores accounting for an additional 3.2% of variance, which are similar to the values of 36.7% and 4.3% in Patterson et al. Correcting for range restriction results in Stage 2 scores accounting for about 44.8% of CSA variance, with Stage 3 accounting for an additional 3.6% of variance.

Overall the analyses of incremental validity find results concordant with those of Patterson et al, **with no real evidence of Stage 3 contributing to the incremental validity of MRCGP AKT, but a small amount of additional variance being contributed by Stage 3 to predicting CSA of the order of 3 to 4% of variance**. Whether the additional predictive benefit provided by Stage 3 is cost-effective is open to dispute, and will be discussed in the next chapters.

5.5.2 Prediction of entering the GP Register after 4 years, and ARCP Outcome 4

The analyses of Patterson et al. do not look directly at the outcome variable of entering the GP Register within 4 FTE years, or of receiving an ARCP Outcome 4, each of which in some sense indicates the completion of training.

5.5.3 Entering the GP Register within 4 years of FTE training

The major predictors of being on the GP Register within 4 years are the marks attained at the first attempts at MRCGP AKT and CSA, neither of which is surprising or worth taking further. Of more interest is the question of whether the Stage 2 and Stage 3 marks contribute additional prediction even after taking MRCGP marks into account. A logistic regression on the raw data found that SJT provided additional prediction of entry to the Register (p=.004, exp(B) = 1.004) whereas CPST did

[15] i.e. the two sets of values are .73 and .745, .49 and .527, .47 and .489, .54 and .527, .31 and .284, and .42 and .381.

[16] Patterson et al fit some models after including age and sex in the model, and also after including place of graduation. Although of sociological interest, such approaches are probably not appropriate for a study of a selection process since age, sex, ethnicity, place of qualification, etc should not be used in selection.

» *Table 5.1 Predictive validity for Stage 2 and Stage 3 measures of MRCGP AKT and CSA results.*

	Raw correlations				Correlations correct for range restriction			
	Patterson et al		Present study		Patterson et al		Present study	
	AKT	CSA	AKT	CSA	AKT	CSA	AKT	CSA
Trainees	All (UK/IMG)	All (UK/IMG)	**All**	**All**	All (UK/IMG)	All (UK/IMG)	**All**	**All**
Stage 2	na	na	**0.716**	**0.609**	na	na	**0.762**	**0.582**
CPST	.73 (.70/.60)	.49 (.37/.23)	**0.745**	**0.527**	na (.78/.63)	na (.58/.52)	**0.790**	**0.668**
SJT	.47 (.29/.27)	.54 (.29/.26)	**0.489**	**0.527**	na (.50/.33)	na (.57/.46)	**0.564**	**0.604**
Stage 3	.31 (.18/.11)	.42 (.25/.17)	**0.284**	**0.381**	na (.31/.20)	na (.39/.43)	**0.401**	**0.516**

not provide additional prediction (p=.699, exp(B)=1.206). However in the imputed datasets, the effects were variable and inconsistent, and the pooled results showed no evidence of additional prediction of either SJT (p=.863) or CPST (p=.535). Stage 3 scores showed no additional predictive value after MRCGP results and Stage 2 scores were in the model, either for the raw data (p=.444) or for the pooled imputation results (p=.671). Overall it seems that **Stage 2 and Stage 3 prediction of entry to the GP Register is mediated entirely via their effect on the MRCGP AKT and CSA scores.**

5.5.4 ARCP Outcome 4

Low marks at MRCGP AKT and CSA are both highly significant predictors of an ARCP Outcome 4. As in the previous analysis, we also assessed whether CPST, SJT or Stage 3 had any additional predictive value beyond that resulting from their prediction of AKT and CSA. In the raw data SJT did show a predictive effect which was just significant (p=.017, exp(B)=.994), lower SJT scores being associated with a higher probability of an Outcome 4, but the result was not significant in the pooled estimates for the imputed data (p=.959). CPST had no additional predictive value in the raw or the imputed data. The Stage 3 score was similar, having a significant additional predictive value in the raw data (p=.001, exp (B) = .951), lower Stage 3 marks having a higher probability of an Outcome 4, but the pooled effect was not significant in the imputed datasets (p=.565). Overall, as with entry to the GP Register, an **ARCP Outcome 4 is related entirely to low MRCGP AKT and CSA scores, with no incremental variance accounted for by Stage 2 or Stage 3 scores**.

5.6 THE CONSEQUENCES OF USING LONGER SELECTION TESTS AT STAGE 2 AND STAGE 3

All selection tests are unreliable, and they can usually be made more reliable if they are made longer. If the Stage 2 CPST and SJT and the Stage 3 selection centre were to be made longer then they would undoubtedly be more reliable, although there are diminishing returns with increasing length, reliability generally improving as the square of the length of a test[17]. As a test gets longer, so its measures become closer and closer to the 'true' or 'latent' variable which it is measuring, with less and less error variance in the test result.

A test which is twice as long takes twice as long to administer, although that need not necessarily mean that the cost is twice as high, as there are many fixed overheads in the production of any psychometric test. A key question for GP Selection is whether selection may be more cost-effective if the CPST or the SJT or Stage 3 were made longer. It is not any easy question to answer, but neither is it impossible. The imputed datasets described in this chapter allow estimates of the correlations

[17] The relationship is defined by the Spearman-Brown formula, which strictly says that the measurement error is halved when the test length is quadrupled.

A test which is twice as long takes twice as long to administer, although that need not necessarily mean that the cost is twice as high, as there are many fixed overheads in the production of any psychometric test. A key question for GP Selection is whether selection may be more cost-effective if the CPST or the SJT or Stage 3 were made longer. It is not any easy question to answer, but neither is it impossible. The imputed datasets described in this chapter allow estimates of the correlations between the various selection and outcome measures corrected for restriction of range in the selected groups. In chapter 3 of the report we also reported estimates of the reliability of CPST, SJT and Stage 3, and the reliabilities of MRCGP AKT and CPA are reported in the literature (Wakeford, Denney, Ludka-Stempien, Dacre, & McManus, 2015). Using the correlation matrix and the reliability estimates it is straightforward to calculate the true, disattenuated correlations between the selection measures and the outcome measures, the disattenuated correlations being those if the pure latent traits had been measured without error variability. The disattenuated correlation matrix between the selection and outcome measures can then have error variance included within it of the size which would be expected were CPST, SJT and Stage 3 to be of differing lengths from those present used. That correlation matrix, or particularly the covariance matrix on which it is based, along with the mean scores for applicants, can then be used in a Monte Carlo analysis in a program such as *Matlab* to generate multiple cases of the multivariate normal data for a particular set of test lengths. The different selection models can then be run and prediction and the cost-effectiveness of different combinations of outcomes assessed, and compared to the most cost-effective.

Within the analyses in this report there are tantalising suggestions of differences between the CPST and SJT of Stage 2, with recurrent hints that the SJT, is predicting somewhat different outcomes than the CPST. The CPST as a knowledge test correlates more strongly with MRCGP AKT, whereas the SJT correlates more strongly with MRCGP CSA, the implication being that the SJT and CSA may be sharing in assessing some form of socio-cultural, communicative expertise. The SJT is quite a lot less reliable than the CPST, but it is also only half the length. Knowing the consequences of increasing its length is therefore highly desirable. Likewise, Stage 3 has a poor reliability at present, but in effect it is only four stations in length. Were Stage 3 to be doubled or tripled in length (and CSA is thirteen stations), then might its increased reliability make it a cost-effective method of selection, perhaps in a subset of candidates? Sadly we have not been able to answer such interesting questions in the time available. It still remains a tantalising prospect, and no doubt needs doing in the future, but for the present we have no idea what results it would show.

5.7 CONCLUSIONS

Multiple imputation provides a practical solution to several problems in studying selection. It provides a straightforward solution to the problem of range restriction, and therefore allows estimation of predictive validities for candidates across the entire range of abilities, and not merely for those who have been selected into training. Imputation has the advantage that it can account for the predictive validity of multiple variables in a sophisticated way (as, for instance, in hierarchical regressions).

Multiple imputation also allows 'virtual selection systems' to be run, putting all candidates through a range of possible selection models to compare the outcome across a range of different variables. We have done this for 11 different models, and have compared their effect both on immediate outcome measures, but also on other measures such as demographic factors. The analyses suggest that the Stage 2 measures are more effective in predicting outcomes than the Stage 3 measures. That conclusion will be looked at again in the later chapters which consider the economics and cost-effectiveness of different selection models.

The imputed data also allow, in principle, an analysis of the effects on predictability and cost-effectiveness of using selection tests of different length, which would have important practical implications in the design of an assessment.

Chapter 6

Costs of GP selection and training

Chapter 6.

Costs of GP selection and training

6.1 INTRODUCTION

This chapter firstly shows how the cost of selecting GP trainees, using the 2014 selection process, can be estimated. As an illustration, we estimate the total selection cost for trainees who applied to start GP training in August 2014. 2014 was the second year of selection following an increase to the pass mark for the Stage 2 tests (reducing the number of candidates progressing to Stage 3 compared with pre-2013), which was maintained in 2015, and the fifth year following the change to the scenario types used in Stage 3. However for the first time, a third, nationally-run round of recruitment was required in 2014 to increase the fill rate.

Secondly, we show how the cost of GP training based on a trainee completing in three years can be estimated, together with associated additional costs for other selection and training outcomes (not filling a post, a trainee requiring an extension or a trainee failing to obtain GP Registration).

6.1.1 Why are estimares required?

Since there is no comprehensive process by which the costs of selection or training are collected for each individual candidate/trainee, it is not possible to use actual costs and hence we must use estimates. The subsequent sections of this chapter provide detail of how we have done this and the sources of data we have used.

6.1.2 Summary of literature on selection costs and economic evaluation of selection

There is little evidence on the cost of selection or recruitment into or within medicine and any evidence that does exist needs to be carefully scrutinized in terms of the costing perspective used and the types of costs included, if fair comparisons between studies are to be made. Considering the perspective of the selecting organization, shortlisting using written responses to application form questions scored by GP assessors was reported to be 2.5 times as costly per candidate than using machine marked tests (Patterson et al., 2009), although it is not clear whether the costs associated with developing and maintaining an item bank for the machine marked tests have been included. Similarly while the cost of selection per appointed GP trainee in 2005 was reported to be £400-£450 (in 2005 prices; approximately £550 in 2014 prices), it is not known which costs were included in this estimate (Patterson and Lane, 2007).

Thomas and colleagues developed a costing model that aimed to allow those in charge of selection processes – in any specialty - to estimate total selection costs using the opportunity cost basis applied in the current study (Thomas et al., 2010). Example selection processes evaluated using the model clearly showed that the design of a selection process could have a significant effect on costs; for example in a large specialty with nationally-coordinated selection and a low competition ratio such as GP, moving from 3 x 10 minute selection centre stations to 3 x 30 minute stations would increase total selection costs by 60%. This study also suggested that investments in specialty and GP selection would have a positive pay-off over the duration of a training programme, even with what would be considered a poor correlation (coefficient of 0.25) between selection scores and training performance (Thomas et al., 2010). For example, up to around £500 (in 2008/09 prices) could be spent on selection, per applicant in Core Medical Training for selection to have a positive pay-off, over just two years before further selection into higher specialty training (Thomas et al., 2010).

The finding that investment in specialty selection usually pays-off is collaborated with evidence from the field of industrial and organizational psychology (Boudreau, 1988). Here, selection utility analysis is often used to evaluate the cost-benefit

of selection processes (Brogden, 1949), with utility (net monetary benefit) dependent on the competition ratio, predictive validity and the variability in job performance amongst those appointed. An example of how this methodology can be used to evaluate the financial consequences of changing a selection process is reported by Payne and colleagues, in relation to assessment centres used by Ford Motor Company (Payne et al., 1992). The current study uses cost-effectiveness analysis, rather than selection utility analysis, partly because of the difficulties of obtaining estimates of the variability in job performance (i.e. in the **quality** of care provided) but also because of the political imperative to maximize the output of GP training (i.e. the number of trainees obtaining GP Registration).

6.2 SPECIFYING THE COSTING PROCESS

6.2.1 Perspective

The economic evaluation is undertaken from a societal perspective, but only in relation to the provision of GP services. This qualifier merits further explanation. In essence, we have created a virtual UK-wide organization, independent of the Health Services in each of the four devolved nations, which is responsible for providing a pre-determined quantity of GP services and managing the budget required to do so. The budget itself is funded through taxation of the UK population.

We therefore consider the cost of the UK recruitment process managed by GPNRO and its consequences for the provision of health care by GP trainees and those subsequently qualifying as GPs. We have included the costs and outcomes associated with recruitment/selection (including candidate time costs), training/remediation for GP and subsequent primary care service provision but any monetary costs met by the candidates/trainees themselves (e.g. MRCGP examination fees) are excluded.

As noted above, we have assumed that our virtual GP-service provider is responsible for providing a pre-determined **quantity** of health care; for this evaluation the relevant quantity is determined by the number of GP training posts available – which is, in turn, dictated by the number of new GPs required to maintain a pre-determined level of service by qualified GPs. If a GP training post is not filled, short-term financial or accounting savings may be made as no salary or training costs are incurred but less health care is provided to patients and thus health gains/prevention of health losses are forgone. In an economic evaluation, this loss of care is known as an **opportunity cost** and needs to be costed. We assume that the value of care lost is equal to the salary cost of the doctor who would have provided it[1]. Similarly, while most LETBs pay GP assessors to attend training and selection events, to cover or contribute to the cost of a locum to cover their clinical duties, some LETBs reported benefitting from GPs giving up their time for 'free'. Since there is no such thing as a free GP, their opportunity costs need to be included in an economic evaluation, since the GP is not providing primary care while attending these events.

The need to create a virtual provider of GP services, rather than use an unqualified societal perspective, occurs because we exclude the impact of GP recruitment on other specialties. For example, a candidate whose first choice was GP but was rejected during the GP selection process and who enters training in another UK specialty instead still provides health care to UK taxpayers, but the value of such care is excluded from this evaluation. Similarly, a trainee who leaves GP training without obtaining GP Registration may remain working for the NHS and providing healthcare to NHS patients, but our virtual GP-service provider perspective means the value of such care is excluded.

6.2.2 Time frame and discounting

In terms of costing GP training and subsequent service as a GP, we have used a ten year (full time equivalent, FTE) time horizon from the beginning of GP training. At this stage in the evaluation, we assume all trainees train and work at 100% FTE.

We use 2014/15 costs in pounds sterling in the evaluation. For costs to be incurred in future years (e.g. training extensions), 2014/15 values are discounted at 3.5%, as recommended in the Green Book (HM Treasury, 2011).

[1] Another way of looking at this is to imagine that our virtual GP-service provider would be fined if they did not fill all of their training posts and ensured that all trainees obtained GP Registration in three years FTE training time. This would include the care that is provided by trainees working in secondary care during the first 18 months of their GP training.

» *Table 6.1: The 2014 candidate cohort by round of application and country of qualification*

Medical School		Round 1		Round 2		Round 3	
		UK	Overseas	UK	Overseas	UK	Overseas
Round 1 only	Applied	3,385	1,378	-	-	-	-
	Took S2	3,105	1,274	-	-	-	-
	Took S3	2,612	1,067	-	-	-	-
	Accepted post	1,938	788	-	-	-	-
Round 2 only	Applied	-	-	282	113	-	-
	Took S2	-	-	214	83	-	-
	Took S3	-	-	123	53	-	-
	Accepted post	-	-	70	31	-	-
Round 3 only	Applied	-	-	-	-	107	50
	Took S2	-	-	-	-	89	39
	Took S3	-	-	-	-	59	24
	Accepted post	-	-	-	-	33	12
Rounds 1 & 2	Applied	442	220	220	220	-	-
	Took S2*	380	189	22	22	-	-
	Took S3	319	159	136	136	-	-
	Accepted	0	0	63	63	-	-
Rounds 2 & 3, 1 & 3 and 1, 2 & 3 (combined due to small numbers)	Applied	9	3	9	2	-	-
	Took S2*	6	2	1	0	-	-
	Took S3	1	2	1	2	-	-
	Accepted post	0	0	0	0	-	-
TOTAL	Applied**	3,836	1,601	733	335	-	-
	Took S2*	3,491	1,465	258	105	-	-
	Took S3	2,932	1,228	394	191	-	-
	Accepted post	1,938	788	225	94	-	-

* Note that S2 numbers represent unique sittings of Stage 2, i.e. excluding those who took it in a previous round and had their scores carried forward.
** The total exceeds 5,992 due to double and treble applications across Rounds.

6.3 THE 2014 COHORT

Using data provided to us by GPNRO, the 2014 cohort comprised 5,992 unique candidates and 3,090 accepted candidates for 3,559 posts. A summary of the progress of candidates through the selection process, split by application round and country of qualification, is shown in Table 6.1.

6.4 SELECTION COSTS

6.4.1 What costs are incurred during selection?

The GP selection process requires UK-wide coordination by GPNRO (Stages 1 and 2) and Stage 3 selection centres run locally by English LETBs and Wales, Scotland & Northern Ireland (including assessor training). There is also a national (UK) budget for the on-line application form, psychometric evaluation and maintenance of the question bank. This national budget is managed by HEE but appropriate components are recharged to each of the devolved nations according to the number of training posts available. In 2014 a third selection round (for England only) created additional costs.

However, an economic analysis requires that data on **resource use**, rather than simply costs directly incurred, are considered. The most expensive resource used during GP selection but not costed is that of candidate time and thus we consider how this can be valued appropriately and included as part of the calculation of selection costs below.

We also expect that other, usually senior, GPs (e.g. Training Programme Directors) will use their "employed educationalist" time but may also give up their own time to plan and participate in the selection process, for example by making inter-LETB monitoring visits, helping to ensure national calibration of standards and leading moderation in Stage 3 without also participating as an assessor. Without downplaying the importance of such work, we expect the total value of such time to be relatively small compared to the other costs of selection (including the time costs of assessors for training and in the Stage 3 selection centres) and, in the absence of an estimate of such time, it has been excluded from these figures.

6.4.2 Data

Costs data for 2014 incurred nationally by both GPNRO and seven LETBs[2] were submitted to HEE, combined with the national budget and made available to us. However, as described below, these costs were incomplete and required clarification and augmentation. All other English LETBs, Wales, Scotland and Northern Ireland were contacted separately with a request for costs data and responses were received from five[3]. The LETBs not supplying any data are all located in the north of England, which may result in response bias. Missing data and the lack of a consistent response format mean that the data reported here are subject to some uncertainty. As such, while reported to the nearest pound in tables, costs are reported in the text to three significant figures. Other costs, such as salaries, are based on more precise data and are therefore given in full.

6.4.3 Nationally-incurred costs

The selection costs incurred at UK level are shown in Table 6.2 and total £896,000 (to 3 SF). We have incorporated the costs of Stage 1 as "fixed" costs incurred by GPNRO i.e. these costs are included in the GPNRO running costs shown in Table 6.2 and would be incurred in any of the approaches to selection considered in Chapter 7.

6.4.4 LETB-incurred costs for Stage 3

LETB's were asked to submit information on the cost of their Round 1 Stage 3 selection processes in a number of categories (e.g. assessor time[4], venue costs, role-player costs and candidate travel expenses). The key cost categories missing from the HEE proforma were assessor training (see below) and the number of LETB administrative/ management staff attending each day of Stage 3[5]. Incomplete data from the 12 LETBs providing any data were supplemented by contacting LETB Postgraduate

[2] East Midlands, KSS, South West, Thames Valley, West Midlands, Severn and London Recruitment.

[3] Wessex, East of England, Scotland, Wales and Northern Ireland. We did not seek costs from Defence given its small size.

» *Table 6.2: UK-wide costs, £*

Description	Amount	Notes
GPNRO running costs	£100,000	A breakdown of costs was not provided
Funding to LETBs for item-writing events to maintain the Stage 2 & 3 item banks	£50,000	Because the CPST and SJT used in Stage 2 have been running for a number of years, we have not included an amortised value for their initial development.
Fees to Pearson Vue for Stage 2 tests (£56 per candidate)	£305,032	Number of candidates (5,447) taken from Table 6.1. Candidate travel costs for Stage 2 were excluded, since we were informed that these are very rarely claimed and those that are claimed are subject to a £5 maximum.
National Round 3 selection centre (hosted by HE North East)	£83,465	While the Round 3 budget provided to us assumed candidates were entitled to claim up to £100 in travel expenses for Stage 3, the mean claim for Rounds 1 & 2 was only £11. A national process will invariably mean higher travel costs so we have doubled this figure and applied this to the actual number of candidates attending (rather than the planned capacity). No assessor training costs included (see below for details).
Psychometric analysis of scores data	£150,000	Undertaken by Work Psychology Group.
The application IT system	£207,735	No costs were provided, so an estimate was made using the 2015/16 total costs of the Oriel system (used for Medicine, Dentistry and Healthcare Scientists) of £1.38m. We have assumed that GP posts comprise around 30% of all specialty training posts (including those offered at higher levels) and that 50% of the costs of Oriel should be attributed to Medical specialty training.
TOTAL	**£896,232**	

Deans and their management teams or by using mean imputation if data were not provided. Costs for the LETBs providing data are summarized in Table 6.3.

Scaling-up Stage 3 'on the day' costs:

Because not all LETBs submitted costs it was necessary to scale-up the costs to include all UK LETBs. To do this, we chose one selection centre day as the unit for scaling. The data in Table 6.3 give a total cost of £1,360,000 (to 3 SF) across the 12 LETBs; with a total of 46.5 days this gives a mean of £29,000 per day (to 3 SF). The total capacity across the 12 LETBs was 4,452 candidates, or a mean of 96 per day. Thus if all selection centres ran at full capacity, the mean cost per candidate would have been £307 (£29,000/96). Our approach to scaling-up costs to estimate 2014 selection costs for the whole of the UK is based on 'weighted mean imputation', which assumes that daily costs and capacities in the five LETBs for whom we do not have data would also have means of £29,000 and 96 per day respectively.

[4] Both GP and non-GP (Lay) assessors participate in Stage 3. While both incur costs, these are 2-2.4 times higher for GP compared with Lay assessors (the exact multiplier varies by LETB). Where data on the daily rate paid to GP assessors were available, this ranged from £327 to £692. This range is within that of the locum fee for a day's GP service (GP ONLINE. 2014. Locum rates 2012/13: locums booked directly by practices [Online]. Available: http://www.gponline.com/locum-rates-2012-13-locums-booked-directly-practices/article/1227547 [Accessed 25/02/2015.) so could plausibly fall short of, or exceed the locum cost.

[5] We applied the NHS Agenda for Change (2013/14) pay rate at the mid-point of Band 4, increased by 1% to reflect the uplift for 2014/15 (£10.71 per hour + 20% for London), assuming an 8 hour day. One LETB did not provide data on administrative staff attendance at Stage 3, so the weighted mean ratio of admin staff to maximum candidate capacity was used, weighted by the maximum candidate capacity in each LETB providing this information (0.17 admin staff per candidate).

» *Table 6.3: Stage 3 'on the day' costs by LETB, £*

Cost category	East Midlands	KSS	South West	Thames Valley	West Midlands	Severn	London recruitment
Venue	26,624	32,000	7,800	12,007	72,691	11,412	36,800
Assessors (time and travel)	78,595	62,500	30,056	31,551	113,703	58,349	193,498
Administration	7,042	3,816	2,714	2,798	9,158	4,325	19,945
Candidate travel	5,381	5,000	1,000	2,688	5,290	3,992	1,100
Role-players	39,372	14,400	4,550	12,600	56,760	29,136	12,600
Other	0	0	896	1,476	0	0	0
Total	**157,014**	**117,716**	**47,016**	**63,120**	**257,603**	**107,214**	**263,943**
Candidates attending	**206**	**402**	**99**	**166**	**359**	**222**	**710**
Mean per candidate	**762**	**293**	**475**	**380**	**718**	**483**	**372**

* While no venue cost was incurred, an opportunity cost for the use of LETB facilities was used.

The GPNRO website (GP National Recruitment Office, 2015) suggests that on average, there are 1.5 selection centre places for each vacancy in Round 1, which would imply 5,340 places across the UK in 2014 (3,559 posts x 1.5, to 3 SF)[6]. This would require 56 selection centre days (5,340/96, rounded up to the next whole number) and a total cost for Round 1 of £1,640,000 (£29,000 x 56, to 3 SF). In 2014 there were 840 vacancies after Round 1 (personal communication, Jonathan Howes) and at this stage 2-3 selection centre places per post are available (GP National Recruitment Office, 2015). Using 2.5 places per vacancy, 2,100 selection centre places would be needed, or 22 days (2,100/96), giving a total cost for Round 2 of £646,000 (to 3 SF). The **total UK Stage 3 'on the day' cost is therefore £2,290,000 (to 3 SF)** for Rounds 1 and 2.

One reason for the variation in per candidate actual attendance costs between LETBs shown in Table 6.3 (overall mean £420) is the difference in attendance rates compared to selection centre capacity. All 12 LETBs had 'spare capacity' in their Stage 3 selection centres: the mean attendance rate was 73% (3,246 attending/ 4,452 capacity) and this ranged from 47% in West Midlands to 99% in Severn. The range of cost per candidate at full capacity ranged from £189 in Scotland to £477 at Severn. However there is no evidence of economies of scale (i.e. LETBs with a higher candidate capacity do not have a lower mean cost per place).

Assessor training costs:

We consider assessor training as a fixed cost incurred once during each year of selection and therefore assume that Round 2 and 3 assessors would also have participated in Round 1, so no new assessors require training. A major change in the selection process, either in terms of the number of candidates attending Stage 3 (requiring more or fewer assessors) or how the process works (so all assessors require re-training) would impact on the total annual cost of training.

To estimate total assessor training costs, we have used data provided to us by each LETB to construct a UK training costs model. Each LETB provided data in a slightly different format, so a model-based approach was considered appropriate. The unit of scaling applied was half a day, since correspondence with the LETBs suggested that, on average, each assessor attended training lasting half a day. We assumed assessors only attended one day of Stage 3 in Round 1.

[6] Suggesting costs data were obtained for 83% of Stage 3 places (4,452/5,340).

We calculated the mean number of GP (26) and Lay (4) assessors required per selection centre day (GP: 1,209 total/46.5 days; Lay: 174 total/46.5 days) and multiplied this by the number of selection centre days required (56; as shown above) to determine the total number of assessors who required training for the UK (GP: 26 x 56 = 1,456; Lay: 4 x 56 = 224). We applied typical GP (£530) and Lay (£250) assessor daily costs to calculate the total time cost of attending training for half a day (GP: 1,456 x £530/2 = £386,000; Lay: 224 x £250/2 = £28,000). Across the 12 LETBs providing data, a total of 47 training half-days were held; dividing the total number of assessors in these LETBs (1,209 + 174 = 1,383) by 47 gives a mean half-daily attendance rate of 30 assessors/half-day. We then divided the total number of assessors requiring training in the UK by the half-daily attendance rate to find the number of half-days of venue hire required ((1,456 + 224)/30 = 56). We assume that venue costs for assessor training accrue at one-third of the daily rate as for selection, given that mean assessor attendance (30/day) is approximately one-third of mean candidate capacity per day (96/day). The weighted mean half-daily venue cost in these LETBs was £2,775, assuming the same cost as per selection and weighted using the number of selection days, giving a total venue cost of £51,800 (56 x £2,775/3). **The total UK cost of assessor training is therefore estimated at £466,000 (£386,000 + £28,000 + £51,800, to 3 SF).**

6.4.5 Candidate time costs

Candidates working in the NHS are likely to need to take time off work in order to attend selection events (both Stage 2 and Stage 3). Because the candidates are not at work when at a selection event, they are not providing care to patients, so the associated opportunity costs to the NHS need to be taken into consideration (i.e. the value of care not provided or locum costs). We do not have data on whether a candidate is currently working in the NHS, so we have assumed that all UK graduates are currently employed by the NHS. We then assume all UK graduate candidates take one day (paid) leave in order to attend each of Stage 2 and Stage 3, which is not part of the candidates' annual leave or study day entitlements.

We assume that a candidate is paid at the average F2 salary and include wages/salary and salary on-costs (employers' pension and NI contributions), which are combined with hours worked to value this time as described in the 2014 Unit Costs of Health and Social Care (Curtis, 2014) and inflated at 1% to 2014/15 values (NHS Employers, 2014). **Each day is therefore valued at £159**, assuming a five day working week.

We excluded the opportunity cost of candidate time in preparing for selection processes (e.g. completing application forms and taking practice Stage 2 tests) and any mental health consequences relating to selection outcomes (e.g. the cost of treating depression arising from being unsuccessful and the 'value' of health lost by the candidate from their depression).

We subsequently used data on candidate numbers to calculate candidate time costs. Using the data in Table 6.1 above, we estimate that **a total of 7,223 NHS days were spent by candidates in GP selection in 2014, valued at £1,150,000 (to 3 SF).**

6.4.6 Total selection costs

The total selection cost for 2014 was estimated using the sum of the individual cost elements: nationally-incurred costs (£921,000), LETB Stage 3 (£2,290,000) and assessor training (£466,000) costs and candidate time costs (£1,150,000).

- **Total cost of selection: £4,800,000**
- **Mean cost per unique candidate (N=5,992): £801**
- **Mean cost per post filled (N=3,090): £1,580**

Since we have augmented the costs data supplied by LETBs to HEE, obtained costs from other LETBs and the devolved nations, and incorporated candidate time and other opportunity costs, we are mindful that our estimates are not comparable to others that may be available at a national level, either for the cost of GP selection or for selection into other specialties. **We therefore caution others against making such comparisons.**

6.5 COST OF SELECTION AND TRAINING OUTCOMES

6.5.1 Defining outcomes

Progression to CCT:

The minimum training time required for completion of the GP training programme is three years. During training, all trainees have an Annual Review of Competence Progression (ARCP) in which their workplace-based assessments and other evidence relating to their performance, training and development during the year are considered[7]. There are a number of possible outcomes resulting from this review, as detailed in the "Gold Guide" (Department of Health, 2014b). Trainees must receive a satisfactory outcome to progress to the next year in the training programme, or to be eligible for CCT and entry to the GMC's GP Register. Trainees also have to pass the two examinations required for membership of the RCGP, the AKT and CSA, before obtaining CCT. However neither examination has to be passed to progress from year 1 to 2 or from year 2 to 3 of the GP training programme.

Outcomes included in the economic analysis:

The following outcomes for GP selection and training will be included in the economic analysis:

1. A trainee achieving CCT and being listed on the GP Register in three years' FTE training time.
2. A trainee requiring an extension to their training programme but subsequently obtaining CCT and being listed on the GP Register[8].
    a) Following a six month FTE extension.
    b) Following a 12 month FTE extension.
    c) Following an extension of greater than 12 months but less than two years FTE[9].
3. A trainee not obtaining CCT (not listed on the GP Register) within five years' FTE training time[10].
4. Not filling a GP training post.

[7] The ARCP process includes consideration of any investigations into a trainee's conduct, or sanctions imposed, by the GMC which could result in the trainee being directly released from their training programme.

[8] The Gold Guide states that extensions to GP training will normally be for a maximum of six months FTE, but could be an absolute maximum of one year during the total duration of the training programme; and that all trainees should be given a training extension if required, unless they have already used up their permitted extension time, been subjected to disciplinary proceedings or exhausted all attempts at passing an examination. However, the data presented in Chapter 4 (Table 4.11) suggest that six and 12 month extensions are equally likely and that some trainees have longer extensions. In terms of costing extensions, we assume all extensions are given at the end of the third year of training (or equivalent if trainees are less than full time (LTFT) or have had time out of programme (OOP). The data presented in Chapter 4 (Table 4.8) suggested that around 85% of extensions occurred at the end of year 3, making this a reasonable assumption. We assume that all delays (post three years FTE from beginning training) in obtaining CCT and being listed on the GP Register required an extension to training.

[9] Given the rarity of this outcome (around 3% of trainees without any LTFT or OOP from the 2009 cohort, using data from Chapter 4, Table 4.5), we use a single category for long extensions. In terms of costing, we assume a two year extension.

[10] This is a composite outcome which includes trainees given an extension but not obtaining the competencies required to progress and thus leaving the programme, trainees required to leave the programme without being given an extension, trainees voluntarily leaving the programme and trainees taking more than five years' FTE training time to obtain CCT and be listed on the GP Register. This outcome applies to around 9% of the 2009 cohort without OOP or LTFT (Chapter 4, Table 4.5). In terms of costing, we assume trainees complete three years of GP training but then do not provide any service as a qualified GP. This reflects a trainee who fails their MRCGP examinations and is unable to remediate.

The primary outcome to be used in the cost-effectiveness analysis in Chapter 7 is the number of trainees obtaining CCT and being listed on the GP Register within four years' FTE training time. Our focus is on trainees being listed on the GP Register since this is the requirement for subsequent practise as a GP and hence provision of primary care services.

Outcomes not included in the economic analysis:

We cannot incorporate the **quality** of health care provided by selected trainees or those completing GP training. This means that the costs of making up for sub-optimal care (e.g. treating preventable adverse events or not using the most cost-effective therapies) incurred by the health service or the health lost (e.g. Quality Adjusted Life Years) as a result of sub-optimal care incurred by society, have not been included in our evaluation. There is some evidence that qualifying examination scores are linked to quality of primary care in Canada (Norcini et al., 2014, Tamblyn et al., 2002, Wenghofer et al., 2009), but, to our knowledge, there is currently no such evidence for the UK.

6.5.2 Costing outcomes

In order to estimate the long-term (ten-year) impact of the selection process and possible, alternative approaches to selection, a monetary value for not filling a training post and for each of the five training outcomes listed above is required.

GP Training costs for a trainee completing in three years FTE:

PSSRU provide details on total post-qualification training costs for doctors in their 2013 report on the Unit Costs of Health and Social Care (PSSRU, 2014). These costs are in 2012/13 prices and include the trainee's salary, training/tuition costs, infrastructure e.g. libraries, costs/benefits from clinical placements and lost production during time spent training. We are only interested in the cost of delivering GP training in this framework, rather than the total costs incurred post-qualification to the point of qualification as a GP, since these total costs include training costs incurred during Foundation Training. Using the PSSRU data inflated to 2014/15 values and subtracting the cost of training during the Foundation Programme, the **total cost of GP training is £208,600.** We assume this cost is spread equally over the standard three year GP training period.

Unfilled GP training posts:

An unfilled training post leaves a gap in the provision of secondary health care by the trainee during the 18 month period they are in a hospital placement i.e. if a post is not filled, the hospital is required to employ another doctor to undertake the clinical duties the trainee would have undertaken[11]. We assume a staff grade practitioner is employed to cover these duties, at a basic annual salary of £34,441 in 2014/15 prices (NHS Employers, 2014), to which are added the on-costs related to Employers' pension and NI contributions[12]. The Education and Training Tariffs (Department of Health, 2014a) note that salary support for trainees of 50% of salary is provided to Trusts (to cover their time spent training), and we could assume that a replacement doctor would be required to work the other 50% to cover the clinical duties of the trainee. Over 18 months, the discounted service replacement cost of an unfilled post is therefore £31,826. However, it may be more likely that a hospital is required to recruit a full time doctor, the cost of which would be £63,651.

Based on discussions with the Advisory Group for this project, we have assumed that a GP trainee in a GP placement is 'supernumerary' i.e. that they do not provide any health care in addition to what would have been provided by the GP trainer had the trainer not been required to supervise their trainee. Thus an unfilled training post does not result in any lost health care for the 18 month duration of training based in GP.

[11] In other words, our virtual GP service provider is expected to provide a certain number of GP trainees to hospitals and if they are unable to do so they must pay for replacement doctors to ensure no loss of care in the hospital system.

[12] 14% for pension contributions and 13.8% for earnings over £7,956 for NI contributions.

[13] Our virtual provider of GP services has not met the requirement to provide a pre-determined number of newly-qualified GPs and society receives less primary care as a result.

[14] Here we exclude on-costs (Employers' pension and NI contributions) since these do not relate directly to the provision of health care.

[15] The 'break-even point' occurs at an extension length of around 18 months, depending on the % FTE required for the replacement hospital doctor.

A GP trainee qualifying in three years FTE will subsequently deliver seven years of service as a qualified GP within the ten year FTE time horizon. If a GP training post is unfilled, this incurs a cost to society due to the provision of future health care lost as a result[13]. We assume this health care would be valued at the salary cost of a GP[14]. The pay range for a salaried GP employed by a Primary Care Organisation in 2014/15 was £54,863 to £82,789 (Source: NHS Business Services Authority, Pay Circular MD 2/2014). We assume that a GP would progress through this range at a constant rate, reaching the maximum value after ten years of service as a qualified GP (i.e. an increase of £3,103 per annum). All costs in future years are discounted at 3.5% per annum to 2014/15 prices, to give a value of care lost over seven years' GP service of £351,552.

One unfilled training post results in NHS replacement costs and lost health care provision valued at £383,378 (£31,826 + £351,552) with a replacement doctor for 50% FTE or £415,203 with a replacement doctor at 100% FTE. In 2014, 481 posts were unfilled, resulting in a total ten year cost of £184m with replacement doctors at 50% FTE or £200m with replacement doctors at 100% FTE.

Extension costs:

The basic cost of an extension at the end of Year 3 of training is assumed to accrue at the same rate as for the original training. Extension costs related to providing repeated training for extension durations included in the economic evaluation (6 months, 12 months and 24 months) are shown in Table 6.4. Total extension costs are higher than normal training costs, since they include the following extension-specific costs: (1) remediation/educational support costs, (2) LETB administration costs, (3) additional ARCP panel costs, plus a longer ARCP panel cost for those subsequently failing and (4) appeal costs. Estimates of these costs have been derived from correspondence with North East LETB and are summarized in Table 6.4. (Note that the values for each year are not equal due to discounting.)

A trainee requiring an extension but who subsequently achieves GP Registration will provide less than the seven years of GP service provided by a trainee who completes on time, when considering the ten year time horizon for the model. The value of this health care lost therefore needs to be included and is also shown in Table 6.4.

Cost of a trainee who fails to achieve GP Registration:

As noted above, we assume trainees in this outcome category complete three years of GP training but do not provide any service as a qualified GP. We assume such trainees will incur the standard training cost (£208,601), will have one longer ARCP panel at the end of year 3 (£40) and may appeal their ARCP outcome at the end of year 3 (average per trainee of £383). The value of GP care lost is the same as that for not filling a training post (£351,552). The total cost of such a trainee is therefore £560,576.

6.5.3 Summary of selection and training cost consequences

Table 6.5 shows the total cost of each of the selection and training outcomes and the net cost of each alternative outcome when compared to a trainee achieving GP Registration in three years FTE. An important result is that, over ten years, **financial savings would be made by recruiting a trainee who requires an extension of 12 months or less prior to passing compared to leaving a post unfilled[15].** This finding collaborates our choice of outcome measure for the cost-effectiveness analysis, i.e. considering a pass with a 12 month extension as a 'positive' outcome.

6.6 SUMMARY

This chapter has described our approach to estimating the cost of the selection and training processes for GP trainees. Our perspective, that of a virtual provider of GP services across the UK, enables us to isolate those costs pertinent for an evaluation of the GP recruitment process. However, caution is warranted in comparing the costs reported here with other estimates of selection and training costs, given the perspective employed, opportunity costs included and the assumptions and estimates required at some points in the costing process.

We estimated the total UK cost of selection in 2014, at around £4.8M, including the opportunity costs arising when candidates and assessors take time out of NHS jobs to attend selection events. 2014 is not completely representative of the selection process in other years, since a third round of selection was required to fill posts. However the cost of Round 3 was minimal compared to the total selection cost (1.7%). The selection costs need to be compared to those of training and subsequent service as a GP, in particular the costs of 'getting it wrong' and either selecting candidates who require an extension and/or fail, or not selecting candidates who would pass (possibly following an extension) but leaving a post unfilled instead. That, over ten years, it is cheaper to recruit a trainee who requires a 12 month extension prior to achieving GP Registration than it is to leave a post unfilled, highlights the importance of taking a longer-term view with GP training as a link in the chain between Foundation and fully-qualified GPs rather than being evaluated in isolation. If a longer time horizon is considered (i.e. more than a maximum of seven years' FTE service as a GP) then extensions of longer than 12 months will be cost saving to our virtual provider of GP services.

» *Table 6.4: GP training extension costs, £*

Cost	Extension then pass		
	6 Months	12 Months	24 Months
Basic training	208,601	208,601	208,601
Training extension	32,443	67,156	132,041
Remediation/ Educational support*	4,816	9,632	18,938
LETB administration**	508	1,016	1,997
Additional ARCP panel(s) / longer panel(s) ***	199	397	782
Appeals (average per trainee with unsatisfactory outcome)****	383	383	383
Value of health care (GP service) lost	26,046	52,092	103,730
TOTAL	**272,996**	**339,277**	**466,472**

*Costs for a trainee with a medium degree of difficulty or concern (e.g. failing CSA and concerns in one competency). We assume these costs are incurred every six months of extension.

**We assume these costs are incurred every six months of extension.

***We assume a trainee has one ARCP every six months during an extension. Panels that result in a subsequent extension are all assumed to be longer than that for a positive outcome.

****The average LETB appeal cost is estimated at £6,000. GMC data (August 2009-July 2012) suggest a national appeal rate of 6.83% (General Medical Council, 2014). We assume each trainee may only appeal once, at the end of year 3.

Excluded from additional ARCP and appeal costs are the additional effects on lost production as trainees take time out of clinical work to prepare for and attend meetings.

» *Table 6.5: Training cost consequences*

Training outcome	Total cost, £	Net cost compared to pass in 3 years FTE, £
Pass in 3 years FTE	208,601	-
Post not filled (replacement hospital doctor at 50% FTE)	383,378	174,777
Post not filled (replacement hospital doctor at 100% FTE)	415,203	206,602
6 month extension then GP Registration	272,996	64,395
12 month extension then GP Registration	339,277	130,676
24 month extension then GP Registration	466,472	257,871
Failure to obtain GP Registration	560,576	351,975
TOTAL	**272,996**	**339,277**

Chapter 7
Economic evaluation of GP selection

Chapter 7.

Economic evaluation of GP selection

7.1 INTRODUCTION

This chapter uses a cost-effectiveness analysis to show the longer-term consequences of different approaches to selecting GP trainees. The primary measure of effectiveness used is the number of trainees obtaining GP Registration within four years FTE training time. The costs considered are: (1) selection costs, (2) costs of unfilled training posts and (3) additional training and service delivery costs that arise when a trainee does not obtain GP Registration within three years FTE[1]. We compare the cost-effectiveness of the approach to selection used in 2015 (the 'baseline' approach) against six alternative approaches to selection as described in section 7.2.

As noted in Chapter 6, the perspective is that of a virtual provider of GP services for the UK as a whole and a ten year time horizon from the beginning of GP training is used. The evaluation considers two different 'targets' for the number of trainees appointed (and subsequently completing training within four years FTE training time): 3,250 and 3,750. The former is approximately the mean number of posts available for the UK for 2011 to 2014 (the years' data on which the evaluation is based), according to data published by the GPNRO[2]. The latter is based on the more recent target recruitment of 3,250 trainees for England, plus approximately 500 posts available in Scotland, Wales and Northern Ireland combined for 2014 to 2016[3].

Our analyses are based on selection and training outcomes data for the 2011 to 2014 application years[4]. As highlighted earlier in this report, the complete follow-up data required for the economic evaluation (whether the applicant is on the GP Register at five years FTE[5] following the August of the year in which they were selected and, if so, the duration of their training prior to Registration) are only available for 100% FTE trainees in the 2009 and 2010 application years. Furthermore, it is only rarely that we know whether any of those who were rejected during the selection process actually achieved GP Registration and we know nothing of alternative destinations[6]. As a result, it is not possible to use real-life data in the economic evaluation; and imputed outcomes are used based on the approach described in Chapter 5.

We consider three key questions in our analyses in this chapter:

1. What is the relative cost-effectiveness of each of the six alternative models of selection, compared with the 'baseline' approach?

2. (For selected alternative models of selection, and based on a target of 3,750 trainees only) what is the relative impact of (1) increasing the number of posts filled and (2) differences in the effectiveness of different approaches to selection?

[1] These costs include extension costs of trainees who require an extension of 6 or 12 months, but who are still enumerated as having a 'positive' outcome i.e. who obtain GP Registration within four years FTE training time.

[2] See https://gprecruitment.hee.nhs.uk/Resource-Bank.

[3] See https://gprecruitment.hee.nhs.uk/Resource-Bank and https://gprecruitment.hee.nhs.uk/Recruitment.

[4] We do not include data from 2009 and 2010 in the analyses because a different format of Stage 3 was used. 2015 is excluded due to the limited follow-up data available at the time of analysis. However data from 2009 to 2015 were used in the imputation of missing data.

[5] Note that one of the training outcomes included in the evaluation is a trainee requiring a 24 month extension prior to obtaining GP Registration i.e. taking five years. These trainees are not included in the measure of effectiveness used in the economic evaluation, which only counts those trainees obtaining GP Registration within four years FTE training time. Hence both four and five years are referred to as required in this Chapter.

[6] This could occur if an applicant obtained a Certificate of Eligibility for GP Registration (CEGPR) without entering GP training.

3. What is the 'marginal' effect of increasing the target recruitment from 3,250 to 3,750 per year (i.e. what are the selection and training outcomes for the 'last' 500 trainees appointed)?

7.2 APPROACHES TO SELECTION INCLUDED IN THE EVALUATION

Table 7.1 describes the seven approaches to selection considered in the economic evaluation. All approaches assume a common Stage 1 (no application would lead to an offer in any approach unless Stage 1 had been passed). All of the alternative approaches to the baseline (existing) approach should increase the number of GP trainees recruited[7]. Where an approach uses "top-down" selection to fill all posts i.e. the 3,250/3,750 highest scorers on the measure(s) employed, the 3,250/3,750 are selected following adjustment for withdrawal from the selection process and rejection of offers of training posts. We note that some of the approaches considered here might not be considered acceptable by all stakeholders, but are included here for completeness.

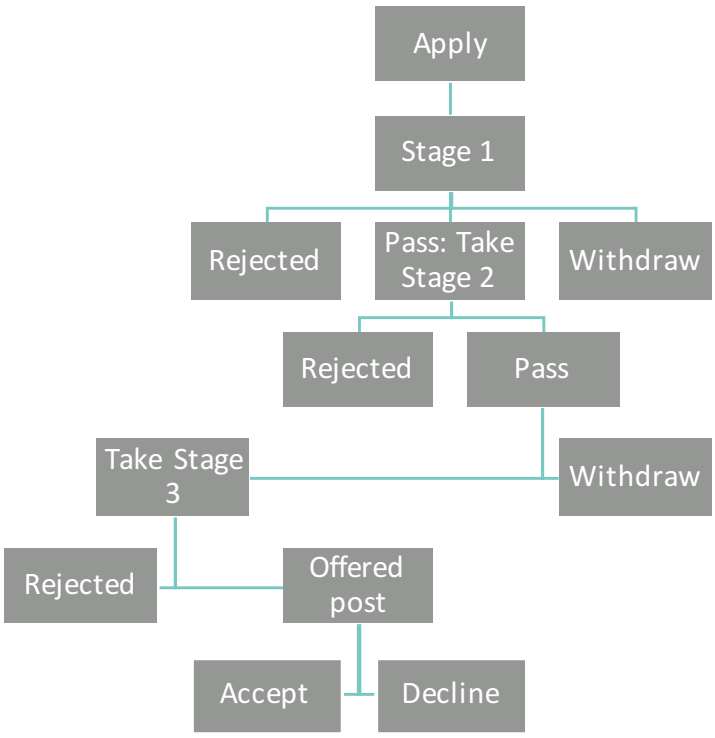
» *Table 7.1: Approaches to selection (all approaches assume a common Stage 1)*

Approach	Stage 2	Stage 3	Rounds required	All posts filled?
Baseline (2015 actual process)	SJT and CPST each with a cut-score of 181	Selection Centre, with moderation	2	No
Random	None	None	1	Yes
Old Stage 2 cut-score	SJT and CPST each with the pre-2013 cut-score of 166	Selection Centre, with moderation	2	No
Stage 3 only	None; all applications invited to Stage 3	Selection centre, no moderation; top scorers selected (no minimum score)	1	Yes
Stage 2 only	Combined SJT and CPST score used to rank applications; top scorers selected (no minimum score)	None	1	Yes
Stage 3 bypass	As baseline, except applications with a combined Stage 2 score of at least 575 are offered posts without attending Stage 3	Selection Centre, with moderation (unless bypassed)	2	No
Equal weight to all	SJT and CPST each with a cut-score of 181.	Selection Centre, no moderation; Stage 3 total scores combined with SJT and CPST (each weighted 33.3%) and top scorers selected	1*	Yes*

* Note: With 3,750 posts not all posts are filled in Round 1. However, the number of posts remaining was low (around 90; 2%) and to maintain consistency it was decided not to include Round 2.

[7] There is a possibility that the Stage 3 by-pass approach may not have any effect if all those by-passing Stage 3 would have passed it anyway; as suggested on the GPNRO website: https://gprecruitment.hee.nhs.uk/Recruitment/Summary-of-Changes, accessed 24/11/15.

» *Figure 7.1: Paths through selection*



7.3 SELECTION AND TRAINING OUTCOMES

7.3.1 Dataset

The aggregate dataset of all applications to GP training between 2009 and 2015 was combined with data relating to ARCP outcomes, MRCGP performance and entry onto the GP Register. The data from different sources were linked using candidates' GMC numbers and thus any candidates with missing GMC numbers were excluded. We only included applications that had passed Stage 1 and who did not drop out prior to Stage 2, i.e. had both SJT and CPST scores. Candidates in Round 3 in 2014 were combined with those for Round 2 and the 2015 'Round 1 advertisement' was considered equivalent to Round 2.

Across the four application years being considered, there were a mean of 5,196 applications per year with Stage 2 scores in Round 1 and 1,145 in Round 2 (included in the analysis if required). The target recruitment was 3,250/3,750 GP trainees per year, or 13,000/15,000 over four years.

7.3.2 Tracking candidates through the selection process

Figure 7.1 shows the paths a candidate could take through the selection process in the economic evaluation, assuming that both Stage 2 and Stage 3 are retained. The starting point for the economic evaluation is the Pass Stage 1 and take Stage 2 box.

7.3.3 Multiple imputation of missing selection and outcomes data

As described in Chapter 5, a process of multiple imputation was used to predict Stage 3 scores and/or training outcomes for up to five years FTE for all applications rejected and/or for whom complete follow-up data were not available[8]. Each

[8] For the economic evaluation, the imputation was undertaken at the level of the application, rather than at the level of the individual candidate.

application therefore had one real or imputed possible training outcome from the following five categories: obtain GP Registration in 3, 3.5, 4 or 5 years FTE, or not obtaining GP Registration.

We know which applications – from those progressing sufficiently far in the process – withdrew having passed Stage 2 (i.e. did not attend Stage 3) and who declined an offer of a post having demonstrated competency at Stage 3. However the applications for whom we needed to impute Stage 3 scores (i.e. those who failed Stage 2) may also have withdrawn or declined an offer. Because these withdrawal or declining offer decisions may not be independent of selection scores, we have also imputed these outcomes where required and used these to adjust for withdrawals and declined offers. Thus in approaches to selection where all posts are filled, posts continue to be offered until 3,250/3,750 have been accepted; we assume that there is no pass-mark, cut-score or level of performance at which competence is demonstrated.

We calculated an "expected time to GP Registration" which was three years for applications with no time out of programme (OOP) or less than full time training (LTFT)[9]. To adjust for any OOP or LTFT, we adjusted the expected time to GP Registration appropriately, assuming that trainees with LTFT trained at 60% FTE, as in previous chapters.

Because the imputation process is subject to statistical variability, it was repeated ten times so that the consistency of results could be examined.

For each imputation and approach to selection, we used the available and imputed data for each application to identify, across the four application years used in the analysis, the total number of unfilled posts and the total number of trainees in each of the five training outcome categories as noted above. In order to reflect a typical application year when reporting results, each value was divided by four.

We then calculated, for each of the ten imputations, the **incremental** number of selected trainees achieving GP Registration within four years FTE training time for each of the alternative approaches to selection compared with the baseline approach[10].

7.4 COSTING EACH APPROACH TO SELECTION AND SELECTION AND TRAINING OUTCOMES

The cost of each approach to selection was estimated as detailed in Table 7.2, with the various cost elements taken from Chapter 6. Selection costs are not affected by the results of the imputation. We decided not to include selection costs as part of the sensitivity analysis because these costs were a very small proportion of total selection and training cost consequences (approximately 0.4% in the baseline approach). Thus selection costs could double and still contribute less than 1% of the total ten year costs.

The cost consequences associated with each of the selection and training outcomes (i.e. the number of unfilled posts and the number of trainees in each of the five training outcome categories) estimated in Chapter 6 (Table 6.5) were applied to the results of each of the ten imputations for each approach to selection (mean results across the ten imputations are shown in Table 7.3). We assume that for unfilled posts, a replacement hospital doctor is required to work at 100% FTE.

We then calculated, for each of the ten imputations, the total **incremental** costs (including selection costs and selection and training cost consequences) for each of the alternative approaches to selection compared with the baseline approach.

7.5 RESULTS

7.5.1 Selection and training outcomes

A summary of the results for each approach to selection, in terms of the number of applications with each of the key selection outcomes and the number of unfilled posts across the 10 imputations are presented in Table 7.3A (3,250 posts) and

[9] We added a 'grace period' of two months to each expected time to GP Registration to allow for any delays in processing Registration applications.

[10] For all analyses that used incremental effects and/or costs, we always compared by imputation number, i.e. imputation 1 for baseline was compared with imputation 1 for all other approaches; followed by imputation 2 for baseline compared with imputation 2 for all other approaches etc.

» *Table 7.2: Costing approaches to selection (£ 2014/15 values)*

Cost	Details
GPNRO running costs	£100,000 for all approaches
Stage 2 & 3 item bank maintenance	£0 in random selection £25,000 if only Stage 2 or Stage 3 used £50,000 if both Stage 2 and Stage 3 used
Stage 2 variable costs	£56 x number of applications (Round 1 or Rounds 1 & 2, as appropriate), where Stage 2 used
Psychometrics	£0 in random selection £75,000 if only Stage 2 or Stage 3 used £150,000 if both Stage 2 and Stage 3 used
IT System	£207,735 for all approaches
Stage 3 'on the day' costs	£29,351 x days needed (places needed / places per day, rounded up to nearest integer) Places needed: 1.5 per post in Round 1 (or number of applications in Stage 3 only) 2.5 per post in Round 2 (where required) 0 in Stage 2 only Places per day: 96 in all approaches with Stage 3 except "Stage 3 only" and "Equal weight to all" (128 places/day) as no moderation means one additional cohort per day
Assessor training costs	Calculated based on the method described in Chapter 6 based on the number of assessors required for each selection approach
Candidate time	£159 x number of applications for each included Stage/Round x proportion of applications from UK medical school graduates at each Stage/Round

B (3,750 posts). Four approaches (random, Stage 2 only, Stage 3 only and Equal weight to all) fill all 3,250 posts since posts are offered (based on relevant scores for all except random selection) until 3,250 applications have accepted them (i.e. if an offer is rejected a post is offered to the next application on the list). This is also the case for three of these four approaches (the exception is Equal weight to all) when there are 3,750 posts; for Equal weight to all the Stage 2 cut-score of 181 on each test means that not quite all of the additional 500 posts are filled (with only one Round of selection). Reducing the Stage 2 cut-score and the Stage 3 by-pass increase the number of posts filled compared with baseline, but not all posts are filled for either 3,250 or 3,750 posts.

With both 3,250 and 3,750 posts, all other approaches increase the number of trainees achieving GP Registration within four years FTE training time. **The largest increase is with the Stage 2 only approach, with a mean of an additional 239/550 GP Registrations compared to the baseline approach for 3,250 and 3,750 posts respectively** (approximately 10%/24% for 3,250 and 3,750 posts respectively).

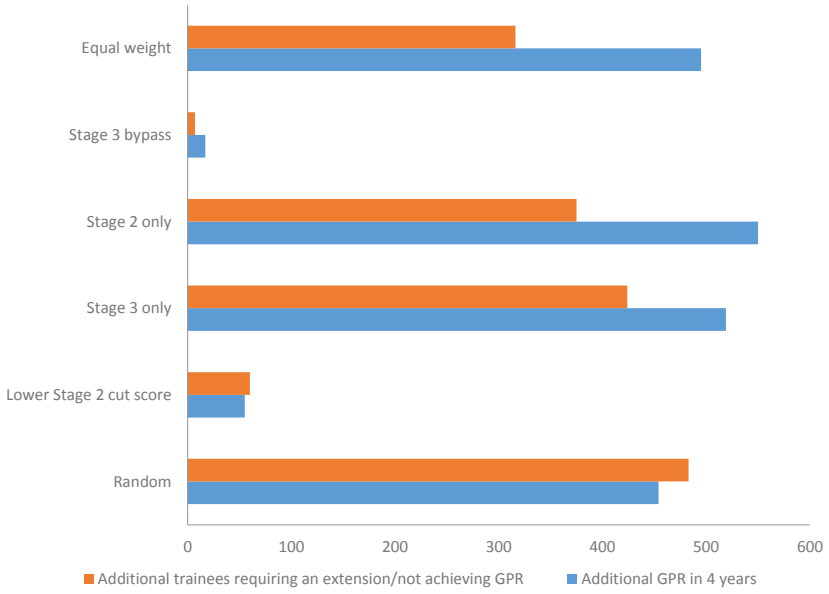
With the Stage 2 only approach, the minimum combined SJT and CPST score amongst those appointed is 424 with 3,250 posts and 311 with 3,750 posts. The latter means that some of those currently rejected at Stage 2 (with a cut-score of 181 on each test) would be appointed.

Although all alternative selection approaches increase the number of posts filled compared with the baseline approach, more trainees means more extensions as well as more GPs. Thus there is a trade-off between the resulting increase in the number of GP Registrations within four years FTE with the increase in the number of trainees requiring an extension (of any length) and/or failing to achieve GP Registration. **Outside of the economic analysis results which will be presented later in this chapter, the short-term patient safety concerns and the practical and emotional consequences to GP trainers, Programme**

Directors and the trainees themselves of additional extensions must be weighed against the long-term consequences of an increase in qualified GPs. In terms of numbers, the trade-off required is summarised in Figure 7.2, based on the means reported in Table 7.3B. All trainees requiring an extension (of any length) and/or not obtaining GP Registration are grouped together. For example, with Stage 2 only selection of 3,750 trainees, compared to baseline selection of 3,014, there would be approximately 550 more GP Registrations within four years FTE, but also an additional 103 extensions of six months, 87 of 12 months and 30 of 12-24 months, as well as 155 more trainees not obtaining GP Registration within five years FTE. The proportion of the incremental 736 trainees selected obtaining GP Registration within four years FTE (74.7%) is just below that of the 76.8% across all 3,014 trainees in the baseline approach. The implication is that there is little difference in the extension and failure rates between baseline selection of 3,014 (25.2% and 18.3% respectively) and Stage 2 only selection of 3,750 (26.2% and 18.9% respectively) although the absolute number of extensions and failures is higher and resources to address any additional patient safety problems and to support these trainees and their trainers must be made available.

All of the alternative models of selection increased the number of GP Registrations compared with baseline, but this effect was due to the increase in trainee numbers as well as the increased effectiveness of the alternative approaches. Table 7.3C shows, for selected approaches, the selection and training outcomes where only 3,014 posts are filled i.e. the number filled with the baseline approach. By comparing these results with those where up to 3,750 posts are filled (Table 7.3B) enables us to separate out the impact of **(1)** increasing the **number of posts filled** and **(2)** differences in the **effectiveness** of different approaches to selection i.e. differences in the **proportion of trainees achieving GP Registration within four years FTE.** Based on the means reported in Table 7.3C, Figure 7.3 summarises the incremental results compared to the baseline approach to selection. For example, with the Stage 2 only approach there would be an additional 550 GP Registrations within four years FTE compared with baseline (Table 7.3B). Had only 3,014 posts been filled using the Stage 2 only approach, there would have been an additional 77 GP Registrations (Table 7.3C), as shown by the orange part of the bar in Figure 7.3. Thus for the same number of trainees, Stage 2 only is more effective than baseline, but the majority of the additional 550 GP Registrations (550 – 77 = 473) are due to this approach filling more posts than baseline, as shown by blue part of the bar in Figure 7.3. Random selection of the same number of trainees would lead to fewer GP Registrations within four years compared to baseline selection (the orange bar is negative), while the other alternative approaches considered here are more effective than baseline selection. The blue, trainee numbers sections for all approaches are considerably larger than the orange, effectiveness sections and this implies that **increasing the number of trainees appointed has a greater impact on the number of GP Registrations within four years FTE than the relative effectiveness of an alternative selection model.**

» Figure 7.2: Additional number of trainees requiring extensions/not achieving GP Registration and the number of additional GP Registrations within four years FTE training time compared with baseline (1,315 extensions and 2,316 GP Registrations), by approach to selection, 3,750 posts.



» *Table 7.3A: Selection and training outcomes by selection approach (3,250 posts available): Mean (SD) across the 10 imputations.*

	Baseline	Random	Old Stage 2 cut-score	Stage 3 only	Stage 2 only	Stage 3 by-pass	Equal weight to all
Posts filled	3,014 (0)	3250 (0)	3,108 (0.5)	3,250 (0)	3,250 (0)	3,034 (1.5)	3,250 (0)
Posts not filled	236 (0)	0 (0)	142 (0.5)	0 (0)	0 (0)	216 (1.5)	0 (0)
Pass in 3 years FTE	1,699 (20.7)	1,692 (50.8)	1,733 (21.1)	1,805 (23.4)	1,892 (24.4)	1,713 (21.1)	1,849 (22.2)
6 month extension then GPR	296 (14)	352 (22.6)	308 (14.8)	343 (14.1)	307 (15.5)	297 (14.0)	325 (15.3)
12 month extension then GPR	321 (25)	358 (37.8)	330 (25.9)	353 (22.3)	354 (23.7)	324 (25.0)	352 (23.1)
24 month extension then GPR	144 (10.3)	158 (21.0)	151 (10.8)	155 (10.8)	143 (8.4)	144 (10.3)	149 (9.7)
No GPR within 5 years FTE	554 (14)	691 (29.7)	586 (15.1)	594 (16.0)	554 (17.7)	557 (13.8)	575 (16.0)
Total GPR in 4 years or less	**2,316 (19.1)**	**2,401 (36.2)**	**2,371 (19.5)**	**2,501 (24.4)**	**2,553 (22.5)**	**2,333 (19.6)**	**2,527 (21.1)**
Incremental GPR in 4 years or less cf. baseline	**N/A**	**85 (34.0)**	**55 (2.5)**	**185 (8.9)**	**237 (10.9)**	**17 (1.8)**	**210 (10.9)**
Incremental GPR in 4 years or less cf. baseline, using means across the 10 imputations	**N/A**	**3.67%**	**2.37%**	**7.99%**	**10.2%**	**0.73%**	**9.07%**

Note: GPR: On GP Register

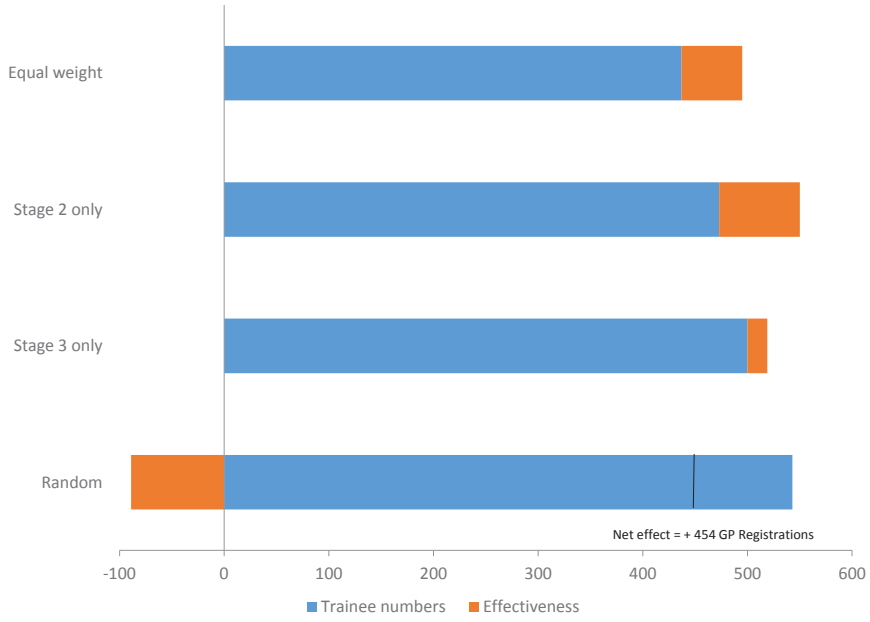
» *Table 7.3B: Selection and training outcomes by selection approach (3,750 posts available): Mean (SD) across the 10 imputations.*

	Baseline	Random	Old Stage 2 cut-score	Stage 3 only	Stage 2 only	Stage 3 by-pass	Equal weight to all
Posts filled	3,014 (0)	3750(0)	3,108 (0.5)	3,750 (0)	3,750 (0)	3,034 (1.5)	3,661 (25.5)
Posts not filled	736 (0)	0 (0)	642 (0.5)	0 (0)	0 (0)	716 (1.5)	89 (25.5)
Pass in 3 years FTE	1,699 (20.7)	1,952 (58.6)	1,733 (21.1)	2,010 (29.5)	2,059 (30.5)	1,713 (21.1)	2,029 (28.6)
6 month extension then GPR	296 (14)	406 (26.1)	308 (14.8)	418 (18.7)	399 (18.2)	297 (14.0)	383 (17.8)
12 month extension then GPR	321 (25)	413 (43.6)	330 (25.9)	407 (27)	408 (28.7)	324 (25.0)	399 (28.2)
24 month extension then GPR	144 (10.3)	182 (24.2)	151 (10.8)	180 (13.8)	174 (12.3)	144 (10.3)	170 (11.4)
No GPR within 5 years FTE	554 (14)	797 (34.2)	586 (15.1)	734 (22.8)	709 (24.6)	557 (13.8)	680 (22.5)
Total GPR in 4 years or less	**2,316(19.1)**	**2,771 (41.8)**	**2,371 (19.5)**	**2,836 (30.7)**	**2,867 (31.6)**	**2,333 (19.6)**	**2,811 (33.0)**
Incremental GPR in 4 years or less cf. baseline	**N/A**	**454 (38.8)**	**55 (2.5)**	**519 (18.7)**	**550 (23.3)**	**17 (1.8)**	**389 (37.5)**
Incremental GPR in 4 years or less cf. baseline, using means across the 10 imputations	N/A	19.6%	2.37%	22.41%	23.7%	0.73%	16.8%

» *Table 7.3C: Selection and training outcomes by selection approach (3,750 posts available but only 3,014 filled): Mean (SD) across the 10 imputations.*

	Baseline	Random	Stage 3 only	Stage 2 only	Equal weight to all
Posts filled	3,014 (0)	3,014 (0)	3,014 (0)	3,014 (0)	3,014 (0)
Posts not filled	736 (0)	736 (0)	736 (0)	736 (0)	736 (0)
Pass in 3 years FTE	1,699 (20.7)	1,569 (47.1)	1,699 (22.8)	1,797 (23.5)	1,762 (20.9)
6 month extension then GPR	296 (14)	326 (21)	308 (12.2)	268 (13.8)	285 (14.4)
12 month extension then GPR	321 (25)	332 (35)	329 (21.4)	328 (21.7)	327 (20.9)
24 month extension then GPR	144 (10.3)	147 (19.5)	142 (10.7)	129 (8.5)	134 (8.9)
No GPR within 5 years FTE	554 (14)	641 (27.5)	537 (16)	492 (15.7)	506 (13.4)
Total GPR in 4 years or less	**2,316 (19.1)**	**2,227 (33.6)**	**2,335 (23.9)**	**2,393 (19.8)**	**2,374 (18.8)**
Incremental GPR in 4 years or less cf. baseline	**N/A**	**-89 (31.7)**	**19 (8.4)**	**77 (8.0)**	**56 (6.3)**
Incremental GPR in 4 years or less cf. baseline, using means across the 10 imputations	N/A	-3.84%	0.82%	3.32%	2.42%

» Figure 7.3: The relative impact of increasing trainee numbers and different selection approach (up to 3,750 posts filled) effectiveness in producing additional GP Registrations within four years FTE compared with baseline (2,316 GP Registrations/3,014 posts filled).



7.5.2 Selection costs

Total selection costs by approach are summarised in Tables 7.4A (3,250 posts) and B (3,750 posts). As would be expected, **selection at random is the cheapest option, with total costs of £308,000 (to 3 SF) compared with £4,340,000/£5,020,000 for the baseline approach for 3,250 and 3,750 posts respectively.** With the exception of the approach lowering the Stage 2 cut-score, all other alternative approaches are cost-saving compared to baseline. There are three reasons for this: (1) a reduced selection process (e.g. Stage 2 only), (2) higher daily capacity at Stage 3 by not using moderation reducing the number of Stage 3 days required (and also the number of assessors) and/or (3) not requiring a second selection round to fill posts. **Using Stage 2 only – the cheapest alternative approach that would feasibly be adopted - reduces selection costs by around 75%.** Increasing the number of posts from 3,250 to 3,750 has no effect on total costs for the random, Stage 2 only and Stage 3 only approaches and increases total costs by 6% for Equal weight to all and between 15 and 18% for the remaining three approaches. (If a second Round was added to fill the remaining 90 posts for Equal weight to all then costs would increase by more than 6%.)

7.5.3 Cost-effectiveness

Our primary measure of cost-effectiveness is the incremental total cost per incremental trainee achieving GP Registration within four years FTE training time, compared to the baseline approach to selection. This incremental cost/incremental GP Registration has been calculated for each imputation for each alternative approach to selection. Figures 7.4A and B plot the results of the cost-effectiveness analysis for 3,250 and 3,750 posts respectively, showing one point for each imputation for each approach to selection. Note that the two plots have different axis scales, but certain points from Figure 7.4A (3,250 posts) have been reproduced on Figure 7.4B (3,750 posts) so that the difference can be ascertained.

The blue horizontal and vertical lines split the graph into four quadrants. Any points in the top left quadrant would be dominated by the baseline approach: incremental total costs would be higher and fewer trainees would obtain GP Registration (although there are no points in this quadrant). Points in the bottom right quadrant dominate the baseline approach: incremental total costs are lower and more trainees obtain GP Registration. For the remaining two quadrants,

[11] Points in these two quadrants are not cost-saving but may still be cost-effective.

» *Table 7.4A Selection costs for 3,250 posts, £'000 2014/15.*

	Baseline	Random A/B	Old Stage 2 cut-score	Stage 3 only*	Stage 2 only	Stage 3 by-pass	Equal weight to all
Running costs (including Stage 2 costs)	863	308	863	408	699	863	699
Stage 3 "on the day" costs	1,849	0	1,849	1,203	0	1,556	1,145
Assessor training costs	424	0	424	341	0	349	324
Candidate time costs	1,202	0	1,287	589	589	1,107	1,114
Total cost	**4,338**	**308**	**4,423**	**2,541**	**1,288**	**3,875**	**3,281**
Incremental cost cf. baseline	**N/A**	**-4,030**	**85**	**-1,797**	**-3,050**	**-463**	**-1,057**

» *Table 7.4B Selection costs for 3,750 posts, £'000 2014/15*

	Baseline	Random A/B	Old Stage 2 cut-score	Stage 3 only*	Stage 2 only	Stage 3 by-pass	Equal weight to all
Running costs (including Stage 2 costs)	863	308	863	408	699	863	699
Stage 3 "on the day" costs	2,466	0	2,466	1,203	0	2,172	1,291
Assessor training costs	491	0	491	341	0	416	374
Candidate time costs	1,202	0	1,287	589	589	1,107	1,114
Total cost	**5,021**	**308**	**5,106**	**2,541**	**1,288**	**4,558**	**3,478**
Incremental cost cf. baseline	**N/A**	**-4,713**	**85**	**-2,480**	**-3,733**	**-463**	**-1,543**

* Note for Stage 3 only we assume application numbers are known sufficiently far in advance such that the number of Stage 3 places required (and number of assessors requiring training) equals the number of applications passing Stage 1 (i.e. spare capacity is not required).

top right and bottom left, a trade-off between incremental costs and effects is required; e.g. in the top right quadrant, costs are higher but so too is the number of trainees obtaining GP Registration[11]. The blue diagonal line on the graph represents a plausible cost-effectiveness threshold. The gradient of the line is the present value of care one GP would provide in seven years FTE service as a GP (approximately £350,000 after discounting at 3.5% per annum), assuming as previously, that the value of care is proxied by the GP's salary. We take this as the willingness to pay for a qualified GP. Points below this line can be considered cost-effective compared to the baseline approach: the incremental cost per incremental GP Registration is less than the value of care that the GP would provide.

For 3,250 trainees, all of the points for the five non-random alternative selection approaches dominate the baseline approach, plus three of the ten points for random selection. A further five points for random selection of 3,250 trainees are in the top right quadrant but below the cost-effectiveness threshold. The remaining two points are above the threshold, suggesting that the incremental costs are above the value of care that would be provided by the incremental GPs. **However in none of the ten imputations is random selection of 3,250 trainees dominated by the baseline selection process (in which 3,014 trainees are selected).**

Stage 2 only appears to be the most cost-effective alternative, at the extreme bottom right of the graph. **For 3,750 trainees, all alternative approaches to selection dominate the baseline approach, including random selection.** Stage 2 only still appears to be the most cost-effective alternative, although the results are less clear-cut than for 3,250 posts, with Stage 3 only also a plausible alternative. The points for Equal weight to all have shifted slightly to the left because not all of the 3,750 posts are filled.

Figure 7.4 also shows the variability in incremental total costs and GP Registrations across the ten imputations (by considering the vertical and horizontal spread of points for each imputation respectively). However **its usefulness in comparing approaches is limited as it is not clear which specific point for a given selection approach should be compared to a specific point for another approach (i.e. in some cases, the points for one approach overlap with those of another, so it is not obvious which approach is the most cost-effective).** Table 7.5 therefore provides overall cost-effectiveness (average cost per GP Registration) and incremental cost-effectiveness ratios (additional cost per additional GP Registration when compared with baseline) across the ten imputations.

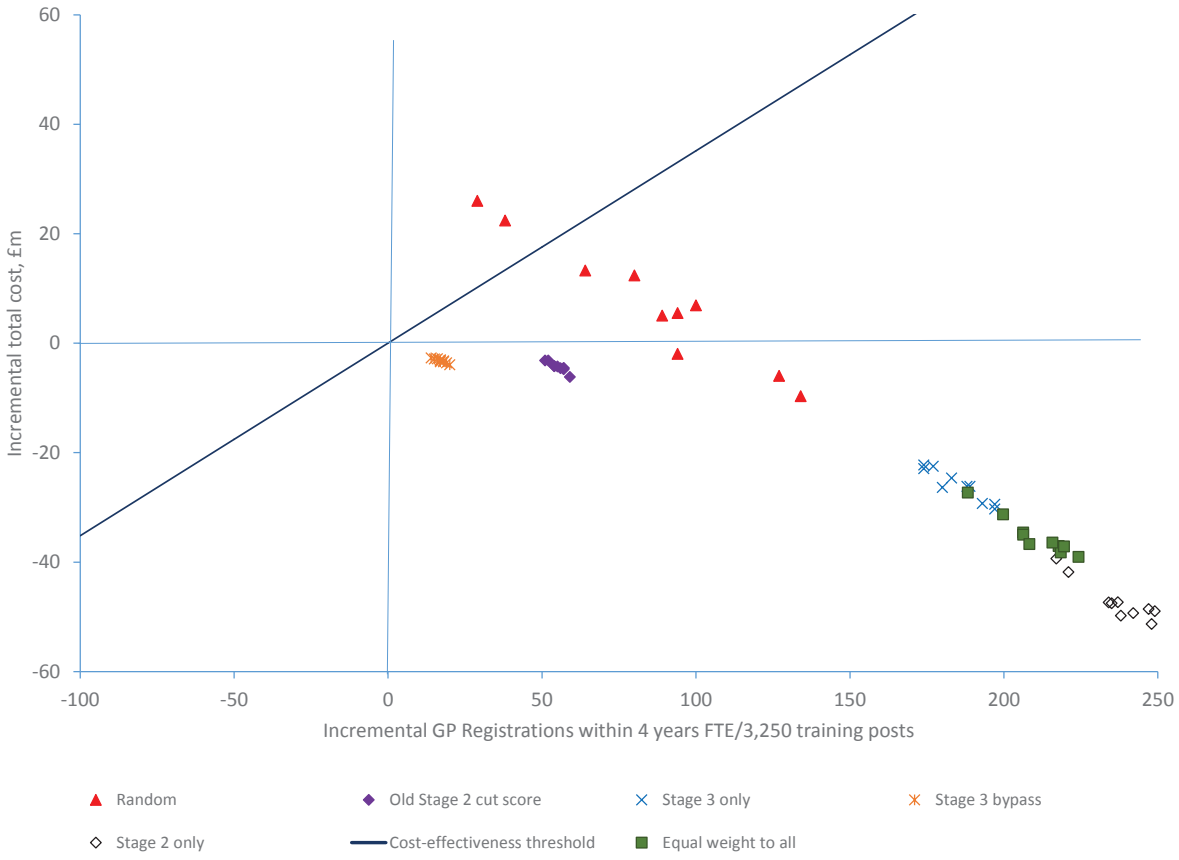
With 3,250/3,750 posts respectively, the overall cost-effectiveness ratio – the average cost per GP Registration using the baseline approach - is £442,000/£532,000, compared with £383,000/£404,000 using Stage 2 only. The average cost per GP Registration is higher with 3,750 posts for all approaches except random selection since the 'marginal' trainees appointed will be more likely to require an extension and/or less likely to obtain GP Registration than the initial 3,250 selected (see below for further explanation). **The Stage 2 only model has the lowest mean cost per GP Registration in four years FTE training time in all of the ten imputations for both numbers of posts and is thus the most cost-effective approach to selection considered in the economic evaluation.**

In terms of incremental cost-effectiveness compared with baseline selection, the best incremental cost-effectiveness ratios – the cost-savings per additional GP Registration - are a cost-saving of £199 per additional GP Registration in four years with the Stage 2 only approach and a cost-saving of £188 per additional GP Registration in four years with the Stage 3 bypass approach with 3,250 posts. While random selection of 3,250 trainees results in more GP Registrations in four years than baseline selection of 3,014, this comes at an average additional cost (over ten years) of £87 per additional GP Registration.

With 3,750 posts the Stage 3 bypass approach has the highest cost-saving per additional GP Registration in four years (£188). This approach is incrementally cost-effective because those selected through the bypass (who would have been rejected at Stage 3) have a relatively high probability of obtaining GP Registration within four years. However, with both 3,250 and 3,750 posts the additional number of GP Registrations from the Stage 3 bypass approach is low (17), so it is unlikely to be sufficient on its own to meet workforce requirements and is therefore when considering the overall cost-effectiveness ratio it is not the best option due to the number of unfilled posts.

With 3,750 posts all alternative approaches to selection produce more GP Registrations AND save money over ten years, even random selection.

» *Figure 7.4A: Scatter diagram to show the results of the incremental cost-effectiveness analysis for 3,250 posts, compared with baseline selection (i.e. baseline approach is at the origin); note Figures 7.4A and B have different x and y axis scales.*



The key drivers of these results are:

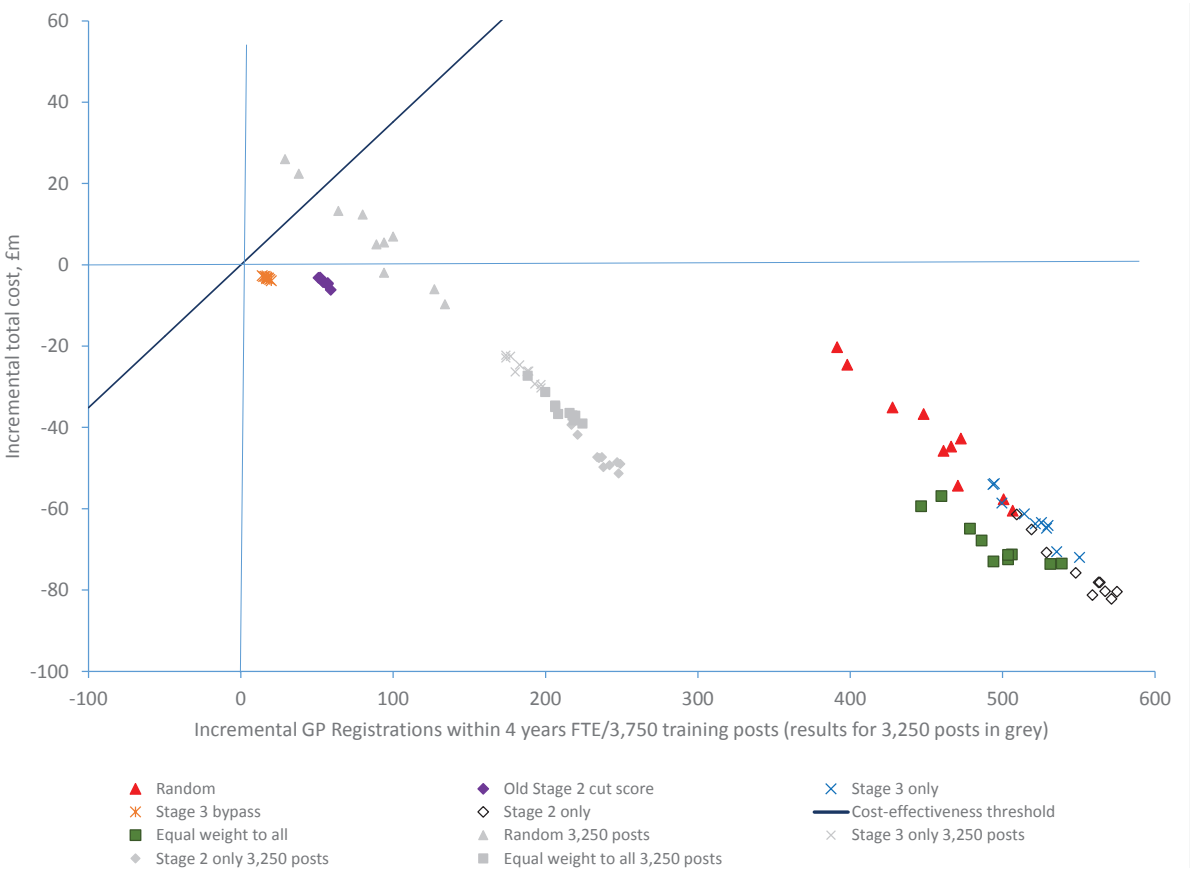
- Increasing the number of posts filled by using an alternative approach to selection compared with baseline has more of an impact on cost-effectiveness than the increased effectiveness of those alternative approaches in producing more GPs, as shown in Figure 7.3.
- The low variation in the proportion of trainees who would obtain GP Registration in four years FTE in each of the selection approaches, as shown in Table 7.5A/B.
- The significant negative impact on cost-effectiveness of the number of selected trainees who do not obtain GP Registration within five years FTE.

Selection costs, which make up less than 1% of total costs, do not impact on cost-effectiveness.

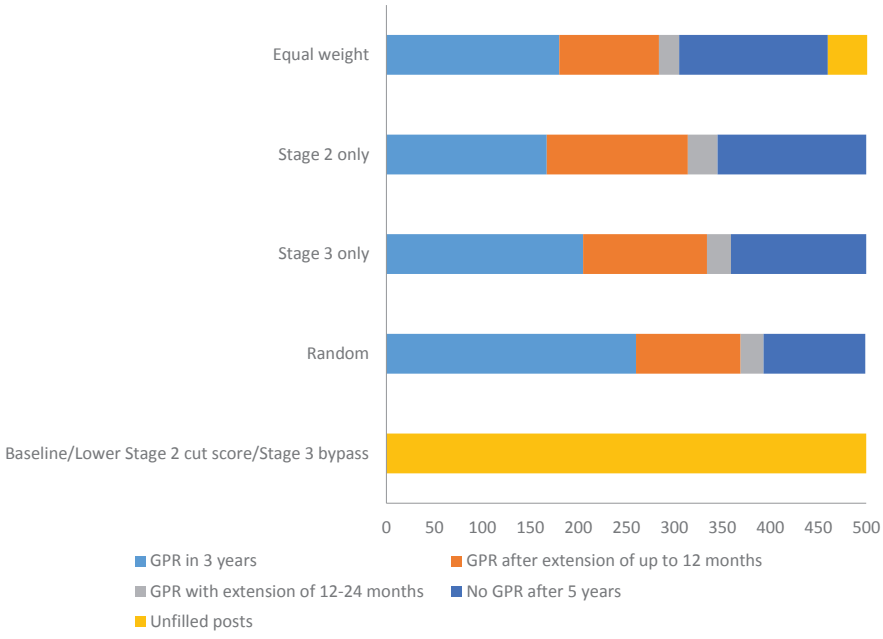
7.5.4 The 'marginal' 500: increasing the number of posts from 3,250 to 3,750 per year

Figure 7.5 summarises the selection and training outcomes for the 'marginal' 500 posts when the target recruitment is increased from 3,250 to 3,750 per year. For simplicity, some of the outcome categories have been combined. The results are based on the mean differences at imputation level for each approach to selection (approximately equal to the differences in

» *Figure 7.4B: Scatter diagram to show the results of the incremental cost-effectiveness analysis for 3,750 posts, compared with baseline selection (i.e. baseline approach is at the origin); note Figures 7.4A and B have different x and y axis scales.*



» *Figure 7.5: Selection and training outcomes when target recruitment is increased from 3,250 to 3,750 per year.*



» *Table 7.5A: Cost-effectiveness analysis results, 3,250 posts, £'000 2014/15: Mean (SD) across the 10 imputations.*

	Baseline	Random	Old Stage 2 cut-score	Stage 3 only	Stage 2 only	Stage 3 by-pass	Equal weight to all
Selection and training cost consequences (including unfilled posts but excluding selection costs)*	1,019,830 (4,968)	1,031,260 (12,842)	1,015,396 (5,291)	995,104 (6,654)	975,762 (6,813)	1,017,086 (5,052)	985,515 (6,029)
Increase in selection and training cost consequences compared with 'optimal' selection	50%	52%	50%	47%	44%	50%	45%
Incremental total cost cf. baseline (including selection costs)	N/A	7,399 (11,541)	-4,349 (850)	-26,523 (2,966)	-47,119 (3,707)	-3,208 (375)	-35,371 (3,548)
Change in total cost cf. (using mean across the 10 imputations)	N/A	0.72%	-0.42%	-2.59%	-4.60%	-0.31%	-3.45%
Total cost per GPR in 4 years or less (cost-effectiveness ratio)	442 (6)	430 (12)	430 (6)	399 (6)	383 (6)	438 (6)	391 (6)
Incremental total cost per incremental GPR in 4 years or less compared with baseline** (incremental cost-effectiveness ratio)	N/A	87	-78	-143	-199	-188	-168
% of posts filled with GPR in 4 years or less	76.8 (0.6)	73.9 (1.1)	76.3 (0.6)	77.0 (0.8)	78.6 (0.7)	76.9 (0.6)	77.7 (0.7)

* The figures in this row are large. Note that the minimum selection and training cost consequences, which would occur if 3,250/3,750 posts were filled and all 3,250/3,750 trainees obtained GP Registration in three years FTE training time (the 'optimal' selection process with zero cost), would be £677,953,000/£782,253,750. Incremental total costs (including selection costs) are shown as a percentage change from baseline in the fourth row and highlight that large absolute changes are actually small when considered in percentage terms.

** Care is required when interpreting this outcome. A negative value could occur if (1) incremental costs are lower and incremental GPRs higher (a good outcome, shown in red) or (2) incremental costs are higher and incremental GPRs are lower (a bad outcome, but no instances here). A positive value could occur if (3) both incremental costs and incremental GPRs are higher (requires consideration of the cost-effectiveness threshold or willingness to pay for an additional GPR, shown in amber) or (4) both incremental costs and incremental GPRs are lower (again requires consideration of the cost-effectiveness threshold, but no instances here). This outcome is calculated as the mean incremental cost divided by the mean incremental GPRs (rather than the mean incremental cost-effectiveness ratio) and hence no standard deviation can be calculated.

» *Table 7.5B: Cost-effectiveness analysis results, 3,750 posts, £'000 2014/15: Mean (SD) across the 10 imputations.*

	Baseline	Random	Old Stage 2 cut-score	Stage 3 only	Stage 2 only	Stage 3 by-pass	Equal weight to all
Selection and training cost consequences (including unfilled posts but excluding selection costs)*	1,227,432 (4,968)	1,189,916 (14,818)	1,222,997 (5,291)	1,167,293 (9,152)	1,155,825 (9,548)	1,224,688 (5,052)	1,160,571 (8,327)
Increase in selection and training cost consequences compared with 'optimal' selection	57%	52%	56%	49%	48%	57%	48%
Incremental total cost cf. baseline (including selection costs)	N/A	-42,230 (13,386)	-4,349 (850)	-62,619 (6,046)	-75,340 (7,197)	-3,208 (375)	-68,406 (6,086)
Change in total cost cf. (using mean across the 10 imputations)	N/A	-3.43%	-0.35%	-5.08%	-6.11%	-0.26%	-5.55%
Total cost per GPR in 4 years or less (cost-effectiveness ratio)	532 (6)	430 (12)	518 (6)	413 (8)	404 (8)	527 (6)	414 (8)
Incremental total cost per incremental GPR in 4 years or less compared with baseline** (incremental cost-effectiveness ratio)	N/A	-93	-79	-121	-137	-188	-138
% of posts filled with GPR in 4 years or less	76.8 (0.6)	73.9 (1.1)	76.3 (0.6)	75.6 (0.8)	76.4 (0.8)	76.9 (0.6)	76.8 (0.7)

* The figures in this row are large. Note that the minimum selection and training cost consequences, which would occur if 3,250/3,750 posts were filled and all 3,250/3,750 trainees obtained GP Registration in three years FTE training time (the 'optimal' selection process with zero cost), would be £677,953,000/£782,253,750. Incremental total costs (including selection costs) are shown as a percentage change from baseline in the fourth row and highlight that large absolute changes are actually small when considered in percentage terms.

** Care is required when interpreting this outcome. A negative value could occur if (1) incremental costs are lower and incremental GPRs higher (a good outcome, shown in red) or (2) incremental costs are higher and incremental GPRs are lower (a bad outcome, but no instances here). A positive value could occur if (3) both incremental costs and incremental GPRs are higher (requires consideration of the cost-effectiveness threshold or willingness to pay for an additional GPR, shown in amber) or (4) both incremental costs and incremental GPRs are lower (again requires consideration of the cost-effectiveness threshold, but no instances here). This outcome is calculated as the mean incremental cost divided by the mean incremental GPRs (rather than the mean incremental cost-effectiveness ratio) and hence no standard deviation can be calculated.

the means reported in Tables 7.3A and B). What might initially be surprising is that, amongst the last 500 trainees appointed, the best outcomes occur with random selection. This is because – as there is a positive correlation between selection scores and outcomes - the trainees with the highest likelihood of obtaining GP Registration have already been selected in the other approaches, leaving those at the lower end of the ability distribution to fill the marginal posts. With random selection, those not in the initial 3,250 have the same likelihood of obtaining GP Registration as those in that initial 3,250, so the overall likelihood of obtaining GP Registration is unchanged (as shown in the final row in Tables 7.5A and B).

While the Stage 2 only approach gives the highest number of GP Registrations within four years FTE with 3,750 posts, amongst the last 500 appointed, approximately two-thirds will require an extension (of any length) and/or fail to achieve GP Registration (compared to just under half of the last 500 (or any) appointed with random selection and 42% of the first 3,250 appointed based on Stage 2 only).

Table 7.6 shows the selection and training cost consequences for just the marginal 500 posts for each approach to selection. The marginal selection costs, being such a small proportion, are excluded. As above, as the number of posts to be filled increases, random selection becomes more cost-effective (although remains dominated by alternative models as shown in Figure 7.4B). Similarly, the cost-savings for the marginal 500 are smallest with Stage 2 only because it is the most effective approach for the initial 3,250 trainees. **Compared to leaving all 500 posts unfilled, every alternative model that increases recruitment is cost-saving, even amongst the last 500 trainees to be appointed.**

Clearly, determining the target number of qualified GPs (and, in turn, the number of GP trainees) is a workforce planning decision made outside of the recruitment process and therefore of this evaluation. Yet such decisions – i.e. whether the current target of 3,750 trainees per year is appropriate - can be informed by data such as those presented here. If the future care forgone when a qualified GP is not 'produced' by the system does have financial value (based on a GP's salary) as assumed here, then financially, it makes sense to fill all 3,750 posts. However, as with the earlier discussion of Figure 7.2, the financial savings need to be weighed against the practical and emotional consequences of increased extensions, numbers failing to obtain GP Registration, and potential increase in patient safety issues. For example, with Stage 2 only, approximately 17% of the top 3,250 selected will not obtain GP Registration within five years FTE, compared with 31% of the 'marginal' 500.

It is also important to highlight that decisions regarding which selection process to use should not be based on marginal analysis but by consideration of the overall impact across all trainees appointed. Comparing the results for 3,250 and 3,750 posts suggests that Stage 2 only will be the most cost-effective option for any number of posts within this range. Equal

» *Table 7.6: Selection and training cost consequences for the 'marginal' 500 posts.*

Approach to selection	Selection and training cost consequences, £'000	Incremental cost compared with baseline, £'000
Minimum possible cost consequence if all 500 pass in 3 years FTE	104,301	
Baseline/Old Stage 2 cut-score/Stage 3 bypass*	207,602	N/A
Random	158,655	-48,946
Stage 3 only	172,189	-35,413
Stage 2 only	180,064	-27,538
Equal weight to all	175,056	-32,546

* This is the cost associated with leaving all 500 posts unfilled.

7.6 IMPLICATIONS OF THE ASSUMPTIONS REQUIRED IN THE ECONOMIC EVALUATION

In this section we discuss the possible implications of the key assumptions required in the economic evaluation.

We did not incorporate the variability in selection costs between LETBs and how missing data from three LETBs may have biased the results. This was primarily due to the very small proportion of total costs (less than 1%) accounted for by selection costs. A UK-wide approach was taken, rather than modelling each LETB separately, which would have been impossible given the time available and the interaction between LETBs in the Clearing process. It may be the case that some posts could not be filled in any model if there were insufficient applications preferencing a LETB.

In costing an unfilled post, we assumed a replacement hospital doctor would be required at 100% FTE during the 18 month hospital placement for a GP trainee, although it is assumed that such a trainee would only spend 50% of their time undertaking clinical duties (Department of Health, 2014a). Repeating the analysis for 3,250 posts based on the cost of not filling a post associated with a 50% replacement cost (total for each unfilled post of £383,378 rather than £415,203) did not change the majority of the results, although in only one of the ten imputations was random selection cost-saving compared to baseline selection.

We used a ten year time horizon for the economic evaluation. A longer time horizon would increase the cost of not filling a post, since more than seven years of GP service would be lost. A longer time horizon would also increase the cost-effectiveness threshold, making the diagonal line on Figure 7.4A/B steeper and thus random selection to fill posts more likely to be cost-effective compared to baseline selection of 3,014.

The assumption regarding the use of salary to proxy the value of care provided by a GP is used in determining the cost-effectiveness threshold (the willingness to pay for a qualified GP). Here, the effect of an increase in this value is to increase the cost-effectiveness threshold, making the diagonal line on Figure 7.4 steeper and thus random selection of 3,250 trainees more likely to be cost-effective compared to baseline selection of 3,014.

We also assumed that a post unfilled or trainee not obtaining GP Registration left a gap in the provision of primary care in the future. In reality, existing GPs may work unpaid overtime to see additional patients (and thus avoid foregone health outcomes), although the opportunity cost of doing so - the value of their leisure time foregone plus the additional risk of burnout – should still be included in an economic evaluation.

We were unable to consider the quality of care provided by trainees. We might reasonably assume that trainees who do not obtain GP Registration would be more likely to provide unsafe, unprofessional or disrespectful care than those who do. Although the frequency of such poor care is likely to be low, the health and financial costs (including litigation) of any contingent preventable adverse events could exceed the care foregone by not filling a post. Thus further work to explore the errors made by trainee and qualified GPs, and relating their frequency and consequences to selection scores, would be valuable in order to improve the economic model used here. This work would enable a potential 'bottom line' for appointment in a revised approach to selection be identified: we have focused on filling posts, but there may be a point at which further appointments are inadvisable. **While still the most cost-effective approach overall when 3,750 posts were filled, two-thirds of the last 500 trainees appointed using Stage 2 scores only would require an extension or fail to achieve GP Registration. Whether the additional GP Registrations achieved justifies the patient safety, practical and emotional costs of such outcomes is a decision to be made outside of this evaluation.**

7.7 SUMMARY

This chapter has reported our cost-effectiveness analysis of the current, baseline approach to selecting GP trainees, compared with six alternative approaches to selection. While not all alternatives may be considered acceptable to all stakeholders, their inclusion is illustrative; for example the results for random selection demonstrate the importance of filling posts based on the ten year time horizon under the perspective of a virtual provider of GP services in the UK that is used in the economic evaluation.

With this time frame and perspective in mind, **the most cost-effective approach to selection from those considered in this chapter would be to select trainees on the basis of their combined SJT and CPST scores only.** This would provide just over 200 additional GP Registrations within four years FTE with 3,250 posts and around 550 with 3,750 posts. All trainees would be selected in Round 1 and all posts would be filled. With 3,250 posts, the adverse training consequences of this approach to selection are low, since the mean number of trainees selected but not obtaining GP Registration across the ten imputations is the same in this Stage 2 only approach as in the baseline approach (around 500/3,250); however there are slightly more extensions (around 40 per year). This is an important finding, since increased failures might reduce the morale or standing of the profession as a whole, something which cannot easily be valued and included in an economic evaluation. However, more of a trade-off is required if 3,750 posts are filled: compared to baseline there are more extensions (220 per year) and trainees not obtaining GP Registration (around 150 per year).

~ This page is intentionally left blank ~

Part 2

Broader Recruitment and Selection Perspectives

~ This page is intentionally left blank ~

Chapter 8

The numbers of GPs, and the influences of medical school and other factors on choosing to specialize in General Practice

Chapter 8.

The numbers of GPs and influences before and during medical school on choosing to specialize in General Practice

8.1 INTRODUCTION

Becoming a GP is a choice on the part of an individual medical student, and like all choices it will be preceded by intentions which in turn will be influenced by a student or doctor's background, by the training they receive, both generally, and specifically by particular medical schools. There are also historical shifts in the popularity of general practice as a specialty, in part influenced by changes in the status of the specialty and the opportunities it offers.

In the post-war period, general practice had reached the nadir of its status. The Royal College of General Practitioners was founded in 1952, in part as a response to the Collings Report, the RCGP's own website describing how[1].

> *"In 1950 The Lancet (Collings, 1950) published a report, made by a visiting Australian doctor on his personal survey of British general practice ... . He had come prepared to admire and to learn, but was appalled by what he found. In his report, which was given prominence by the Lancet, he painted a dramatic picture of exhausted and demoralised doctors, hurried work and low standards. His report made it impossible for the medical establishment to ignore the impending crisis."*

The 30-page report, accompanied by a hard-hitting three-page editorial (Anonymous, 1950) initiated the slow processes of change. General practice was on a cusp, and was believed to have a meagre future – if it had a future at all - even by some of its most effective practitioners:

> *"Some of the best family doctors whom I met were actively discouraging their sons and daughters from succeeding them in general practice: as I have already explained, they believed that the eclipse of general practice is at hand – that it is to be supplanted by more and more institutional and specialist medicine" (p.572)[2].*

As with most of medical education in the last half century, the major influence on General Practice training has been the Todd Report of 1968 which also discussed the generally low status of general practice and the absence of proper training, with many suggestions on how undergraduate and postgraduate training in GP should be organised. A particular problem was seen to be that:

> *"undergraduate clinical teaching has been based almost entirely on patients referred to hospital [... with no attempt to introduce...] the wider problems of sickness in the community" (Royal Commission, 1968)para 277).*

The changes in the quantity and type of undergraduate general practice curricula and teaching post-Todd have been well described by Harding et al (Harding et al., 2015). Curriculum time for the teaching of general practice rose from <1% in 1968 to 13% in 2008, but then did not change until 2013, while numbers of clinical sessions for teaching in fact decreased

[1] http://www.rcgp.org.uk/about-us/history-heritage-and-archive/history-of-the-college.aspx

[2] In passing it has to be said that the phrase "and daughters" belies many common stereotypes of the period. Less than 30% of medical students at the time were female, with the proportion being less than 20% before the Second World War, with a rise to nearly 30% during the war, and then a subsequent falling back to 22% by 1965, after which began the steady climb to the modern situation with a clear majority of medical students being female McManus, I.C. and Sproston, K.A. (2000) Women in hospital medicine: Glass ceiling, preference, prejudice or cohort effect? Journal of Epidemiology and Community Health, 54: 10-16.
The cohorts labelled in Figures 1A and 2A relate to cohort studies described later in this chapter

from 2002 to 2013. The number of practices involved in teaching doubled between 1986 and 2013, and while departments of general practice did not exist prior to 1968, all medical schools had them by 2002, although that number subsequently fell to under 50%.

While all students are now exposed to GP teaching, in contrast to the situation in the 1960s where barely any students experienced GP, Lancaster (Lancaster, 2015) has emphasised, taking the long view into account, that while undergraduate teaching had undoubtedly increased dramatically since the 1960s, that same period overall saw a relative **decline** in the numbers of doctors becoming GPs. This inevitably raises some doubt about the specific influence of GP teaching on long-term trends, which may reflect other changes in the profession or wider society.

Lancaster's finding raises many questions about the nature of the influence of undergraduate GP teaching on deciding to become a GP. In particular: might GP teaching only have an influence on those who are already interested in GP as a career; might GP teaching be putting off some students who might otherwise have considered it later; and are specific schools particularly adept at encouraging medical students to become GPs, perhaps through their teaching? In a separate chapter we report details of the questionnaire study of applicants to the 2015 specialty selection round, which does suggest that experience of GP in medical school and particularly after medical school are important in influencing the decision to apply for GP training.

This chapter will look at a number of different issues:

1. **Historical trends:** How has the proportion of doctors becoming GPs changed in recent years, with 'recent' being defined in terms of the life-time of the older doctors working in the NHS? If proportions are stable then they may be hard to change, whereas if they are more labile then there may be more opportunity to influence them (but those forces may also put people off becoming GPs). These questions can mainly be answered by looking at doctors on the Medical Register.

2. **Recent trends in applications for and completion of GP training:** For cohorts graduating since about 2000 it has become possible to monitor when doctors enter the GP Register (introduced in 2006). Applicants for GP training can also be monitored for the selection rounds for 2009 to 2015 (graduation cohorts from 2007 to 2013), giving an assessment of the time of the first serious statement of intent to become a GP.

3. **Attitudes and intentions towards a career in General Practice.** When do students and doctors develop an intention to become a GP, and how likely is it that intentions eventually manifest as actually becoming a GP? Occasional suggestions in the literature that it might be better if some medical schools were only to train GPs implicitly assumes that would-be GPs can reliably be identified before entry to medical school. A number of studies bear on this question, including the work of the Oxford Medical Careers Research Group (MCRG), a series of three large cohort studies originating in St Mary's in the 1980s and 1990s, and smaller studies, particularly of 16-year olds and 11-year olds.

4. **The general influences of background and medical training on becoming a GP:** How do demographics (sex, ethnicity) and family background (doctors as parents or in the family) influence becoming a GP? How do intentions change through medical training, and what is the influence of perceived quality of GP training?

5. **The particular influences of specific medical schools:** Some medical schools produce more GPs than others, but why that is so is less clear; do the students differ when they enter medical schools (differential selection); are they influenced differently by their medical schools (differential training); and do students from different medical schools fare equally well during GP training? Once again, it is an issue going back to the Todd report of 1968:

   "First year students at Oxford and Cambridge differed greatly from students in other schools, opting ... considerably less often for general practice [with final year students showing the same pattern]. ... Students in provincial schools in both years showed greater preference for general practice... " (pp 359-360).

However, the Report also commented:

> *"Our studies do not suggest that a major effect on the part of a medical school to interest students in general practice results in a larger number choosing that career." (Royal Commission, 1968) p.358, our emphasis).*

None of the studies or the datasets to be looked at below has all of the necessary information to answer all of the questions asked above, but together, as a mosaic, they allow the broader picture to be seen.

8.2 HISTORICAL TRENDS

Since 2006, all doctors working independently as GPs have been required to be on the GMC's GP Register (and doctors working as other specialists have been required to be on the GMC's Specialist Register since 1996). The GMC's LRMP (List of Registered Medical Practitioners), which includes all doctors, whether or not they are on the GP and/or Specialist Registers, , therefore makes a good starting point for looking at the proportion of GPs and Specialists, organised by the year (cohort) in which the doctors graduated. Each doctor can be identified by their unique GMC registration number. The present analysis is based on the LRMP of May 2015, and includes both doctors who are living and practising, as well as those who no longer have a licence to practise, or who are deceased. Figure 8.1 shows the **numbers** of doctors in each graduation cohort, whereas Figures 8.2 and 8.3 show the **proportions** of doctors in each graduation cohort, broken down in each case by whether they are on the GP or Specialist Registers.

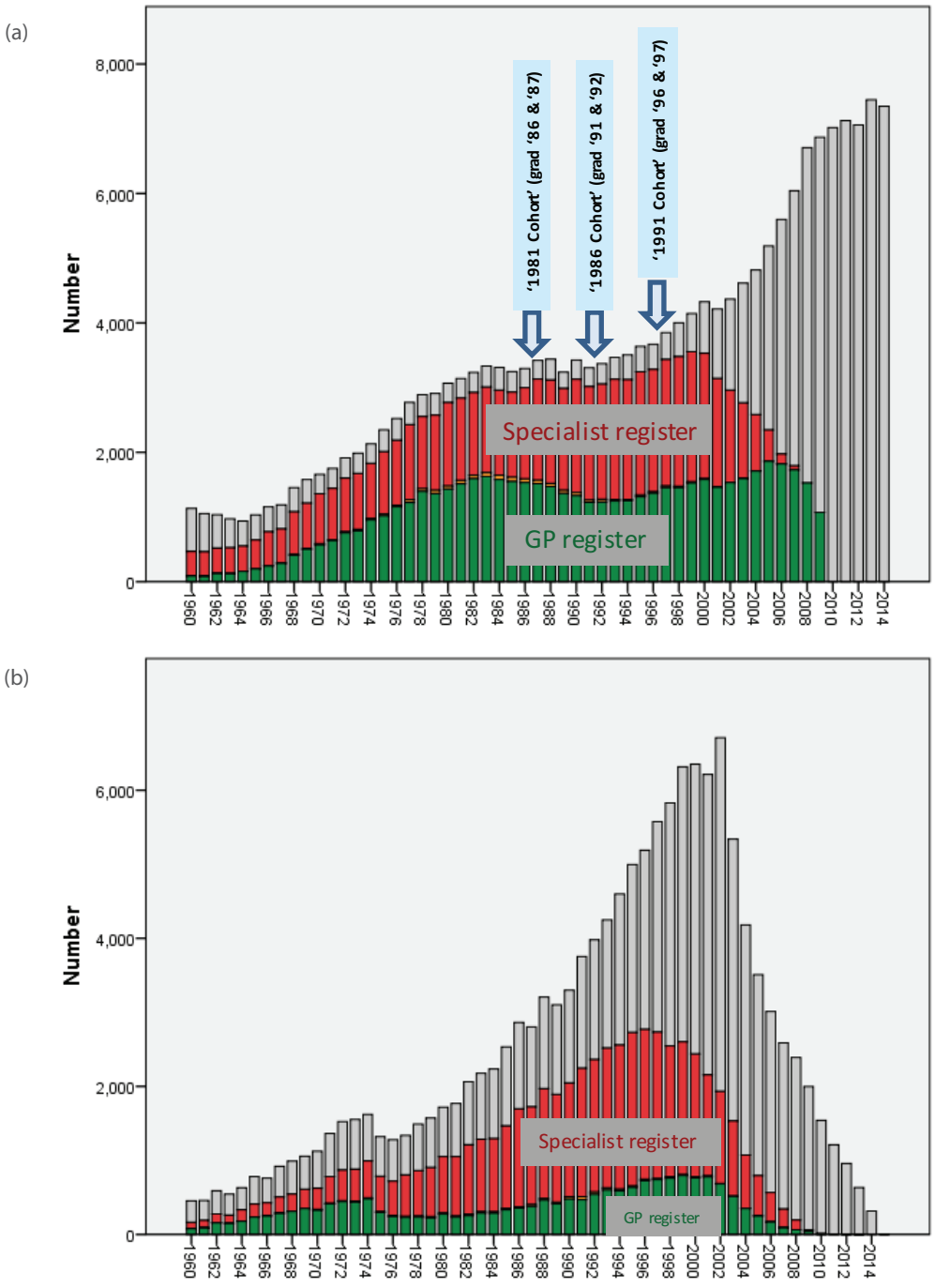
Figure 8.1 shows the **numbers** of doctors on the LRMP from each graduation cohort, separated into UK and non-UK graduates. The various stages of expansion of UK medical schools are readily seen. Numbers of non-UK graduates are more complex as doctors only come onto the GMC LRMP when they are permitted to practise in the UK, which may not be until they start specialty training (or later). Doctors on the GP Register are shown in green, those on the Specialist Register in red, a very small number on both Registers in orange, and those on the LRMP but on neither Register in grey.

Figures 8.1a and 8.1b are not easy to interpret, not least because the number of doctors from each cohort varies, typically increasing with year. Figures 8.2a and 8.2b show the same data but with the **proportion** of doctors on the GP and Specialist registers shown. Considering UK graduates, doctors graduating before about 1970 are less likely to be on either Register, probably due to them retiring before the registers were introduced, and they can be ignored for present purposes. Proportions on the Registers after about 2000 for the Specialist Register, and about 2006 for the GP register are confused by doctors not always having had time to get onto the Register by 2015, the date of the recent LRMP which was used. However from 1970 to 2000 several patterns are very clear:

1. **Those on neither Register:** About 15% of UK graduates are not on either Register, suggesting that their practice or careers did not involve seeing patients or making clinical decisions, or were practising in roles that did not require GP or Specialist Registration. This proportion is surprisingly constant, but highlights the importance of correctly interpreting the DoH[England]'s target that "[I]n future **at least half of doctors going into specialty training will be training as GPs.**" [2008] (Department of Health, 2008) (p.15, para 36, our emphasis). However Peile's interpretation of the target as "... 50% of new medical graduates should be recruited to general practice" (Peile, 2013) (p.565) would increase the target by 7.5 percentage points.

2. **The GP register:** For graduates from about 1974 to 1987 about 46% of all graduates were on the GP Register, and indeed 54% of those taking any form of specialist training were on the GP Register. **From 1974 to 1987, therefore, the DoH [England]'s post-2008 target was being met.** The proportion of GPs then falls quite quickly for four or five years from 1987 to 1991, and then again is stable for graduates from about 1991 to 2005, averaging about 36% of all graduates entering the GP Register (and they are 42% of all those completing GP or specialty training). That change can also be seen clearly in the longitudinal data from the MCRG for the 1983 and 1988 graduation cohorts (Lambert et al., 2002).

» *Figure 8.1: The numbers of doctors in each graduation cohort on GP and specialist registers[3]: (a) UK graduates, total numbers; (b) non-UK graduates, total numbers[3].*



[3] The cohorts labelled in Figures 8.1 (a) and 8.2 (a) relate to cohort studies described later in this chapter.

5. **Sex differences:** The proportion of women entering general practice is higher than that of men (see below for a more detailed discussion). Figure 8.3 shows the proportions of UK graduates on the GP and Specialist Registers. Male doctors are more likely to be on one of the Registers; just over 90% on average, compared to around 82% for females. It is clear that the broad picture for men and women is similar across the cohorts, with the drop in the proportion of doctors on the GP register occurring at much the same time in both sexes. However the fall was much larger for males than it was for females.

Perhaps the most striking feature of Figures 8.2 and 8.3 is the **relative stability of the proportions of graduates over time who are on the GP Register.** The proportion did fall quite quickly for UK graduates from 1987 to 1991, and for non-UK graduates rather earlier, for the 1974 to 1982 cohorts. The cause of neither of these drops is obvious, but may reflect either changes in general practice or changes in hospital practice. The pattern is also very similar in the two sexes, despite the proportion of women entering medical school increasing from about 20% in the 1960s to about 65% in the first decade of the present millennium. The big picture overall is of relative stability, with some lability. **The implication may be that producing large and rapid changes in the proportions of doctors becoming GPs may not be straightforward.**

8.3 RECENT TRENDS IN DOCTORS BECOMING GPS OR APPLYING FOR SPECIALIST TRAINING IN GP

The previous section has shown that the long term trends, at least for UK doctors graduating from about 1991, is of relative stability, with about 36% of each UK cohort going onto the GP Register. Little information is available however on the timing of that process. Timing can be assessed in more recent cohorts by looking at the time after qualification when a GP goes on to the Register and the time after qualification when a doctor applies for specialist training. Information can also be gained from ARCP on transfers between different specialty training programmes.

8.3.1 Time from qualification to entering the GP Register:

The analyses of figure 8.2 are relatively limited, because generally there is little information on when doctors became GPs, be it soon after leaving medical school or many years afterwards. The GP Register began in 2006, and at that point almost all practising GPs would have gone on to it. Since 2006 therefore, it is possible to look at the rate at which new doctors in each cohort have gone on to the Register. The denominator in these calculations is the total number of doctors qualifying from all UK medical schools in a calendar year, and the numerator is the number of those doctors who are on the GP Register in the years after they graduate. Figure 8.4a shows, for each cohort from 2000 to 2010, the proportion who are on the GP Register 4, 5, 6 years etc. after qualification. A very small proportion of graduates are on the Register in their 4th year after qualification[4], and then the rate rises with each year, topping out at 36.9% of the 2000 cohort of doctors being on the GP Register, much the same as the long-term figure calculated earlier. More revealing though is that **it is not until eight or nine years after qualification that the vast majority of those who are going to become GPs are on the Register.** In particular that is significantly longer than expected under a 'standard training model' of two years foundation followed by three years of GP training, so that GPs could go onto the Register in the fifth year after graduation. In fact that applies to only about a half of GPs. The separate graduation cohorts are shown in different colours in figure 8.4a: Black for 2000-2002, Blue for 2003-2005, Green for 2006-7 and red for 2008 to 2010. Without doing any detailed statistical comparisons, it is clear that the lines for the various cohorts overlie one another, suggesting that the process was stable for the graduation cohorts from 2000 to 2010[5].

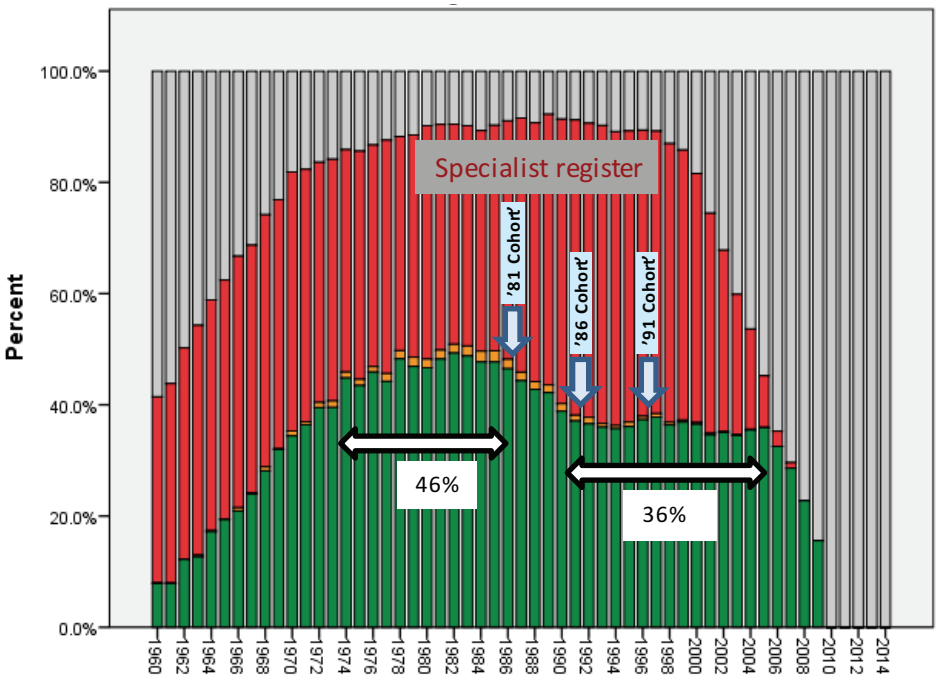
8.3.2 Time from qualification to applying for GP training:

Doctors can only go onto the GP Register after being successful in their assessments in a recognised training programme. An early indication of future GP numbers can therefore be obtained by looking at numbers of UK graduates applying for GP training. Data were provided for the selection programmes in 2009 to 2015 (i.e. for beginning training in 2009 to 2015, doctors applying in the previous autumn of 2008 to 2014). The UK graduate cohorts of 2007 to 2013 were studied[6], as these would have been at the end of FY2 in 2008 to 2014, and hence applying for the 2009 to 2015 programmes. As in the previous analysis, the denominator is the number of doctors on the LRMP who graduated from UK medical schools in each calendar year, and the numerator is the number of doctors in that cohort who applied for GP training in each year after graduation.

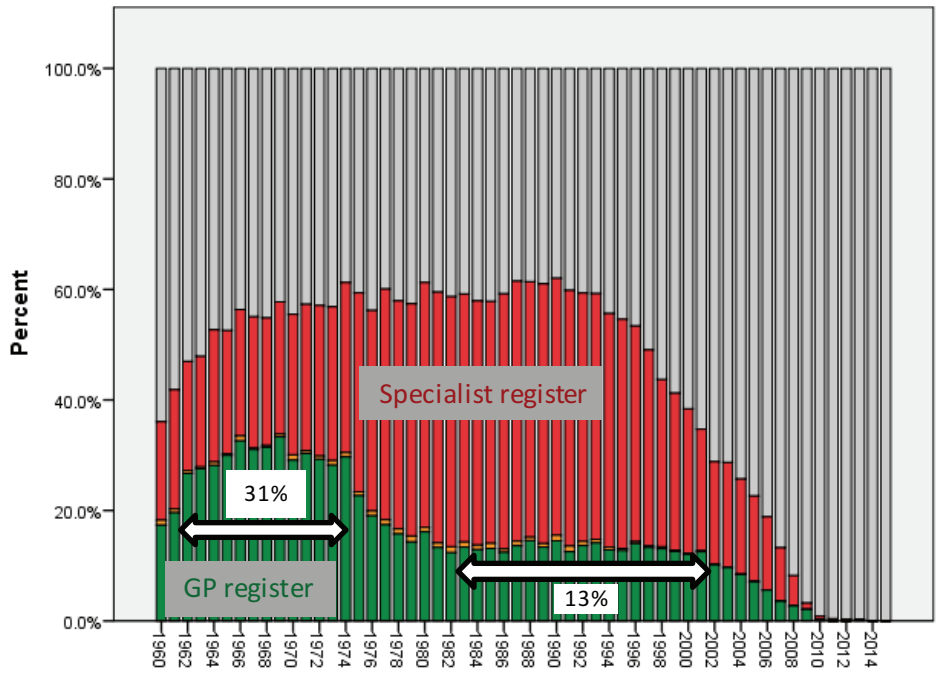
[4] Given that the minimum training time is five years, these doctors may have been slightly late obtaining their initial registration but were able to 'make up' for lost time.

» *Figure 8.2: The percentage of doctors in each graduation cohort who are on the GP and specialist registers: (a) UK graduates, total numbers; (b) non-UK graduates, total numbers.*

(a)

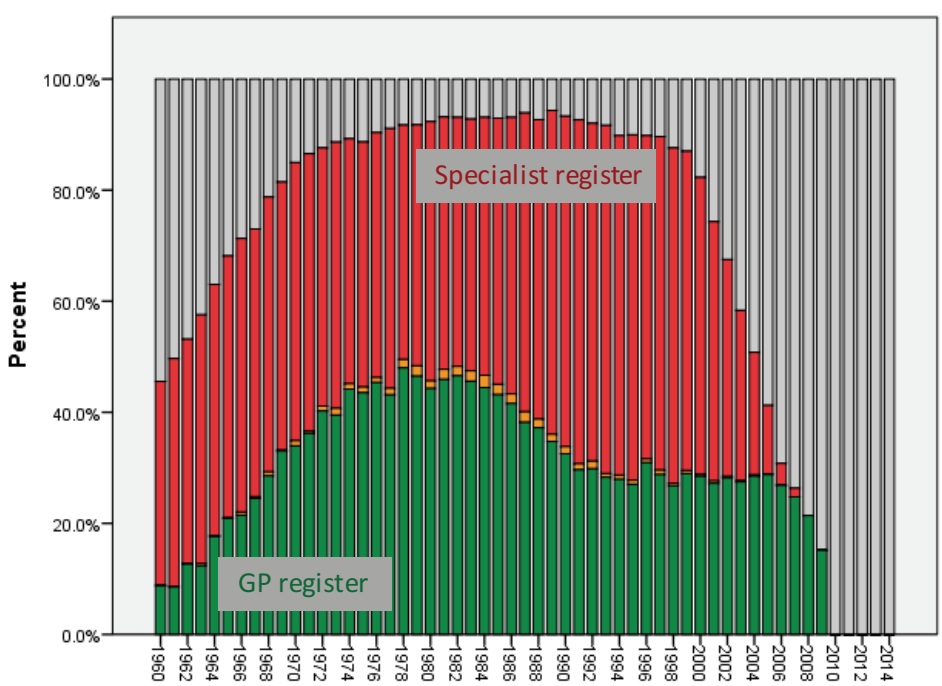


(b)



» *Figure 8.3: The percentage of male and females in each graduation cohort on GP and specialist registers: (a) male UK graduates; (b) female UK graduates..*

(a)



(b)

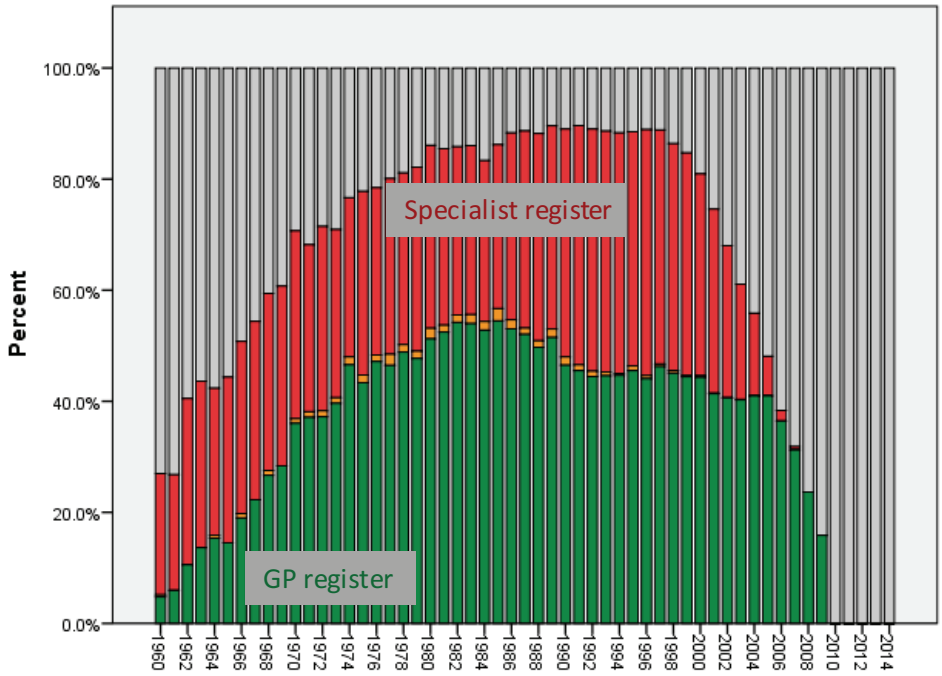


Figure 8.4b shows the proportion of graduates in the 2007 to 2013 cohorts who have applied to the GP training programme at various intervals after qualification. Notice that the 2007, 2008 and 2009 cohorts had also appeared in figure 8.4a, where their rate of entry onto the Register was similar to that of earlier cohorts. The 2007 cohort provides the longest span of data, with 38% applying in the first year possible, rising to a total of 52% having done so by eight years after qualification. This is an important result as it means that in **this cohort over a half of the UK graduates had applied for GP training at some time.** The questionnaire study of 2015 applicants (see Chapter 9) suggests that of those applying for GP, about 69% have done so as their first choice, and in 2015 83% accepted their GP offer in round 1 (see Chapter 10). Overall therefore, perhaps 30% to 36% of all graduates are likely to be serious about entering GP training at some point after graduation, with more potentially doing so after being unable to obtain a post in their first choice specialty. The 2008 cohort shows a very similar pattern to the 2007 cohort. However **as the years continue it is clear that the proportion applying in their second year post-graduation is dropping, particularly rapidly for 2012 and 2013 graduates,** with just 24% of 2013 graduates applying in 2015. Some evidence suggests that doctors may just be taking 'a gap year', but it is also possible that there is a sustained reduction in applications to general practice.

8.3.3 Transfers in and out of General Practice:

Many doctors do not apply for GP training immediately after Foundation, and the data suggest that many begin training in another specialty and later transfer to GP. A sense of the flows of doctors between specialties can be gained by looking at the specialties for which doctors are training in ARCP data. Data for ARCP are available for 2010 to 2014 (i.e. the years August 2009 to July 2010 through to August 2013 to July 2014). The data are not easy to interpret as the data are both left-censored (many trainees come into the data in 2010 already having nearly completed training) and right-censored (many trainees have data only for 2014 or 2013 and 2014).

To assess flows between specialties we have looked at the first post-foundation specialty for which there is an ARCP record, and the last post-foundation specialty for which there is an ARCP record. This almost certainly under-estimates the proportions of doctors changing specialty, but it should still a good sense of the direction of the flows. Table 8.1 shows the flows between specialities within these ARCP data. The rows show the initial specialty (i.e. in the first ARCP record available), and the columns show the last specialty in the data available. Of 67,710 doctors with ARCP data, 19,723 are in GP training at the beginning and the end of the data. **However 199 doctors have left GP (shown by the red shaded cells), whereas 1513 (shaded green, almost eight times as many) have transferred into GP.** Losses from GP are mainly to Medicine (54) and Psychiatry (52), with others shifting to Paediatrics (18), Emergency Medicine (18), O&G (11) and Anaesthetics (10). In the other direction, doctors transferred into GP mainly from Medicine (598), Surgery (285), Paediatrics (149), Anaesthetics (125), O&G (115) and Psychiatry (102). The data are not good enough for telling **when** these transfers are occurring but they do give a good indication of where the transfers are coming from and going to. Perhaps most surprising is that entrants into GP are not only 'the usual suspects' (Medicine, Paediatrics, Psychiatry), but there are also large numbers coming from Surgery, Anaesthetics, O&G and ACCS, none of which would traditionally be looked to for possible GP recruits. These results – that **GP is often a later career choice** - support those reported above which highlighted that some doctors are applying to enter GP training up to eight years (or even more) after graduation. However the results do not explain the change over time in the proportion of UK graduates applying for GP training in their second year following graduation.

8.3.4 Explaining the changes:

It is difficult to predict what will happen in the next few years to the percentage of UK graduates applying to GP training. In Figure 8.4b, compared with 2007, fewer graduates from the 2008 to 2010 cohorts applied in Year 2, but more applied in Years 3 and 4, so that the level of applications caught up with that of the 2007 cohort. That 'catch up' effect though is less obvious for the 2011 cohort, and less so still for the 2012 cohort, although there is still a higher percentage of applications

[5] The red lines suggest a small drop in the most recent year where the rate does not go up as quickly as expected. That is an artefact of these analyses being carried out using the LRMP of August 8th 2015. Although the majority of doctors go on to the GP Register in the first week of August (which is about week 31), there is still a small proportion, typically about 10% of the entrants for that year, who enter the Register after about week 32. Since the analysis in the figure is by calendar year the very final year is not 100% complete and the actual gain in that year will be slightly higher.

[6] There is no record in the GP selection data of whether a candidate has applied before, and hence the only data which make sense to use are those for cohorts graduating from 2007 onwards, as they could not have applied before 2009.

in Year 3 after graduation for 2012 than for the 2007 cohort i.e. the difference in the cumulative percentages is reduced. There are two possible mechanisms underlying the current low level of GP applications in the second post-graduate year:

1. **Model A:** Doctors qualifying since 2011 are becoming less and less likely to want to go into General Practice. Fewer apply in Year 2, and applications in subsequent years parallel that of the 2007 curve, topping out at about 38% applying at one time or another for GP. That is the scenario shown in Model A with the black line with small dashes.

2. **Model B:** Graduates in recent years are as keen on a career in general practice as previously, but they are happy to delay a year or two before applying for GP training. The implication is that by 8 years after graduation there will be a similar proportion of doctors who have applied to GP training as with the 2007 graduates. That is Model B, shown by the large black dashed line. That is not such a pessimistic scenario, as the number of doctors applying for training is similar to previous years However the 'fallow years', say one to three for typical doctors, would be one to three life-time years of GP practice that would be lost. If the fallow years are years spent out of the UK or out of medical work, then the years would be entirely lost to the NHS.

3. **Model A or Model B?:** Deciding between Model A and Model B is not easy, and there is little in these particular data to provide a compelling answer one way or the other. The 2008 to 2010 cohorts suggest Model B, whereas 2012 and to a lesser extent 2011 cohorts lean more towards Model A, which might indicate a trend in that direction. Historically, the relative stability of preferences of UK doctors for GP would suggest that Model B may be correct. However, a previous 'step-change' in the popularity of GP occurred between about 1987 and 1991 (see the previous section), and there is no clear explanation for that earlier seismic change in the popularity of general practice, and hence it might be repeated.

A question which cannot be answered from Figures 4a and 4b is the extent to which the decline in applications to GP in FY2 is **specific to general practice, or is more general and across the specialties.** The aggregated data published generally on specialty selection cannot provide answers to such questions. Moreover, since specialty training is typically longer than GP training, looking at LRMP data on the Specialty Register will be less accurate than for GP.

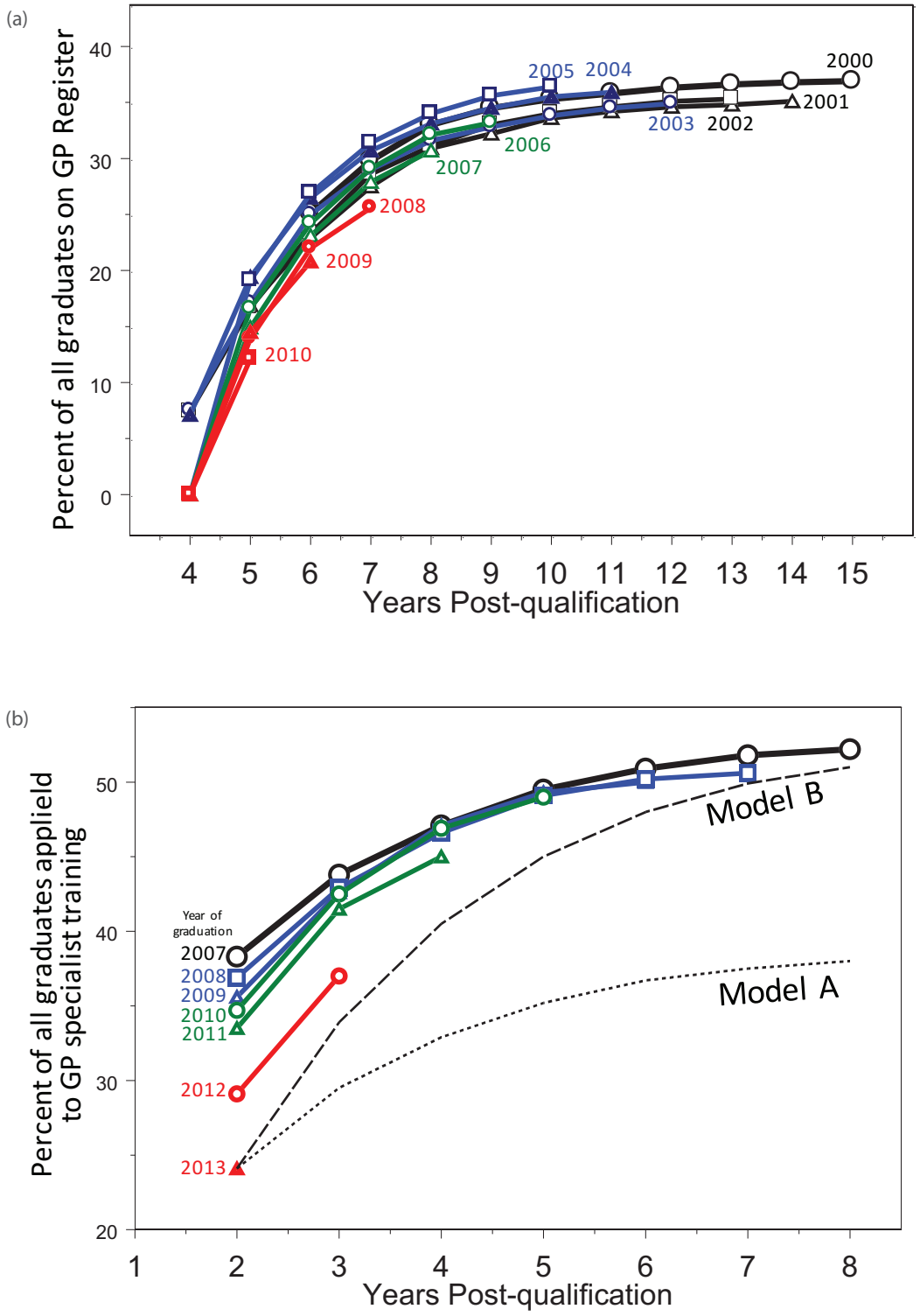
Figure 8.4a suggests that 35% or so of UK graduates in recent years have entered the GP Register. However the data of Figure 8.4b suggests that over 50% of UK doctors have applied for GP training at some time. Even accounting for those applying to GP as a 'back up' choice, the difference of 15 percentage points is large and raises the question of **why such a high proportion of graduates of UK medical schools, whose courses are overseen by the GMC, are unsuitable to enter or complete GP training.** In 2015, for example, of 3452 UK graduate applicants for GP training in round 1, 381 (11%) were either not shortlisted or not considered appointable at Stage 3 (see Chapter 10 for further details). There are multiple possible answers, with the standards set for medical school assessments being one of them, the stringency (or a lack of reliability) of the criteria used for selection into GP training being another.

8.4 ATTITUDES AND INTENTIONS TOWARDS A CAREER IN GENERAL PRACTICE BEFORE AND DURING MEDICAL SCHOOL

A commonplace in social psychology is that the best predictor of behaviour is intentions, and the best predictors of intentions are attitudes. Stated intentions of medical students to enter general practice have therefore been looked at as indicators of future workforce numbers. That process was started in the UK a half-century ago with the studies carried out by ASME (Association for the Study of Medical Education) in 1966 for the 1968 Royal Commission on Medical Education, the Todd Report (Royal Commission, 1968). The results of this and other studies are summarised below. Key questions concern not only the proportions of students and doctors interested in careers in general practice, but also the validity of those stated intentions for predicting actual careers in general practice.

In this brief review of some of the results from the various studies, we will start with the few studies of schoolchildren and teenagers before application to medical school, and we will then look at the Parkhouse and MCRG studies of new

» *Figure 8.4: (a) The proportion of UK graduates for the various cohorts from 2000 to 2010 who were on the GP Register by 4, 5, 6, etc. years after graduating. Percentage of UK graduates a) on GP register and b) applying for GP training; (b) The proportion of UK graduates for the most recent cohorts, from 2007 to 2013, who had applied to GP training by 1,2,3, etc. years after qualifying. For descriptions of Model A and Model B see the text.*



» *Table 8.1: Flows in and out of GP for trainees in ARCP from 2010 to 2014.*

	Anaes.	EM	GP	Med	O&G	Oph	Paed.	Path.	Psych.	PubH.	Radiol.	Surg.	ACCS	BBT	Total
Anaesthetics	5102	31	125	77	0	0	5	3	3	1	3	16	8	0	5374
Emerg. Med	24	1148	34	55	0	0	2	0	1	1	0	4	12	0	1281
GP	**10**	**18**	**19723**	**54**	**11**	**3**	**18**	**8**	**52**	**6**	**9**	**9**	**1**	**0**	**19922**
Medicine	95	24	598	13640	6	5	12	76	24	5	62	8	5	0	14560
O & G	1	2	115	6	2748	0	0	1	5	2	2	0	0	0	2882
Ophth.	0	0	6	2	0	914	0	0	0	0	0	0	0	0	922
Paediatrics	9	1	149	52	0	0	4653	2	8	2	5	4	1	0	4886
Pathology	0	0	7	36	1	0	0	977	1	2	0	0	0	0	1024
Psychiatry	0	0	102	13	0	0	3	2	4533	0	3	0	1	0	4657
Public Health	0	0	9	4	0	0	0	1	1	307	1	0	0	0	323
Radiology	0	0	12	8	0	1	0	0	1	0	1491	0	1	0	1514
Surgery	20	7	285	22	9	5	10	16	3	1	90	7991	3	0	8462
ACCS	446	443	71	128	0	0	3	2	1	0	4	2	163	0	1263
BBT	0	0	0	0	0	0	0	0	0	0	0	0	0	40	40
Total	5707	1674	21236	14097	2775	928	4706	1088	4633	327	1670	8034	195	40	67110

graduates. Finally we consider the three cohort studies which have detailed longitudinal data, which will be described here only in terms of the proportions showing an interest in general practice, but in the following section will be analysed longitudinally.

The various surveys have used different response measures. The Todd Report questionnaires asked for a 'first choice of career'[7]. The Parkhouse/MCRG questionnaires have asked intentionally vague and open-ended questions of the form, "What is your career choice? List up to three choices in order of preference. Bracket together any choices that are equal. You can be as specific or as general as you like" (Parkhouse, 1991) (p.315), and whilst allowing flexibility, the format also has inevitable problems in the coding and analysis. The Cohort studies with which one of us (ICM) has been involved have asked for five-point ratings of each of a range of specialties (and for such data the totals of 'definite intentions' may sum to more than 100%), and the results have also been supplemented by a first choice request as used by the Todd Report.

We begin with two studies before medical school, as well as some data on the age at which medical school applicants decided to study medicine.

8.4.1 Pre-medical school:

Age 11

The 2013 STARS study (McManus et al., 2015), which was a broadly representative study of 10- and 11-year old-children in ten state secondary schools in Southern England, took place in three stages. In each stage, children were asked an open-ended question about what they would like to do when they grew up, and about 10% spontaneously mentioned 'doctor'. In the third stage, specific information was also asked about 33 different careers and jobs, including GP, Hospital doctor, Surgeon and Psychiatrist. Table 8.2 shows that while about 10% of 10-11-year old potential-medics would like to be a GP, many more would prefer to be a Hospital Doctor, suggesting that Hospital Practice is preferred to GP even at quite a young age. It is also the case that 66% of the children wouldn't like a career as a GP at all[8].

The Medlink studies at age 17

The Medlink studies recruited young adults attending large sixth form conferences for students thinking of applying to medical school (McManus et al., 2006)[9]. The questionnaire asked for ratings of 20 different medical specialties, with the responses for GP shown in Table 8.3. Although the results for specialties other than GP are not shown, many are more popular than GP.

Studies of medical students and junior doctors

The first modern study of medical students was carried out by ASME for the Royal Commission on Medical Education of 1968 'The Todd Report' (Royal Commission, 1968), and it set the scene for all future studies. The survey looked at first and final year students in 1966, and built on an earlier study of 1961. Questions asked about first career choices, both broadly and then specifically[10], with the results for general practice shown in Table 8.4.

[7] They also asked, "Which of these is the **least** attractive?", but presumably the question failed miserably as there is no further discussion of it. What it probably shows is that there is a need to rank or to rate all of the specialties, but while most attractive is informative, least attractive is not.

[8] The study also found no relation between Cognitive Ability Test scores and wanting to be a doctor, suggesting that the majority of those with aspirations to be doctors are unlikely to achieve that goal given the high entry requirements for medical courses.

[9] Data from the 2002 study were not published, but are included here.

» *Table 8.2: Percentage of 10-11-year old children in the 2013 STARS study who, in open-ended questions, indicated that they would like to be a doctor, and their interest in four different medical specialties (which were part of a structured set of 33 different careers and jobs).*

	Study	Year studied	N	Job	Wouldn't like it at all	Not sure	Would like it a lot
11-year olds	STARS*	2013	1301	GP	66.3%	23.4%	10.2%
"	"	"	"	Hospital doctor	50.3%	23.2%	26.5%
"	"	"	"	Surgeon	66.0%	19.1%	14.8%

* "Here is a list of different jobs that people do. Say how much you might like to do each one by ticking one of the smiley faces next to it. Tick ☺ if you would like it a lot, ☹ if you wouldn't like it at all, and 😐 if you are not sure".

» *Table 8.3: Attractiveness of General Practice as a career for two large groups of young adults attending a sixth-form conference on medical careers.*

	Study	Year studied	N	Job	Definitely NOT	Not very attractive	Very At-tractive	Definite intention
17-year olds	Medlink 2002*	2002	1395	GP	9.4%	23.0%	54.8%	12.8%
"	Medlink 2003	2003	2817	GP	13.5%	30.8%	42.0%	13.7%

* "Below is a list of possible medical careers […]. How attractive do you find each of them?"

» *Table 8.4: Results from the 1968 Todd Report indicating the attractiveness of GP as a career choice.*

	Study	Year studied	N	No interest	Broad interest in GP*	GP as first choice**
First Year	Todd Report	1966	2371	79.4%	4.3%	16.3%
Final Year	Todd Report	1966	1788	71.4%	5.1%	23.5%

* "What is your first preference among the following types of medical work?" [6 options]. ** "Below is a more detailed list of specialties in which a medical career can be pursued. … What at present is your first choice?". Per cent with Broad interest calculated as answer to first question minus answers to second question (i.e. Broad Interest – First Choice).

[10] The question about broader preferences was also asked in the Cohort Studies of 1981, 1986 and 1991, but they are not described here.

8.4.2 Studies of UK medical graduates:

The Parkhouse/ UK Medical Careers Research Group studies

James Parkhouse carried out his first national study of the career preferences of UK medical graduates in 1975, studying doctors who had graduated in 1974 at the end of their first PRHO year[11]. Subsequent studies over varying time intervals followed, firstly at yearly intervals, and then three-yearly and other intervals. The study migrated between universities, and is now based at Oxford as the UK Medical Careers Research Group (MCRG), with many studies having been published under the leadership of Michael Goldacre and Trevor Lambert (www.uhce.ox.ac.uk/ukmcrg/publications.php). The broad structure of the surveys has been relatively constant, asking doctors to indicate up to three careers, in order of preference, but with the use of 'bracketing' if required. Doctors themselves write in the names of the careers, and these may be very specific or very general. A problem in interpreting and comparing results is that GP may be first choice, second choice or third choice or it may be a tied choice. Although there are many publications, not all methods of how these choices have been used to generate the results are provided for all cohorts in all forms. Table 8.5 summarises some findings from a number of publications and reports over the years, and gives a flavour of the results. A major strength of the data is that doctors are followed up over many years (and that is particularly well seen in the paper of Lambert et al (2002) (Lambert et al., 2002)). A weakness is the lack of background information prior to the PRHO/F1 year[12], the lack of other detailed information after qualification, and the relative lack of correlational cross-sectional and longitudinal causal modelling. Nevertheless, the MCRG provides some of the most important data into preferences for general practice as a career across forty years.

Table 8.5 summarises statistics for all of graduation cohorts which have been studied, mostly looking at the first year Post Qualification (Yr 1 PQ), which for much of the life-time of the studies was at the end of the PRHO year, but subsequently was at the end of the first Foundation Year (FY1). Somewhat different measures are used at different times, but the broad picture is relatively clear. The proportion of Year 1 first choices for GP is higher for 1974 to 1983 graduates than it is for graduates from 1993 onwards. That mirrors the proportions of doctors on the GP Register (see figure 8.2a), but sadly there was a hiatus in first year data collection between the 1983 and 1993 graduates, just when, as was shown in figure 8.2a, the substantial drop in GP Register entrants occurred.

The 1981, 1986 and 1991 Cohort Studies

One of us (ICM), in conjunction with the late Professor Peter Richards, then Dean of St. Mary's Hospital Medical School, London, initiated three cohort studies of medical student selection and training, which are still being followed up today, the first of which looked at applicants in 1980 for admission in 1981[13], and the other larger ones looked at applicants in 1985 and 1990 for admission in 1986 and 1991 (McManus, 1983, McManus et al., 2013a, McManus et al., 2011, McManus et al., 1999, McManus et al., 1995, McManus et al., 2013b); see http://www.ucl.ac.uk/medical-education/medical-education-studies for further details. The studies are properly longitudinal and they complement the Oxford MCRG studies in that instead of having a little data across a very large number of individuals over many years, they have much deeper and broader information about a somewhat smaller, although still large, number of students and doctors in just three cohorts but with a much wider range of longitudinal data, and therefore causal modelling of change and predictions of outcomes can be carried out. In particular there is detailed information on doctors' career perceptions before they arrived at medical school. The most recent of the cohorts entered medical school in 1991, graduated in 1996/7, and therefore have been graduates for 18 years or so, and hence may not be equivalent to current graduates. That is a problem if one believes that everything has significantly changed in medical education since 1997, and there is no overlap in the experience of recent and earlier graduates. That is probably unlikely (and an alternative rhetorical position would argue that nothing of serious substance in medical education and practice has changed since the time of Hippocrates). To reject the data from the 1996/7 graduates

[11] There had been earlier studies in just Manchester and Sheffield from 1971 to 1973 which set the scene for the national surveys Parkhouse, J. (1991) **Doctors' careers: aims and experiences of medical graduates.** London: Routledge.(p.9) Parkhouse, J. and Howard, M. (1978) A follow-up of the career preferences of Manchester and Sheffield graduates of 1972 and 1973. Medical Education, 12: 377-381.

[12] The PRHO (pre-registration house officer) year was a compulsory year in hospital practice, carried out after a doctor had graduated, in which they usually had two six-month appointments as a house physician and a house surgeon. These were later replaced by the Foundation Programme, with the first year (F1) being pre-(full)registration, and the second year (F2) after full registration had been granted.

» *Table 8.5: Summary of the popularity of General Practice as a career in the various studies carried out by Parkhouse, and then the UK Medical Careers Research Group, into career choice at the end of the first year of clinical practice.*

Year PQ= Post-Qualificaiton	Study	Graduation cohort	Year studied	N	GP not chosen	GP third choice	GP second choice	GP first choice[a]
Yr 1 PQ	(Parkhouse, 1991)	1974	1975	2,022	30.5%	16.4%	19.7%	33.4%
Yr 1 PQ	(Parkhouse, 1991)	1975	1976	2,217	32.0%	13.2%	19.3%	35.5%
Yr 1 PQ	(Parkhouse, 1991)	1976	1977	2,536	37.1%	12.6%	18.0%	32.3%
Yr 1 PQ	(Parkhouse, 1991)	1977	1978	2,666	35.4%	13.8%	17.5%	33.3%
Yr 1 PQ	(Parkhouse, 1991)	1978	1979	2,632	35.4%	12.9%	17.2%	34.5%
Yr 1 PQ	(Parkhouse, 1991)	1979	1980	2,778	33.8%	11.4%	18.0%	36.8%
Yr 1 PQ	(Parkhouse, 1991)	1980	1981	2,855	39.3%	9.6%	14.3%	36.8%
Yr 1 PQ	(Parkhouse, 1991)	1983	1984	3,368	33.5%	7.5%	14.3%	44.7%
Yr 1 PQ	(Lambert et al., 2002)	1988	1989c	-				-
Yr 1 PQ	(Lambert et al., 2002)	1993	1994	2,621				25.8%
Yr 1 PQ	(Lambert et al., 2002)	1996	1997	2,926				20.0%
Yr 1 PQ	(Lambert et al., 2002)	1999	2000	2,727				25.0%
Yr 1 PQ	(Lambert et al., 2002)	2000[b]	2001	2,975				25.5%
					GP not chosen	GP Mentioned not first	GP tied first	GP untied first
Yr 1 PQ	(Lambert and Goldacre, 2011)	2000b	2001	2,978	53.5%	16.9%	7.4%	22.2%
Yr 1 PQ	(Lambert and Goldacre, 2011)	2002	2003	2,778	56.6%	17.1%	6.1%	20.2%
Yr 1 PQ	(Lambert and Goldacre, 2011)	2005	2006	3,128	47.4%	19.6%	9.8%	23.2%
Yr 1 PQ	(Lambert and Goldacre, 2011)	2008	2009	3,302	50.5%	18.5%	9.7%	21.3%
Yr 1 PQ	(Lambert and Goldacre, 2011)	2009	2010	2,917	50.1%	19.3%	10.2%	20.4%
Yr3 PQ	(Lambert and Goldacre, 2011)	2000	2003	2,968	57.7%	11.5%	2.9%	27.9%
Yr3 PQ	(Lambert and Goldacre, 2011)	2002	2005	2,748	58.9%	11.2%	3.8%	26.1%
Yr3 PQ	(Lambert and Goldacre, 2011)	2005	2008	2,709	55.4%	6.3%	3.2%	35.1%

[a] Parkhouse (Parkhouse, 1991) gives second and third career choices but these do not seem to be strictly comparable to the categories used later by the Oxford group.

[b] Note that data for 2000 have different percentages, probably due to a different way of handling ties. c This cohort was not studied immediately after graduation.

as of any relevance would require a) clear evidence of major substantive structural change in the nature of medical students and medical training, and b) the discarding of the experiences of anyone discussing medical education who graduated before 1996. In the absence of better information (and there is an urgent need for better information), these cohorts are likely to be informative about general processes in medical training and career choice. As was shown earlier, the patterns of GP choice seem remarkably similar across the decades.

The three cohort studies began by looking at applicants for admission in 1981, 1986 and 1991, but for consistency with the MRCG and other studies, these could also be called the 1986/7, 1991/2 and 1996/7 cohorts, since this was when most doctors graduated. The location of the three cohorts is shown on Figures 8.1a and 8.2a above. Of interest is that the 1981 cohort just precedes the drop in numbers entering the GP Register which occurred from about that year onwards.

The questions in the three cohort studies were based on the career preference questionnaires developed for the Todd Report (Royal Commission, 1968), although as well as the first choice being asked for, each of the various specialities was rated on a five point scale. That allows perceptions of all specialties to be assessed (and elsewhere these data have been used to generate a map of specialties using multidimensional scaling (Petrides and McManus, 2004)). Asking only for the first or the first three preferences means that most specialties are simply not available for discussion and with a few specialties dominating the first choices the analyses that can be carried out are limited.

For all three cohorts, data are available at application (i.e. within a few weeks of submitting the UCCA/UCAS form), and in the final year of the medical course[14]. In the 1991 cohort, data were also available for a subset of students in the third year of the course (i.e. the first clinical year, not an intercalated degree), and for all doctors in the study at the end of the PRHO year (Yr 1 PQ). The 1991 cohort was also followed up in 2002, and all three cohorts were followed up in 2009, when almost all doctors would have been in specialties, and therefore the question was couched in terms of, "… if you were starting a medical career again, how attractive would you find each of these … specialties?".

Table 8.6 summarises the results, both on the five point scale of attractiveness of general practice, and the percentage of first choices for GP. Of particular interest is that the proportion of GP first choices for the 1991 cohort of 1996/7 graduates at the end of the PRHO year (Yr 1 PQ) is 22.0%, which is very similar to the 20.0% of GP first choices in the MCRG data for the 1996 graduates in 1997. The Cohort Studies and the MCRG studies therefore appear to be measuring at equivalent levels.

The cohort studies also provide transverse comparisons across the cohort years and longitudinal comparisons by year of training. The proportions of applicants with different levels of interest in GP are similar across all three cohorts. However by the final year, each successive cohort reports GP as less attractive than the previous cohort. For the 1991 cohort the effect is such that the attractiveness of GP had fallen during medical school, compared to the 1981 cohort when it increased and the 1986 cohort when it was relatively stable. The reasons for this are unclear and may reflect changes within the medical school or external changes in the NHS or something else.

What is interesting, however, is the sustained increase in the attractiveness of GP for the 1991 cohort following graduation and that this cohort has the highest proportion on the GP Register by 2012 (almost 40%, compared to just under one-third for the two earlier cohorts). Of course, a doctor may train as a GP because they were unable to obtain a training post in their preferred specialty, but these results suggest that to a certain extent **career preferences are malleable both during medical school and following graduation.**

The cohort studies allow more detailed studies of the relation between the various measures of interest in GP at various stages and the gold standard outcome measure, which is being on the GP Register of the GMC's LRMP.

[13] As with all larger studies, it had been preceded by a series of smaller studies from 1974 onwards which allowed testing of the methods and the questionnaires, with some results published Cruickshank, J.K. and McManus, I.C. (1976) Getting into medicine. **New Society**, 35: 112-113, McManus, I.C. (1976) Archive for research data [letter]. **Lancet**, i: 1188-1188, McManus, I.C. (1985) **Medical Students: Origins, selection, attitudes and culture**. University of London: MD thesis (see http://www.ucl.ac.uk/medical-education/publications/md).

[14] This was the final year for individual students, and therefore could be five, six, seven or more years after medical school entry, which might also have been deferred either by choice or by retaking examinations, a feature which complicates most longitudinal studies.

» *Table 8.6: Summary of interest in General Practice as a career in the three cohort studies.*

PQ= Post Qualification	Study	Grad'n cohort	Year studied	N	Definitely NOT[a]	Not very attractivea	Moderately attractive[a]	Very Attractive[a]	Definite intention[a]	GP 1st Choice[b]
Applicants[a,b]	1981 Cohort	1986/7	1980	1,107	6.4%	15.4%	32.6%	38.2%	7.9%	-
	1986 Cohort	1991/2	1985	1,983	7.0%	15.4%	30.3%	38.8%	8.5%	23.7%
	1991 Cohort	1996/7	1990	5,222	8.4%	15.3%	31.5%	36.5%	8.3%	17.6%
3rd Year[a,b]	1991 Cohort	1996/7	1993	632	10.6%	14.7%	37.7%	33.4%	3.6%	14.2%
Final Year[a,b]	1981 Cohort	1986/7	1986/7	335	8.1%	9.6%	23.3%	44.8%	14.3%	36.5%
	1986 Cohort	1991/2	1991/2	375	12.0%	11.5%	28.3%	40.3%	8.0%	20.9%
	1991 Cohort	1996/7	1996/7	1,608	15.2%	16.0%	31.3%	31.0%	6.3%	20.6%
PRHO (yr1 PQ)[a,b]	1991 Cohort	1996/7	1997/8	1,395	20.9%	17.5%	28.2%	20.9%	12.5%	22.9%
					Very unattractive[c]	Fairly unattractive[c]	Fairly attractivec	Very attractive[c]	Extremely attractive[c]	
Yr 6 PQ	1991 Cohort	1996/7	2002	1,629	17.1%	21.3%	25.4%	17.9%	18.4%	-
Yr 13 PQ	1991 Cohort	1996/7	2009	1,689	9.4%	18.9%	29.5%	27.6%	27.0%	-
Yr 18 PQ	1986 Cohort	1991/2	2009	490	12.9%	19.4%	31.6%	24.7%	11.4%	-
Yr 23 PQ	1981 Cohort	1986/7	2009	272	15.8%	18.8%	27.9%	27.2%	10.3%	-
									On GP register	
GP Register	1981 Cohort	1986/7	2012	510	67.1%	-	-	-	32.9%	-
	1986 Cohort	1991/2	2012	919	69.9%	-	-	-	30.1%	-
	1991 Cohort	1996/7	2012	2,762	60.4%	-	-	-	39.6%	-

[a] "Below is a detailed list of specialties in which are medical career can be pursued. Please indicate your attitude towards the specialties as a possible career."

[b] "If you were forced to choose just one of the above categories as your future career, which would it be?"

[c] "Irrespective of your current career post, if you were starting a medical career again, how attractive would you find each of these eleven broad areas of medical practice as a specialty?" [question slightly different for 2002].

The relationship of first choices and the five-point scale of interest in GP:

There is a good correlation at all times of measurement between interest in GP on the five-point scale and putting GP as first choice of career (Table 8.7). However not all of those with 'Definite' intentions put GP as their first choice, and a few putting it as their first choice had weak interests in GP. Perhaps most crucial is that a third to a half of those putting GP as their first choice saw it as only "very attractive" rather than having a definite intention towards it. That latter means that the five point scale is overall a better predictor than first choice, and in part can account for differences between studies using the two different methods of measurement.

» *Table 8.7: Relationship of interest in a career in General Practice in the three Cohort Studies on a five-point scale, and that the stated intention that General Practice is the 'first choice'.*

Interest in General Practice	Applicants		Final Year		PRHO (91 cohort only)	
	N	First Choice	N	First Choice	N	First Choice
Definitely NOT	551	1.5%	304	3.0%	289	0%
Not very attractive	1,047	2.3%	319	1.6%	240	0.4%
Moderately attractive	2,104	3.3%	654	8.0%	380	3.7%
Very attractive	2,498	32.9%	758	40.5%	283	49.1%
Definite intention	581	70.1%	172	81.4%	173	89.6%
N	**6,781**		**2,207**		**1,365**	

ROC curves for predicting being on the GP Register from previously-expressed levels of interest in GP:

A key question is the extent to which previously-expressed levels of interest in GP subsequently predict being on the GP Register, i.e. are statements of intent valid? Figure 8.5 shows ROC curves for the prediction of being on the GP Register in relation to the five-point scale at Application, Year 3, Final Year and PRHO year (and for completeness, the relationship of attractiveness of GP as a future career for those in 2002 who mostly have chosen their careers). Areas under the ROC curves are shown in Table 8.8.

Of those students who at entry to medical school say they have definite intentions to become a GP, only around half become GPs; equally a quarter of applicants who are definitely NOT interested in GP end up on the GP register. As can be seen in Table 8.9, however, there is undoubtedly some predictive value. Prediction increases through the third year, final year and then the end of the first post-qualification year, as can also been seen as the ROC curves shift outwards over time. Nevertheless, in comparison with the attractiveness of GP at the end of the fifth or sixth year post-qualification, when most choices will have been made, earlier career preferences are not strong predictors of becoming a GP. That has important implications for workforce planning, as it implies that **in the past there were many GPs who made their mind up about entering general practice only a number of years after graduation** (Lambert et al., 2013). That has important implications for the current specialty selection programmes where doctors typically are making career choices in FY2 or the year after, although Goldacre and colleagues (Goldacre et al., 2010) note how many still 'deviate' from this path:

> *"Now, as in the past, doctors' medical career trajectories do not invariably take a straight and relentless course from qualification through specialist training to consultant or general practitioner principal posts" Michael Goldacre (2010).*

» *Figure 8.5:*

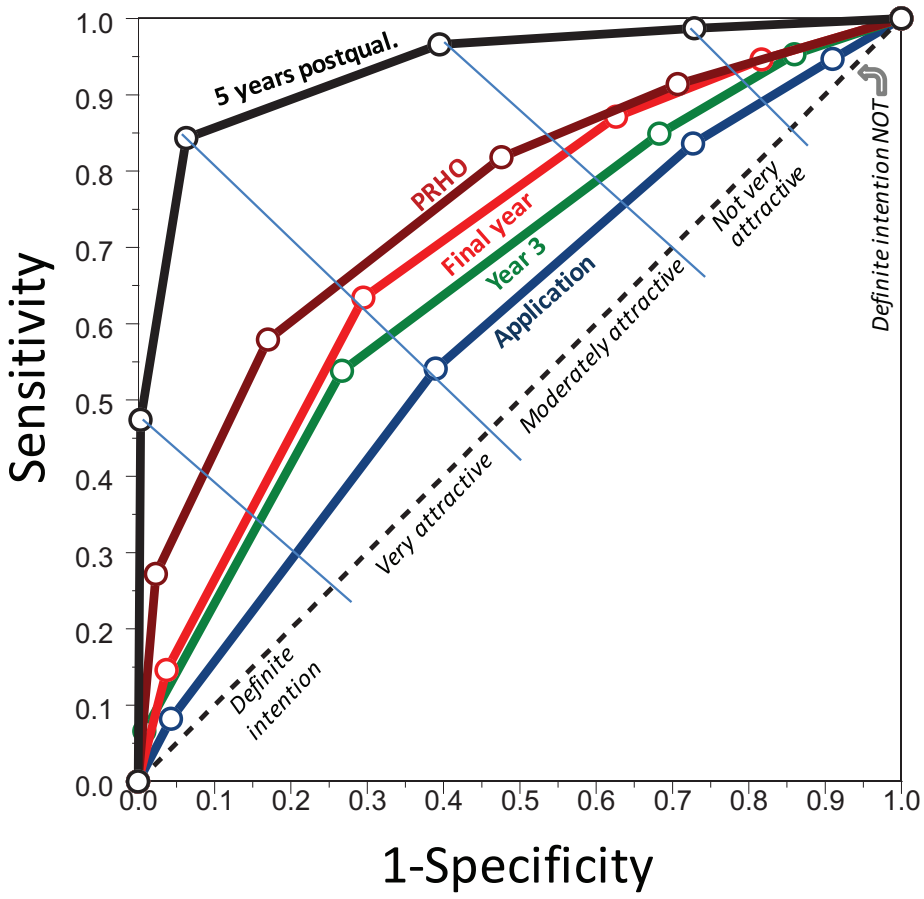


Figure legend: The y-axis plots sensitivity: the true positive rate, or the proportion of those on the GP register who stated each level of intention or a higher one. For example, of those who actually entered the GP register, 27% gave a rating of 'definite intention' towards GP in their PRHO year, and around 58% gave a rating of either 'definite intention' or 'very attractive'. The x-axis plots 1-specificity: the false positive rate, or the proportion of those not on the register who stated each level of intention or a higher one. For example, For example, of those not on the register, around 2% gave a rating of 'definite intention' towards GP in their PRHO year, and around 18% gave a rating of either 'definite intention' or 'very attractive'. The predictive power of the 'test' can be estimated by the area under the ROC curve: an area of 0.5 (equivalent to the dotted diagonal line through the middle of the plot) suggests a completely uninformative test (i.e. equivalent to chance), while an area of 1 (a right-angled line through the points (0,0), (0,1) and (1,1)) suggests a perfect test.

» *Table 8.8: Area under ROC curve for expressed interest in General Practice on both a five-point scale, and eventually ending up on the GP Register.*

Area under ROC curve (0.5 = chance)	Five-point scale
Applicants	.593
Year 3	.663
Final Year	.704
PRHO (Yr 1 PQ)	.760
Yr 5 PQ (2002; 91 cohort only)	.936

8.5 THE FOUNDATION YEAR: GP EXPERIENCE AND CAREER DESTINATIONS

8.5.1 Career intentions during the Foundation Programme

The Foundation Programme carries out useful surveys of doctors at the end of their Foundation Training. In the various surveys of career destinations at the end of F2 (U. K. Foundation Programme Office, 2012b, 2013a, 2014a), questions were asked retrospectively about intentions at the beginning of F1. Table 8.10 shows the percentages of UK graduates intending to enter GP, the percentage who did, and the percentage reporting that they had changed their career intention.

Particularly striking is that:

1. About a quarter of doctors at the beginning of Foundation said they intended to become GPs, and at the end of GP about a quarter had entered GP specialty training, but,

2. Over a third of Foundation Doctors changed their career intentions during Foundation, and

3. Only about 60% of Foundation doctors entered specialty training at the end of Foundation. Nevertheless,

4. Of those entering Specialty Training, about 35-36% entered GP specialty training.

Overall therefore these figures are remarkably like those from the various studies described above, and in the cohort described below, with about 35% of those making choices entering GP, but there being **much fluidity of career intentions in the years immediately after qualification.**

8.5.2 Experience of General Practice in the Foundation Programme.

Table 8.11 shows the proportion of Foundation Doctors experiencing a GP rotation during F1 and F2. F1 doctors cannot prescribe and therefore cannot strictly have GP rotations, although a tiny percentage did during 2011, 12 and 13. A median of 42.6% of F2 doctors had experienced a General Practice rotation, with the proportion hardly changing from 2009 to 2014.

Although over forty percent of F2 doctors experienced general practice, only one third of those doctors would have experienced GP in the first third of F2, and hence **only about 14% could have been influenced by that experience before applying for specialty training.** If, as the questionnaire study suggests, it is the experience of working in GP which particularly influences applying for specialty training in GP, then the vast majority of F2 applicants to specialty training will not have experienced working in GP[15].

The Round 1 questionnaire (see Chapter 9), in May 2015, would have included F2 doctors who had been in GP rotations from August-November and December-March, and some would have been doing their April-July rotation. 32% of those still in F2 reported in the questionnaire that the experience of working in GP had influenced them towards applying for GP, which is compatible with the various numbers.

A major absence from official data is a record of which doctors did which specialties and in which order during each of the rotations of F1 and F2. Without that it is difficult to assess the extent to which those rotations had influenced specialty choice. Likewise, in the absence of proper prospective information from before entering foundation, it is difficult to know to what extent choice of foundation rotation is influenced by a desire to become a GP and to what extent the desire to become a GP is causally influenced by experience in the rotation[16].

[15] A small proportion of foundation doctors will also have experienced 'taster' sessions in GP, although it is difficult to have a precise idea of how frequent is that.

[16] In retrospect it would have been useful to have asked the questionnaire respondents to say what rotations they had done and in what order, and that question should be asked in future questionnaires. Likewise a question on the interest in GP at the end of undergraduate training would have been useful. It would also of course have been useful if anyone had been collecting that information on an official basis.

» *Table 8.9: Percentage of doctors ending up on the GP Register in relation to stated interest in a career in General Practice at different career stages in the three cohort stud-*

Interest in General Practice	Applicants		Third Year		Final Year		PRHO (91 cohort only)		Attractiveness of GP if starting medical career again	2002	
	N	% on GP Register	N	% on GP Register	N	% on GP Register	N	% on GP Register		N	% on GP Register
Definitely NOT	289	25.6%	61	16.4%	292	15.1%	277	16.6%	Very unattractive	272	2.9%
Not very attractive	597	26.1%	86	25.6%	320	19.1%	234	22.2%	Fairly unattractive	339	3.8%
Moderately attractive	1,223	33.9%	217	30.4%	642	30.2%	369	34.7%	Fairly attractive	400	19.0%
Very attractive	1,476	43.6%	194	51.5%	747	53.3%	281	58.7%	Very attractive	285	79.6%
Definite intention	218	52.8%	17	82.4%	169	70.4%	164	89.0%	Extremely attractive	295	99.0%
N	3,803	36.9%	575	36.9%	2,170	37.6%	1,325	40.5%	N	1,591	38.7%
GP not first choice	2,589	34.4%	489	31.5%	1,950	28.9%	1,026	28.1%		-	-
GP first Choice	621	47.8%	80	68.8%	584	59.1%	301	83.7%		-	-
N	3,210	37.0%	569	36.7%	2,534	35.9%	1,327	40.7%		-	-

» *Table 8.10: Percentages of doctors intending to enter GP at various stages during Foundation training.*

	2012	2013	2014
Beginning of Foundation Training			
Intention to enter GP*	-	23.7%	23.4%
End of Foundation training			
Entered Specialty training*	67.8%	64.7%	58.7%
Entered GP training*	24.4%	23.7%	20.8%
% entering GP training**	36.1%	36.6%	35.4%
Change during Foundation Training			
Per cent changing career intention during Foundation*	-	32.6%	31.5%

* Percent of all doctors in foundation

** Percent of doctors entering specialty training

» *Table 8.11: Percentages of doctors reported as having a GP rotation during F1 and F2.*

Year	Percentage of doctors experiencing a GP rotation during F2	Percentage of doctors experiencing a GP rotation during F1
2009	48.9%[17]	0%
2010	41.4%	0%
2011	42.0%[18]	0.1%
2012	43.8%	0.1%
2013	40.7%	0.1%
2014	43.3%	0%

8.5.3 Destinations of F2 doctors from the Foundation Programme.

Information on destinations of F2 doctors is in the UKFPO Career Destination Reports (U. K. Foundation Programme Office, 2012b, 2013a, 2014a), and, for 2010, is in the Annual Report (U. K. Foundation Programme Office, 2010). A summary is in Table 8.12, below.

[17] The method of calculation was changed from 2009 and 2010 to 2011 onwards, and we quote the latter method here, which was reported in Table 8.13 of the 2011 report (and is simply three times the values in the 2009 and 2010 reports).

[18] In the main report for 2011 U. K. Foundation Programme Office (2011) **Foundation Programme Annual Report 2011: UK Summary.** Cardiff: UK Foundation Programme Office (available at http://www.foundationprogramme.nhs.uk/download.asp?file=Foundation_Programme_Annual_Report_Nov11_FINAL.pdf)., it is said that the proportion of F" doctors rotating through GP is 42.0% (Table 8.11, p.11)., and that same figure is quoted in Table 8.13 (p.12). However Table 8.13 of the reports for 2012 U. K. Foundation Programme Office (2012a) **Foundation Programme Annual Report 2012: UK Summary.** Cardiff: UK Foundation Programme Office (available athttp://www.foundationprogramme.nhs.uk/download.asp?file=UK_Foundation_Programme_Annual_Report_2012_FINAL.pdf)., 2012 U. K. Foundation Programme Office (2013b) **Foundation Programme Annual Report 2013: UK Summary.** Cardiff: UK Foundation Programme Office (available at http://www.foundationprogramme.nhs.uk/download.asp?file=UK_Foundation_Programme_Annual_Report_2013_FINAL.pdf. and 2014 U. K. Foundation Programme Office (2014b) **Foundation Programme Annual Report 2014: UK Summary.** Cardiff: UK Foundation Programme Office (available at http://www.foundationprogramme.nhs.uk/download.asp?file=FP_Annual_Report_2014_-_FINAL_Nov_2014.pdf). reports the value as 35.6%. Here we have used the value 42.0% since it is more compatible with the years before and after, and is quoted as that value in the main table of the 2011 report.

» *Table 8.12: Destinations of doctors at the end of Foundation training.*

PQ= Post Qualification	2010	2011	2012	2013	2014	2015
Specialty training in UK - run-through training programme	42.9%	34.0%	33.5%	29.9%	29.5%	24.0%
Specialty training in UK - core training programme	37.4%	34.0%	30.5%	29.6%	26.8%	26.0%
Specialty training in UK - academic programme	1.4%	1.5%	1.6%	1.5%	1.6%	1.3%
Specialty training in UK – FTSTA	0.8%	1.1%	0.8%	0.2%	0.2%	0.1%
Specialty training in UK - deferred for higher degree	0.3%	0.1%	0.1%	0.2%	0.1%	0.0%
Specialty training in UK - deferred for statutory reasons	0.3%	0.5%	0.5%	0.5%	0.3%	0.5%
Sub-total for specialty (incl. GP) training in UK	**83.1%**	**71.3%**	**67.0%**	**64.4%**	**58.5%**	**52.0%**
Locum appointment for training (LAT) in UK	0.5%	0.4%	0.7%	0.6%	0.5%	0.5%
Specialty training outside UK	1.6%	0.8%	1.1%	0.6%	0.3%	0.4%
Service appointment in UK	2.1%	2.3%	3.3%	3.5%	5.6%	9.2%
Other appointment outside UK	4.0%	7.4%	6.6%	4.8%	3.9%	0.4%
Still seeking employment as a doctor in the UK	3.4%	6.3%	7.4%	7.6%	8.4%	8.6%
Still seeking employment as a doctor outside the UK	-	3.7%	5.5%	6.5%	5.1%	4.3%
Not practising medicine - taking a career break	**4.9%**	**4.6%**	**6.1%**	**9.4%**	**11.3%**	**13.1%**
Not practising medicine - permanently left profession	0.4%	0.1%	0.2%	0.3%	0.3%	0.3%
Other (e.g. anatomy demonstrator, higher education)	-	3.0%	1.9%	2.3%	6.1%	5.5%
Other (i.e. NOT specialty training in UK)	**16.9%**	**28.7%**	**33.0%**	**35.6%**	**41.5%**	**48.0%**

The table shows several very important features. However the 2010 data must be treated with some care, as two of the later categories are not present in the 2010 data. For 2011 to 2015:

1. The proportion going into Specialty Training, be it run-through training or core training, has declined substantially from 71% in 2010 to 52% in 2015, a decrease of 19%. That is a very large change over five years.

2. The proportion not going into ST has increased concomitantly from 29% to 48%. That change of 19% is the result of several factors[19]. The largest component is for a "career break", which increased from 5% in 2011 (almost identical to that for 2010) to 13% in 2015, a difference of +8.5%. There is also a change of +6.9% in service appointments in the UK, +2.5% of 'Other' (e.g. anatomy demonstrating and higher education), +2.3% for those still seeking employment in the UK and +0.6% for those still seeking employment outside the UK. Those increases total +20.8%, more than the net change of +19.3%, mostly due to differences of -7.0% in those with other appointments outside the UK. Although just under half of the apparent change is explicitly due to career breaks, in practice it could well be that several of the other categories also fit into that group, reflecting a disinclination to go straight into formal training programmes (with more formal exams, etc).

8.6 LONGITUDINAL MODELLING OF AN INTEREST IN A CAREER IN GENERAL PRACTICE

The cohort studies also allow longitudinal modelling, at the level of the individual, of an interest in general practice in medical students. Particular interest concerns sex differences, ethnicity, the influence of a medical background (particularly in GP), academic attainment, the teaching of GP at medical school, and the role of PRHO posts in surgery and medicine. Inevitably all of those measures are intercorrelated to a greater or lesser extent, and therefore causal path modelling is the best way to approach the issue. Essentially the variables in the analysis are put into a causal order, which usually can be in terms of their temporal ordering, and then variables are regressed on measures which are causally prior to them[20].

8.6.1 Causal ordering

The variables can be divided into sets:

Main variables, expressing an interest in a career in General Practice:

1. The four measures on a five-point scale indicating interest at Application, Year 3, Final Year and in the PRHO year (labelled **Interest@App'n, Interest@Y3, Interest@FY and Interest@PR**. These were placed in their temporal order. There were also four measures, **1st@App'n, 1st@Y3, 1st@FY and 1st@PR** derived from each, indicating whether GP was a first choice on those occasions. These four 1st variables were placed immediately after the 5-point scales to which they applied.

2. Postgraduate outcome measures, consisting of **On GP Register**, and attitudes towards GP as a career were they starting medicine again, **Choice Yr6 PQ** and **Choice Yr 13+ PQ**. The latter two have to be at the right hand side of the model, but it was arbitrarily decided to place **On GP Register** between PRHO and Year 6.

Endogenous variables: Nine process variables endogenous to the model. In causal order they are:

1. **Age of deciding to become a doctor.** There are often suggestions that doctors who decide earlier on a medical career are different from those who apply later. The questionnaires asked applicants:

    a) **The age at which they first thought of becoming a doctor.** The median was 12, quartiles = 12 to 14.

[19] Not all of the rounded percentages sum to 100% in the original table.

[20] Path modelling was carried out by multiple regression; all variables being assumed to be manifest rather than latent. Multiple regression was carried out in SPSS, with backwards removal of variables which were not statistically significant. Because of the large sample size and the number of variables being considered, and because very small effects were of little practical significance, the alpha level was set at 0.0001. Missing values were handled using mean substitution which although not optimal is also unlikely to have biased any of the key findings.

b) **The age at which they first definitely decided to become a doctor.** Median = 16, quartiles=14 to 16.

These two variables were placed immediately to the right of the exogenous variables, occurring relatively early in life.

2. **A-level grade.** The mean A-level grade was calculated for candidates who took A-levels. It was placed immediately to the right of the age variables. The median grade amongst applicants was 3.5 (equivalent to BBC/BCC) with an inter-quartile range of ABB to CDD. Note that there was less grade inflation in A-levels when these studies were done and only about 10% of applicants had the maximum score equivalent to a grade of AAA. Amongst entrants, the median grade was equivalent to ABB (inter-quartile range AAB to BCC) with 18% of entrants scoring AAA. A-levels were taken before the candidate completed our questionnaires and therefore put prior to Interest@App'n.

3. **Perceptions of GP teaching at medical school.** In the final year questionnaire the students were asked a series of questions about the teaching of 9 basic medical science subjects and 16 clinical subjects, one of which was general practice. For each they were asked:

   a) "How interesting you found [the subject]" (**GP Interesting**); scored as Very (3: 46.2%), moderately (2: 42.8%) or not interesting (1: 11.1%).

   b) "How difficult you found [the subject]" (**GP Difficult**); scored as Very difficult (3: 1.8%), moderately difficult (2: 43.1%), or not difficult (1: 55.2%).

   c) "How useful you think it will be in your future practice" (**GP Useful**); scored as Very useful (3: 55.7%), moderately useful (2: 38.0%), or not useful (1: 6.3%).

   d) "Whether you think more or less time in the curriculum should be devoted to the subject (**GP More time**); scored as More time (3: 33.0%), Same time (2: 55.1%) or Less Time (1: 11.9%).

   High scores therefore indicate that students found GP teaching to be interesting, useful, difficult and needing more time. These variables were placed between Year 3 and Final Year.

4. **Ratings of PRHO jobs in Medicine and Surgery.** Almost all pre-registration house officers did one job in medicine and one job in surgery. Each of the posts was rated on a 7-point scale (7: Excellent, 6: Very Good, 5: Good, 4: Adequate, 3: Not very good, 2: Poor, 1: Bad), and a mean score for the medicine and the surgery posts calculated. Further details are given elsewhere (McManus et al., 2002). The mean score for medicine jobs (5.65; median 6: inter-quartile range 5 to 6) was higher than for surgery jobs (5.30; median 5.5; inter-quartile range 4.5 to 6).

Exogenous variables: There were seven variables which are strictly exogenous to the model, and so causal relationships between them cannot be analysed, and hence they are notionally to the far left of the model and can influence all other measures. All are shown at the left-hand side, but to avoid confusing the model too much, direct lines are not drawn between them and the variables they influence, but instead small flags indicate their presence. The variables are:

1. **Female.** Women are more positive to most aspects of general practice and therefore it is easier to interpret the model in terms of the variable labelled as female. 47.1% of participants were female.

2. **BME.** This was a flag for a student or doctor being Black or Minority Ethnic. This was given by self-description at application in the 86 and 91 cohorts, and by a combination of self-description at FY and photos at application in the 81 cohort. 30.9% of participants were BME.

3. **Cohort.** The three cohorts of 81, 86 and 91 were included as a linear term.

4. **Medical family.** Applicants were asked in detail about a range of health care professions in their immediate and more extended family. Four variables were derived:

    a) **Doctor Parent.** 17.8% of participants said that one or both parents was a doctor.

    b) **GP Parent.** 6.4% of participants said that one or both parents were GPs.

    c) **Doctor in family.** 37.4% of participants reported that at least one relative was a doctor (parent, grandparent, uncle or aunt, sibling, cousin or spouse).

    d) **GP in family.** 20.7% of participants reported that at least one relative was a GP.

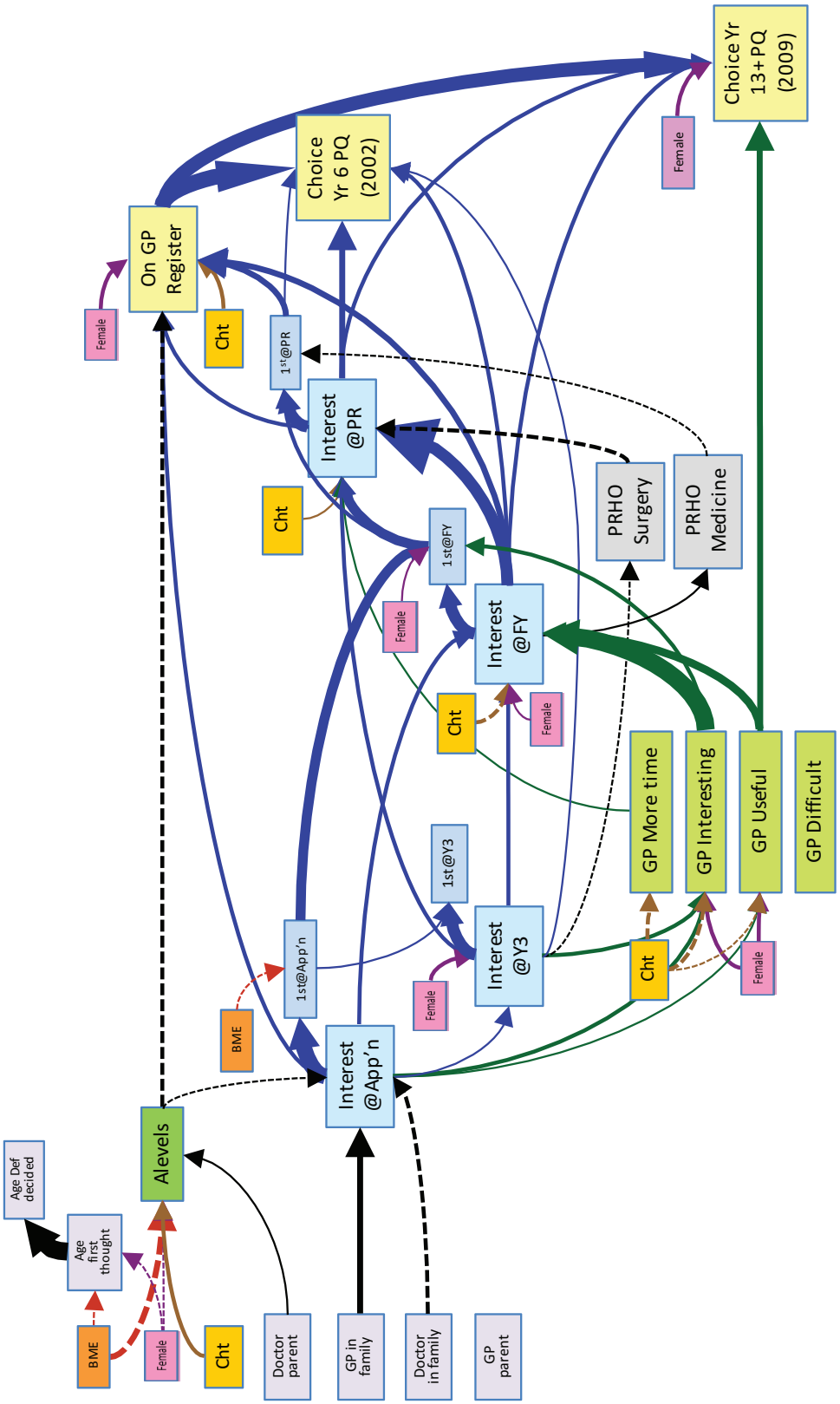
8.6.2 Results

Figure 8.6 shows the fitted path model. At first sight it looks complicated, but it summarises a lot of complex statistical information in a relatively intuitive graphical format, for which most of the statistical subtleties can be ignored. Nevertheless it is worth reading carefully, concentrating at first on the most important components of the model. The arrows indicate 'path coefficients' which are equivalent to the standardised beta coefficients in multiple regression. Positive coefficients are shown as solid lines, and negative coefficients as dashed lines. Lines are only shown if they reach a $p<0.001$ level of significance, and it should be said that the analysis began with a 'saturated model', any arrow which runs from a leftward variable to a more rightward variable being considered for inclusion. Colours are used to help clarify aspects of the model (see below).

1. **The 'backbone' of the model:** Across the middle of the model are four large blue boxes each of which is accompanied by a small blue box. These represent a participant's interest, on a five-point scale, in a career in general practice at Application, Year 3, Final Year and at the end of the PRHO year. The small blue boxes indicate that a participant put GP as their first choice. The blue arrows between the boxes show that later attitudes are caused by earlier attitudes with varying strengths. Interestingly there are direct paths between some of the '1st choice' boxes, suggesting that this is differentiated from merely rating GP very highly. The other part of the backbone consists of the three large yellow boxes at the right, which indicate being on the GP Register, and having positive attitudes towards GP as a possible career if starting out in medicine again, either at year 6 or year 13+. The yellow boxes are also connected to the blue boxes by the blue lines, and it can be seen that earlier attitudes influence later attitudes. This part of the model is both important and essentially unsurprising since it is showing that there is relative consistency in a participant's attitudes, those preferring GP at the beginning also tending to prefer it later on.

2. **The endogenous, process variables:** These can be considered in turn.

    a) **Perceptions of GP teaching.** The four measures of the perception of GP teaching are shown in pale green at bottom centre, and paths leading to or from them are shown in green. The most important is finding GP to be Interesting, which has a large effect on interest in GP as a career in the final year, and it also has a smaller effect on putting GP as a first choice. An important point to note is that those who have an interest in a career in GP at application or in year 3 find GP teaching to be more interesting and useful, although the effect is relatively small. However, and it is a fundamental strength of path modelling, the influence of interesting GP teaching on rating GP high as a career in the final year has taken into account that those finding the teaching interesting had a greater interest in a GP career at application to medical school. **Teaching which makes GP seem interesting and useful seems therefore to have an important role in encouraging students to consider careers in general practice.** There are also other minor effects of teaching, although a particularly interesting one is that **having seen GP teaching at medical school as useful means that doctors rate general practice higher as a career thirteen or more years later.** Such long-lasting effects of perceptions of good teaching are important, and while often reported anecdotally, it is rarer to find solid statistical evidence for them.



» *Figure 8.6: The fitted path model*

b) **PRHO posts.** Participants' ratings of the quality of medicine and surgery house-jobs are shown in grey at the lower right, with black arrows going in and out. The effects are small and limited. There is a sense in which doctors rating their PRHO posts as good are less likely to consider a career in general practice (or to turn it around, those who dislike their medicine and surgery jobs are more likely to consider general practice). Those who already had an interest in GP as a career also showed small effects of liking their surgery job less, but seemed to rate their medicine jobs more highly. Overall, PRHO posts seem to have had little influence on career intentions towards GP.

c) **A-level grades.** A-level grades are seen as the dark green box at top left. There are two effects, students with higher A-level grades tending to be less interested in GP at application to medical school, and there is an additional effect that those with higher A-level grades are less likely to go onto the GP Register[21]. Although it still remains controversial, there are other data suggesting that GPs tend to have lower A-level grades than hospital doctors and also scored somewhat lower on a cognitive ability test (McManus et al., 2003).

d) **Age at deciding to study medicine.** This is shown in the top left of the diagram, and simply has no effect on any later measures, those deciding early and late appearing to be equivalent in their interest in careers in general practice.

3. **The exogenous variables:** These are female (shown as purple boxes and lines), BME (shown as dark orange boxes and lines), and Cohort (shown as light orange boxes and brown lines). In addition there are four measures describing whether there is a history of medicine in the family.

a) **Female.** Being female has a number of influences throughout the model (and there is little doubt that women are more likely to go into careers in general practice than are men). Women in these cohorts had slightly lower A-level grades than the men, although that is probably not true of modern cohorts). There is however no difference in interest in careers in GP in the male and female applicants to medical school. That is interesting and important, and Table 8.13 shows the raw data for applicants, along with the increasing difference between men and women as they pass through medical school, which is clearly seen.

The differences between males and females in the path diagram are shown by the purple boxes and arrows dotted all over it. Even by Year 3, women are more interested in GP, and they then find GP teaching to be both more interesting and useful. Those help to increase their interest in GP, but there is then an additional effect on the final year rating of GP, and an additional effect on putting GP first in the final year. Finally, women are more likely to be on the GP Register, even taking all of the earlier effects into account, and then 13 or more years after graduation they rate their interest in general practice as higher, once again taking all of the earlier differences into account. Although women see GP teaching as better, that it is far from being the only explanation of their greater interest in GP, since it has taken off by year 3, and continues to grow through to being on the GP Register and beyond. There seems to be no single effect but the cumulative effect of a myriad of effects at different times.

b) **Cohort differences.** Cohort effects are shown as light orange boxes, and like the effects of female are distributed over the diagram, indicating multiple separate effects. A-level grades show cohort effects, rising steadily over the 1981, 1986 and 1991 cohorts (a positive effect). GP teaching is seen as less interesting, less useful and deserving less time by the later cohorts (a negative effect), and the later cohorts are less interested in GP careers by their final year (negative effect). However, later cohorts then shown a relative increase in an interest in a GP career at the end of the PRHO year (a positive effect), again by the end of the PRHO year, and again an increase in interest in GP for the later cohorts in going onto the GP Register. Some of the subtlety of the effects can be seen in Table 8.6.

[21] Note that these data cannot distinguish between wanting to go on the GP Register but failing the exams which are necessary, and not wanting to go onto the Register in the first place.

» *Table 8.13: Male-female differences in interest in a career in General Practice at various career stages in the Cohort Studies.*

Interest in General Practice	Applicants		Third Year		Final Year		PRHO (91 cohort only)	
	Male	Female	Male	Female	Male	Female	Male	Female
Definitely not	7.4%	8.2%	14.1%	7.2%	17.0%	10.7%	26.1%	16.8%
Not very attractive	15.8%	14.9%	19.0%	10.6%	16.5%	12.4%	18.9%	16.4%
Moderately attractive	31.5%	31.1%	38.9%	36.4%	32.2%	27.4%	28.5%	27.9%
Very attractive	36.8%	37.7%	26.0%	40.5%	28.8%	39.7%	18.5%	22.7%
Definite intention	8.5%	8.1%	1.9%	5.3%	5.5%	9.9%	7.9%	16.1%
N	4285	4027	311	321	1100	1218	620	774
χ^2 (4 df)	3.38, p=.496		29.9, p<.001		62.3, p<.001		37.2, p<.001	
GP first Choice	19.7%	18.8%	10.4%	17.9%	18.3%	25.9%	15.2%	28.8%
N	3448	3475	307	318	1883	1856	611	784
χ^2 (1 df)	.904, p=.342		7.2, p=.007		31.4, p<.001		36.0, p<.001	
Odds ratio (95% CI)	1.03 (.97-1.09)		1.43 (1.07-1.91)		1.27 (1.16-1.38)		1.65 (1.38-1.98)	

» *Table 8.14: Influence of medical parents and relatives on interest in a career in general practice.*

Interest in General Practice at application	Parents			Relatives		
	No doctors	1+ Doctor(s) but no GP(s)	1+ GP(s)	No doctors	1+ Doctor(s) but no GP(s)	1+ GP(s)
Definitely not	7.8%	7.5%	8.6%	7.8%	10.1%	6.2%
Not very attractive	14.6%	18.2%	18.9%	15.0%	19.2%	13.4%
Moderately attractive	31.2%	32.1%	31.5%	31.1%	33.6%	30.0%
Very attractive	38.0%	34.1%	33.6%	38.2%	31.3%	39.6%
Definite intention	8.4%	8.2%	7.4%	8.1%	5.8%	10.8%
N	6755	964	593	4836	1501	213
%	81.3%	11.6%	7.1%	58.2%	18.1%	23.8%
χ^2 (4 df)	19.4, p=.013			84.3, p<.001		
GP first Choice	19.3%	19.8%	17.8%	19.9%	14.2%	21.4%
N	5639	794	490	4129	1180	1614
%	81.5%	11.5%	6.5%	59.6%	17.0%	23.3%
χ^2 (1 df)	.904, p=.342			25.5, p<.001		

c) **BME.** Black and Minority students show few differences in the model, having a somewhat lower average A-level grade, and being a little less likely to put GP down as a first choice at application, but otherwise showing no other differences.

d) **Medical families.** Coming from a medical family has little effect on a GP career, with one or two interesting exceptions. The children of medical parents have somewhat higher A-level grades, and that can indirectly reduce interest in general practice. More interesting is that having doctors in the wider family has a medium-sized effect on the level of interest in general practice in applicants (and see the raw data below). Of those with no medical relatives, 19.9% put GP as a first choice at application. For those who have a GP as a relative, there is a slightly higher proportion who put GP first (21.4%), although the difference is not statistically significant. However amongst those having at least one relative who is a doctor but **not** a GP, there is a **decreased** rate of putting GP first (14.2%), which is statistically very significant (Odds ratio = 0.72, 95% CI = .62 to .84). It would seem, therefore, that **doctors who are not GPs have an effect in putting off their younger relatives from considering becoming a GP.**

8.6.3 Educational attainment in FPAS and choosing General Practice

As mentioned earlier, those interested in general practice in the cohort studies had somewhat lower A-level grades, and that was also found in a previous study, along with somewhat lower scores on a cognitive ability test (McManus et al., 2003). Differences in educational attainment can also be looked at in the data provided to us in the current study, merging together data from the Oriel application system for 2015 and FPAS data for the years from 2012 onwards. The Oriel file records the specialties applied to, with in most cases there being a single specialty, but in about 30% of cases there are applications to two or more specialties (and all of those have been included as "2+").

FPAS (Foundation Programme Application System) collects data about all final year medical students as a part of their applications for foundation programmes. The data come in two sorts, a Situational Judgement Test (SJT) sat by candidates in their final year, and an Educational Progress Measure (EPM). The EPM is complicated as it consists of three separate components: a decile of attainment of the student within their own medical school; additional points for any extra degree, including intercalated degrees or PhDs, with higher points for higher classed degrees; points for up to two peer-reviewed scientific publications[22].

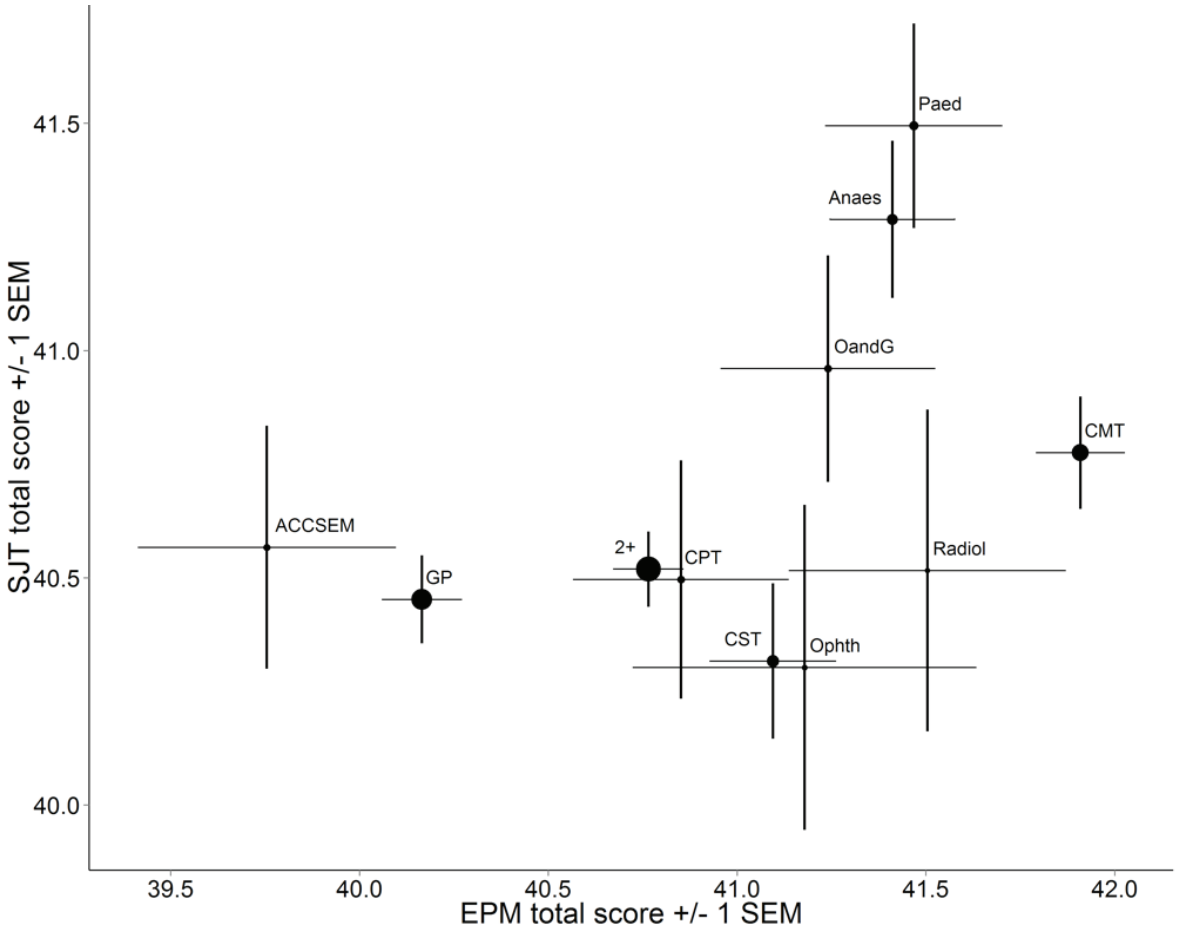
Figure 8.7 shows the average EPM and SJT scores for applicants to different specialties. For the EPM, the highest average scores are for those applying to Core Medical Training (CMT), with GP applicants having rather lower scores, the difference having an effect size of about .36 (i.e. about one third of a standard deviation), which is reasonably large, and highly significant. The Oriel/FPAS data therefore support the conclusions from the earlier studies of local average attainment in those applying for general practice. The practical consequences of those differences are, of course, an entirely separate matter.

8.7 DIFFERENCES BETWEEN MEDICAL SCHOOLS

UK medical students usually spend four to six years at medical school. Those years are often in the late teens and the early to mid-twenties, and are likely to make an impression upon students. Ex students undoubtedly think so (Abse, 1978), although large-scale academic analyses of such effects are relatively rare, with it not being clear whether, as one of us asked previously, the differences between medical schools are "beneficial diversity or harmful deviations" (McManus, 2003). Of course becoming a doctor does not start at medical school, the process of application occurring a year or two earlier, and the decision to apply sometimes occurring long before that, as the studies reported above clearly indicate (McManus et al., 2006). As the path model shows, newly arrived medical students are not therefore a tabula rasa, with newly formed wax ready to be written on as a medical school sees fit, but already come with ideas, prejudices, attitudes, insights and motivations.

[22] The three separate measures are difficult to interpret since, a) the deciles assume that the 5th decile in the University of X is the same as the fifth decile in the University of Y; b) in some schools it is compulsory to take an intercalated degree, and intercalated students almost always get a first or a II.i; and c) the most likely reason for having publications is that an intercalated degree has been taken. Nevertheless the EPM is clearly measuring something relevant to educational attainment.

» *Figure 8.7: FPAS EPM and SJT in relation to Specialty Choice in 2015. The horizontal axis shows the EPM score, ± 1 SEM, and the vertical axis show the SJT score, ± 1SEM, for candidates whose applied for training in the larger specialties in 2015. Labels are mostly self-explanatory, although "2+" refers to candidates applying to two or more different specialties.*



Studies from the Todd Report to the recent Plint Report (Plint, 2014) have found that graduates of some medical schools are more likely to become general practitioners than others, with the latter particularly emphasising that the 'new medical schools' were more likely to have graduates who went into general practice, a result also found by the UK Medical Careers Research Group for graduates in the years 2000, 2002, 2005, 2008 & 2009 (Lambert and Goldacre, 2011). That there are differences seems indisputable. **Much more problematic are the reasons for such differences.** Different types of students go to different medical schools for multiple reasons, and medical schools differ in many ways. Which of those factors, including differences at entrance, explain the differences in proportions of GPs is unclear. We will begin by looking at the longer-term picture, and then look at the more recent data in more detail.

8.7.1 The longer-term picture: 1974 to 2014

As shown earlier in Figures 8.1 and 8.2, the proportion of UK-qualified doctors on the GP register showed two periods of stability, for graduates from about 1974 to 1988 and then again for graduates from about 1991 until 2004. (LRMP data after 2004 are not complete due to not all doctors having yet gone on to the GP register.) Table 8.15 shows the nineteen medical schools during these two periods (London graduates at that time were not distinguished in the GMC Register by medical school). Table 8.15 also shows data on preferences for GP from the MCRG study for 2000, 2002, 2005, 2008 & 2009 (Lambert

and Goldacre, 2011), with London schools merged and only 'old' schools (and fuller data are presented later in Table 8.16). Table 8.15 also shows the averaged data for graduates going into GP training immediately after F2 in 2012 and 2014 (U. K. Foundation Programme Office, 2012b, 2014a), which are shown in more detail in Table 8.16, as well as the percentage of graduates from the three cohort studies on the GP Register. It should be remembered that: a) the overall proportion of doctors going onto the GP register was lower in 1991-2004 than in 1974-1988 (37% vs. 49% across all UK medical schools), b) that Lambert et al data on intentions are classified in three different ways, and c) for 2012 & 2014 the percentages are of doctors going into GP training immediately after F2. The 2012 and 2014 data are only for those doctors from medical schools which existed in the earlier period, with the five separate London schools being combined. The data for Table 8.15 are sorted on the percentage of graduates on the GP Register for 1991-2004.

There is a strong correlation between the percentage from each medical school going onto the GP Register in 1974-1988 and in 1991-2004 (r=0.92) with both sets of values being based on very large numbers of doctors (35597 and 54966). The three categories for the MCRG data correlate highly with one another (0.99, 0.94 and 0.95), with the 'any' category correlating most highly with GMC 1991-2004 data (r= 0.92) and the 1974-88 data (r= 0.82). The correlations with the percentages of trainees in 2012 and 2014 are somewhat lower (r= 0.79 for GMC 1974-1988, r= 0.79 for GMC 1991-2004, and r=0.74 for the MCRG 'any' data), but are still high and highly significant (p<.001 in all cases). Similarly the correlations with the percentage of those on the GP Register for the three cohort studies is a little lower, but still high and statistically significant (r= 0.79 for GMC 1974-1988, r= 0.85 for GMC 1991-2004, and r=0.77 for the MCRG 'any' data; all p<0.001). A useful measure of the similarity of the seven sets of data is the average correlation between them, which is 0.82 (SD 0.087; range 0.70 to 0.99)[23]. The key messages are that a) there is general stability in the 'GP-productivity' of a medical school over time, relative to that at other schools and b) reported interest in GP correlates with becoming a GP at medical school level.

Oxford and Cambridge are to some extent outliers, with around a ten percentage point difference in the percentage of both 1974-1988 and 1991-2004 graduates on the GP Register compared with the next lowest school (as the Todd Report had noted). Removing Oxbridge reduces the average correlation from 0.82 to 0.67 (SD 0.145; range = 0.49 to 0.99), although all remain statistically significant, particularly for percentages on the GMC Register for 1991-2004 with GMC 1974-88 (r= 0.79, p<.001); with MCRG 'any' (r=0.81, p<.001) and with 2012 and 2014, (r=0.66, p=.004).

Taken overall it is clear: a) medical schools undoubtedly differ in the proportion of their graduates who become GPs, b) that the difference is not merely between Oxbridge and the rest, but is present across all schools, and c) the differences appear to be surprisingly stable across time, suggesting deep internal or external structural reasons for their origins, rather than the effects of minor shifts in curricula or other short-term factors.

8.7.2 More recent data

The Plint Report (Plint, 2014) quoted data from the 2012 UKFPO career outcome survey (UK Foundation Programme Office, 2012b), and data are also available for the 2014 survey (U. K. Foundation Programme Office, 2014a), but no data seem to be available for the 2013 survey. The published data are less than perfect, and indicate only the percentages of F2 doctors intending to go straight into GP specialist training, with no indication of sample sizes. The data are shown in Table 8.16 and the unweighted correlation between 2012 and 2014 is 0.66, the variability mainly being due to relatively small sample sizes and hence sampling variation[24]. For most purposes the average of the 2012 and 2014 data is used in the calculations; these averages range from 12% at Oxford to 34% at Keele. Table 8.16 also shows percentage choices for GP from the UK MCRG data (which were shown earlier in an abbreviated form in Table 8.15, and here also include the Numbers on which the data are based). The column marked 'N app' gives the total numbers of applicants to GP training for the years 2009 to 2015 and gives an approximate indication of the relative numbers on which the various percentages are based[25]. Correlations for the

[23] The value of 0.82 could be regarded as a 'typical' test-retest correlation, although some of the measures are not, of course, properly independent.

[24] The same tables in the original source also includes the proportions of F2 doctors going into Core Psychiatry Training, which is only about 3% of trainees (i.e. about 6 doctors from a medical school with 200 undergraduates) and the consistency across years is minimal, with a correlation which is actually negative with a value of -.217. Such results should be treated with great care.

[25] Lancaster students were presumably subsumed within Liverpool, Swansea had only 50 (earlier data presumably being subsumed within Cardiff), and Keele had only 95 (previously being subsumed within Manchester).

» *Table 8.16: Percentages of 2012 and 2014 F2 doctors exiting to GP training, career preferences in the 2000, 2002, 2005, 2008 and 2009 graduates in the UK MCRG data, and the mean scores of graduates from the different medical schools for those applying for GP selection in 2009 to 2015.*

Medical School	Percent F2s on GP training programmes			UK Medical Careers Research Group Preferences for General Practice 2000, 2002, 2005, 2008 & 2009 qualifiers				N app; number applying for GP training	Mean scores at GP selection			
	2012	2014	Average 2012-14	Untied 1st choice	Any 1st choice	Any choice	N		Stage 2 total	Stage 2 CPST	Stage2 SJT	Stage 3 total
Aberdeen	25.7%	28.7%	27.2%	24.0%	33.5%	50.9%	475	874	522.7	255.6	266.7	53.2
Barts and The London	29.4%	23.3%	26.4%	22.2%	30.8%	49.0%	555	1732	502.5	248.4	254.1	53.8
Birmingham	32.7%	29.8%	31.3%	25.5%	34.3%	55.2%	776	1661	529.3	265.5	263.8	54.6
Brighton and Sussex	20.4%	12.8%	16.6%	29.0%	35.5%	57.9%	107	388	539.7	269.8	270.0	54.9
Bristol	15.5%	16.5%	16.0%	18.0%	27.3%	45.6%	528	830	543.2	274.1	269.2	54.5
Cambridge	11.2%	18.5%	14.9%	13.0%	21.1%	34.5%	432	335	560.8	282.8	278.0	56.4
Cardiff	24.5%	17.2%	20.9%	23.5%	33.9%	55.6%	707	1291	534.1	268.0	266.1	55.3
Dundee	19.9%	18.5%	19.2%	27.0%	34.8%	51.1%	411	569	525.3	260.5	264.7	53.2
East Anglia	36.8%	28.1%	32.5%	19.1%	34.8%	61.8%	89	523	520.7	256.5	264.1	55.9
Edinburgh	20.8%	11.7%	16.3%	14.9%	23.2%	39.2%	719	820	546.8	272.6	274.19	55.1
Glasgow	18.6%	24.4%	21.5%	18.9%	28.6%	46.1%	644	845	523.1	254.2	268.9	53.9
Hull York	31.4%	28.9%	30.2%	30.4%	43.8%	65.2%	112	490	520.6	259.0	261.6	53.6
Imperial College	16.0%	14.9%	15.5%	18.1%	23.3%	41.3%	774	1203	531.4	269.4	262.1	54.6
Keele	38.5%	28.8%	33.7%	-	-	-	-	95	508.0	249.8	258.3	54.2
King's College London	25.7%	25.5%	25.6%	17.5%	24.9%	44.9%	952	1712	529.3	263.4	266.0	55.0
Leeds	29.8%	23.9%	26.9%	27.8%	37.6%	56.0%	598	1073	524.6	259.4	265.2	54.9
Leicester	28.3%	27.4%	27.9%	20.9%	29.7%	47.9%	512	1054	521.6	262.5	259.2	53.4
Liverpool	25.9%	20.0%	23.0%	26.0%	36.1%	53.6%	535	1357	517.0	254.0	262.9	53.8
Manchester	25.1%	19.6%	22.4%	22.3%	30.5%	50.6%	984	2021	526.2	260.6	265.7	54.6
Newcastle	23.8%	20.5%	22.2%	22.6%	31.6%	50.9%	658	1289	531.9	264.0	267.8	54.3
Nottingham	24.6%	21.1%	22.9%	22.2%	31.2%	48.9%	650	1091	536.3	266.2	270.1	55.1
Oxford	12.8%	11.3%	12.1%	10.9%	17.7%	36.2%	423	303	565.8	286.7	279.2	56.3

» *Table 8.16 continued.*

Medical School	Percent F2s on GP training programmes			UK Medical Careers Research Group Preferences for General Practice 2000, 2002, 2005, 2008 & 2009 qualifiers				N app; number applying for GP training	Mean scores at GP selection			
	2012	2014	Average 2012-14	Untied 1st choice	Any 1st choice	Any choice	N		Stage 2 total	Stage 2 CPST	Stage2 SJT	Stage 3 total
Queen's University Belfast	16.9%	18.5%	17.7%	23.7%	33.5%	47.5%	541	855	522.6	257.0	265.7	51.0
Sheffield	33.5%	20.6%	27.1%	24.1%	33.6%	50.4%	673	1057	525.4	261.5	263.9	54.3
Southampton	23.9%	19.2%	21.6%	24.3%	33.1%	51.6%	498	1009	522.1	257.1	265.0	53.3
St George's, London	27.0%	22.7%	24.9%	21.0%	30.3%	48.9%	524	1220	522.5	261.5	261.1	53.7
Swansea	20.0%	30.8%	25.4%	-	-	-	-	50	540.7	267.0	273.8	56.7
University College London	17.9%	12.3%	15.1%	19.8%	26.1%	45.0%	874	1326	538.7	272.3	266.3	54.0
Warwick	27.5%	30.9%	29.2%	27.1%	43.0%	64.3%	207	697	531.0	266.2	264.9	55.0
Peninsula	26.1%	24.7%	25.4%	31.4%	40.7%	55.9%	118	489	529.4	262.2	267.2	54.3
Lancaster	-	31.6%	31.6%	-	-	-	-	-	-	-	-	-

remainder of this section are simple correlations based on the means and percentages in Table 8.16, with significance levels based the number of pairs[26].

8.7.3 Exploring differences in 'GP-productivity' between medical schools

There are clear and sustained differences between medical schools in their 'GP-productivity'. There are a number of possible reasons for this, which is explored in this section by looking at associations between data from various sources, aggregated by medical school. The possible reasons are: a) GP-producing schools **attract and/or select applicants and entrants** who are already more interested in GP as a career, b) GP-producing schools are more likely to **increase interest in GP** and c) would-be GPs from GP-producing schools perform better in the **GP selection process**. Our measures of GP-productivity for our analyses of each of these possible reasons have been selected to be as contemporaneous to each reason being explored as possible. It is also plausible that d) trainee GPs from GP-producing schools perform better in MRCGP and ARCP during training. However we currently do not have sufficiently robust data to enable us to explore this hypothesis.

Attracting and selecting would-be GPs:

There is little recent information on the students who choose to go to each medical school before they arrive. That is an important omission, since it may well be that those in particular who go to the 'new' schools, which produce more GPs, are systematically different from those going to Oxbridge or older schools which produce fewer GPs. Cohort studies do however allow such issues to be addressed. The last two columns of Table 8.15 show the average rating of GP as a career for entrants to the various schools (and the data were collected soon after application), as well as the average A-level grades of those entering each of the schools. The latter, as expected, show predictable differences, with Oxford and Cambridge in particular having entrants with higher academic attainment.

Here, our measure of GP-productivity is the percentage of each school's graduates between 1991 and 2004 who are on the GP Register (LRMP data; Table 8.12). There is a positive correlation between the average rating of GP at application to medical school and GP-productivity ($r=0.72$, $n=19$, $p<.001$) and a negative correlation between the average A-level grade of medical school entrants and GP-productivity ($r= -0.77$, $n=19$, $p<.001$).

The two explanatory variables are also negatively correlated ($r= 0.56$, $n=19$, $p=.013$), consistent with the effect found in individuals (see path analysis in Figure 8.6, where medical students interested in GP show lower A-levels at application to medical school). A key question then is whether the differences in GP-productivity are due to differences in **interest at application** or differences in **academic attainment?** We have attempted to answer this question using a multiple regression, with GP-productivity (at the medical school level), regressed on interest in GP at application and academic attainment at entry. Interestingly **both** effects are statistically significant, higher production of GPs being independently associated with higher interest in general practice at application (beta=+0.43, $t(16)=2.676$, $p=.017$) and with lower academic attainment at entry (beta= -0.54, $t(16)=-3.377$, $p=.004$). The effects of prior interest and lower academic attainment are therefore statistically independent[27].

[26] Calculating the correlations is far from straightforward, as the pairs of values have different Ns (and hence different standard errors), with the added problem that the Ns and errors can be different for the X and the Y values making up the correlation (and the range of Ns can be large, as can been seen in Table 8.13. There is no straightforward way of resolving that problem, and therefore a simple correlation has been used. Exploration suggests that in practice it makes little or no difference to the broad findings.

[27] This analysis used the GMC proportion of doctors on the GP Register as the dependent variable (GP-productivity) as it is statistically more robust than the proportion of doctors on the GP Register in the Cohort Studies themselves. The measure inevitably is a percentage, and hence smaller sampling sizes can result in statistical fluctuations (and that problem is less obvious with mean scores at A-level or in preference for general practice since the measures are continuous). Repeating the analysis with the proportion of doctors in the cohorts on the GP Register gives broadly similar results, GP-productivity correlating with average rating of GP at application ($r=0.61$, 19 df, $p=.006$) and average A-level grades ($r=-0.66$, 19 df, $p=.002$), and in the regression using both predictors, A-level grade is statistically significant (beta=-0.47, $t=-2.257$, $p=.038$), whereas rating of GP is not (beta=0.35, $t=1.675$, $p=.113$), although the effect is similar in size to that reported in the main text.

The consequence of these results is that **differences in GP-productivity between medical schools are due, at least in part, to differences in early interest in GP and in academic attainment between the cohort of students entering each school.** This could be due to student choice and/or medical school selection processes (e.g. a focus on communication skills in a school with high GP-productivity and academic attainment in one with low GP-productivity), but we cannot distinguish between these possibilities here.

Increasing interest in GP:

The next question is whether differences in GP-productivity are affected by what happens **during** medical school. If a medical school has an influence on encouraging their students to become GPs then the attractiveness of GP to students as a career at the end of medical school, or at the end of the PRHO year, should predict the proportion of doctors on the GP Register **after taking differences at entry into account.** Inevitably the statistical power is limited, but a regression analysis can consider this question by regressing the proportion of doctors on the GP Register (GP-productivity) on final year or PRHO interest in GP, after including interest at entry and previous academic attainment into account. The simple answer is that **there is no effect (Final Year: beta= -0.07, p=.717), and neither are there effects of PRHO ratings.** However, it is important to recognise that the path model described above did suggest that positive perceptions of GP teaching and experience in medical school did increase interest in GP as a career choice and thus further study to determine whether there are differences in such perceptions between schools would be useful. A resolution of the discrepancy may be that all medical schools influence students' perception of GP as a career, but they do so in broadly similar ways, so that there are small or non-existent differences between medical schools[28].

Performance at GP selection:

Our measure of medical school GP-productivity for this analysis is the average percentage of F2s going into GP training programmes across 2012 and 2014; as shown above, there is a strong positive correlation at medical school level between this variable and the proportion of graduates from earlier cohorts being on the GP Register. In Table 8.13, these averages range from 12% at Oxford to 34% at Keele.

For each medical school in Table 8.16, the mean score at Stage 2, CPST, SJT and Stage 3 total (i.e. the score based on just the selection centre stations) was calculated for selection years from 2009 to 2015. Data for 2009 and 2010 at both stages 2 and 3 had been calculated in a different way and on a different scale than for later years, and therefore scores for 2009 and 2010 were individually rescaled for Stage 2, CPST, SJT and Stage 3 total so that their mean and SD were the same as that for the average of 2011 to 2015, making them comparable[29].

There is a **strong negative correlation between schools' mean Stage 2 scores and GP-productivity** ($r=-0.67$ for the CPS, -0.61 for the SJT and $r=-0.68$ combined, all $p<0.001$, $N=30$). However there is no relationship between mean Stage 3 scores and GP-productivity ($r=-0.07$, $p=0.70$, $N=30$). Taken together these results **do not suggest that GP-producing schools are producing doctors who are in some sense more fit to enter GP training** (and indeed the data on Stage 2 suggests the converse)[30].

The conclusion has to be that **medical schools DO differ in their production of GPs, but it is unclear as to whether that is due to schools encouraging students to become GPs, but it does NOT seem to be due to schools with high GP-productivity making their students more likely to succeed in the GP selection process, but rather would-be GPs are more likely to enter some medical schools than others, be it through students' selection of medical school or medical schools' selection of students. This effect is partially mediated by academic attainment.** It is also unlikely to be due to an independent medical school effect on the

[28] If this seems unlikely, it is worth remembering that all medical schools influence medical students by teaching them the Krebs cycle, but the Krebs cycle taught and learned is the same across all medical schools.

[29] As elsewhere, this is the score just for the selection centre measures, and is not the eventual score, which incorporates the banded Stage 2 scores, and which is used for ranking candidates.

[30] Ideally a further analysis would be undertaken that includes interest in GP and previous academic attainment as a predictor, but a lack of contemporaneous data (to 2009-2015 GP selection scores) and changes in the nature of A-Level grades since the cohort studies makes this problematic. We might expect, for example, that differences in Stage 2 scores are at least partially mediated by differences in academic attainment at entry to medical school.

likelihood of passing the MRGCP (Wakeford et al., 1993), but this hypothesis has not been tested formally, and it is probable that Deaneries/LETBs also have influences on MRCGP attainment (Wakeford et al., 1993). Data on success in specialty training applications and postgraduate examinations by medical school have recently started being reported by the GMC (see http://www.gmc-uk.org/education/25496.asp) and will provide a further set of data for analyses of the type reported in this chapter in due course. The present datasets also have information on training Deaneries/LETBs, and analysis of those is also possible.

8.8 SUMMARY

This chapter has shown that the proportion of UK medical graduates entering the GP Register has shown remarkable consistency over time, with the exception of a large, unexplained reduction from 46% for pre-1987 graduates to 36% for post-1991 graduates. Data for post-2004 graduates are unreliable since many would-be GPs are still in training. Although the absolute number of non-UK graduates on the GP Register was fairly stable, this is due to an increasing number of doctors entering the UK, as the proportion of non-UK graduates has decreased to around 13%. Of those on either GP or Specialist Registers, non-UK graduates are less likely to be on the GP Register compared with UK graduates (approximate percentages are 25% and 42%) and therefore it is worth exploring further why those who qualified overseas do not enter GP.

The data on time since qualification to entering GP training indicate that many doctors do not follow the direct graduation -> 2 year Foundation Programme -> entry into GP training model but that many enter GP training (for whatever reason) later in their careers, some from what might be considered unusual specialties such as surgery. However, recent data suggest that either more doctors are delaying entry or that GP is becoming less popular, with a very large fall in the proportion of 2012 and 2013 graduates applying for GP in their second year post-graduation than previously.

The remainder of this chapter focused on UK medical schools and pre- and during- medical school influences on becoming a GP. Previous academic attainment and pre-medical school interest in GP appear to be highly influential, with being female (positive) and having non-GP doctors as family members (negative) also important. There is mixed evidence as to the effect of medical school experience; the path model suggesting that positive perceptions of GP teaching do influence entry into GP, but there was no evidence for differences between medical schools in the size of that effect. These seemingly discrepant results may however simply be due to there being no difference in how GP teaching is perceived across schools. That students from medical schools with lower levels of GP-productivity do better in the Stage 2 GP selection tests suggests that entry into GP for students at schools with high GP-productivity is not being facilitated by their choice of school.

We have only looked a little at attitudes towards general practice after medical school, but it seems likely that Foundation experience has an influence on specialty choice decision. More contemporaneous data would also enable analyses using more historical data to be repeated; although we have shown reasonable stability in the relative likelihood that graduates from each school will enter GP over time. Finally, we note that our analyses identify associations rather than causations and we therefore end by quoting a 2011 report from the Centre for Workforce Intelligence:

> *"Further work is needed to understand why we are not yet attracting the right number of doctors in training into [general practice] ... Although our recommendations bring us closer to the level of training needed to balance hospital and general practice specialties, we have not been able to do this."* (Centre for Workforce, 2011) p.28, our emphasis).

APPENDIX 8.1

A note on UK Medical Schools

Appendix 8.1

A note on UK medical schools

Medical schools here are primarily classified using the GMC's LRMP, and using their names for institutions, which may not be the same names as used by the Medical Schools Council and other organisations. The LRMP makes clear when medical schools began awarding degrees, and that year is used here, although occasionally it may only be a single graduate who for some reason is ahead of the rest. When medical schools stop awarding degrees is less clear as in part the awarding institution is then at the choice of the student. Medical schools and awarding bodies can be divided into several groups. Dates are the years of graduation:

First-generation medical schools:

The 'original' medical schools, some of which are ancient, and all of which have awarded degrees since before the Second World War, and sometimes long before: Oxford University, Queens University of Belfast, University of Aberdeen, University of Birmingham, University of Bristol, University of Cambridge, University of Edinburgh, University of Glasgow, University of Leeds, University of Liverpool, University of London, University of Manchester, University of Sheffield and the University of Wales.

The first wave of expansion:

The first wave of expansion. A group of medical schools formed in the 1960s and 1970s, as part of general expansion of universities: University of Newcastle upon Tyne (1963), University of Dundee (1968), University of Nottingham (1975), University of Southampton (1976) and University of Leicester (1980).

The second wave of expansion:

The second wave of expansion. The second wave of expansion occurred in the 1990s, and resulted in both entirely new medical schools, and also medical schools allied to existing medical schools which then separated off from them.

1. **Entirely new medical schools:** Five medical schools were created in universities or from groups of medical schools where no medical school previously existed: Universities of Exeter and Plymouth ('Peninsula Medical School', 2007), University of East Anglia (2007), The University of Brighton and the University of Sussex (2008), and the University of Hull and the University of York ('Hull-York', 2008).

2. **Twinned medical schools:** These are complicated and are listed separately.

    a) Keele University (2012). For its earlier years, graduates of Keele University took Manchester medical degrees.

    b) The University of Warwick (2007). From 2004 to 2007 degrees were awarded by Leicester Warwick Medical School.

    c) Swansea University (2014). Originally a part of the University of Wales, Swansea awarded its first degrees in 2014.

3. **Independent degree awarding powers:** Several medical schools which were originally part of larger organisations gained their own degree awarding powers.

a) University of London. The federal university of London broke up in the 2000s, and separate degree awarding powers were granted to Imperial College London (2008), King's College London (2010), University College London (2010), and St George's Hospital Medical School (2012). At present, Barts and The London School of Medicine and Dentistry remains the only London medical school awarding University of London degrees.

b) University of Wales. As mentioned above, the University of Wales split into Cardiff University (2011) and Swansea University (2014).

c) Peninsula Medical School. Peninsula Medical School split into the separate medical schools of the University of Exeter and University of Plymouth, but neither body has yet awarded its own degrees.

Non-medical schools and historical medical schools:

Historically, organisations such as Royal Colleges and the Society of Apothecaries have had the ability to grant recognised qualifications, and they are omitted. Likewise the University of St Andrews and the University of Durham have awarded degrees, but these are also omitted here.

Medical school mergers:

In London, mainly as the spin-off of various official reports, there was a string of mergers in the 1980s and onward. These resulted in the five medical school conglomerates that now exist:

1. **Imperial College School of Medicine**. The Westminster Hospital Medical School and Charing Cross Hospital Medical School merged in 1984 to form Charing Cross and Westminster Medical School (CXWMS). St. Mary's Hospital Medical School had merged in 1988 with Imperial College, which at that time had no medical school, and in 1997 it merged with CXWMS and the Royal Postgraduate Medical School at Hammersmith to form Imperial College School of Medicine.

2. **UCL Medical School:** University College Hospital School of Medicine merged in 1987 with the Middlesex Hospital School of Medicine to form University College and Middlesex School of Medicine (UCMSM). UCMSM merged in 1998 with the Royal Free Hospital School of Medicine to form the Royal Free & University College Medical School (RFUCMS), which in 2008 was renamed as UCL School of Medicine.

3. **King's College London School of Medicine:** In 1982 Guy's Hospital Medical School and St Thomas's Hospital Medical School merged to form the United Medical and Dental Schools (UMDS). UMDS subsequently merged in 1998 King's College School of Medicine to form GKT School of Medicine, which was renamed in 2005 as King's College London School of Medicine.

4. B**arts and The London School of Medicine:** St Bartholomew's Hospital Medical College and the London Hospital Medical College merged in 1995 to form Barts and the London School of Medicine and Dentistry, as a part of Queen Mary University of London (QMUL).

5. **St. George's, University of London:** The only medical school not to have merged, St George's Hospital Medical School migrated in 1980 from Hyde Park to Tooting, and subsequently changed its name to St. George's, University of London (SGUL).

Interpreting medical school differences is often far from straightforward, particularly when medical schools have split, and graduates continue to name the previous institution as their Place of Qualification. The University of London is particularly vexed in this respect.

~ This page is intentionally left blank ~

Chapter 9

Specialty Training Candidate Questionnaire

Chapter 9.

Specialty training candidate questionnaire:

What predicts whether doctors apply for specialty training in general practice?

9.1 INTRODUCTION

A questionnaire was distributed to all those applying for CT1/ST1 specialty training posts in Round 1 2014/15. All candidates were emailed information about the questionnaire and asked to complete the questionnaire online, using Survey Monkey. Two reminder emails were sent following the initial invitation, approximately two weeks apart. This chapter focuses on responses relevant to this study as a whole; i.e. recruitment and selection into General Practice (GP).

9.2 RESPONSE RATE, DEMOGRAPHICS AND APPLICATION NUMBERS

3,838 responses to the questionnaire were received following a total of three invitation emails, **a response rate of 32.6%** from the 11,782 round 1 candidates for CT1/ST1 specialty training posts. **Just under one third of GP candidates responded** to the questionnaire (N=5,112 using GPNRO data and 4,837 using Oriel data). These figures can be compared to the mean response rate of 38% for online surveys of health professionals reported in a meta-analysis (Cho et al., 2013). Table 9.1 below compares the demographics of respondents with the candidate pool as a whole (using data from the Oriel online application system); as well as the demographics of respondents applying for GP training with the GP candidate pool (using data from GPNRO and from Oriel). These comparisons enable us to evaluate the representativeness of the respondent sample and thus the generalizability of the findings to the candidate pool. The data also highlight some inconsistencies, since the data for candidates for GP from GPNRO and Oriel should be identical, but are not.

Comparing questionnaire respondents who applied to GP with the population of GP candidates, **males and, to a certain extent, non-EU trained and Asian/Asian British candidates were less likely to respond to the questionnaire** and are thus under-represented in the results.

As the data are based on candidares in Round 1 of 2014/15 recruitment only, we cannot generalise these results to candidate cohorts in other Rounds or years. All data presented in other sections of this chapter only include questionnaire respondents unless otherwise specified.

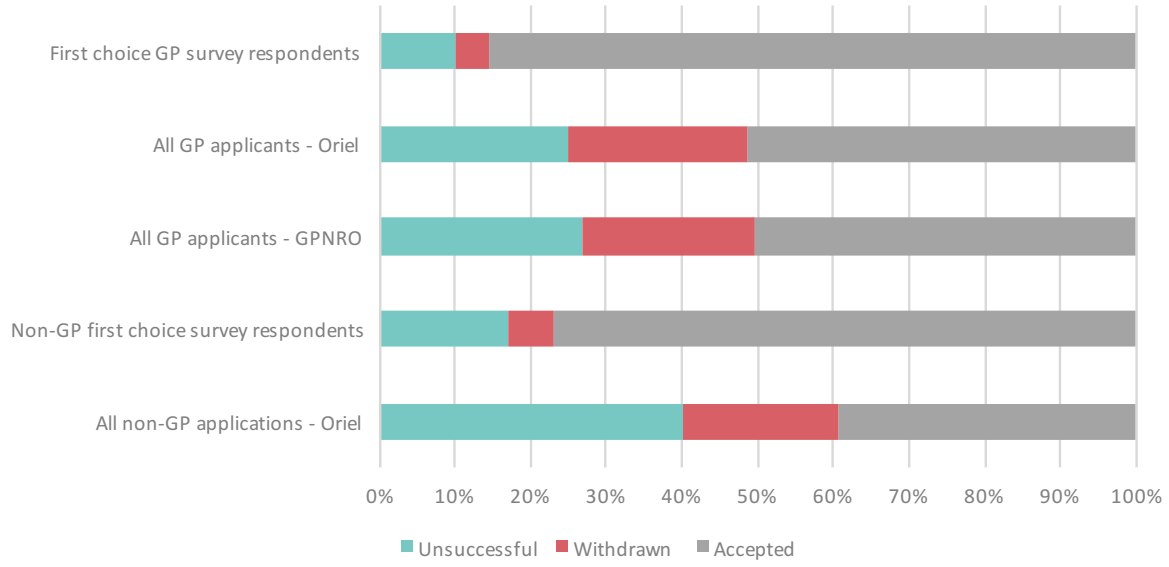
Figure 9.1 presents data from Oriel, GPNRO and from questionnaire respondents on specialty applications and outcomes. The data from Oriel represent candidates rather than applicants, while those from all questionnaire respondents only reflect respondents' status for their first choice specialty. Comparing the status of questionnaire respondents with that of all candidates suggests that **those responding to the questionnaire were more likely to have accepted a post.** The potential response bias for other aspects of this analysis therefore needs to be noted: respondents are more likely to give positive feedback regarding GP since they are more likely to have accepted a post compared to non-respondents. A more extensive analysis of progression through the selection process is described in a later chapter.

» *Table 9.1: Demographic comparison of respondents and candidates*

	All respondents	All candidates Oriel	All respondents applying for GP	All GP candidates GPNRO	All GP applicants Oriel
	N (%)	N (%)	N (%)	N (%)	N (%)
N	3,838	11,782	1,594	5,112	4,837
Employment status on application:				N/A	
Accredited Foundation post	2,024 (52.7)	5,172 (43.9)	805 (50.5)		1,987 (41.1)
Completed Foundation but not in a training post	568 (14.8)		242 (15.2)		
Overseas	202 (5.3)		67 (4.2)		
Non-training post	658 (17.1)		325 (20.3)		
Training in another specialty	81 (2.1)		31 (1.9)		
Not known	305 (7.9)	6,610 (56.1)	124 (7.8)		2,850 (58.9)
Years since qualification:	(N=3.814)		(N=1,585)	N/A	
2	1,825 (47.9)		699 (44.1)		
3	762 (20.0)		276 (17.4)		
4	366 (9.6)		149 (9.4)		
5+	861 (22.6)		461 (29.1)		
Place of initial training:					
UK	3,048 (79.4)	8,845 (75.1)	1,184 (74.3)	3,696 (72.3)	3,452 (71.4)
Other EU country	263 (6.9)		85 (5.3)	288 (5.6)	
Non-EU Country	492 (12.8)		312 (19.6)	1,128 (22.1)	
Did not respond	35 (0.9)	348 (3.0)	13 (0.8)		0 (0.0)
Gender:					
Male	1,376 (35.9)	5,147 (43.7)	477 (29.9)	1,964 (38.4)	1,857 (38.4)
Female	2,055 (53.5)	6,487 (55.1)	962 (60.4)	3,080 (60.3)	2,925 (60.5)
Not disclosed	407 (10.6)	148 (1.3)	155 (9.7)	68 (1.3)	55 (1.1)
Age in years (at 1st January 2015):	(N=3,321)		(N=1,376)	29 (27 to 33)	
Median (IQR)	28 (26 to 31)		28 (26 to 32)		
Ethnic group:					
White British	1,680 (43.8)	5,221 (44.3)	598 (37.5)	1,996 (39.0)	1,868 (38.6)
Other White	410 (10.7)	1,343 (11.4)	144 (9.0)	457 (8.9)	426 (8.8)
Asian or Asian British	727 (18.9)	2,770 (23.5)	404 (25.4)	1,581 (30.9)	1,523 (31.5)
Black or Black British	166 (4.3)	731 (6.2)	95 (6.0)	403 (7.9)	391 (8.1)
Chinese	173 (4.5)	438 (3.7)	62 (3.9)	158 (3.1)	148 (3.1)
Mixed	110 (2.9)	442 (3.8)	53 (3.3)	150 (2.9)	146 (3.0)
Other / Undisclosed	572 (14.9)	837 (7.1)	238 (14.9)	367 (7.2)	335 (6.9)

Percentages may not sum to 100% due to rounding.
N/A: Not available in dataset.

» *Figure 9.1 Applications and outcomes by specialty.*



9.3: SPECIALTY CHOICE: THE 'COMPETITIVENESS' OF GENERAL PRACTICE

Table 9.2 shows the specialty choices of questionnaire respondents. 1,202 (31%) applied to GP as their first choice and 546 (14%) applied as a non-first choice; a further 15% of respondents stated that they considered GP, but did not apply, with 41% of respondents not considering GP at all. In terms of whether GP is a definite choice, **about two-thirds of those applying to GP stated that GP was their first choice specialty.** This proportion was slightly higher amongst non-UK graduates (67% UK vs. 73% non-UK).

The proportion of GP candidates choosing GP as their first choice is on a par with Paediatrics, Obstetrics & Gynaecology and Clinical Radiology and exceeded only by highly specialised choices (Cardiothoracic Surgery, Neurosurgery, Oral & Maxillo Facial Surgery and Ophthalmology. Almost one quarter of those applying to GP as their non-first choice specialty applied to Core Medical Training as their first choice (column 4 in Table 9.2). Looking at these data 'the other way round' (column 5 in Table 9.2), shows how many candidates to a specialty as their first choice also apply to GP (presumably as a back-up). For example, just over two-thirds of those applying for Broad Based Training as their first choice also applied to GP.

Table 9.3 shows the number of specialty applications made by different groups of candidates. These statistics suggest that survey respondents and those applying to GP tended to make more applications in total (that respondents tended to make more applications may partially explain why they were more likely to have been successful, but this is a one tentative hypothesis). In turn, either 'first choice' GP candidates are more likely to have a 'back-up' choice (or choices), or when **those with other first choices have a 'back-up' it is likely to be GP.** An interrogation of the data suggests that the latter explanation is more likely: 'first choice' GP candidates apply to fewer specialties on average than those with other first choices (means 1.42 vs. 1.76), while 38% of those with non-GP first choices and who make more than one application also apply to GP. (Note that no significance testing has been undertaken since these were post-hoc rather than planned analyses.)

9.4: SPECIALTY CHOICE: WHY DO DOCTORS CHOOSE GENERAL PRACTICE?

Table 9.4 considers whether candidates choosing GP as their first choice specialty do so for different reasons than candidates to other specialties. Respondents were asked to indicate all reasons for their first choice of specialty, hence the total number of responses exceeds the total number of candidates. Reasons selected by over 50% of respondents are shown in bold. The final column shows the relative likelihood that a 'first choice' GP candidate gave each reason compared to a candidate to another specialty. Values greater than 1 indicate that reason was more frequently stated by those applying to GP as their

» *Table 9.2: Specialty applications*

	Applied as first choice	% of all candidates to specialty doing so as first choice	First choice specialty if applied to GP as non-first choice	First choice specialty if applied to GP as non-first choice
	N (% of respondents)		N (% of 546 non-first choice GP candidates)	N (% of first choice candidates to that specialty)
General Practice	**1,202 (31.3)**	**68.8**	**N/A**	**N/A**
ACCS Anaesthetics	180 (4.7)	39.7	17 (3.1)	17 (9.4)
ACCS Emergency Medicine	199 (5.2)	58.5	32 (5.9)	32 (16.1)
Acute Medicine	50 (1.3)	25.3	10 (1.8)	10 (20.0)
Anaesthetics	266 (6.9)	60.3	31 (5.7)	31 (11.7)
Broad Based Training	49 (1.3)	24.5	33 (6.0)	33 (67.3)
Cardiothoracic Surgery	25 (0.7)	89.3	0 (0)	0 (0)
Clinical Radiology	144 (3.8)	68.2	44 (8.1)	44 (30.6)
Community, Sexual and Reproductive Medicine	14 (0.4)	23.3	6 (1.1)	6 (42.9)
Core Medical Training	607 (15.8)	64.6	132 (24.2)	132 (21.7)
Core Surgical Training	180 (4.7)	59.6	53 (9.7)	53 (29.4)
Core Psychiatry Training	250 (6.5)	70.4	36 (6.6)	36 (14.4)
Histopathology	45 (1.2)	62.5	12 (2.2)	12 (26.7)
Neurosurgery	48 (1.3)	92.3	1 (0.2)	1 (2.1)
Obstetrics & Gynaecology	177 (4.6)	74.7	41 (7.5)	41 (23.2)
Oral & Maxillo Facial Surgery	13 (0.3)	100	1 (0.2)	1 (7.7)
Ophthalmology	84 (2.2)	95.5	27 (4.9)	27 (32.1)
Paediatrics	222 (5.8)	77.6	54 (9.9)	54 (24.3)
Public Health	83 (2.2)	70.9	16 (2.9)	16 (19.3)

Percentages may not add to 100% due to rounding.

» *Table 9.3: Specialty applications*

	All candidates - Oriel	GP candidates - Oriel	All survey respondents	All survey respondents applying for GP
Median	1	1	1	2
Mean	1.39	1.59	1.65	1.83
Range	1 - 15	1 - 15	1 - 13	1 - 2

» *Table 9.4: Reasons for first choice of specialty: GP as first choice (N=1,202) vs. non-GP as first choice (N=2,636)*

	GP as first choice	GP applicants Oriel	All survey respondents
	N (%)	N (%)	
My seniors/consultants advised me that I would be well suited to it	229 (19.1)	811 (30.8)	0.62
I believe that working within the specialty will allow me to provide good continuity of patient care	**822 (68.4)**	865 (32.8)	2.08
The financial rewards associated with specialty	161 (13.4)	169 (6.4)	2.09
Good work/life balance in specialty	**1,087 (90.4)**	906 (34.4)	2.63
Intellectual challenge of specialty	346 (28.8)	**1,823 (69.2)**	0.42
Positive experience in clinical posting in specialty	387 (32.2)	**1,737 (65.9)**	0.49
Positive experience at Medical School	429 (35.7)	**1,344 (51.0)**	0.70
Information received at Medical School	81 (6.7)	229 (8.7)	0.78
Prestige associated with the specialty	24 (2.0)	305 (11.6)	0.17
My personality is well suited to the specialty	**768 (63.9)**	**1,865 (70.8)**	0.90
Highly competitive specialty	35 (2.9)	345 (13.1)	0.22
Less competitive specialty – I was more likely to get a specialty training post	104 (8.7)	85 (3.2)	2.68
Less competitive specialty – I would have the option to choose where I wanted to work (geography)	212 (17.6)	96 (3.6)	4.84

Percentages may not sum to 100% due to rounding.

» *Table 9.5: Sources of careers advice: GP as first choice (N=1,202) vs. non-GP as first choice (N=2,636)*

Q13	GP as first choice	Other specialty as first choice	Relative likelihood
	N (%)	N (%)	
Senior trainees and/or consultants	**623 (51.8)**	**1,933 (73.3)**	0.71
Educational supervisor	398 (33.1)	1,110 (42.1)	0.79
Postgraduate clinical tutor	62 (5.2)	213 (8.1)	0.64
College tutor	21 (1.7)	136 (5.2)	0.34
Careers websites	305 (25.4)	830 (31.5)	0.81
Foundation School events	249 (20.7)	582 (22.1)	0.94
None	325 (27.0)	366 (13.9)	1.95

Percentages may not sum to 100% due to rounding.

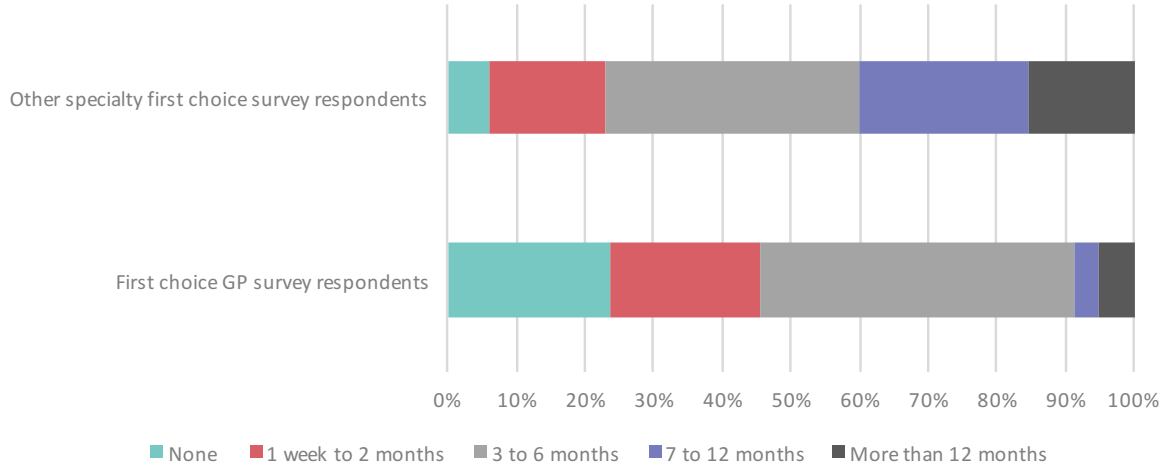
first choice specialty and vice-versa. Relative differences with a four-fold or higher magnitude are shown in pink. Although the absolute numbers are fairly small, **those applying to GP as their first choice compared to those choosing other specialties are much less likely to do so because of the prestige and (high) competitiveness of GP, and more likely to do so because they are more likely to get a training post in their preferred LETB.** There are smaller, but still important differences for a number of other reasons, including **work-life balance (more likely amongst 'first choice' GP candidates)** in particular.

Figure 9.2 considers whether a lack of previous post-qualification exposure to GP could be hindering applications. Previous experience in candidates' first choice specialty is compared between candidates choosing GP as their first choice specialty and candidates to other specialties. **Those applying to GP as their first choice have had less experience in GP than those applying to other specialties as their first choice have had in that specialty** (the Kendall's tau-b correlation coefficient is fairly low but statistically significant: -0.287, $p<0.001$). There is however a strong effect of place of qualification: 14% of UK first choice GP candidates report no GP experience compare to 47% of non-UK first choice GP candidates (the corresponding figures for candidates with other first choices are 6% UK and 9% non-UK).

This result may explain why the proportion of 'first choice' GP candidates giving "positive experience in clinical posting in specialty" as a reason for choosing GP is half that as for those applying to other specialties as their first choice (Table 9.4). (The alternative is that such clinical posting experience is less positive in GP than in other specialties.)

Finally, Table 9.5 compares who provided careers advice to those choosing GP as their first choice with candidates to all other specialties. Respondents were asked to indicate all sources of advice received, so the totals sum to more than the number of respondents. Sources stated by more than 50% of respondents are shown in bold. The final column shows the relative likelihood that a 'first choice' GP candidate gave each source of advice compared to a candidate to another specialty. Values greater than 1 suggest that source of advice was more frequently stated by those applying to GP as their first choice specialty and vice-versa. Relative differences with a two-fold or higher magnitude are shown in pink. What is clear from Table 9.5 is that **those applying to GP as their first choice specialty are less likely to receive advice from every source compared to candidates to other specialties** and thus a lack of positive careers advice could be reducing the size of candidate pool.

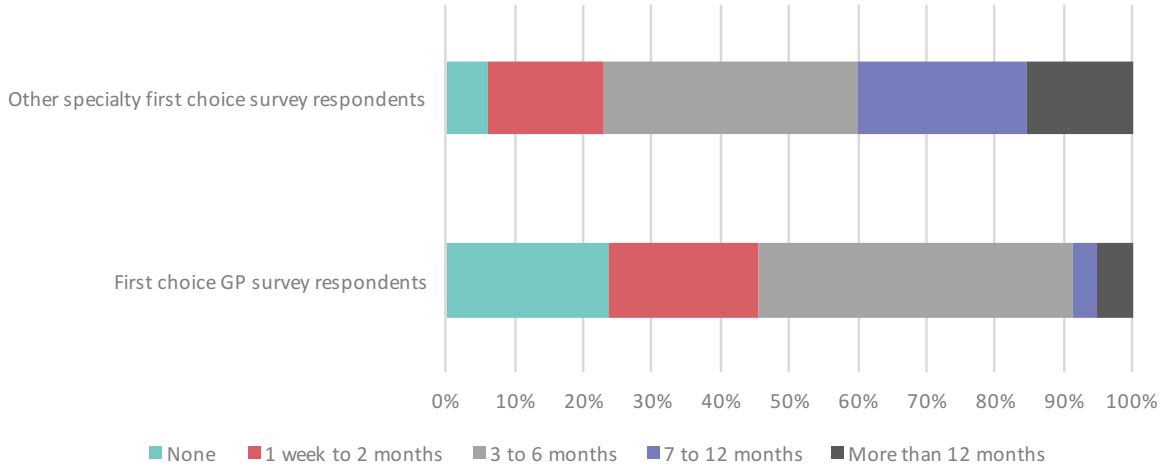
» *Figure 9.2: Previous experience in first choice specialty*



9.5: LETB PREFERENCES

Figure 9.3 considers the relative importance of specialty and LETB between those choosing GP as their first choice specialty and those choosing other specialties. **Those choosing GP are more likely to consider location important when making their application decisions** (the Kendall's tau-b correlation coefficient is low, but statistically significant: 0.178, $p<0.001$).

» *Figure 9.3: Relative importance of specialty and LETB: GP as first choice (N=1,202) vs. non-GP as first choice*



The primary driver behind LETB choice appears to be candidates' current location: in total, **70% of candidates stated that their current LETB was their first choice LETB.** Those applying to GP as their first choice were more likely to want to remain in their current LETB, with 76% giving their current LETB as their first choice LETB, compared with 68% amongst those with another specialty preference.

Table 9.6 firstly considers whether those applying to GP as their first choice have different LETB preferences to those applying to other specialties and secondly whether any particular LETBs stand out as being selected as first preference due to their general and/or training reputations for GP. Respondents could select multiple reasons for their first choice of LETB. **London appears more popular amongst those applying to other specialties as their first choice** (27% stated London as their first choice LETB) compared with those applying to GP as their first choice (22%).

Columns 4 and 5 of Table 9.6 consider reasons for LETB preferences amongst first choice GP candidates. **The North East LETB stands out as frequently being chosen due to its general and training reputation.** To a lesser extent, South West, Thames Valley and Wessex LETBs are also selected due to their reputations. Not shown in the table is a comparison with those not choosing GP as their first choice specialty, but in general **LETBs are less likely to be chosen on account of their general or training reputations amongst those applying to GP as their first choice** compared with those applying to other specialties, with proportions stating these reasons of 38% vs. 52% for general reputation and 26% vs 44% for training reputation.

Data from GPNRO (Round 1 only) were used to determine the effect of whether the LETB where a candidate was offered an interview and/or post differed from the LETB of application on candidates' acceptance decisions. Only two candidates were offered an interview in a LETB that was not their LETB of application, so no further analysis was undertaken. 85 candidates were offered a post in a LETB that was not their initial LETB of application (2.8%). Of these candidates, only 44 (51.8%) accepted their offer, compared to 83.9% of those offered a post in the same LETB as their initial LETB of application. This difference in acceptance rates is statistically significant (chi-squared = 60.283, $p<0.001$). However, given the small numbers involved, **the potential number of GP trainees 'lost' through not meeting LETB preferences is very low.**

9.6: REASONS FOR/AGAINST APPLYING TO GENERAL PRACTICE

The survey respondents' reasons given for and against applying to General Practice are of central importance. This section provides a simple analysis of the reasons given in response to question 14. In the subsequent section a broader, more complex path analysis is undertaken. Findings from the two analyses are generally in agreement.

» *Table 9.6: LETB preferences and reputations: GP as first choice (N=1,084) vs. non-GP as first choice (N=2,432)*

LETB stated as first choice	First choice LETB		LETB general reputation stated as reason for choice	LETB training reputation stated as reason for choice
	GP as first choice	Other specialty as first choice		
	N (%)	N (%)	N (% of first choice GP candidates with LETB as first choice)	N (% of first choice GP candidates with LETB as first choice)
Defence	6 (0.6)	6 (0.3)	0 (0.0)	0 (0.0)
East Midlands	52 (4.8)	112 (4.8)	16 (30.8)	4 (7.7)
East of England	79 (7.3)	136 (5.8)	27 (34.2)	14 (17.7)
Kent, Surrey and Sussex	75 (6.9)	106 (4.5)	25 (33.3)	13 (17.3)
London	235 (21.7)	642 (27.4)	109 (46.4)	80 (34.0)
North East	51 (4.7)	97 (4.1)	37 (72.5)	30 (58.8)
North West	97 (8.9)	223 (9.5)	24 (24.7)	16 (16.5)
Northern Ireland	31 (2.9)	73 (3.1)	3 (9.7)	2 (6.5)
Scotland	103 (9.5)	216 (9.2)	35 (34.0)	29 (28.2)
South West	78 (7.2)	176 (7.5)	42 (53.8)	27 (34.6)
Thames Valley	55 (5.1)	103 (4.4)	29 (52.7)	15 (27.3)
Wales	20 (1.8)	49 (2.1)	3 (15.0)	2 (10.0)
Wessex	32 (3.0)	75 (3.2)	16 (50.0)	15 (46.9)
West Midlands	99 (9.1)	165 (7.0)	23 (23.2)	18 (18.2)
Yorkshire & The Humber	71 (6.5)	163 (7.0)	20 (28.2)	17 (23.9)

» *Table 9.7: Reasons for/against applying to GP: GP as first choice (N=1,202) vs. GP as non-first choice (N=546) vs. those not applying to GP (N=2,090)*

	GP as first choice		GP as non-first choice		GP not applied to	
	Influence for GP	Influence against GP	Influence for GP	Influence against GP	Influence for GP	Influence against GP
	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
Advice from my seniors/consultants	484 (40.3)	213 (17.7)	134 (24.5)	111 (20.3)	185 (8.9)	348 (16.7)
The patient care that I could provide	944 (78.5)	45 (3.7)	272 (49.8)	94 (17.2)	458 (21.9)	712 (34.1)
The financial rewards associated with specialty	417 (34.7)	66 (5.5)	181 (33.2)	50 (9.2)	406 (19.4)	189 (9.0)
Work/life balance	1,127 (93.8)	14 (1.2)	449 (82.2)	35 (6.4)	1,107 (53.0)	250 (12.0)
Intellectual challenge of specialty	682 (56.7)	66 (5.5)	165 (30.2)	120 (22.0)	295 (14.1)	812 (38.9)
Experience working in GP	**659 (54.8)**	**40 (3.3)**	**185 (33.9)**	**104 (19.0)**	**330 (15.8)**	**781 (37.4)**
GP experience at Medical School	**672 (55.9)**	**83 (6.9)**	**220 (40.3)**	**106 (19.4)**	**463 (22.2)**	**869 (41.6)**
Information, advice and comments received at Medical School	**372 (30.9)**	**115 (9.6)**	**100 (18.3)**	**88 (16.1)**	**202 (9.7)**	**416 (19.9)**
Prestige associated with GP	**161 (13.4)**	**283 (23.5)**	**34 (6.2)**	**193 (35.3)**	**49 (2.3)**	**698 (33.4)**
How my personality is suited to GP	960 (79.9)	21 (1.7)	230 (42.1)	139 (25.5)	316 (15.1)	1,164 (55.7)
Highly competitive specialty	115 (9.6)	127 (10.6)	26 (4.8)	113 (20.7)	63 (3.0)	281 (13.4)
Less competitive specialty – I was more likely to get a specialty training post	315 (26.2)	41 (3.4)	199 (36.4)	37 (6.8)	310 (19.6)	165 (7.9)
Less competitive specialty – I would have the option to choose where I wanted to work (geography)	440 (36.6)	40 (3.3)	247 (45.2)	29 (5.3)	485 (23.2)	134 (6.4)
Image of General Practice portrayed in the media	**59 (4.9)**	**515 (42.8)**	**24 (4.4)**	**243 (44.5)**	**33 (1.6)**	**835 (40.0)**
Family expectations	268 (22.3)	110 (9.2)	107 (19.6)	94 (17.2)	160 (7.7)	268 (12.8)
The requirement to take computer-based tests	**100 (8.3)**	**98 (8.2)**	**47 (8.6)**	**60 (11.0)**	**49 (2.3)**	**243 (11.6)**
Other aspects of the GP selection system	**156 (13.0)**	**94 (7.8)**	**49 (9.0)**	**63 (11.5)**	**44 (2.1)**	**271 (13.0)**

Table 9.7 compares three groups of respondents in terms of whether different factors influenced them for or against applying to GP: (1) those applying to GP as their first choice, (2) those applying to GP as their non-first choice and (3) those not applying to GP. Factors that are 'tied' to GP as a specialty are shown in bold, with the most frequent reasons in each column shown in pink. No single threshold has been applied for determining 'frequent' across the columns. **Applications to GP are positively influenced by work-life balance provided in the specialty and the patient care that can be provided.** All three groups reported that **how GP is represented in the media influenced them against applying.** The **nature of the GP selection process did not influence many respondents either for or against applying.** Those not applying to GP may also have had poor previous experiences of GP, either in Medical School or post-graduate posts. A significant determinant of specialty choice, as shown in Table 9.4, was candidates' views of **how their personality 'fitted' with the specialty** and this is again evident in Table 9.7.

9.7: WHAT PREDICTS WHETHER DOCTORS APPLY FOR SPECIALTY TRAINING IN GENERAL PRACTICE? ANALYSIS USING PATH MODELLING

A central question in studying recruitment/selection for general practice concerns the factors influencing whether or not a doctor considers applying for GP training, and actually does so. This section looks at how an overall measure of interest in general practice training is related to perceptions of general practice, experience of it, and demographic factors such as year of qualification, sex, ethnicity and place of graduation.

9.7.1 Interest in GP as a career:

A doctor may have no interest at all in GP, they may consider it, and they may apply for it, although in the latter case it may be their first choice or a secondary choice.

- In the questionnaire we compared four categories of doctor, based on two questions:

  i.) **Q4** ("Please indicate your first choice specialty in Round 1 in 2015 recruitment", with General Practice as option 12. Candidates could only indicate a single specialty even if they had made multiple applications).

  ii.) **Q8/12** ("Did you consider applying to these specialties: General Practice", scored as 1 'Did NOT consider', 2: Considered but did not apply', and 3: 'Applied').

- We created a variable called *InterestInGP*, which had four levels:

  i.) **0: No interest.** Answered "Did NOT consider" to Q8/12.

  ii.) **1: Considered.** Answered "Considered but did not apply" to Q8/12.

  iii.) **2: Applied but not first choice.** Answered "Applied" to Q8/12 but did not have General Practice as the answer to Q4.

  iv.) **3: GP was first choice.** Answered "Applied" to Q4.

- Of the 3,838 respondents to the questionnaire, 1556 (40.5%) were in group 0 (No interest) and can be considered as extremely likely to be interested in a career in general practice under any circumstances. 534 (13.9%) were in group 1 and said they had considered general practice, and it is possible that some might be recruited into GP. 546 (14.2%) were in group 2, and had applied for GP training, but it was not their first choice and must therefore represent a pool of doctors who might be influenced into becoming GPs. Finally, 1202 (31.3%) candidates had put GP as their first choice, and therefore presumably are enthusiastic and show commitment to the specialty.

9.7.2 What predicts interest in GP as a specialty choice?:

The main interest here will be in Q14 which had 17 possible influences on becoming a GP. The rubric said, "As you will be aware, the NHS needs 50% of trainees to become GPs. Please indicate the extent to which the following factors influenced your decision whether or not to apply for GP training." Each item could be answered as "Influence AGAINST applying" (1), "No influence" (2) and "Influence towards applying" (3). The detailed items and the range of answers were shown earlier in Table 9.7.

9.7.3 Path modelling (causal structural equation modelling):

The 17 separate reasons in Q14 are difficult to interpret, in part because they are almost certainly correlated, so that those who answer in a particular way to one will generally answer in the same or opposite way to another. Some of the reasons are also 'prior' to others, with 'Experience working in GP', and 'GP experience at medical school' being likely to cause some of the later reasons. As an example a doctor may have been influenced positively by their experience in medical school, which caused them later to have more or a better experience of working in general practice after graduation, and each of those in turn may influence, say, an awareness of the intellectual challenges of GP. It may also be the case that any or all of the factors may be influenced by background demographics, with UK training, year of graduation, ethnicity and sex perhaps being of importance. The solution to problems such as these is to fit a path model.

- **Method.** A variety of methods of model fitting are possible, including multiple regression and other techniques (Cohen and Cohen, 1983; Kenny, 1979; Maruyama, 1998). Here for simplicity we are considering only measured (and not latent) variables and therefore multiple regression is adequate.

- **Causal ordering.** A key aspect of path modelling is the causal ordering in which variables are presumed to occur. Although it is a commonplace of introductory statistics classes that "correlation does not imply causation", in advanced statistics and most science the intention is precisely to infer causation. Sometimes inference of causality is straightforward, especially when events are ordered in time (Davis, 1985). As an obvious example, there is a correlation between height and being male. It is hard to argue compellingly that some people are male because they are taller, and the straightforward causal explanation is therefore that it is being male (having a Y chromosome or whatever) which has a causal influence on a person being taller. We can use the same logic here. Demographic variables can be treated as prior to all other variables. One key measure, a positive experience of GP at medical school, will occur before a positive experience of working in GP and therefore is causally prior to it. And both measures are prior to the majority of other reasons which are current perceptions of GP, rather than referring to past events. Finally, the actual decision to apply or not to GP is the most recent event, and the one of interest[1].

9.7.4 The path model for applying for specialty training in GP:

Figure 9.4 shows an example of the saturated path model which has been used, which for simplicity has only one reason and two demographic factors, as well as the two measures of experience of GP, in medical school and later. Two key principles apply:

- **The measures are shown in rectangular boxes.** These are the actual measurements collected in the study.

- **Causal paths are shown as arrows.** The arrows have a single head indicating that A causes B and not vice-versa.

- **Measures (variables) are laid out in causal order from left to right.** A variable can have a causal effect on any variable to its right, but not vice-versa. If there are two variables one above the other then no causal link between them is possible.

[1] Of course it is always possible to construct convoluted arguments that perhaps, say, having decided to apply for General Practice one then retrospectively views one's earlier experience of GP in medical school as being better. No data from any transverse study can resolve such issues, and that is why longitudinal studies are so powerful in research as they include what people said they felt at a previous time. In the present case we have no appropriate longitudinal data for considering such hypotheses. There is a strong argument that bigger and better such cohort studies are urgently needed, but that is a separate issue.

- **Non-causal correlations are shown by double-headed arrows.** The double-headed arrow between Demographic 1 and Demographic 2 indicates an association between the measures but its causal direction is unclear. These double-headed arrows are only used with what are called 'exogenous' variables.

- **In a saturated model, all variables have causal influences on all variables to their right.**

- **Models are fitted by starting with a saturated model and then removing non-significant variables.**

- **Each path has a strength indicated by the path coefficient.** Path coefficients are not shown in Figure 9.4 but are equivalent to regression coefficients. In the model shown in Figure 9.7 the path coefficients are standardised (beta coefficients), and indicate the change in the second variable in standard deviation units which would be expected for a one standard deviation change in the first variable. They are dimensionless and hence can be compared across different types of variables, and are equivalent to effect size measures such as Cohen's d.

- **Measures can have direct and indirect effects.** Variables A and B can both effect C, but it also can be that A causes B. As a result A can have a direct effect on C as well as an indirect effect via B, with both processes being estimated independently and included in the final model.

» *Figure 9.4: Saturated path model example*

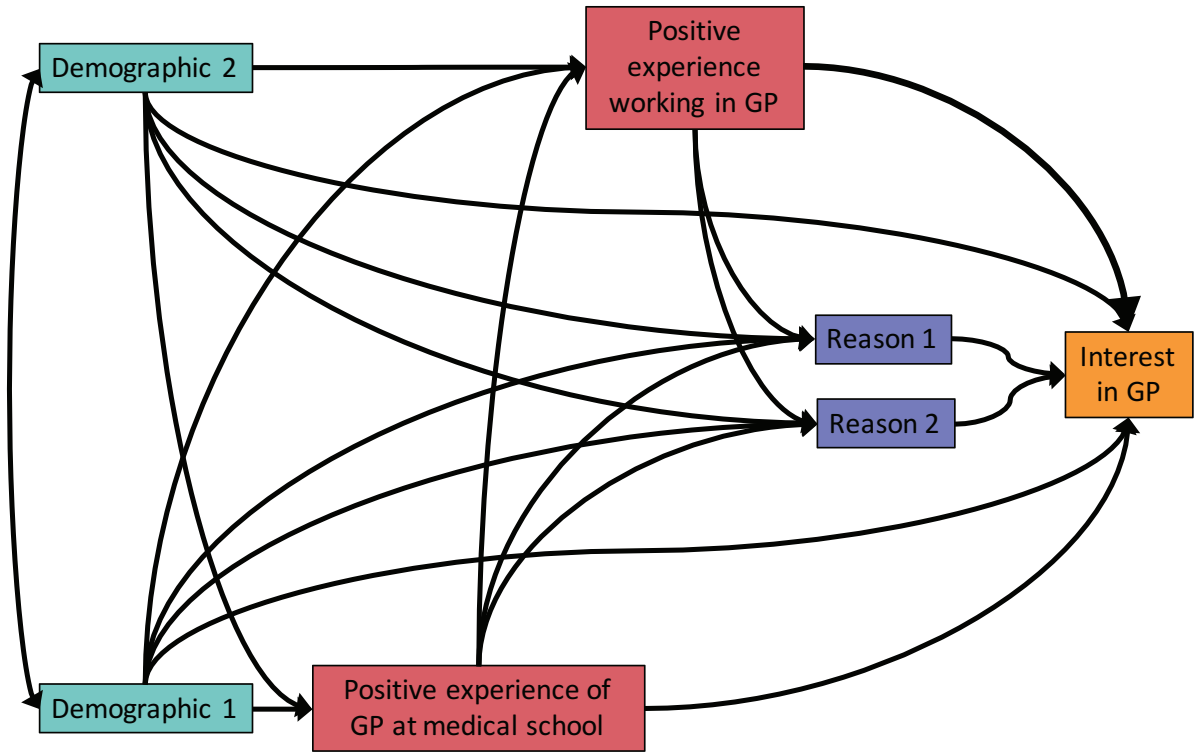


Figure 9.4 shows all potential relationships between background (demographic) measures, experiences of GP at medical school and in work, current reasons for wishing to be a GP, and interest in becoming a GP.

9.7.5 Preliminary analyses:

Before fitting the path model it is useful to carry out some preliminary analyses using conventional statistics.

- **Reasons vs Demographics as predictors.** Multiple regression was used to compare the relative role of demographics and reasons as explanations of *InterestInGP.*

  i.) **Demographics then Reasons.** At step 1 the four demographic variables were entered, giving R=.261, with 6.8% of the variance accounted for. Adding in the 17 reasons increased R to .735, accounting for 53.8% of the variance, accounting for a further 47.3% of the variance in InterestInGP.

  ii.) **Reasons then Demographics.** In the second analysis, the 17 reasons were entered at step 1, accounting giving R=.719, accounting for 51.4% of the variance. Adding in the four demographic variables increased R to .735, as before, with 53.8% of the variance, and increasing the accounted variance by 2.5%.

  It is clear that reasons are a far more powerful predictor of an interest in a career in General Practice than are demographics, although the latter do have some impact.

- **Are all reasons important?** A preliminary multiple regression regressed InterestInGP firstly on all of the 4 demographic measures as well as the two reasons involving previous experience of GP, and then using forward entry regression on the remaining reasons. Variables were entered hierarchically and only five reasons, along with the two experience measures, accounted for reasonable quantities of variance, and therefore only those variables were included in the model. Five reasons accounted for an extra 1% or more of the variance, taking other variables into account, and there were included in the final models. The variables were, in order of entry:

  i.) "How my personality is suited to GP"

  ii.) "Work/life balance"

  iii.) "The patient care that I could provide"

  iv.) "Image of General Practice portrayed in the media"

  v.) "Intellectual challenge of specialty"

  In addition the model included "Experience working in GP" and "GP experience at medical school". Overall therefore seven reasons are looked at in detail.

- **Multiple regression vs binary logistic and ordinal logistic regression.** Technically, using multiple regression to analyse variables potentially violates the assumption of multiple regression that residuals are normally distributed. The analyses using the four demographic variables and just the seven important reasons were re-run using different modelling methods. Binary logistic regression (using applied to general practice as first choice as the dependent variable), showed similar levels and estimates of significance as did multiple regression. Ordinal logistic regression was only possible using the four demographic variables, but that showed the important conclusion that the thresholds for the four categories of *InterestInGP* were at -.378 (SE .071), .212 (SE .070) and .860 (SE .072), with intervals of .590 and .648 suggesting that the scale is equal interval to a first approximation. Finally, to confirm that the assumptions of multiple regression were not being seriously violated, a multiple regression was run with 1000 bootstrap replications, and the effects and their significance were similar to those in a simple regression.

- Interactions of demographic factors. An analysis of variance was carried out which included the three binary demographic variables (UKgrad, BME and sex) and their interactions, as well as Year of Qualification as a covariate. All interaction terms were significant. However repeating the model but also including the seven reasons for being a GP

showed that none of the interaction terms were now significant. Interaction terms can therefore be safely excluded from the path analysis.

9.7.6 The fitted path model:

Figure 9.5 shows the fitted model which used multiple regression, regressing each variable on all of those which are causally prior to it (i.e. to the left). A significance level of 0.001 was used throughout to take account of the repeated significance testing and the large sample size. Sex and BME had occasional missing values, and since these were rare (respondents with no missing data = 3428/3838, 89%) mean substitution was used in analyses. Note that the width of the arrows in figure 9.5 is proportional to the size of the path coefficient, and that negative effects are shown as red dashed lines.

The diagram is complicated, but then there is little reason to expect that complex social phenomena will be simple. Having said that, there are clear patterns visible in the results, all of which are relevant to the question of why some doctors choose to apply for specialist training in general practice.

- **The importance of personality and work-life balance.** The two main drivers of an interest in GP are feeling that one's personality is suited, and the work-life balance. The ability to provide patient care and an intellectual challenge also have small additional effects, and the image of GP portrayed in the media has a small negative effect. There are also small positive effects of working in GP or experience of GP at medical school. There are small effects of demography, with BME doctors being more interested in GP, and male doctors and more recent graduates being less interested. But of all of these, **it is perceived suitability of personality which overrides all else.**

- **The role of experience of GP, both in work and at medical school.** Although positive experiences of GP have small direct effects on an interest in GP, they have large indirect effects. Perceiving that one has a suitable personality for GP is driven strongly by positive experiences of working as a GP, and also by positive experiences at medical school, the two effects being statistically independent. Positive experiences also drive other factors as well, but less than in the effect on personality. Not surprisingly, having a positive experience of working in GP itself is strongly driven by having positive experiences of GP at medical school. The indirect effects can be calculated by multiplying the path coefficients, so experience of working in GP has an indirect effect via personality of .335 x .369 = .123, which is larger than all other direct effects on Interest in GP except Work-Life balance.

- **Demography has only small effects.** There are small effects of demography; the sweeping red lines from year of graduation and being male show that males report lower interest on almost all reasons for an interest in GP than females, as do recent graduates compared with earlier graduates. UK graduates show almost no differences from non-UK graduates, but BME graduates, be they UK or non-UK, are somewhat more positive about GP, particularly on work-life balance. There are direct effects of these demographics on Interest in GP, independent of the reasons (blue boxes); Earlier graduates and BME graduates indicate higher direct interest in GP, as do females to a lesser extent.

- **The image of GP in the media has only small effects.** There is a small influence of "Image of General Practice portrayed in the media" upon interest in a GP career, with those indicating that this was important being less likely to be interested in a GP career (i.e. the path is negative, indicated by a dashed red line in figure 9.5). In retrospect this question could have been better worded[2] as it does not distinguish between a negative images of GP in the media and positive images of GP in the media, and there are both. It seems reasonable in context to presume that it is being interpreted by the respondents as a negative (i.e. critical) image of GP in the media, and we assume that in our interpretation.

9.7.7 Predicting which candidates will apply for GP:

The implications of the path model are that, based on their responses to Q14 in the questionnaire, the best predictors of whether a doctor applies for a training post in GP are their responses to the following statements:

[2] If the questionnaire is to be repeated in a future year, and it would be useful to do so, then this question either needs modifying or supplementing with an additional question.

» *Figure 9.5: The fitted path model*

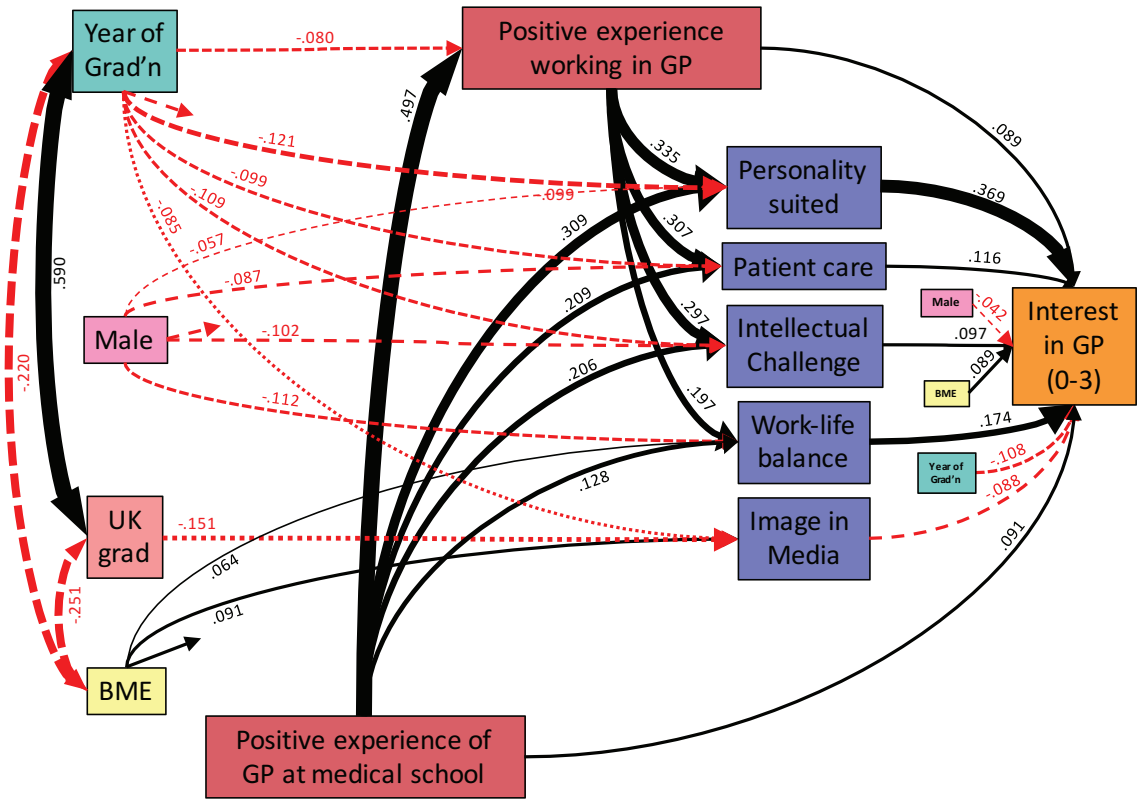


Figure 9.5 shows the causal paths between the various measures. Arrows show the fitted direction of causation, wider arrows indicating stronger effects, and dashed, red arrows indicating negative effects. Standardised path coefficients are shown alongside arrows.

- How my personality is suited to GP
- The patient care that I could provide
- Intellectual challenge of specialty
- Work-life balance
- GP experience at Medical School
- Experience working in GP

A sum of the scores on these six measures (1=Influenced against applying, 3=Influenced towards applying), should predict actually having applied. A summed score (*'GPscore'*) will be in the range 6 to 18. Figures 9.6a and 9.6b shows that *InterestInGP* shows clear differences in a total 'GPscore'[3]. Those who did not consider GP generally had a GPscore of 12 or lower and those with GP as their first choice mainly had scores of 15 and higher. However, there is considerable overlap between 'considered' and 'applied but not 1st choice' candidates.

Figure 9.6c shows an ROC curve for distinguishing those applying for GP as first choice from the other 3 categories; there is a good relationship, allowing reasonable prediction, with the area under the ROC curve being .847 (SE.007). A majority of the effect is due to feeling one has the right personality for GP, with an ROC based on it alone having an area under the curve of .839 (SE .006). A closer look at Figure 9.6a suggests that there is good discrimination between those who put GP as their

[3] We note that in an ideal world we would use a proportion of the data set to identify the factors predicting interest in GP and the remainder to validate the results. However given the large sample size we would expect the results from both samples to be broadly similar i.e. the model would be validated.

first choice and those who never considered it (and that is shown in the ROC curve in Figure 9.6d where the area under the curve is .949 (SE .004). As expected, perceiving one has the correct personality alone predicts very well, with an area under the curve of .914 (SE .006).

It is much more difficult to distinguish between are those who have considered GP or those who have applied but not as a first choice. The range of scores is much greater, as can be seen in the box plot of Figure 9.6b, with individuals covering the entire range of possible scores. What is influencing the decision in those cases is not clear from the present analyses, but may well include factors such as geography, career plans of a partner, family constraints, etc. And of course it could be that some doctors indeed enjoy their experience of GP, feel they would be good at being a GP etc, but despite all of that have a burning interest in ophthalmology or public health or whatever[4]; that happens, and means that prediction will always be less than perfect. It is **possible** that relatively small changes in these most important factors would persuade some candidates to move from 'considered' to having GP as a back-up. In terms of system changes, work-life balance and positive experiences of GP, both at medical school and Foundation training may be the most amenable to change.

9.8: EXPLORATION OF ADDITIONAL VARIABLES AFFECTING INTEREST IN GP

9.8.1 Screening for other possible factors in an interest in GP:

If there are other factors which are important, then a screen through the other variables in the questionnaire might provide some insights. The following analysis therefore looked a number of other measures in the survey on a purely exploratory basis. This was, inevitably, a fishing expedition and the reader should be aware of that. The entry criterion was $p<.001$, but as will be seen below, 32 measures were put into the analysis.

9.8.2 Other factors that may influence an interest in GP:

At the final stage of a multiple regression we entered into the analysis:

i.) The 14 reasons given for making a first choice of specialty. These reasons to some extent overlapped with the reasons for choosing GP, and were included whether or not the first choice was for GP.

ii.) Previous experience in the first choice specialty. Coded as None, 1 to 2 weeks, 3-6 months, 7-12 months, or more than 12 months.

iii.) The 7 reasons given for the first choice of LETB.

iv.) The relative importance of specialty versus location. On a five point scale.

v.) The 7 types of specialty choice advice which might be been received.

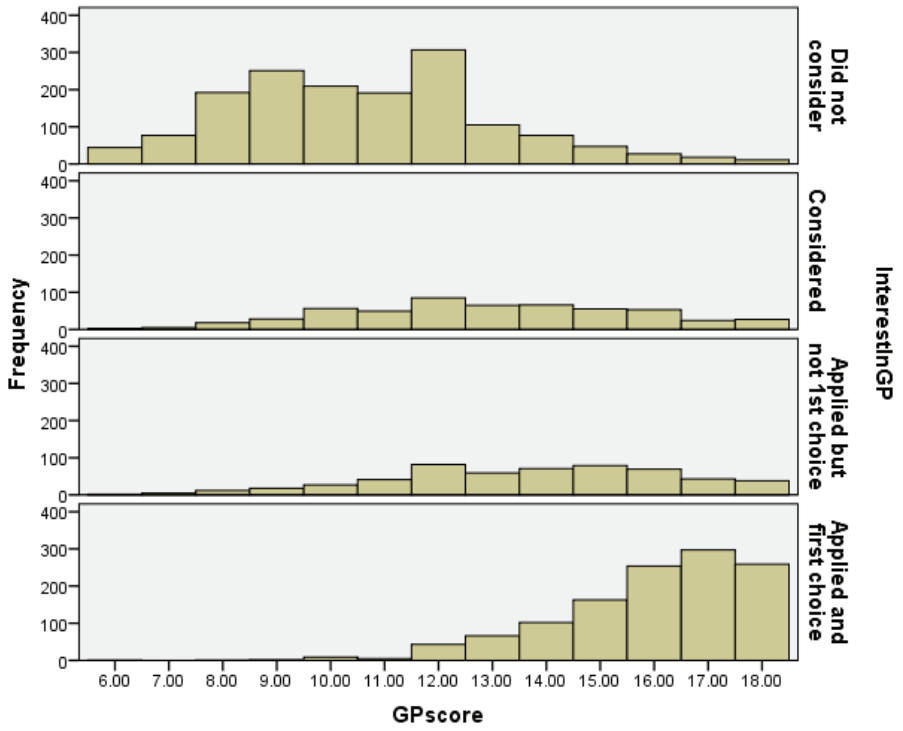
vi.) The number of specialities applied to.

vii.) The number of LETBs applied to.

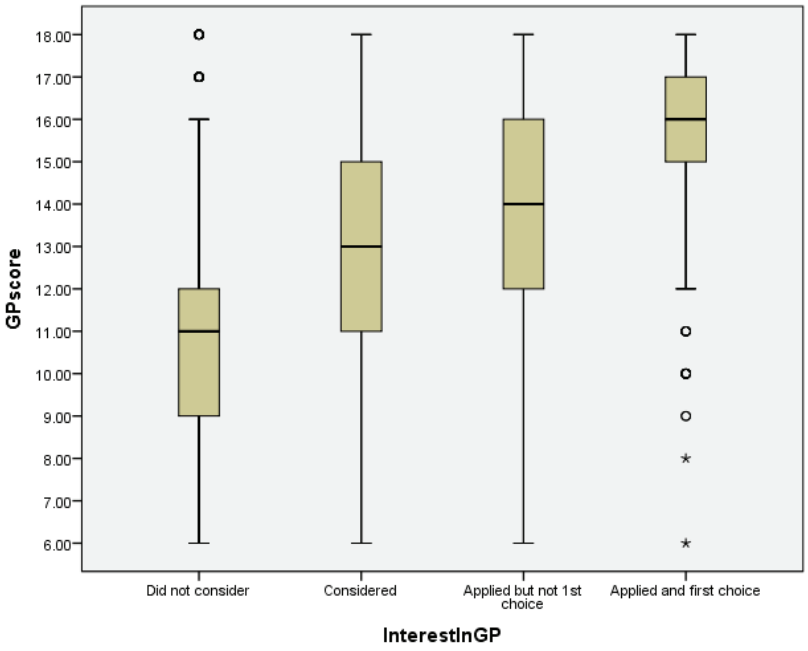
[4]There were 11 doctors who scored 18 on the GP score but did not consider applying for the GP programme, and these had applied to ACCS Anaesthetics, ACCS Emergency Medicine, Acute Medicine, CMT (2), CST, Histopathology, Obstetrics and Gynaecology (2), Paediatrics and Public Health.

» *Figure 9.6 (a) histograms of GPscore for respondents who did not consider GP, considered it but did not apply, applied but not as first choice and those who applied for GP as their first choice; (b) box and whisker plots for the data in (a) the horizontal lines showing the medians, the ends of the boxes the quartiles, the ends of the whiskers 1.5 times the inter-quartile range, and the circles showing outliers; (c) and (d) show standard ROC curves, with (1-Specificity) on the x-axis and Sensitivity on the y-axis, the individual points of which the curves are made up consisting of the thresholds on GPscore from 6/7 to 17/18. (c) contrasts those putting GP as their first choice against all other candidates, and (d) contrasts those putting GP as their first choice against those who never considered GP.*

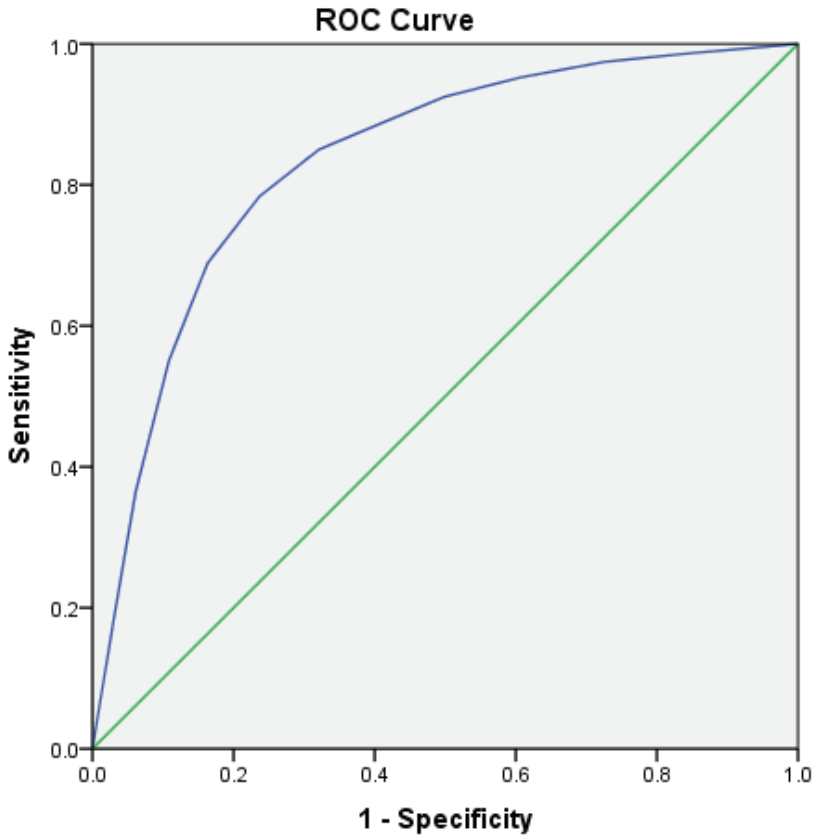
(a)



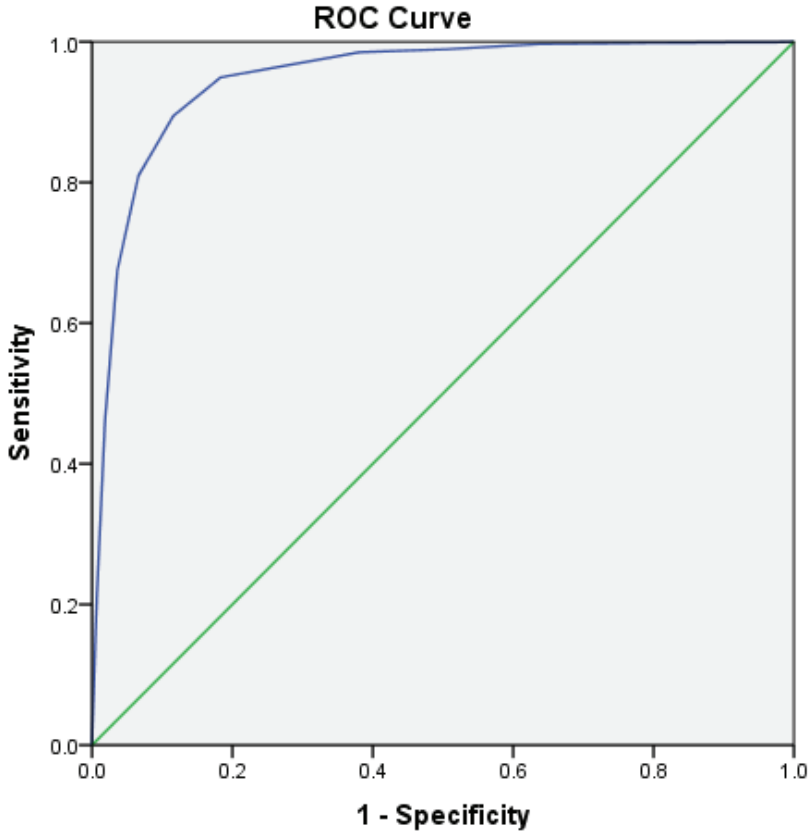
(b)



(c)


ROC Curve
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0

(d)


ROC Curve
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0

9.8.3 The other factors that might increase or decrease an interest in GP:

In order of entry into the regression equation, significant factors at the p<.001 level (but see the caveats above) were:

i.) **Candidates with a greater amount of experience in their first choice were less likely to have an interest in GP (beta= .161).** Most doctors are applying for specialities other than GP, and the more experience they have in their chosen specialty then the less their interest in GP.

ii.) **Being interested more in the location of a training post than in the specialty was associated with more interest in GP (beta=.071).** If location is important, for whatever reason, then GP is more likely to allow a post in a particular region.

iii.) **If the first choice was chosen because it was in a less competitive area of the country then GP was of more interest (beta=.078).**

iv.) **If a specialty was chosen specifically because it was highly competitive, then GP was of less interest (beta= -.061).**

v.) **The more specialties which had been considered then the greater the interest in GP (beta= .071).**

vi.) **If personality was felt to be particularly suited for the first choice, then overall GP was less likely to be of interest (beta=-.048).** As mentioned earlier, the majority of first choices are not for GP, and therefore being suited to a first choice means one is less likely to be interested in GP.

vii.) **The more LETBs considered then the less the interest in GP (beta= .042).** It is not quite clear what is going on here, and the effect size is small.

Together the seven measures accounted for an additional 3.1% of variance, whereas the 5 reasons in Figure 9.5 accounted for 51.4%. However as emphasised above, this part of the analysis is very exploratory.

9.9: FREE-TEXT COMMENTS REGARDING GP APPLICATION

9.9.1 Comments made by applicants:

The question on reasons for applying to GP included an open box for comments. Overall, 218 wrote comments with a mean length of 107 characters (median=58, quartiles 33 to 129, range 1 to 1479). There were many more 'useable'[5] comments from those who did not consider GP (91), than those who considered it but did not apply (33), applied for GP but not at first choice (24) and those who put GP as their first choice (33).

All comments were coded into 11 themes; comments covering multiple themes were counted in each theme. The number of comments in each theme, by each group of respondents, is shown in Table 9.8. Care is required in comparing the number of comments between groups, due to the low proportion of respondents who responded to this question and the different number of doctors in each group.

The full text of these comments can be found in the Appendix, and provide a rather different perspective from the statistical approach adopted in the main text. Many respondents who did not apply for GP simply had a preference for another specialty; although such preferences were occasionally strongly voiced: "I would rather retrain to become an accountant or desk monkey…". The nature of a GP's work was also mentioned by a number of respondents who did not apply to GP, with GPs perceived as working in isolation under the threat of litigation. Workload issues were focused on the short consultation

[5] By 'usable', we mean comments that provided additional reasons for or against applying for GP. Non-usable comments included 'good', 'none', and those indicating that the respondent was not eligible for GP training as they were applying for Public Health training as a non-medic.

time and uncertainty over the future role of GPs in the NHS, and there was a perception that previous benefits of a good work-life balance may soon be eroded.

Comments relating to the training and assessment processes for GP included the short training time as both a positive and negative, with those including this as a negative suggesting that three years was not sufficient to gain the knowledge and skills required to work as a qualified GP. Some respondents commented on the GP selection process, including those who found it difficult and/or stressful to prepare for and take the computer-based Stage 2 tests. While some highlighted that they thought the 'OSCE-style' format of Stage 3 was fairer than the interview-based processes of other specialties, others would have preferred more of a "chance to shine". Again it is important to note that only a minority of respondents provided free text comments, so the extent to which these findings can be generalised is unclear.

» *Table 9.8: Comments by theme*

Theme	Applied to GP as first choice	Applied to GP but not as first choice	Considered GP but did not apply	Did not consider GP
Careers advice and experience of other doctors in GP	1 (-)	2 (-)	0	2 (-)
The working environment	1 (+)	1 (-)	1 (+)	4 (-)
Previous experience in GP	1 (+)	0	2 (-)	2 (-)
Interest in GP and career prospects	7 (+)	1 (+)	1 (+)	37 (-)
		3 (-)	3 (-)	
Match with own skills and abilities	0	0	2 (-)	5 (-)
Prestige of GP and portrayal of GP in the media	1 (-)	4 (-)	2 (-)	5 (-)
The GP selection process	9 (+)	2 (-)	3 (-)	5 (-)
	5 (-)	1 (+)		
GP training and assessment	4 (+)	3 (-)	3 (-)	6 (-)
	1 (-)	1 (+)		
GP workload and contract	0	4 (-)	13 (-)	22 (-)
Clinical duties and risk	1 (+)	3 (-)	2 (-)	5 (-)
Work-life balance	2	0	0	5 (-)

9.10: SUMMARY

The questionnaire, despite it being anonymous and completed online, with little in the way of background data, nevertheless provides some useful insights into specialty choice in general, and in particular into the choice of GP as a specialty of interest as a career. It should be remembered that of course the data were collected when Round 1 was almost completed, and most candidates knew the outcome of Round 1 for them. Nevertheless there was a good response rate, particularly in absolute terms with over 3,800 junior doctors completing it (and we know of no other comparable study). The number of participants means that complex multivariate statistics can be applied with a reasonable chance of them being informative and the effects not representing chance or artefactual associations. The corollary is that such a large N means that effects can be statistically significant but having very small effect sizes. Broad conclusions that can be drawn include:

- Believing that one's **personality is suited to GP** is a major predictor of an interest in a career in GP. It should also be said that while 64% of GP candidates believe their personality is suited to GP, 69% of first choice candidates to other specialities also believe that those specialties are suited to their personality (although presumably they are different personalities for different reasons).

- **Work-life balance** is probably of particular importance in increasing an interest in GP.

- **Positive work experiences of GP or experiences of GP at medical school** increase an interest in GP as a specialty, usually mediated via increasing awareness that one's personality is suited to GP. However positive experiences in other specialties are likely to mean that there is a lower interest in GP.

- **The media image of GP** has little effect on interest in GP careers, although there is a somewhat greater effect in males, and those who graduated more recently, and less of an effect in BME doctors.

- **Demographics** also have effects on interest in GP, with males and recent graduates being less positive in general about all aspects of GP, despite reporting similar work and medical school experience of GP.

- Some candidates regard **location of a training program** as particularly important, and for them GP is of more interest; as it is also for those who do not want highly competitive training programs 12.

- Free text comments in the Appendix, from those with differing degrees of interest in GP, provide a different perspective on what makes GP attractive or unattractive.

APPENDIX 9.1

Specialty recruitment questionnaire

This questionnaire was distributed electronically by HEE to all those applying for CT1/ST1 specialty training posts in Round 1 2014/15 using Survey Monkey. Three emails were sent out to increase the response rate.

1. Specialty Recruitment Feedback Questionnaire – for all applicants

Specialty Applications

***1. At the time of submitting your application, what was your employment status?**

- ◯ In an accredited Foundation Programme Post
- ◯ Completed Foundation previously with FACD, but not working in a training post
- ◯ Working overseas
- ◯ Working in a non-training post
- ◯ Already in specialty training, in another specialty. Please specify below
- ◯ Other (please specify)

***2. In what year did you complete your initial training (e.g. MBChB, MBBS)?**

3. Where did you undertake your initial medical training (e.g. MBChB, MBBS)?

- ◯ UK
- ◯ other EU country
- ◯ non-EU country

***4. Please indicate your first choice specialty in Round 1 in 2015 recruitment.**

***5. Please state reason(s) for your first choice of specialty. (Please select all that apply)**

- [ ] My seniors/consultants advised me that I would be well suited to it
- [ ] I believe that working within the specialty will allow me to provide good continuity of patient care
- [ ] The financial rewards associated with specialty
- [ ] Good work/life balance in specialty
- [ ] Intellectual challenge of specialty
- [ ] Positive experience in clinical posting in specialty
- [ ] Positive experience at Medical School
- [ ] Information received at Medical School
- [ ] Prestige associated with the specialty
- [ ] My personality is well suited to the specialty
- [ ] Highly competitive specialty
- [ ] Less competitive specialty – I was more likely to get a specialty training post
- [ ] Less competitive specialty – I would have the option to choose where I wanted to work (geography)
- [ ] Other (please specify)

***6. Please state your previous experience in your first choice specialty.**

- ◯ 1 week to 2 months
- ◯ 3 to 6 months
- ◯ 7 to 12 months
- ◯ More than 12 months
- ◯ No previous experience

***7. Please state the current status of your first choice specialty application.**

- ◯ Offered and accepted
- ◯ Offered but rejected
- ◯ Withdrew application
- ◯ Not offered - at long-listing (eligibility)
- ◯ Not offered - at short-listing
- ◯ Not offered – after selection centre/interview
- ◯ Other (please specify)

*** 8. Did you consider applying to these Specialties?**

	Did NOT consider	Considered, but did not apply	I applied
ACCS Anaesthetics	◯	◯	◯
ACCS Emergency Medicine	◯	◯	◯
Acute Medicine	◯	◯	◯
Anaesthetics	◯	◯	◯
Broad Based Training	◯	◯	◯
Cardiothoracic Surgery	◯	◯	◯
Clinical Radiology	◯	◯	◯
Community Sexual and Reproductive Health	◯	◯	◯
Core Medical Training	◯	◯	◯
Core Psychiatry Training	◯	◯	◯
Core Surgical Training	◯	◯	◯
General Practice	◯	◯	◯
Histopathology	◯	◯	◯
Neurosurgery	◯	◯	◯
Obstetrics and Gynaecology	◯	◯	◯
Ophthalmology	◯	◯	◯
Oral and Maxillo Facial Surgery	◯	◯	◯
Paediatrics	◯	◯	◯
Public Health	◯	◯	◯

	Did NOT consider	Considered, but did not apply	I applied

***9. Did you consider applying to these LETBs/Deaneries?**

	Did NOT consider	Considered, but did not apply	I applied
Defence	◯	◯	◯
East Midlands	◯	◯	◯
East of England	◯	◯	◯
Kent, Surrey and Sussex	◯	◯	◯
London	◯	◯	◯
Mersey	◯	◯	◯
North East	◯	◯	◯
North West	◯	◯	◯
Northern Ireland	◯	◯	◯
Scotland	◯	◯	◯
South West	◯	◯	◯
Thames Valley	◯	◯	◯
Wales	◯	◯	◯
Wessex	◯	◯	◯
West Midlands	◯	◯	◯
Yorkshire and the Humber	◯	◯	◯

10. Please indicate your first choice LETB/Deanery.

[ ]

***11. Please state reason(s) for your first choice LETB/Deanery. (Please select all that apply)**

- ☐ Good reputation
- ☐ Family/ friends
- ☐ It's where I currently live
- ☐ Interesting/ enjoyable location
- ☐ I have strong personal reasons for needing to be in this region
- ☐ Training reputation
- ☐ Other (please specify)

[ ]

12. How important were specialty and location on deciding which applications to make?

Specialty only	Specialty Slightly	Equal	Location Slightly	Location only
◯	◯	◯	◯	◯

***13. Did you receive any careers advice to help inform your decision on which specialty/specialties to apply for? (Please select all that apply)**

- [ ] Senior trainees and/or consultants
- [ ] Educational supervisor
- [ ] Postgraduate Clinical Tutor
- [ ] College Tutor
- [ ] Careers websites
- [ ] Foundation School events (careers workshops)
- [ ] None
- [ ] Other (please specify)

***14. As you will be aware, the NHS needs 50% of trainees to become GPs. Please indicate the extent to which the following factors influenced your decision whether or not to apply to GP training.**

	Influence AGAINST applying	No influence	Influence towards applying
Advice from my seniors/consultants	◯	◯	◯
The patient care that I could provide	◯	◯	◯
The financial rewards associated with specialty	◯	◯	◯
Work/life balance	◯	◯	◯
Intellectual challenge of specialty	◯	◯	◯
Experience working in GP	◯	◯	◯
GP experience at Medical School	◯	◯	◯
Information, advice and comments received at Medical School	◯	◯	◯
Prestige associated with GP	◯	◯	◯
How my personality is suited to GP	◯	◯	◯
Highly competitive specialty	◯	◯	◯
Less competitive specialty – I was more likely to get a specialty training post	◯	◯	◯
Less competitive specialty – I would have the option to choose where I wanted to work (geography)	◯	◯	◯
Image of General Practice portrayed in the media	◯	◯	◯
Family expectations	◯	◯	◯
The requirement to take computer-based tests	◯	◯	◯
Other aspects of the GP selection system	◯	◯	◯

Other (please specify)

2. Application Process

Oriel

***15. Please rate how you found the Oriel system.**

	Strongly agree	Agree	Neither agree or disagree	Disagree	Strongly disagree
Oriel is easy to use and navigate	◯	◯	◯	◯	◯
Oriel performed well and I experienced no problems with speed and/or performance	◯	◯	◯	◯	◯
The registration process was useful and made the application process easier	◯	◯	◯	◯	◯
The user guidance provided on the Oriel Resource Bank was clear	◯	◯	◯	◯	◯
I found the Oriel Frequently Asked Questions helpful	◯	◯	◯	◯	◯
Overall, my experience with applying for specialty training through Oriel has been positive	◯	◯	◯	◯	◯

***16. What device(s) did you use to access Oriel? (Please select all that apply)**

- ☐ Mobile phone
- ☐ Tablet computer
- ☐ Laptop
- ☐ Desktop computer
- ☐ Other (please specify)

***17. Which web browser(s) did you use to access Oriel? (Please select all that apply)**

- ☐ Google Chrome
- ☐ Microsoft Internet Explorer
- ☐ Mozilla Firefox
- ☐ Safari
- ☐ Other (please specify)

***18. What area of the system did you like most? Please tick only one.**

- ◯ Registration
- ◯ Frequently Asked Questions
- ◯ Document Upload functionality
- ◯ Dashboard and Alerts
- ◯ Online interview booking
- ◯ Other (please specify)

***19. What area(s) of the system do you think needs improvement? Please select all that apply.**

- [ ] Registration
- [ ] Application form
- [ ] Uploading of supporting documentation
- [ ] Online interview booking
- [ ] Frequently Asked Questions
- [ ] Dashboard and Alerts
- [ ] Speed and performance
- [ ] Other (please specify)

20. If you have any other comments you would like to make about Oriel, please add them below.

3. Personal information

***21. Where do you currently live?**

- ◯ Non-EU country
- ◯ Other EU country
- ◯ East Midlands
- ◯ East of England
- ◯ Kent, Surrey and Sussex
- ◯ London
- ◯ Mersey
- ◯ North East
- ◯ North West
- ◯ Northern Ireland
- ◯ Scotland
- ◯ South West
- ◯ Thames Valley
- ◯ Wales
- ◯ Wessex
- ◯ West Midlands
- ◯ Yorkshire and the Humber
- ◯ Other

22. Gender

- ◯ Female
- ◯ Male

23. Please state your year of birth.

24. What is your ethnic group? Choose one option that best describes your ethnic group or background.

Thank you for your time. Your feedback is greatly appreciated and will help to improve the Specialty Selection system.

APPENDIX 9.2

Survey data open comments

The comments listed here were provided by survey respondents when given an "other, please state" option to the question that asked them to consider a list of 17 pre-specified reasons that may had influenced them for or against applying for GP training. 'Useable' comments are provided below (an 'unusable' comment could not be interpreted, e.g. "good"). Comments are sorted by the themes generated by the study authors, but not sorted within themes. Apart from correcting obvious typos where the meaning of the comment is clear, comments have been presented verbatim.

Applied to GP as first choice	Applied to GP but not as first choice	Considered GP but did not apply	Did not consider GP
Careers advice and experiences of other doctors:			
Attitudes of colleagues in other specialities.	My family member completed training,he is very hard working, sincere and committed to GP specialty, He was removed from it as he could not clear CSA. However NHS and patients lost sincere and valuable services. I have known lot of GP trainees and qualified GP doctors, Worked along with them and also know them as a person as they are all in my friend circle.everybody is shocked about his failure. I have seen how much distressing thing for his family.		Family member a GP and miserable.
			Husband advised me against it after his experience working in GP.
The working environment:			
Fed up of management in Hospital Service.	Against - practice can feel isolated/lonely compared to the teams you work with in hospital.		Lifestyle...lots of sitting down.
			relative isolation/solo work of a GP, compared to working within a hospital clinical team or MDT - I prefer the team.
			Don't like working in an office.
			Lonely.
Previous experience in GP (positive or negative):			
Previous experience working at hospital GP Supersurger.		Most strong fact against was that I had yet to work as a doctor in GP, my GP F2 rotation was last, and didn't want to apply for and accept a GP training job before having worked in GP as a doctor, as no idea if this would suit me.	I had a negative experience of GP as a medical student. Having done a GP placement since applying an accepting a CMT post, I would strongly consider it in the future.

Applied to GP as first choice	Applied to GP but not as first choice	Considered GP but did not apply	Did not consider GP
Previous experience in GP (positive or negative) continued:			
			My experience in A&E Minors was similar to a GP placement. Minors basically became a GP surgery with people coming in for prescriptions, foot and skin problems, chronic problems (not acute on chronic) and other things primary care is supposed to deal with and it really put me off, bearing in mind that this is only the tip of the iceberg and GPs have to deal with even more of these problems on top of all the paperwork and QoF. As good as the lifestyle and the money may be, I cannot pursue a career in a specialty purely for financial and lifestyle reasons if I don't have any clinical interest. This may sound a bit harsh, I would rather quit as a doctor than become a GP. By no means am I however belittling primary care, it is arguably one of the toughest specialties these with immense time and organisational pressures.
Interest in GP as a career and career opportunities (positive or negative):			
Ability to have portfolio career.	Academic GP, to become a GPSY.	I would rather retrain to become an accountant or desk monkey, completely unintested in general practice.	I prefer dealing with acutely unwell patients.
Becoming a partner in business. Also the ability to do things outside clinical practce (commissioning etc).	I am already a GP trainee who applied to CMT.	Pefer to work in the hospital rather than primary care.	Really didn't want to be a GP.
GP is a more deployable role within the military.	I already am a GP but could not imagine working full-time in GP for the next 30 years. My application to Psychiatry represents a career change.	I want to be an anaesthetist.	The non-medical aspects of general practice would be infuriating, and are the sole reason I would never consider this as a career.
GPs appear to be the future of integrated care provision.	Wider opportunities in public health engagement.	I decided against applying for GP because I would like to be a specialist in some area of medicine.	Don't want to work as a GP.
Multiple career paths on training completion.		Too much similarity to current career - psychiatry.	I have never had an interest in becoming a GP!
Opportunity for special interest training.		I really enjoy hospital medicine.	Only wanted anaesthetics. Did not to apply other career specialties.

» *Appendix 9.2 continued*

Applied to GP as first choice	Applied to GP but not as first choice	Considered GP but did not apply	Did not consider GP
Interest in GP as a career and career opportunities (positive or negative) continued:			
Variety and opportunity (I get bored easily).		Opportunity to work with children and the variety in GP.	Type of patients that would present to GP. Lack of interest.
		Ultimately, I preferred a different specialty.	I genuinely have no interest in GP as a career. Influenced by lack of patient respect for the healthcare system.
		The flexibility of a GP career (in terms of both time and scope for other career interests), and the degree to which this is imbued in the culture compared to other specialities, is to me the largest positive by some margin.	I hate GP.
			Already a GP and wishing to retrain.
			More interested in other specialties that I find more stimulating.
			Have another specialty that I am more keen to pursue.
			I had an excellent GP placement in foundation years. Nonetheless in some cases simply personality and interest of the trainee will not be best suited to GP training long term.
			I would like to specialise in clinical oncology, not general practice.

Applied to GP as first choice	Applied to GP but not as first choice	Considered GP but did not apply	Did not consider GP
Interest in GP as a career and career opportunities (positive or negative) continued:			
			Prospect of work abroad.
			I want to be a Psychiatrist - which is a different specialty.
			I've been in Psychiatry long enough.
			Did not wish to be a GP, holds no interest for me, wish to specialise and work in hospital envirnoment.
			GP is not for me.
			I don't want to be a GP.
			I find surgery more interesting and enjoyable.
			I wanted to be a surgeon.
			Never wanted to do this career choice.
			I want to be a surgeon.
			I was set on Ophthalmology.
			No GP system in my home country if I choose to return in the future.
			I want to be a surgeon.
			I do not think I would enjoy being a GP.
			It's not that I couldn't see the benefits of being a GP - I just wanted to do paediatrics more. And as GP training is so competitive now I didn't see the point in applying as a back up.
			Just not a career that i am interested in. future of gp career is questionable . i like hospital medicine especially Paediatrics.
			Already working as a GP when applied.
			I am already a qualified GP.

» *Appendix 9.2 continued*

Applied to GP as first choice	Applied to GP but not as first choice	Considered GP but did not apply	Did not consider GP
Interest in GP as a career and career opportunities (positive or negative) continued:			
			Not interested.
			The type of work a GP manages does not appeal to me in any way at all.
Match with own skills and abilities:			
		Wanted more experience in hospital before applying. Don't feel 3 years training is enough.	After spending years studying surgery and obtaining postgraduate degree I was concerned that GP was not suited to my skills set, particularly as minor surgery and practical procedures are now referred to hospital.
		I thought that I would not be suited to situtations with a lot of diagnostic uncertainty. Ophthalmology allows rapid diagnosis in many cases, from exam and investigations.	I would make a terrible GP. I am far too slow. I would hate it.
			I had some experience working as a GP and not felt suited for this job.
			I feel better suited to hospital medicine at present.
Lack of prestige of GP and how GP is portrayed in the media:			
Some consultants advised me against it, others for it. The media was very negative at the time I was applying which was quite disheartening. Lots of stories of GPs quitting due to stress!	Seen as a reject specialty by 'lazy' medical students/doctors. 'Anybody can get in' No prestige.		Media portray GP work and satisfaction very poorly. This, plus lack of prestige, puts people off.
	GP are constantly in the media - always being put down/criticised. Morale of colleagues is very low.		I was put off by the very negative comments from GPs themselves in media (and NHS staff / GPs I spoke to who backed these comments up) advising against GP and reports of stress and burnout etc in media and high drop out rate of GPs (many considering retirement and moving overseas) . The reports that GPs were expected to do more and more for less and very low morale among GPs were very off putting.
			There seems to be a desire by government to scapegoat GPs for all problems in the NHS.

Applied to GP as first choice	Applied to GP but not as first choice	Considered GP but did not apply	Did not consider GP
The GP recruitment and selection process:			
Exam based and fair recruitment system.	Only withdrew my application as unable to sit computer selection exam due to on call rota and this was not changeable.	Applied last year and despite scoring highly in online test and feeling stations had gone well I was deemed unacceptable. The feedback I received was vague and general and didn't allow me to see where I ha done wrong. The communication was poor throughout and I felt after some time that it was a blessing and i would enjoy a more acute specialty more.	I feel that the shortage of GP's is mostly the fault of the GP college in artificially not selecting enough candidates. I know dozens of amazing doctors who applied for GP but were not offered a post or deemed not appointable. In retrospect, this is not their loss, but the loss of GP training who due to a flawed, perhaps even racist selection process, do not now count them as doctors in the GP ranks.
Fairly straight-forward application & selection system.	Interview process based on one type of assessment, does not show broad enough view of applicant. If someone does not feel comfortable in an acting role with zero natural interaction with the examiner then will be at a disadvantage and lacks the opportunity to show many skills required for general practice other than communication (which is understandably the most important but this should be included in the process but with a different style of station also).	I previously applied and despite scoring well in computer tests was deemed unacceptable and given no clear or specific feedback despite asking. As such felt let down by the application process. Most would agree I would make a good gp, am clinically sound and have good bedside manner.	SJT negative.
GP application process does not really allow you to shine - the GP ACF which was what I wanted was a much better process given the opportunity for interview etc.	30 min essay during selection process stage 3.	The computer based tests is the worse that has been introduced to GP selection I was so busy as an FY2 in an academic post that I did not have the time to revise for the test.	The extra test is putting people off.
Less importance placed on points e.g. of publication/ presentation etc.	Was refused GP option to apply due to difficulties getting alternative foundation competency forms signed. Given my previous wide experience and creditials Ifound this very frustrating and sad given the current demand for GPs.		Entrance exam seemed stressful.
Limited account taken of previous experience of training in other specialties - both specialist knowledge and transferable skills.	It was difficult to understand what was required at stage 2 assessment.		Previously did one year and changed entry requirements so can no longer reapply.
Marking should be linear in computer based test.	GP recruitment process was an absolute nightmare therefore I withdrew - offered interview in Newcastle which I could not attend due to work.		

» *Appendix 9.2 continued*

Applied to GP as first choice	Applied to GP but not as first choice	Considered GP but did not apply	Did not consider GP
The GP recruitment and selection process continued:			
No eportfolio required for interview.			
No portfolio needed for application process.			
'OSCE'-style assessment seemed fairer than the panel-based system of other specialties.			
Reliance on objective tests rather than "peacocking" with publications, audits etc.			
Straightforward examination/interview process and less intimidating interview process.			
Working as a locum means my portfolio isn't as competitive as it would have been when I left specialty training 4 years ago. The alternative competences certificate, computer-based tests and selection centre scenarios gave me an opportunity to compete on my ability and aptitude rather than a completely form filling evidence basis.			
Written task prioritisation was the biggest hurdle to cross. Although it seems a fair evaluation process but can be very off putting subjectively if your hand writing is slow or not very neat under such time pressure.			
GP selection far too simplistic and did not feel that it presented an opportunity to sell myself (and by proxy explore my talent and suitability for it) as much as CMT selection did.			
The GP training process, including assessment:			
Friendly and supportive nature of training.	Very well organised in GP comparatively.	Cost of sitting exams.	Poor training in general for GP. 3/4 year program with EU time directive. Not sufficent training for GP!
Portfolio negatively affected choice.	The nature of MRCP, MRCGP exams, for example, where overseas competent trainees fail more than pass easily.	Overreliance on communication skills and holistic skills during programme than on furthering academic and clinical pursuits.	Exams, volume of reflective practice demanded of trainees.

Applied to GP as first choice	Applied to GP but not as first choice	Considered GP but did not apply	Did not consider GP
The GP training process, including assessment:			
Short training.	The reflective learning nonsense.		GP training programme structure - against applying.
Shorter training length.			Very short training programme, I do not feel I could become a competenet GP in 3 years.
Shorter training programme was influence for applying.			Existing exam is difficult to pass for non UK graduate .
Pressures associated with being a GP: work load and contract:			
	The chronic underfunding, damaging top down reorganisation and unrealistic cancer/dementia diagnosis expectations from the government. The ageing population which has not been prepared for, and the general public's complete lack of understanding or appreciation of the huge task facing GPs and the fantastic work which they do.	Expectation to see patient in 8 minutes. Pressure on primary care system.	Likely future of GP working conditions.
	AGAINST: carry sole responsibility for patient, high time pressures i.e. big decisions made within 7min, professional isolation (no team work for fun, feedback or discussion and hence personal development), having to stay late and being unable to hand over jobs, too much paper work,uncertainty over change in working hours.	GP is a political football and the majority of GP's I have met are passionate about the job but do not get the funding or support to allow them to provide the best quality care for their patients. Paperwork involved in GP is too much. Far too stressful and the illusion of a better work life balance.	Patient load.
		GP contract is worsening - this is a negative.	The magnitude of responsibility and paperwork involved in general practice.
		Increasing strain on GPs and large proportion of admin influenced against.	The treatment of the medical profession by the government and governing bodies . I would only training in Anaesthetics, or leave back to Australia. GPs have remarkably little autonomy in the practice as compared to Australia, the pay is significantly less, and the work life balance is much worse. Doctors are sick of being treated as a numbers game.

» *Appendix 9.2 continued*

Applied to GP as first choice	Applied to GP but not as first choice	Considered GP but did not apply	Did not consider GP
Pressures associated with being a GP: work load and contract continued:			
		Nature of GP work - less supported.	I never really considered applying for GP training. Current working conditions are massively impacting on / restricting the potential / making it difficult to deliver good quality care in primary care (10min consultation slots, no 7 day working, not enough / clearly defined referral pathways for each practice, breakdown / delays in comunication hospital - GP - hospital, etc).
		The intense pressures GPs are put under including time management means that to be a good GP who runs on time, you need to be ruthlessly efficient, and that does not fit with the way I think. I would also feel wrong not having the time to show compassion when needed.	A general sense that general practice in the UK is in trouble and many GPs are wanting to leave the profession.
		My impression is that GPs are inadequately supported and resourced to provide the level of care expected/required of them.	increasing workload and bureaucracy in specialty!
		GPs are too busy / unable to provide satisfactory or safe care within 10 minute consultations.	Running a business, having control over your income and budget. Competing in a free(er) market. These are all attractive benefits that hospital medics aren't allowed access to.
		The fact that the demands on GPs time are only increasing with decreasing financial rewards. Ten minutes is not sufficient to see a patient.	Too much bureaucracy and time restraints.
		Patient expectations and limited resources.	Workload, burnout, dealing with non-medical issues outside remit of a medical practitioner.
		The main reason I didn't apply was I feel GPs are too pressured to provide good patient care at the moment and that is something I would find difficult.	10 minute appointments - influence AGAINST.
			Against applying - Heavy workload and responsibility - stress, ensuing reduced rights and pay under conservative government.

Applied to GP as first choice	Applied to GP but not as first choice	Considered GP but did not apply	Did not consider GP
Pressures associated with being a GP: work load and contract continued:			
			Huge time pressure, massive public expectation, constant difficulties with government changes to policy.
			Uncertain future of GP role stability (hours & pay) influenced against applying.
			Why in the name of Yahweh would I want to be a GP, I will do foreign exams, & get lost to the US & earn hundreds of thousands of dollars in my uncles private hospitals. . And use these earnings to help the needy in the developing world. Unless filthy western intervention or anti-BDS movement block this.
			NHS and government-induced pressures, responsibility and potential for legal issues.
			Although I loved my GP FY2 placement and believe I would enjoy GP the changes in the system since 2013 and the current unmanageable GP workload makes it a very unattractive specialty currently.
			Increased demand and too few staff.
Pressures associated with being a GP: clinical nature of work:			
Working in a community based setting.	Managing clinical risk.	APPEARS TO STRESSFUL AND UNREALISITIC EXPECTATIONS FROM PATIENTS.	Lower acuity of patients, less team work involved, lack of procedural experience.
			Being sued for missing something.
			Litigation potential and lack of support from BMA / GMC.
			Patient Cohort and demographics with associated expectations and demands are challenging.

» *Appendix 9.2 continued*

Applied to GP as first choice	Applied to GP but not as first choice	Considered GP but did not apply	Did not consider GP
Work-life balance:			
More realistic for LFTT in the future.			I always feel (maybe wrongly) that I can turn to GP when my hospital career stalls/ I want to live a more balanced life.
No night duty mainly which is quite intimidating and weekend duty-much less extent.			The threat of 7 day working for GPs meaning the work/life balance factor may soon no longer be better for GPs than other doctors.
			I am already MRCGP and am retraining in another specialty - GP has NO work life balance at the moment and you treat patients on a conveyor belt.
			The future of GPs - working hours.
Multiple reasons:			
	The GP representation in the media is why I didnt want to do it. Also I think you are very vulnerable as a GP. I dont want to take on the risk, the targets and family abuse. Also I hate the way hospital doctors talk about GPs put me off. I don't want to have to argue with hospital doctors to get my patient into hospital.	Too much paperwork/portfolio requirements for GP but mainly it is the negative portrayal in the media and poor financial compensation.	I do not find General Practice interesting, and there are inordinate pressures on GPs at present which make this an unattractive career proposition.
	Difficult to know what GP's job will be in the future. Unsure how well I would do at GP exams. Unsure what specialities I would have to do in training.	I was discouraged by the concept of learning ONLY AND ALL of medicine and NICE guidelines in order to pass tedious medical exams, the majority of patients wanting secondary gains from presenting, the extremely long hours as a partner, the idea of also running a business and needing to budget, the lack of specialty interests (only being about 4/5) that a GP can do, the pressures of the CCG, patient expectations, lack of resources in the NHS etc. I did like having the time to discuss concerns with senior collegues, better environment for my health in terms of labour (as a junior doctor we do about 20,000 steps a day running around the hospital, many of my collegues now have sciatica due to this).	Over-stretched system, want to be a hospital doctor for specialist intellectual challenge. GP challenges not the same as previously.

Applied to GP as first choice	Applied to GP but not as first choice	Considered GP but did not apply	Did not consider GP
Multiple reasons continued:			
	GP unfairly bare the brunt of any failings - at all levels. Failure of adequate discharge planning, social care/support and especially central government. They represent excellent value for money - but are portrayed to be poor value, by media and senior hospital doctors.	Isolated position in clinical periods with ever increasing workload, bureaucracy and pressure. Work life balance was the main draw at first but this is only likely to get worse.	I feel I would be poorly suited to GP and although my university GP placement was good, I really didn't enjoy the nature of the work.
	To be honest, I would not have applied for GP had I not been put in a position by the previous Tory government with regard to having impending visa problems and being unable to apply for my specialty of choice, even though I had done all of my training and undergraduate education in the UK. It was well and truly a Plan C backup (and I had 3 back-up plans). In my opinion, the legislation imposed upon GPs at present is making the specialty more and more unpalatable. I have had 2 friends who were set on GP from undergraduate studies change their mind in FY1/FY2 after either working in GP or speaking to GPST trainees. The UK government should seriously reconsider certain aspects of the current contract, such as 24/7 service and 10 min consultations. From personal experience, 10 min consutation is not enough time to deal with anything in GP except maybe a straighforward chest infection. But those who are drawn to GP appreciate the holistic side of care and therefore rushing patients out the door is unpalatable to them. Re 24/7 care - if the patient is unwell or suspicion of unwell, we need to send them to A&E regardless; and patients are going to attend A&E regardless if they believe they are seriously unwell. Better to improve recruitment and working conditions in A&E. The impending 24/7 GP provision of care is what is turning prospective trainees away from the specialty and steering current GPs to emigrate to other countries or take early retirement.	One third of GPs are thinking of retiring in the next 5 years and I struggled to find articles portraying positive aspects of the job. I am not very good at leaving work at work and, reading reports of working in GP, it seems that's a real problem, because there's so much time pressure you're always asking yourself: did I miss something? I don't think I could cope with that. I didn't get a GP post in F2 which may have given me a better idea.	Do not think there is much in the way of work/life balance benefits with GP as previously. Think General Practice is changing in that more is expected on their work load. Less GPs, therefore more burden on the ones remaining. Really dislike the amount of reflective work that is required in general practice training.

» *Appendix 9.2 continued*

Applied to GP as first choice	Applied to GP but not as first choice	Considered GP but did not apply	Did not consider GP
Multiple reasons continued:			
			Poor media coverage and poor long term work outlook.
			The portrayal in media and social media of how difficult life is as a GP - how stressed and unappreciated they are. Also, you don't have access to resources like CT scans, blood tests and specialist advice very easily at all.

Chapter 10

Applications to specialty training 2015

Chapter 10.

Applications to Specialty Training 2015

10.1 INTRODUCTION

This chapter describes the population of candidates to specialty training in 2015, their application decisions and their progress through the selection processes in each specialty, including GP. The data include all ST1/CT1 specialties, plus Trauma and Orthopaedic Surgery (ST3), Chemical Pathology (ST3) and Emergency Medicine (ST4). We focus on exploring applications to GP, considering in particular candidates' progression through the selection process in each round, candidates' preferences when two offers are received and appointability decisions when two applications are made. Unless otherwise stated, the data used in this chapter are from the 'Oriel' applications dataset.

8.2 OVERVIEW OF SPECIALTY APPLICATIONS

Table 10.1 shows the number of applications per candidate[1]. The majority of those who applied in each round just applied to one specialty: 8,572/1,1782 = 73% in round 1 and 1,628/ 2,064 = 79% in round 2. Only 4.2% of candidates made 4 or more applications across the 2 rounds. Multiplying the frequency by the number of applications per candidate indicates there were 1,6340 applications from the 1,1782 round 1 candidates i.e. 1.39 applications per candidate. For round 2, this figure is 1.33 applications per candidate.

Figure 10.1 shows the number of Round 1 candidates per post for ST1/CT1 specialties (blue bars; left axis) and fill rate (orange markers; right axis), sorted by the number of candidates per post. Data on the number of candidates use the Oriel dataset; data on the number of posts are taken from the HEE website[2]. GP had the lowest number of candidates per post (1.3) and almost the lowest fill rate (69%). (Medians across all ST1/CT1 specialties 2.6 candidates per post and 99% fill rate.) In general, the specialties with the highest number of posts were least competitive and had the lowest fill rates.

Table 10.2 shows the number of applications to each specialty in each round according to outcome. Oriel does not distinguish between candidates rejected at long listing (GP Stage 1) from those rejected at shortlisting (GP Stage 2). GP is one of three specialties with an offer acceptance rate (out of the number of candidates) over 50% for round 1 (the others are Paediatrics and CMT). This is, however, more a consequence of competitiveness (the number of applications per place) and/or the 'height of the bar' (the probability of being offered a place), than of the probability that an offer, if made, is accepted. In Round 1, the offer acceptance rate ranges from 49% in BBT to 100% in Cardio-thoracic surgery and Community Sexual and Reproductive Health. The low rate for BBT combined with its low fill rate shown in Figure 10.1 suggests that this may be a 'back up' choice for many. For GP, the offer acceptance rate was 83%, higher than that for CMT (78%) but below that for Paediatrics (87%). Across all non-GP specialties combined, the offer accept rate was 79%.

10.3 GP APPLICATIONS

General Practice had the highest number of applications across all specialties, with 41% of first round candidates applying; the next most popular specialty was CMT with 21% of candidates. In round 2, 64% applied for GP and 28% for CMT. These higher percentages in round 2 are however partly due to a reduced number of specialties with vacancies. As noted above, GP was the least popular specialty in terms of the number of candidates per post.

[1] Specialties such as neurosurgery that only have one application round are included in Round 1.

[2] http://specialtytraining.hee.nhs.uk/specialty-recruitment/competition-ratios/2015-competition-ratios/

» *Table 10.1: Number of applications per candidate*

	Round 1		Round 2		Total	
Applications	**Frequency**	**Percent***	**Frequency**	**Percent***	**Frequency**	**Percent***
0	(898)	(7.1)	(10,616)	(83.7)	(204)[3]	(1.6)
1	8,572	72.8	1,628	78.9	8,282	66.4
2	2,334	19.8	275	13.3	2,847	22.8
3	623	5.3	102	4.9	819	6.6
4	156	1.3	42	2.0	296	2.4
5	48	0.4	10	0.5	117	0.9
6	18	0.2	5	0.2	49	0.4
7	11	0.1	1	0.0	22	0.2
8	10	0.1	-	-	14	0.1
9	7	0.1	-	-	13	0.1
10	-	-	1	0.0	5	0.0
11	1	0.0	-	-	4	0.0
12	-	-	-	-	2	0.0
13	1	0.0	-	-	3	0.0
14	-	-	-	-	2	0.0
15	1	0.0	-	-	-	-

Looking more closely at GP round 1 in Table 10.2[4], 4,837 applied, 51% (2477) accepted offers; 24% withdrew (13% before interview and 11% after offer made). The remaining 25% either were not shortlisted (8%), not appointable (12%) or appointable but still no offer made (4%)[5]. Clearly to increase the number of accepted offers would require encouraging more applications, persuading candidates not to withdraw and making offers to everyone who is appointable.

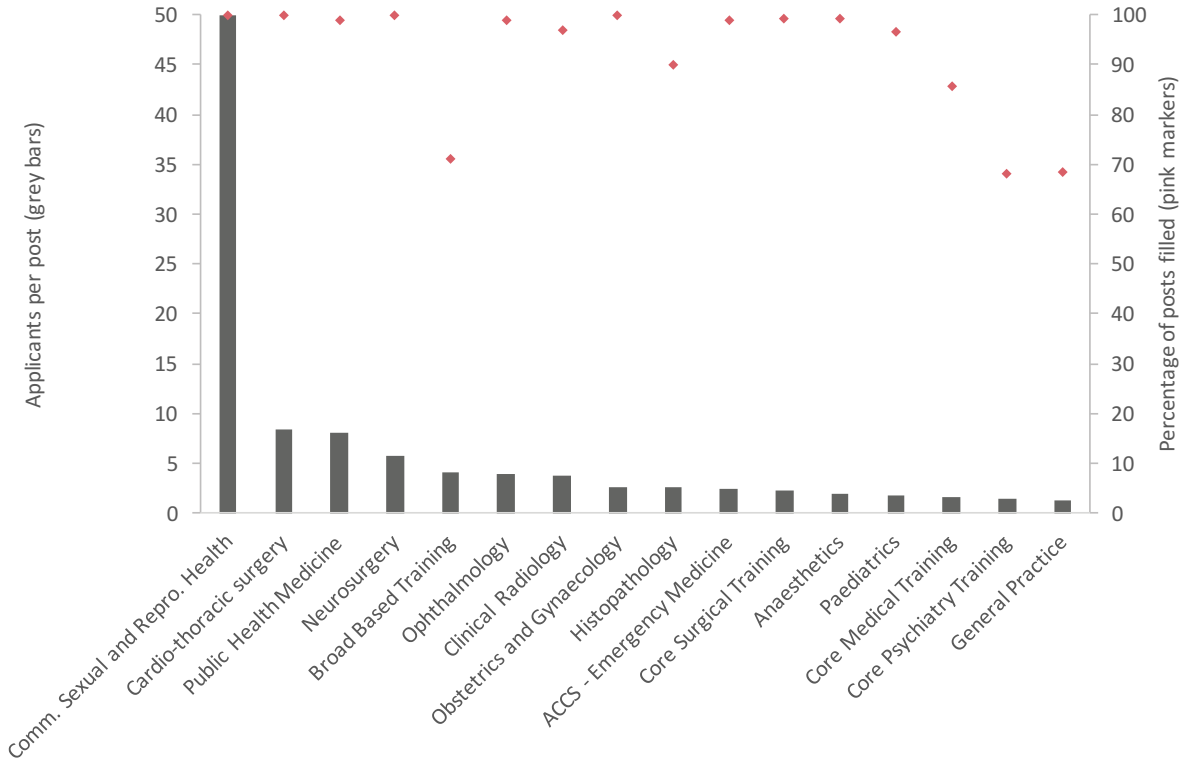
As shown in Table 10.3, 814 doctors applied to GP in both rounds[6]; 310 of these were made an offer in round 2 (38%) and 269 (33%) accepted an offer. Excluding candidates who were shortlisted but withdrew from round 2, the likelihood that re-application will be successful (in terms of being made an offer) was 66% if attended Stage 3 in round 1 and 48% if not shortlisted in round 1. **Re-applications should therefore be encouraged, as there is a fairly high probability of success.**

[3] These 204 candidates have no recorded applications, although a few of them have shortlisting scores, suggestive that there are missing data in Oriel.

[4] These figures from Oriel do not match those that we received directly from GP NRO, which indicated that there were 2559 accepted offers from 5112 applicants (50%) in Round 1.

[5] Presumably due to a lack of available posts in LETBs where these applicants were willing to work.

» *Figure 10.1: Round 1 candidates per post and percentage of posts filled (ST1/CT1).*



10.4 MULTIPLE APPLICATIONS INVOLVING GP

Of those applying to GP in Round 1, substantial percentages of those applying to GP also applied in round 1 to BBT specialties, as shown in Table 10.4: Core Medical Training (13%), Psychiatry (6%), Broad Based Training (5%) and Paediatrics (4%). Other specialties selected by over 2% of GP candidates were: Ophthalmology (2%), Obstetrics and Gynaecology (3%), Clinical Radiology (6%), Anaesthetics (4%), Public Health (3%), Acute Core Common Stem Emergency Medicine (4%) and Core Surgical Training (4%). In Round 2, even higher percentages also applied to most BBT specialties: BBT (8%), CMT (15%) and psychiatry (10%), but not paediatrics (1%).

10.5 OFFER ACCEPTANCE DECISIONS BETWEEN GP AND OTHER SPECIALTIES

For those candidates receiving more than one offer, their offer acceptance decisions provide some insight into candidates' relative specialty preferences. In total, **454 candidates chose between GP and non-GP offers in round 1: 281 (62%) chose the non-GP specialty and 173 (38%) GP.**

Figure 10.2 shows the percentage of candidates in each dyad accepting their GP offer, using the number of candidates accepting either offer as the denominator in each dyad (i.e. candidates declining both offers are excluded). Only specialty dyads where at least 10 candidates accepted one of the two offers are included to preserve anonymity. Across those specialties included in Figure 10.2, the percentage of candidates accepting their GP offer ranged from 12.5% when paired with Clinical Radiology to 57% when paired with Obstetrics and Gynaecology; the only specialty where GP had an accept rate over 50%.

[6] The data suggest that 3 people applied to round 2 despite already accepting an offer. 26 are reported as not being shortlisted in round 2, but had been in round 1. Using GP NRO data 842 applied to both Rounds 1 and 2, with 276 (33%) accepting a round 2 offer. Therefore, it seems that the Oriel data are missing a few applicants and there are other inconsistencies in application outcomes between the two datasets.

[7] Chemical Pathology, Public Health Medicine Round 2 and Acute Care Common Stem - Emergency Medicine Round 2 all deemed some candidates appointable, but did not make any offers. This could be due to errors in the data set.

» *Table 10.2: Specialities applied to, with outcomes[7].*

	Applied but not Shortlisted		Shortlisted but withdrawn		Interviewed but not Appointable		Appointable but no offer made		Offer made but withdrew/ declined/offer elsewhere		Offer accepted		Total*	
General Practice, R1	407	8%	638	13%	591	12%	210	4%	514	11%	2,477	51%	4,837	41%
Core Medical Training, R1	125	5%	105	4%	350	14%	251	10%	372	15%	1,328	52%	2,531	21%
Core Surgical Training, R1	49	3%	49	3%	464	32%	77	5%	194	14%	600	42%	1,433	12%
General Practice, R2	126	10%	416	31%	312	24%	14	1%	65	5%	393	30%	1,326	64%
Anaesthetics	42	3%	22	2%	211	17%	287	23%	88	7%	623	49%	1,273	11%
Clinical Radiology, R1	88	9%	134	14%	237	25%	189	20%	49	5%	239	26%	936	8%
Acute Care Common Stem - Emergency Medicine, R1	75	8%	23	3%	165	19%	97	11%	165	19%	359	41%	884	8%
Paediatrics, R1	47	6%	10	1%	89	11%	174	21%	64	8%	430	53%	814	7%
Public Health Medicine, R1	115	16%	351	49%	96	13%	51	7%	12	2%	87	12%	712	6%
Core Psychiatry Training, R1	56	8%	13	2%	89	13%	132	19%	83	12%	317	46%	690	6%
Obstetrics and Gynaecology, R1	45	7%	26	4%	114	18%	178	28%	31	5%	238	38%	632	5%
Core Medical Training, R2	76	13%	79	13%	172	29%	6	1%	31	5%	223	38%	587	28%
Ophthalmology, R1	18	5%	85	23%	137	37%	23	6%	15	4%	94	25%	372	3%
Broad Based Training, R1	46	13%	72	21%	46	13%	59	17%	62	18%	59	17%	344	3%
Core Psychiatry Training, R2	19	9%	9	4%	79	37%	34	16%	7	3%	67	31%	215	10%
Histopathology, R1	25	12%	5	2%	80	39%	5	2%	17	8%	71	35%	203	2%
Trauma and Orthopaedic Surgery	24	12%	87	43%	12	6%	61	30%	2	1%	16	8%	202	2%
Neurosurgery	16	9%	72	41%	9	5%	34	19%	14	8%	30	17%	175	1%
Broad Based Training, R2	35	21%	68	40%	22	13%	18	11%	10	6%	16	9%	169	8%

» *Table 10.2: continued.*

	Applied but not Shortlisted		Shortlisted but withdrawn		Interviewed but not Appointable		Appointable but no offer made		Offer made but withdrew/ declined/offer elsewhere		Offer accepted		Total*	
Emergency Medicine, R1	28	24%	11	9%	31	27%	31	27%	1	1%	14	12%	116	1%
Core Surgical Training, R2	74	70%	-	0%	20	19%		0%	1	1%	10	10%	105	5%
Community Sexual and Reproductive Health	11	11%	73	73%	-	0%	14	14%	-	0%	2	2%	100	1%
Histopathology, R2	1	1%	3	3%	55	61%	4	4%	5	6%	22	24%	90	4%
Cardio-thoracic surgery	10	15%	39	58%	-	0%	10	15%	-	0%	8	12%	67	1%
Obstetrics and Gynaecology, R2	11	19%	23	40%	11	19%	1	2%	5	9%	6	11%	57	3%
Ophthalmology, R2	3	6%	31	61%	12	24%	2	4%	-	0%	3	6%	51	2%
Clinical Radiology, R2	5	10%	37	74%	5	10%	2	4%	-	0%	1	2%	50	2%
Emergency Medicine, R2	21	55%	-	0%	11	29%	1	3%	4	11%	1	3%	38	2%
Paediatrics, R2	5	19%	14	52%	2	7%	5	19%	-	0%	1	4%	27	1%
Acute Care Common Stem - Emergency Medicine, R2	12	48%	11	44%	-	0%	2	8%	-	0%	-	0%	25	1%
Chemical Pathology	-	0%	14	74%	1	5%	4	21%	-	0%	-	0%	19	0%
Public Health Medicine, R2	-	0%	6	67%	-	0%	3	33%	-	0%	-	0%	9	0%

*% applications to that round; n=11782 for round 1 and 2064 for round 2
R1= round 1, R2= round 2

» *Table 10.3: Cross-tabulation of Round 1 and Round 2 applications for GP, Oriel data*

		Applied but not Shortlisted	Shortlisted but withdrawn	Interviewed but not Appointable	Appointable but no offer made	Offer made but withdrew/ declined/offer elsewhere	Offer accepted	Total*
Outcome: General Practice, round 1	Applied but not Shortlisted	19	78	55	1	5	64	222
	Shortlisted but withdrawn	13	138	46	0	5	43	245
	Interviewed but not Appointable	12	25	101	2	26	148	314
	Appointable but no offer made	0	1	1	0	1	4	7
	Offer made but withdrew/ declined/offer elsewhere	1	2	7	0	4	9	23
	Offer accepted	0	0	2	0	0	1	3
	Total	45	244	212	3	41	269	814

» *Table 10.4: Multiple applications with GP*

	General Practice, round 1, n=4837		General Practice, round 2, n=1326		Total
	N	%	N	%	
Paediatrics, round 1	189	3.9%	37	2.8%	814
Paediatrics, round 2	7	0.1%	8	0.6%	27
Trauma and Orthopaedic Surgery	23	0.5%	11	0.8%	202
Ophthalmology, round 1	109	2.3%	33	2.5%	372
Ophthalmology, round 2	11	0.2%	9	0.7%	51
Cardio-thoracic surgery	8	0.2%	4	0.3%	67
Emergency Medicine, round 1	45	0.9%	19	1.4%	116
Emergency Medicine, round 2	8	0.2%	15	1.1%	38
Neurosurgery	18	0.4%	10	0.8%	175
Obstetrics and Gynaecology, round 1	165	3.4%	45	3.4%	632
Obstetrics and Gynaecology, round 2	15	0.3%	11	0.8%	57
Community Sexual and Reproductive Health	57	1.2%	19	1.4%	100
Chemical Pathology	6	0.1%	12	0.9%	19
Histopathology, round 1	68	1.4%	31	2.3%	203
Histopathology, round 2	30	0.6%	38	2.9%	90
Clinical Radiology, round 1	277	5.7%	105	7.9%	936
Clinical Radiology, round 2	11	0.2%	12	0.9%	50
Anaesthetics	168	3.5%	42	3.2%	1273
Public Health Medicine, round 1	127	2.6%	35	2.6%	712
Public Health Medicine, round 2	2	0.0%	3	0.2%	9
Acute Care Common Stem - Emergency Medicine, round 1	212	4.4%	73	5.5%	884
Acute Care Common Stem - Emergency Medicine, round 2	5	0.1%	4	0.3%	25
Broad Based Training, round 1	241	5.0%	52	3.9%	344
Broad Based Training, round 2	72	1.5%	108	8.1%	169

» *Table 10.4: Continued*

	General Practice, round 1, n=4837		General Practice, round 2, n=1326		Total
	N	%	N	%	
Core Medical Training, round 1	644	13.3%	123	9.3%	2531
Core Medical Training, round 2	121	2.5%	201	15.2%	587
Core Psychiatry Training, round 1	292	6.0%	77	5.8%	690
Core Psychiatry Training, round 2	90	1.9%	127	9.6%	215
Core Surgical Training, round 1	202	4.2%	48	3.6%	1433
Core Surgical Training, round 2	21	0.4%	28	2.1%	105
Total	**4837**	**-**	**1326**	**-**	**12680**

Note candidates applying to two or more other specialties will be included more than once in this table.

Fewer candidates received multiple offers (including GP) in Round 2; of those accepting one offer, 14/30 accepted their GP offer (47%). The numbers are too low to make comparisons between specialties.

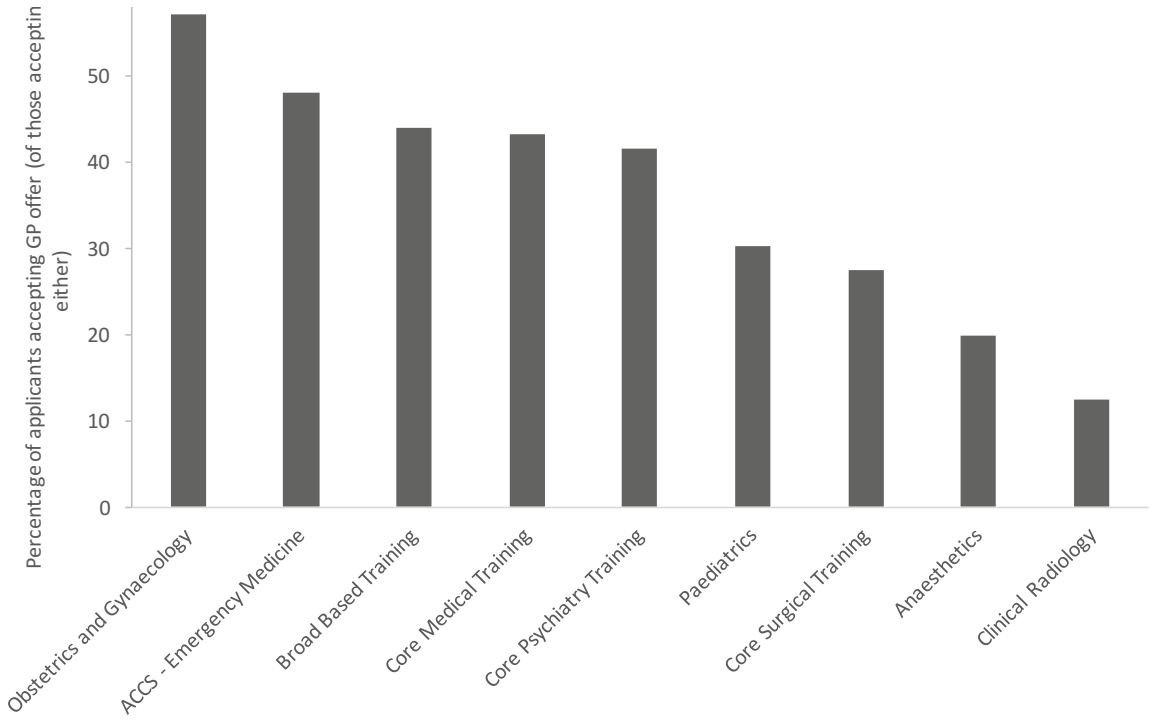
10.6 "HEIGHT OF THE BAR": COMPARING APPOINTABILITY IN GP AND OTHER SPECIALTIES

Above we considered candidates' preference for GP compared with other specialties when they had been given offers for both. Here we consider whether GP or other specialties are more likely to judge the same candidate to be appointable. Note that in competitive specialities, some appointable candidates will not be offered posts: candidates were deemed appointable if coded: "Appointable but no offer made", "Offer made but withdrew/declined/offer elsewhere" or "Offer accepted". They were not appointable if "Applied but not Shortlisted" or "Interviewed but not Appointable". If they didn't apply or "Shortlisted but withdrawn", they were treated as missing data.

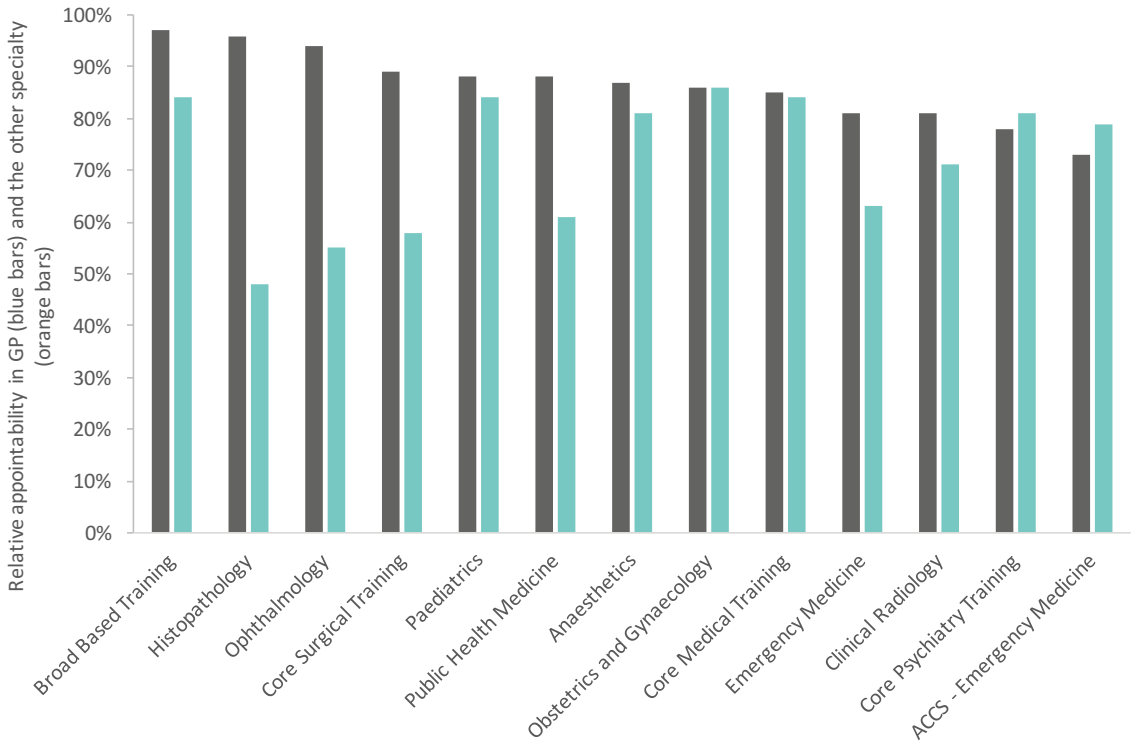
We calculated the relative appointability for each Round 1 GP/other specialty dyad as the percentage of candidates considered appointable in at least one specialty who were appointable in GP and its specialty pair. For example, for the GP-Paediatrics dyad, 98 candidates were considered appointable in both, 22 in GP only, 17 in Paediatrics only and 20 in neither. The relative appointability to GP was (22+98)/(22+98+17) = 88% and to Paediatrics was (17+98)/(22+98+17) = 84%. The relative appointabilities within each GP/other specialty dyad are shown in Figure 10.3, excluding any dyads with less than 10 candidates considered appointable in at least one specialty.

Appointability in GP is only lower than Acute Core Common Stem – Emergency Medicine (absolute difference -6%) and Core Psychiatry Training (-3%). The GP/other specialty dyads with the largest absolute differences (fewer candidates considered appointable in the other specialty than in GP) are Histopathology (absolute difference 48%), Ophthalmology (39%), Core Surgical Training (31%) and Public Health (27%).

» *Figure 10.2: Percentage of candidates accepting a GP offer by specialty declined*



» *Table 10.3: Relative appointability by GP/other specialty dyad (round 1)*



Across all specialties, 85% of candidates considered appointable by GP and/or another specialty were considered appointable by GP and 79% by the other specialty. **This very crude analysis suggests that overall 'the bar' is a little lower in GP than other specialties i.e. it is easier to be deemed appointable.** Alternatively, candidates may have higher levels of the competencies assessed during GP selection than those assessed during selection for other specialties (i.e. based on their skills and abilities, are more suited to GP). Across all specialties, the overall level of agreement regarding appointability, as measured using Kappa, was 0.40 (45% of the maximum Kappa possible given differences in the total number of candidates considered appointable/not appointable in each specialty)[8]. Care is required in interpreting this statistic, since it is possible that a candidate is included twice, jeopardising the assumption of independence. However this result – of fair to moderate agreement according to Landis and Koch (1977) or Fleiss (1981) - suggests that decisions regarding appointability do vary between specialties i.e. different specialties have different requirements, or that one or both selection processes in a dyad is unreliable.

10.7 OFFERS ACCEPTED BY CANDIDATES WHO WITHDRAW THEIR GP APPLICATION

In Round 1, 638 candidates withdrew their GP application after shortlisting and 514 after they had been offered a training post (a total of 1152). In Round 2, 416 after shortlisting and 65 after receiving an offer (total 481).

Table 10.5 shows that 438 (38%) of those withdrawing from GP in Round 1 accepted an offer in another specialty. The main specialty was Core Medical Training (132 candidates), followed by Psychiatry (57), Broad Based Training (34), Anaesthetics (34), Paediatrics (32) and Core Surgery (30). However, note that 52 accepted an offer from GP round 2! For GP round 2 withdrawers, just 69 (14%) accepted offers elsewhere, notably Core Medical Training (30) and Psychiatry (19). Therefore, we do not know the destination of the majority of those who withdraw their GP application. It would be useful to know the proportions who are working abroad, working in non-training clinical posts in the UK, working in non-clinical posts, leaving the medical profession etc. Elsewhere in this report, we indicate that many people turn to GP several years after Foundation training, so we do not know what percentage of these withdrawers are ultimately 'lost' to GP or the wider medical profession.

» *Table 10.5: Offers accepted by candidates who withdrew from GP*

	General Practice, round 1, n=1152		General Practice, round 2, n=481	
	N	%	N	%
Paediatrics, round 1	32	7.3%	3	4.3%
Paediatrics, round 2	1	0.2%	0	0.0%
Trauma and Orthopaedic Surgery	1	0.2%	0	0.0%
Ophthalmology, round 1	6	1.4%	0	0.0%
Ophthalmology, round 2	0	0.0%	0	0.0%
Cardio-thoracic surgery	0	0.0%	0	0.0%
Emergency Medicine, round 1	1	0.2%	0	0.0%
Emergency Medicine, round 2	0	0.0%	0	0.0%
Neurosurgery	0	0.0%	0	0.0%

[8] The maximum possible value of Kappa would only be 1 when both specialties (examiners/assessors) each gave the same number of appointable and not appointable decisions. In this case, disagreement would reflect which candidates were considered appointable, not how many were.

» *Table 10.5: Continued*

	General Practice, round 1, n=1152		General Practice, round 2, n=481	
	N	%	N	%
Obstetrics and Gynaecology, round 1	11	2.5%	1	1.4%
Obstetrics and Gynaecology, round 2	1	0.2%	0	0.0%
Community Sexual and Reproductive Health	0	0.0%	0	0.0%
Chemical Pathology	0	0.0%	0	0.0%
Histopathology, round 1	5	1.1%	0	0.0%
Histopathology, round 2	4	0.9%	3	4.3%
Clinical Radiology, round 1	17	3.9%	2	2.9%
Clinical Radiology, round 2	0	0.0%	0	0.0%
Anaesthetics	34	7.8%	1	1.4%
GP, round 2	52	11.9%	0	0.0%
Public Health Medicine, round 1	4	0.9%	0	0.0%
Public Health Medicine, round 2	0	0.0%	0	0.0%
Acute Care Common Stem - Emergency Medicine, round 1	16	3.7%	6	8.7%
Acute Care Common Stem - Emergency Medicine, round 2	0	0.0%	0	0.0%
Broad Based Training, round 1	33	7.5%	0	0.0%
Broad Based Training, round 2	1	0.2%	2	2.9%
Core Medical Training, round 1	119	27.2%	3	4.3%
Core Medical Training, round 2	13	3.0%	27	39.1%
Core Psychiatry Training, round 1	47	10.7%	1	1.4%
Core Psychiatry Training, round 2	10	2.3%	18	26.1%
Core Surgical Training, round 1	29	6.6%	2	2.9%
Core Surgical Training, round 2	1	0.2%	0	0.0%
Total	**438**	**100%**	**69**	**100%**

*percent of GP withdrawers accepting an offer elsewhere

10.8 CANDIDATES NOT SHORTLISTED FOR GP

Of the 407 candidates who applied to GP Round 1 but were not shortlisted, 97 (24%) accepted offers in either round; most of these were in GP in round 2 (64; 66%), 16 in CMT (16%) and 10 in psychiatry (10%).

The GPNRO has records for the reasons that 360 Round 1 and 2 candidates did not pass Stage 1, this equates to about 7% of candidates[9]; Table 10.6 indicates that the most common reasons were visa/immigration issues (64%) and not providing satisfactory evidence of foundation competence (32%). As some of those rejected in round 1 accept an offer in round 2, it may be possible to reduce the number of stage 1 rejections by speeding up processing of the necessary paperwork[10]. However, further investigation of Stage 1 rejection was beyond the scope of this project.

» *Table 10.6: Reasons for candidates failing Stage 1*

Decision Reason	Candidates	
	N	%
Employment gaps (failure to explain gaps)	5	1%
Exclusion Policy (Failure to provide Support for Reapplication)	6	2%
Fitness to Practice	3	1%
Foundation Competence (failure to provide satisfactory evidence)	115	32%
GMC Registration	1	0%
Immigration Restrictions	230	64%
Grand Total	360	-

To take a historical view we have traced those who failed GP Stage 1 in Round 1 of 2009. Of 6,477 candidates to General Practice in 2009 Round 1, 595 were rejected at Stage 1 (UK: 195/3,494 (5.6%), non-UK 400/2,983 (13.4%)). A follow up in 2015 of the 195 UK graduates found that 29 were on the GP Register and 16 on the Specialist Register; 40 had passed MRCGP AKT and 33 had passed MRCGP CSA; 142 had ARCP records, 99 for a single specialty and 43 for multiple specialties, with the most recent specialty being GP (50), Anaesthetics (20), Medicine (17), Radiology (13), Paediatrics (12), Surgery (10), Emergency Medicine (6), Psychiatry (4), O&G (3), Ophthalmology (3), Pathology (3) and Acute Core Common Stem (1). Therefore, from this 2009 round 1 cohort, of the 195 UK graduates who failed round 1, at least 50 (26%) entered GP training and a further 92 (47%) entered training in another specialty.

10.9 UNAPPOINTABLE UK GRADUATES

There is a view that all UK graduates ought to be suitable for specialty training: otherwise, why were they passed at medical school? Specialties that have more appointable candidates than places are in a position to reject capable candidates, but for GP which under-recruited in 2015, it could be argued that all UK graduates ought to have been offered a place. In the Oriel database, 9,157 candidates have their Country of Qualification recorded as UK; Table 10.7 shows the outcome of those who applied to GP in Rounds 1 and 2. In round 1, 316 (9%) were interviewed but deemed unappointable and a further 201 (6%) were not offered a place.

[9] An earlier footnote indicated that Oriel data do not distinguish between longlisting (Stage 1) and shortlisting (Stage 2); therefore these GPNRO data may not relate directly to the Oriel data.

[10] GP educators have told us that allowing late submission of paperwork can lead to unfilled posts if candidates later drop out.

» *Table 10.7: Outcome of GP applications from UK graduates*

	Round 1		Round 2	
	N	%	N	%
Applied but not Shortlisted	65	2%	18	5%
Shortlisted but withdrawn	191	6%	55	16%
Interviewed but not Appointable	316	9%	67	19%
Appointable but no offer made	201	6%	7	2%
Offer made but withdrew/ declined/offer elsewhere	474	14%	46	13%
Offer accepted	2205	64%	158	45%
Applied	3452	100%	351	100%

10.10 SUMMARY

This chapter has explored the population of candidates to specialty training in 2015 using the Oriel applications dataset. We identified some likely missing data and inconsistencies when comparing the Oriel data for GP candidates with data provided by GPNRO. Another data issue is that Oriel does not distinguish between longlisting (Stage 1) and shortlisting (Stage 2).

Most candidates only applied to one specialty: 73% in Round 1 and 79% in round 2 – results that concur with the survey findings reported in Chapter 9. GP had the lowest number of candidates per post (1.3) and almost the lowest fill rate (69%) across all ST1/CT1 specialties. However, the specialties with the highest number of posts (and GP has the most) tended to be least competitive and have the lowest fill rates.

Of the 4,837 who applied to GP in Round 1, 51% (2,477) accepted offers, and 24% withdrew (13% before interview and 11% after offer made). The remaining 25% either were not shortlisted (8%), not appointable (12%) or appointable but still no offer made (4%). To increase accepted offers requires encouraging more applications, persuading candidates not to withdraw and make offers to everyone who is appointable. 83% of those receiving a GP offer in Round 1 accepted it, compared with 79% for all other specialties combined. While this suggests that GP is at least as preferable to accepted candidates compared with other specialties, these figures do not control for differences in the number of applications made by candidates to each specialty.

The data on progression through selection for those rejected in Round 1 suggest that re-applications should therefore be encouraged, as there is a fairly high probability of success. Excluding candidates who were shortlisted but withdrew from Round 2, the likelihood that re-application will be successful (in terms of being made an offer) was 66% if attended Stage 3 in round 1 and 48% if not shortlisted in Round 1.

Of the 454 candidates choosing between GP and non-GP offers in round 1, 281 (62%) chose the non-GP specialty and 173 (38%) GP. Fewer candidates received multiple offers (including GP) in round 2; of those accepting one offer, 14/30 accepted their GP offer (47%). Thus when in direct competition with another specialty, GP is less preferable to the alternative.

Across all specialties, 85% of candidates considered appointable by GP and/or another specialty were considered appointable by GP and 79% by the other specialty. This very crude analysis suggests that overall 'the bar' is a little lower in GP than other specialties i.e. it is easier to be deemed appointable. Across all specialties, the overall level of agreement regarding

appointability, as measured using unweighted Kappa, was 0.40 (45% of the maximum Kappa possible given the relative frequencies), suggesting that different specialties have different competency requirements (or unreliable selection processes).

Based on GPNRO data, the most common reasons for rejection at Stage 1 were visa/immigration issues (64%) and not providing satisfactory evidence of foundation competence (32%). It may be possible to speed up processing of the necessary paperwork and thus reduce the number of rejections.

~ This page is intentionally left blank ~

Part 3

Synthesis and recommendations

weight to all and Stage 3 only are the next-best alternatives, with similar average costs per GP Registration within four years FTE with both 3,250 and 3,750 posts.

~ This page is intentionally left blank ~

Chapter 11
Synthesis and Recommendations

Chapter 11.

Synthesis and Recommendations

11.1 INTRODUCTION

As indicated in the introduction (Chapter 1), this evaluation can be viewed in terms of the Utility framework used to evaluate assessments, in which the components of utility are identified as: re liability, validity, educational impact, acceptability of the method to the stakeholders and cost (van der Vleuten, 1996, van der Vleuten and Schuwirth, 2005). An important aspect of the framework is that the authors suggest that if one component has zero value, then overall utility is also zero regardless of the value attached to the other components. We have looked closely at **reliability, validity and cost-effectiveness** of GP selection. Educational impact and acceptability have not been investigated; with high-stakes examinations, it could be argued that these are less important, but that is for policy makers to decide. In addition, we did not want to spend scarce evaluation resources repeating surveys of acceptability to GP applicants that are undertaken by WPG (e.g. Lopes, Ashworth and Tate 2013) and were likely to produce similar results given the similar methodology that we would have employed[1].

We are aware of numerous debates and decisions within GP education circles regarding changes to the GP selection system. Whilst we are mindful of these, we endeavour in this chapter to focus on the data and information presented to us and analysed by us, so that we remain as objective as possible. We delineate what we understand to be key issues and describe the implications of our findings for these issues; however, many of the necessary decisions are inevitably political in the sense that they involve policy, and policymakers will need to weigh up the competing concerns. Of course we are aware that some suggested changes are beyond the remit of those involved in specialty training. We are also aware that some of these recommendations have already been discussed and may have been implemented.

It can be challenging to relate the highly statistical analyses undertaken in this evaluation to the experiences of candidates, assessors etc. engaged in the selection process and to relate these to probable training outcomes. For example, many assessors may be convinced that they can discriminate between the four Stage 3 competencies, but our analyses indicate that there is very weak evidence for this across the whole selection system. Similarly we've heard it said that Stage 3 can spot problem doctors: whilst we cannot discount this possibility completely, the low Stage 3 reliability suggests this is unlikely to happen consistently.

We are also aware that analysis of selection data can only assess what is and what has happened, and that it cannot easily compare radically different selection systems. Precisely the same considerations apply in clinical medicine, and there the gold standard for evaluating therapeutic and treatment interventions is the randomised controlled trial (RCT), particularly when several are combined in meta-analyses. If there is dissent as to the value of a component of a selection system (such as the ability of Stage 3 to spot problem doctors), and that dissent entails genuine equipoise, with the arguments and the evidence finely balanced on both sides, then an RCT is ethically justified and should provide a proper answer to the question. Whether RCTs of selection processes would be acceptable to doctors is another matter, but there is an argument that if RCTs are the appropriate way to decide on the treatment of patients then they are also the appropriate way to decide on the selection of the doctors who will be administering those treatments. RCTs are now becoming the norm in education in general, but they are still rare in medical education, and extremely rare in postgraduate medical education and selection for training. It is perhaps time for that to change, and GP selection would be an ideal environment for RCTs given the large numbers and hence high statistical power.

[1] Note this argument could be applied to the analyses of reliability and validity that have also been undertaken by WPG. However our methods of analysis are different to those used by WPG (we believe they are also more appropriate) and hence replication with different methods was considered essential.

It should be noted that we have modelled the GP selection system using 2009 to 2015 data. Thus we are considering what might have happened with different selection systems with the same candidates. How applicable this is to future selection will depend in part upon the impact of external changes e.g. changes in government policies and changes in behaviour that might make some potential candidates more or less likely to apply. A major potential policy change is to increase GP training to four or five years: the impact of this upon both who applies and on training outcomes is beyond the scope the present study.

11.2 THE GP SELECTION PROCESS

Between 2009 and 2015, Round 1 GP applications for UK graduates rose from 3503 in 2009 to 4318 in 2013, and then declined to 3696 in 2015; non-UK graduate applications have halved in that time. Throughout this period, around 61% of UK graduate applications have led to an offer being accepted; about 22% failed at Stages 1, 2 and 3, and 18% withdrew. For non-UK graduates, the acceptance rate is much lower at about 24%, and perhaps falling during this period; about 71% fail at Stages 1, 2 and 3 and just 6% withdraw. Tackling the fall in numbers of non-UK candidates and their low success rate in Stage 1, when due to delays in paperwork (rather than a lack of suitability for GP training),[2] are therefore areas where relatively 'quick wins' may be possible[3].The RCGP has already investigated similar issues in recent years regarding reasons for higher IMG failure at CSA than AKT (Esmail and Roberts 2013, Roberts, Atkins and Hawthorne, 2014). A similar approach could be taken regarding Stage 3, if it has not already been undertaken, since the data in Chapter 2 suggest that IMGs are more likely to be rejected at Stage 3 than Stage 2, while for UK candidates the rejection rates at each of these stages are similar.

Larger specialties tend to have lower competition ratios and have greater problems filling their training posts. This applies to General Practice, as the largest specialty: that a slightly higher percentage of candidates accepted an offer from GP than for the other specialties combined generally reflects the lower competition for posts rather than preferences for GP. On the latter, when offered two posts, only 38% chose GP in preference to another specialty; however, we do not think this figure is particularly bad as GP may be a common back-up choice (i.e. it is a back-up choice for candidates from almost all other specialties). Candidates applying for GP tended to have lower FPAS scores than for other specialties. When candidates applied to GP and another specialty, a higher proportion were considered appointable by GP than the other specialty, suggesting 'the bar' may be a little lower in GP.

Costs and hence cost-effectiveness was considered from the perspective of a virtual provider of all NHS GP services across the UK. For 2014, the total selection cost was estimated at £4.89m (approximately £820 per candidate and £1,580 per post filled). This cost was made up of: nationally-incurred costs, including Stage 2 test fees (£921,000), LETB Stage 3 'on the day' costs (£2,290,000), assessor training (£466,000) and candidate time (£1,220,000). The cost for three years' GP training is about £210,000; each six-month extension approximately costs £65,000. We estimate the value of healthcare lost for not filling a training post to be around £415,000; £352,000 due to losing seven years post-CCT general practice and £64,000 loss for not providing service in secondary care during GP training. In 2014, 481 posts were unfilled, resulting in a total ten year loss of health provision worth around £200m. These figures indicate that for the health service overall: selection costs are insignificant; even a trainee who requires a 12-month extension is worth training; and that unfilled posts and trainees who fail (and thus never provide service as a GP) are huge costs to the system.

We have modelled likely outcomes for 3,250 and 3,750 UK trainees. As expected, the 'marginal 500' trainees have worse outcomes than the first 3,250 trainees selected e.g. for Stage 2 only selection, 31% are predicted not to enter the GP register within five years, compared with 17% for the first 3,250[4]. However, compared to leaving these last 500 posts unfilled, increasing recruitment is cost-saving over ten years for all approaches.

Using the assumptions and costs in the previous paragraphs, cost-effectiveness modelling indicates that the number of unfilled posts - which results in high future costs of 'missing' primary care - is the main driver of cost-effectiveness. A secondary driver is the predictive validity of the selection processes, so 'Stage 2 only' is better than 'combining CPST, SJT and Stage 3 scores' with equal weight given to each, which in turn is better than only using 'Stage 3 scores'.

[2] In 2016, IMGs subject to the Resident Labour Market test have a longer period to make their application for sponsorship, and all candidates have longer to provide evidence of Foundation competency, such that this issue is already being addressed.

[3] We note the 50% target applies to UK graduates.

[4] Data from Tables 7.3A & B.

We cannot comment from a political perspective on the view that GP training costs are unsustainable. However, from a cost-effectiveness perspective, selection is relatively cheap and it is the 'hidden' cost of leaving GP training posts unfilled is that is unsustainable, particularly given an aging and increasingly multi-morbid population which is increasing its demands for healthcare as new treatments are developed. Of course the challenge is to select more trainees who are trainable. Although time constraints precluded detailed investigation, as Stage 2 scores decrease, the likelihood that trainees are released from training (ARCP Outcome 4) increases from around 2.0% with a combined score of 600 to 13.8% with a combined score of 400. We have estimated system-wide medium to long term costs for trainees who are released, but different stakeholders are likely to be more concerned about specific short term impacts such as the financial, emotional and time burdens for the LETBs and teaching practices that deal with struggling trainees. It is a political task to weigh up these short term negatives against the imperative to train more GPs (when most of those incrementally selected will obtain GP Registration, although perhaps following an extension longer than the six months perceived as the norm in the Gold Guide (The Gold Guide, 2014)).

For many reasons, our economic analyses cannot take into account the potential long-term costs, both human and financial, for patient morbidity and mortality which might well be related to differences in the competency of doctors who are selected using different selection methods. There are few data in existence which relate performance at examinations such as those used in GP selection and those used in MRCGP to hard patient health outcomes. However the most important such study, by Norcini and his colleagues suggested there might be a fairly strong relationship (Norcini et al., 2014). Such studies in the UK are much needed to assess the generalizability of such findings, and the relationship of selection scores to patient outcomes.

Recommendation:

> **R1:** Although not currently being pursued, interpretation of the '50%' target should reflect the percentage of UK graduates who do not go straight from Foundation into Specialty training (currently 48% and rising), and/or who never enter either the GP or Specialist Registers (historically 15% and possibly rising). Similarly, as the '3250' target is for England only, the targets for Scotland, Wales and Northern Ireland should also be published[5]. As GP selection is UK-wide, it would be helpful if such announcements are agreed and coordinated between the four countries.
>
> **R2:** Invest in improving the validity and reliability of GP and specialty selection. In terms of the whole UK healthcare system, this is likely to be highly cost-effective in the long-term.

11.2.1 Stage 1

Around 7% of GP candidates were rejected at Stage 1 for immigration issues or not providing evidence for Foundation competence. About half of those who re-applied for Round 2 were offered a place.

Recommendation:

> **R.3:** Seek to reduce Stage 1 rejections and encourage those who are rejected to re-apply. In 2016, candidates are being given longer to provide evidence of Foundation competency and for IMGs to complete their application for sponsorship so a move to addressing this recommendation is already underway.

11.2.2 Stage 2

We estimated alternate forms reliability to be 0.73 for CPST, 0.57 for SJT, and 0.73 for the Stage 2 total. As the SJT typically has just 50 questions, extending its length would be a simple way to increase its reliability. Currently, the Stage 2 assessments have a relative standard i.e. a candidate's marks are relative to those in that assessment round; a score of 200 in one year is not equivalent to a score of 200 the next.FPAS EPM and SJT scores correlate moderately with CPST and SJT scores from GP selection. It may be that these FPAS scores could be used in Specialty Selection, generally.

[5] Historical 'places' or 'vacancies' are available at https://gprecruitment.hee.nhs.uk/Resource-Bank (accessed 07/01/2016), but we do not know if future targets are agreed or published.

FPAS EPM and SJT scores correlate moderately with CPST and SJT scores from GP selection. It may be that these FPAS scores could be used in Specialty Selection, generally.

Currently, candidates need to score at least 180 in both CPST and SJT to progress to Stage 3, and their SJT Band is used in making the final decision after Stage 3. This is much cruder than using the continuous scores. In terms of quality assurance, it would seem sensible to have an external check on these procedures.

Recommendations:

R.4: Use statistical equating of the CPST and SJT scores across years e.g. by Rasch modelling; this makes Stage 2 an absolute form of assessment.

R.5: Use continuous Stage 2 scores rather than Bands.

R.6: Select to Stage 3 using the total CPST + SJT score rather than a cut-off mark for each; however, we have not investigated this in detail e.g. the optimal way to weight SJT compared with CPST.

R.7: Increase the length of the Stage 2 assessments, particularly SJT, to increase reliability.

R.8: In terms of transparency, it would make sense to separate development and administration of the selection system from its evaluation (this also applies to Stage 3).

11.2.3 Stage 3 and offers of training places

Correcting for restriction in range and using multiple imputation, the Stage 2 selection assessments are better predictors of AKT and CSA than is the Stage 3 Selection Centre assessment. CPST is a much better predictor of AKT (r=0.79) than is the SJT (0.56), whereas the CPST (0.67) and SJT (0.60) are more similar at predicting CSA. The Stage 3 correlations are 0.40 with AKT and 0.52 with CSA. Stage 3 scores account only for about 3-4% of additional variance in CSA scores after taking Stage 2 into account. In short, Stage 3 adds very little to the predictive validity from Stage 2; therefore, a straightforward recommendation might be to stop using Stage 3. Before considering this possibility, potential improvements to Stage 3 will be discussed.

Throughout this report, we have referred to Stage 3 as an 'OSCE-type' assessment. This is because it has different stations with assessors and simulators (role-players), just like an OSCE, which means that the types of statistical analysis that are appropriate to use with OSCEs are also appropriate for Stage 3[6]. It has long been recognized in the OSCE literature that the number of stations is a major determinant of reliability, with 12+ (Swanson et al., 1991) or 7 – 11 (van der Vleuten and Schuwirth, 2005) stations being regarded as desirable for high-stakes examinations. Therefore, there has been a move to multiple mini-interviews (MMIs) for entry to medical school in particular, i.e. admissions OSCEs (Eva et al, 2004, Dore et al, 2010). We recognise that currently specialty selection procedures often do not use this multiple mini-interview approach and we have not compared GP selection with other selection systems. We have not explored this area in detail, but note a systematic review in which the modal student MMI selection system has 10 eight minute stations with "with Cronbach's alpha = 0.69–0.98 and G = 0.55–0.72" and good prediction of OSCE performance (Pau, Jeevaratnam, Chen, et al., 2013, p1027). Therefore, this route may improve reliability and validity, but it is unlikely to produce dramatic improvements without a substantial increase in assessment time. We note that unlike many other specialties, the GP selection process already uses only one assessor per station and thus the almost cost-neutral option of including more stations with one assessor in each does not apply to GP as it does elsewhere.

Although some may disagree that the judgements made by assessors can be interpreted as quantitative data, we have used the numerical scores to undertake an analysis of reliability; noting that such scores are used in practise to determine candidates' total selection scores for post allocation and clearing processes. To enable comparison with other published estimates (as outlined in Chapter 3), we have calculated Stage 3 reliability in several ways, including those that we believe are sub-optimal and could result in artificially inflated estimates of reliability. Cronbach's alpha, comparing the total scores obtained in each of the four stations gives an average internal reliability across 2011-2015 of 0.62. This is lower than the estimates of

[6] The Written station is a little different as there is no simulator, so less OSCE-like, although some other OSCEs include similar stations.

0.87 and 0.89 published by Patterson et al. (Lievens and Patterson, 2011, Patterson et al., 2009a). It is not clear how these alternative estimates were calculated, but the text in the papers is suggestive of reliability within a simulation – which would be inappropriate since all judgements within a simulation are made by a single assessor and are thus not independent. The alternate forms reliability averaged 0.43; this compares scores obtained by re-sitting candidates and, by imputation, those who only sat Stage 3 once. Using a generalizability approach, the alternate forms reliability averaged about 0.50 for several different approaches. All these reliabilities appear to have fallen in the last two years; we are unsure of the reasons for this, but suspect it is due to reduced variance amongst candidates i.e. due to fewer particularly weak or strong candidates (e.g. some weaker IMGs choosing not to apply in light of the 2014 legal ruling that the CSA does not discriminate against BME and IMG candidates). The FACETS analysis suggests that assessors, simulators and cases all contribute to unreliability: we haven't considered this in detail, but it may be that matching case difficulty and training to improve consistency of simulators may be as important as assessor training and calibration.

The FACETS analysis also emphasizes that it is only possible to assess whether, say, assessors are more or less hawkish if those assessors assess on a range of simulated cases and with a range of simulators. If case, assessor and simulator always work together then there is no possibility of teasing apart variation due to each of them. It therefore makes sense to design the system so that different combinations of case, assessor and simulator occur and then in principle the effect of each can be disentangled.

It has been argued that combining an unreliable OSCE-type examination with a more reliable written task is a pragmatic approach when an OSCE is desirable for other reasons such as acceptability (Newble 2004). If this line is taken, then the final decision should be weighted more highly than it is at present towards Stage 2 scores, particularly including the CPST score which currently is not used.

Currently a complex algorithm and moderation are used to decide whether a candidate is offered a place. The SJT Band and the number of competencies with a mean score of 3 or greater are the major factors in determining the initial outcome from the algorithm. To justify keeping the competency scores separate in this way would require evidence that they are distinct. However, the generalisability analyses indicate that only one competency (Empathy and Sensitivity) has any residual variance and that is very low i.e. candidates are not assessed as being stronger or weaker in different competencies in any consistent way across the four stations. Similarly, we found no evidence that the four separate stations are contributing specific variance. The low reliability of the separate skill and station scores means that undue emphasis should not be put upon individual scores or different profiles of scores as the total score contains most of the key information. Perhaps more importantly, as Stage 2 is more reliable and predictive than Stage 3, offer decisions should be based more on performance at Stage 2 than at Stage 3.

From our limited observation of the moderation exercise, those involved treat the process seriously and discuss individual candidates as well as can reasonably be expected. However, the moderation decision depends upon the moderator's interpretation of brief descriptions of what the candidate did, so we are doubtful that this process enhances the reliability of Stage 3. To investigate this further, statistical and sociolinguistic analyses may be appropriate. Furthermore, the previous paragraph indicates no empirical support for using the algorithm and that Stage 2 scores should be prominent in offer decisions (combined with total scores from Stage 3): if these points are accepted, then the logic underlying the moderation exercise is removed.

In summary, Stage 3 has low reliability, as expected given the low number of stations and there seems little justification for the complex algorithm or the moderation exercise. However, although we have not formally investigated the educational impact and acceptability of Stage 3, they are likely to be high.

Recommendations:

> **R.9:** Use a combination of CPST, SJT and the Stage 3 total scores rather than the algorithm to decide which candidates should be offered places. If this is done, weight the Stage 2 CPST and SJT continuous scores more heavily than the Stage 3 score.

R.10: Withdraw the moderation procedure. If uncomfortable with this without further evidence, undertake sociolinguistic and additional statistical analyses of moderation to enable a more comprehensive assessment to be made of what it adds, and whether it does so consistently.

R.11: Investigate the possibility of increasing the number of stations to increase reliability and predictive validity. Removing the moderation session would mean about a third more time is available. There may also be potential to reduce the length of the scenarios. Any such changes would need to be designed carefully.

R.12: The generalizability analysis suggests that simulators and cases contribute roughly similar amounts of variance (error) as assessors; so:

a) Consider ways to reduce differences between simulators and cases, as much as between assessors e.g. more simulator training and feedback on their 'hawkishness' or 'doveishness'; ideally cases would be piloted, but perhaps more realistic is consideration of the difficulty of previous cases when developing new ones (for which Rasch analysis provides a useful methodology).

b) Design the selection centres so that variances due to assessors, simulators and cases can be partitioned (distinguished) i.e. keep careful records of which assessor, simulator and case are seen by each candidate, and ensure each assessor and simulator experiences at least two simulators/ assessors and cases e.g. rather than confounding assessors with simulators by always pairing them in the same way.

R.13: Avoid using Cronbach's alpha as a measure of reliability. Generalisability analyses are well developed and should be used routinely within and between selection centres. Techniques such as the EM algorithm and alternative forms reliability can also be used.

R.14: Investigate why candidates (particularly those who are graduates of UK medical schools) are failing Stage 3 (and Stage 2) and consider enhancing medical school or Foundation training to help applicants address the areas of weakness identified. Publish the results to help prospective candidates understand what is required of them and consider using videos of poor, acceptable and excellent candidates in example scenarios to help applicants prepare for the Selection Centre[7].

14.2.4 Should Stage 3 be abolished?

The estimated annual selection cost for GP is around £4.3 million, primarily consisting of the time of candidates and assessors (and hence loss of provision of healthcare). Abolishing Stage 3 could save around £3 million of this. In addition, Stage 3 has poorer reliability and hence predictive validity than Stage 2; assuming 3250 posts are filled, abolishing Stage 3 is estimated to result in approximately 43 more trainees completing GP training within 3 years, 22 fewer needing extensions of up to two years, and 21 fewer not entering the GP Register within 5 years[8]. Although these numbers are relatively small, there is still a prima facie argument to abolish Stage 3. An initial counter argument is to note that £3 million equates to the service provision of seven trainees[9] and so is actually insignificant when compared with the value of healthcare provided. However, the real debate is concerned with the acceptability and educational impact of losing Stage 3. Here, the reader should be mindful that we are inevitably relying on logic and anecdote, rather than empirical evidence. General Practice is regarded as a person-centred medical specialty. In training, the biggest and most important hurdle to CCT is the CSA. Therefore, retaining Stage 3 emphasizes the importance of face-to-face communication that is central to general practice and meets the expectations of HEE's Values Based Recruitment framework (see https://hee.nhs.uk/our-work/attracting-recruiting/values-based-recruitment). This is likely to help encourage potential trainees who have the right personalities and skills to apply and the educational impact will be to encourage human interaction rather than studying books. In short, selection should

[7] The same logic should apply to all specialties where a significant proportion of UK graduates are considered unappointable.

[8] Comparing Stage 2 only and 'Equal weight to all' columns in Table 7.3.

[9] Including as a qualified GP up to 10 years after selection.

endeavour to improve the measurement of what is important for general practice (Stage 3) rather than simply relying on what is easy to measure (Stage 2). If this argument is accepted, then the recommendations regarding Stage 3 above should be considered. If not, then it may be appropriate to stop using Stage 3. Alternatively there may be arguments for other processes, perhaps involving formal assessment of communicative ability during undergraduate training or during the Foundation Year (and good communication is important in all medical specialties), or early assessments of communication during GP training, perhaps with intensive intervention and remediation for those who are not at an appropriate level.

Recommendation:

R.15: Consider whether the greater reliability, validity and cost-effectiveness of abolishing Stage 3 are beneficial, given potential losses in terms of educational impact and acceptability.

11.3 DATA QUALITY

Overall, we are delighted by the breadth and depth of data made available for this study by GPNRO, HEE, the devolved nations, a number of LETBs, RCGP and GMC. We understand how rare it is to have such access to these datasets – and to be able to link them. However, throughout this report we have highlighted many areas where the data are unclear or inconsistent: there are numerous potential reasons for this. To really understand the processes at work, routine collection, **checking/ cleaning, linkage, analysis and archiving** of such data are of utmost importance. In this recommendation, a few issues are included to illustrate why such knowledge is invaluable.

Recommendation:

R.16: Take a long-term strategic approach to obtaining, checking, using and storing data related to recruitment and training. Ideally this is in all specialties from the beginning of medical school to retirement. This is now underway with UKMED, but it will take many years before it will have sufficient longitudinal data to address many of the issues in this report; consequently, 'bottom-up' improvements in data retention are also desirable.

a) Investigate why in 2014, 5% of candidates appeared to pass Stage 2, but were not invited to Stage 3: if due to LETB preferences, should offers elsewhere still be made?

b) Investigate why in 2015, 224 (4%) GP candidates appear to have been deemed appointable, but were not made an offer (even if to a different LETB).

c) Investigate why some doctors have no GP ARCP records despite accepting offers, and others enter the GP register despite being recorded as not accepting offers (they could be via the CEGPR route, but may be due to data errors).

d) Maintain accurate ARCP records, particularly regarding OOP, TOOT and LTFT, so that FTE lengths of training can be calculated precisely.

R.17: UKMED will also inevitably be limited in the measures it obtains of medical students and their interests and intentions, and separate methods of collecting such data routinely need to be developed.

11.4 MEDICAL SCHOOL, FOUNDATION TRAINING, OTHER SPECIALTY TRAINING AND 'TIME OUT'

The Department of Health's proposal that "at least half of doctors going into specialty training will be training as GPs" (Department of Health, 2008, p.15, para 36) needs to be interpreted carefully as historically about 15% of UK graduates do not enter either the GP or Specialist Registers. In 2015, just 52% of doctors at the end of the Foundation Programme moved straight into specialty training (including GP); so the **immediate target** would then be about 26% going into GP and the **longer-term target** about 43%. Approximately 35% of the 1990 to 2006 UK graduation cohorts are now on the GP specialist register although about 50% of recent cohorts have applied at some time.

Historically, it has taken about eight years for almost all UK graduates to decide whether or not to apply to GP i.e. there are 'late converts' to general practice. Given the rapid recent increase in UK graduates not applying immediately for specialty training, forecasting is extremely difficult, which is why we estimate between 36% and 50% of current UK graduates are likely to **apply** to GP training[10]. Some of these will be as second or third choices and others will not be offered posts; although there is considerable uncertainty, the historical trends suggest the Department of Health target will be very challenging to meet unless the number of non-GP specialty training posts is substantially reduced.

There seem to be deep-seated attitudes held at all levels which mean that hospital careers are often favoured over GP. Some medical schools 'produce' more GPs, perhaps by attracting students interested in GP or by providing particularly positive GP experiences. Because students from these schools do less well at Stage 2 and no better at Stage 3, this increased 'GP production' appears to be the result of student/doctor choice rather than the ability of the schools to develop the GPs of tomorrow. We would counsel against jumping to 'easy' conclusions based solely on the existence of differences between medical schools in the proportion of students who become GPs.

Analysis of the Specialty Training Applicant Questionnaire indicates a lack of both GP experience and positive careers advice amongst those who could, but do not, apply for GP. Whether a doctor is likely to apply for GP as their first-choice depends upon them: believing their personality is suited to GP; wanting to provide GP patient care; seeing GP as providing important work-life balance; finding GP intellectually stimulating; and having positive experiences of GP at medical school and elsewhere.

Given so many doctors take several years to enter GP, long-term systemic changes may be more effective than quick fixes. Although the issues raised here are likely to be beyond the remit of the policy makers reading this report, it is important to acknowledge the context in which GP selection operates. For example, a radical suggestion would be to alter the final medical school assessments so that they are predominantly focussed on General Practice: if implemented successfully, it would be likely to positively alter attitudes towards GP within medical schools.

Recommendations:

> **R.18:** Positive GP experiences at medical school are important influences on applying for GP, particularly as they help students become aware that GP suits their personality. It is possible that improving these experiences and providing positive careers advice towards general practice may increase long-term uptake of GP.
>
> **R.19:** Address the problems that hinder Foundation doctors from gaining GP experience in Year 1.

[10] Of course, situations can change; for example, the junior doctors dispute over their contracts that began in 2015 may be encouraging more to pursue their medical careers abroad e.g. http://www.theguardian.com/society/2015/sep/22/junior-doctors-resist-contract-that-could-cut-pay-by-applying-to-work-outside-uk [accessed 11/12/2015]

R.20: Consider ways to encourage and facilitate applications from those who have been working abroad or in another specialty.

11.5 GP TRAINING AND BEYOND

Although the remit of this report is GP recruitment, we have argued strongly that a 'cradle to grave' perspective is ideal. Here, we note some of the important, relevant post-selection issues:

- About 2.5% of doctors accepting GP training posts appear not to have subsequently taken up training.
- About 23% of trainees are either LTFT or OOP at some point in their training: understanding these groups are important, particularly as those with LTFT appear to have worse outcomes.
- Around 50% of trainees enter the GP register within 3 years of starting GP training; this rises to 81% within 6 years. We didn't investigate reasons (beyond failure to obtain the required competencies, including MRCGP) for not entering the GP register, but it would be good to know why some choose not to continue in GP. The ARCP records provided to us were minimal in form, but the real documents are probably richer in structure, giving percentage LTFT, reasons and type of OOP, and time out of training. Such data need to be available for wider evaluations of GP training (e.g. so that TOOT can be identified rather than using the absence of ARCP, which will always be a fallible and risky process).
- We speculate about changing priorities and attitudes that make GP relatively more attractive to older doctors (not considered in detail and beyond the scope of this project) for example by the increased emphasis on work-life balance it provides compared to other specialties.
- The Plint report highlights capacity issues for GP practices involved with training. Expansion of experience in Medical School and Foundation, coupled with more GP trainees, some of whom are LTFT and some requiring extensions all place increased burden on those responsible for delivering training (Plint 2014).
- We do not comment on measures to improve GP retention (the Centre for Workforce Intelligence, 2014), since retention was beyond the scope of this project.
- Ideally, evaluations of this nature would consider how well trainees perform in practice, once qualified, and their impact upon patient safety. We urge the collation of data that will enable this to happen in the near future.

11.6 FINAL THOUGHTS

Readers who have worked their way through this entire document will be clear that we strongly advocate much stronger data management and linking, better use of statistical techniques and even more time, money and effort to be spent on GP **selection**. Politically this may be unacceptable, but it should be remembered that GP selection is choosing over half the medical workforce of tomorrow. GP **recruitment** is about encouraging capable doctors into the specialty; improving recruitment is at least as important as improving selection as selection can only select from those who choose to apply.

References

References.

Abse, D. (1978) My medical school. London: Robson Books.

Allison, P. D. (2002). Missing data. Thousand Oaks, CA: Sage.

American Educational Research, A., American Psychological, A. and National Council on Measurement in, E. (1999) Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

American Educational Research, A., American Psychological, A. and National Council on Measurement in, E. (2014) Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Anonymous (1950) The Collings Report. Lancet, 255: (6604): 547-549.

Anonymous (1990) Examining the Royal Colleges' examiners [editorial]. Lancet, 335: 443-445.

Anonymous (2014a) GP National Recruitment Stage 3 Standardisation Project.

Anonymous (2014b) Resource to support training for selection and moderation [Assessment and Moderation Manual].

Arthur, W., Glaze, R., Jarrett, S.M., et al. (2014) Comparative evaluation of three situational judgement rest response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. Journal of Applied Psychology, 99: (3): 535-545.

Baghaei, P., Pishghadam, R. and Navari, S. (2009) A new method for standard-setting using the Rasch model. Iranian Journal of Applied Linguistics, 12: (1): 61-85.

Bartlett, F.C. (1946) Selection of medical students. British Medical Journal, ii: 665-666.

Bloch, R. and Norman, G. (2012) Generalizability theory for the perplexed: A practical guide and introduction: AMEE guide no. 68. Medical Teacher, 34: 960-992.

Boudreau, J.W. (1988) Selection utility analysis: A review and agenda for future research. CAHRS Working Paper Series, 423.

Brannick, M.T., Erol-Korkmaz, H.T. and Prewett, M. (2011) A systematic review of the reliability of objective structured clinical examination scores. Medical Education, 45: 1181-1189.

Brennan, R.L. (2001a) An essay on the history and future of reliability from the perspective of replications. Journal of Educational Measurement, 38: (4): 295-317.

Brennan, R.L. (2001b) Generalizability theory. New York: Springer-Verlag.

Brennan, R.L. (2001c) Manual for mGENOVA, version 2.1. Iowa City, IA: Iowa Testing pPrograms (available at http://www.uiowa.edu/~casma/computer_programs.htm).

Brennan, R.L. (2001d) Manual for urGENOVA: Version 2.1. Iowa City, IA: Iowa Testing Programs (available at http://www.uiowa.edu/~casma/computer_programs.htm).

Brogden, H.E. (1949) When testing pays off. Personnel Psychology.

Burt, C. (1943) Validating tests for personnel selection. British Journal of Psychology, 34: 1-19.

Cardinet, J., Johnson, S. and Pini, G. (2010) Applying generalizability theory using EduG. New York: Routledge.

Carr, V., Patterson, F., Burr, B., et al. (2009) Evaluation of situational judgement tests to select postgraduate trainees: Validation studies in two specialties. AMEE conference poster (available at http://www.workpsychologygroup.com/assets/Library/Library-SJTs-to-select-postgrauate-trainees-AMEE.pdf).

Centre for Workforce, I. (2011) Shape of the Medical Workforce: Informing medical specialty training numbers. London: Centre for Workforce Intelligence (available at http://www.cfwi.org.uk/publications/medical-shape-2011/attachment.pdf).

Centre for Workforce Intelligence (2014) "In-depth review of the general practitioner workforce: Final report".

Cohen, J. and Cohen, P. (1983) Applied multiple regression/correlation for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum.

Collings, J.S. (1950) General Practice in England today: A reconnaissance. Lancet, 255: (6604): 555-585.

Crick, J.E. and Brennan, R.L. (1983) Manual for GENOVA: A generalized analysis of variance system. Iowa: American College Testing Program (available at http://www.uiowa.edu/~casma/computer_programs.htm).

Cronbach, L.J. (1951) Coefficient alpha and the internal structure of tests. Psychometrika, 16: 297-334.

Cronbach, L.J., Gleser, G.C., Nanda, H., et al. (1972) The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley.

Cronbach, L.J. and Shavelson, R.J. (2004) My current thoughts on coefficient alpha and successor procedures. Educational and Psychological Measurement, 64: (3): 391-418.

Cruickshank, J.K. and McManus, I.C. (1976) Getting into medicine. New Society, 35: 112-113.

Curtis, J.L. (2014) "Unit Costs of Health and Social care 2014". Canterbury, PSSRU.

Davis, J.A. (1985) The logic of causal order. London: Sage.

DeMars, C. (2015) Item response theory. New York: Oxford University Press.

Department of Health (2008) A High Quality Workforce: NHS Next Stage Review. London: Department of Health (available at http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_085841.pdf).

Department of Health (2013). Delivering high quality, effective, compassionate care: developing the right people with the right skills and the right values. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/203332/29257_2900971_Delivering_Accessible.pdf [accessed 23/12/2015]

Department of Health (2014a) "Education and training tariffs: Tariff guidance for 2014-15". London.

Department of Health (2014b) A reference Guide for Postgraduate Specialty Training in the UK ("The Gold Guide"). Fifth edition. London.

Dore, K.L., Kreuger, S., Ladhani, M., et al. (2010) The reliability and acceptability of the Multiple Mini-Interview as a selection instrument for postgraduate admissions. Acad Med, 85: (10 Suppl): S60-63.

Dormandy, L. and Laycock, K. (2015) Triumph of process over practice: changes to assessment of physicians. BMJ Careers, 28th July 2015: http://careers.bmj.com/careers/advice/Triumph_of_process_over_practice%3A_changes_to_assessment_of_physicians

Eckes, T. (2011) Introduction to many-facet Rasch measurement: Analysing and evaluating rater-mediated assessments. Frankfurt am Main: Internationaler Verlag der Wissenschaften.

Enders, C. K. (2010). Applied missing data analysis. New York: The Guilford Press.

Eva, K.W., Rosenfeld, J., Reiter, H.I., et al. (2004) An admissions OSCE: the multiple mini-interview. Medical Education, 38: 314-326.

Fleiss, J.L. (1981) Statistical methods for rates and proportions (2nd ed.). New York: John Wiley.

General Medical Council (2014) "Progress of doctors in training split by postgraduate bodies". London.

Goldacre, M., Davidson, J. and Lambert, T. (2010) The junior doctor exodus. BMJ Careers, http://careers.bmj.com/careers/advice/view-article.html?id=20001543.

Graham, J.W. (2009) Missing data analysis: Making it work in the real world. Annual Review of Psychology, 60: (549): 576.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological

research. Psychological Methods, 11, 323-343.

Hambleton, R.K., Swaminathan, H. and Rogers, H.J. (1991) Fundamentals of item response theory. Newbury Park: Sage.

Harding, A., Rosenthal, J., Al-Seaidy, M., et al. (2015) Provision of medical student teaching in UK general practices: A cross-sectional questionnaire study. British Journal of General Practice, 65 (DOI: 10.3399/bjgp15X685321): (535): e409-e417.

Harris, B.H.L., Walsh, J.L. and Lammy, S. (2015) Clinical Medicine. Clinical Medicine, 15: (1): 40-46.

Hayes, A.F. (2013) Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. New York: Guilford Press.

Health Education England (2014) "A reference guide for postgraduate specialty training in the UK (the Gold Guide) 5th Edition". London, NHS.

HM Treasury (2011) "The Green Book". London.

Howorth, S. (2014) Stage 3 (Selection & Assessment Centre) Manual. Manchester: North Western Deanery.

Kenny, D.A. (1979) Correlation and causality. New York: John Wiley.

Lambert, T. and Goldacre, M. (2011) Trends in doctors' early career choices for general practice in the UK: Longitudinal questionnaire surveys. British Journal of General Practice, 61: (588): e397-e403.

Lambert, T., Smith, F. and Goldacre, M. (2013) GPs' job satisfaction: Doctors who chose general practice early or late. British Journal of General Practice, 63: (616): e726-e733.

Lambert, T.W., Evans, J. and Goldacre, M.J. (2002) Recruitment of UK-trained doctors into general practice: findings from national cohort studies. British Journal of General Practice, 52: 369-372.

Lambert, T.W., Goldacre, M.J., Edwards, C., et al. (1996) Career preferences of doctors who qualified in the United Kingdom in 1993 compared with those of doctors qualifying in 1974, 1977, 1980, and 1983. British Medical Journal, 313: 19-24.

Lambert, T.W., Goldacre, M.J. and Turner, G. (2003) Career choices of United Kingdom medical graduates of 1999 and 2000: questionnaire surveys. British Medical Journal, 326 194-196.

Lancaster, T. (2015) Editor's Choice [Letter on undergraduate time in general practice and recruitment to General Practice]. British Journal of General Practice, 65: (636): 340-340.

Landis, J.R. and Koch, G.G. (1977) The measurement of observer agreement for categorical data. Biometrics 33 (1): 159-174.

Levin, J. (1972) The occurrence of an increase in correlation by restriction of range. Psychometrika, 37: (1): 93-97.

Lievens, F. and Patterson, F. (2011) The validity and incremental validity of knowledge tetss, flow-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes seleciton. Journal of Applied Psychology, 96: (5): 927-940.

Linacre, J.M. (2004) FACETS Rasch measurement computer program. Chicago: Winsteps.com.

Little, R. J. A. & Rubin, D. (1987). Statistical analysis with missing data. Wiley.

Little, R. J. A. & Rubin, D. B. (2002). Statistical analysis with missing data. Second edition. Hoboken, NJ: John Wiley.

Lopes, S., V. Ashworth and R. Tate. (2013). "GP Selection QA Review 2013: Stage 2 Candidate Evaluation Final Report." from http://gprecruitment.hee.nhs.uk/Portals/8/Documents/Annual%20Reports/GP%20Stage%202%20evaluation%20report%202013.pdf [accessed 30/12/2015]

Maruyama, G.M. (1998) Basics of structural equation modelling. Thousand Oaks, California: Sage.

McManus, I.C. (1976) Archive for research data [letter]. Lancet, i: 1188-1188.

McManus, I.C. (1983) Bimodality of blood pressure levels. Statistics in Medicine, 2: 253-258.

McManus, I.C. (1985) Medical Students: Origins, selection, attitudes and culture. University of London: MD thesis (see http://www.ucl.ac.uk/medical-education/publications/md).

McManus, I.C. (2003) Medical School differences: beneficial diversity or harmful deviations? Quality and Safety in Health Care, 12: 324-325.

McManus, I.C. (2006) Supertext [Review of Bannister and Fransella's Inquiring Man]. Times Higher Education Supplement, May 26th, Textbook Guide: I-I.

McManus, I.C. (2012) Is it true that left-handed people are smarter than right-handed people? Scientific American Mind, May 2012: http://www.scientificamerican.com/article.cfm?id=is-it-true-that-left-handed-people.

McManus, I.C. (2013) The marking and standard setting of PLAB Part 2. London: Report to the PLAB Review, April 2013.

McManus, I.C., Davison, A. and Armour, J.A.L. (2013) Multi-locus genetic models of handedness closely resemble single locus models in explaining family data and are compatible with genome-wide association studies. Annals of the New York Academy of Sciences, 1288: 48-58.

McManus, I.C., Davison, I. and Taylor, C. (2015) Using the EM algorithm for handling restriction of range in reliability and validity coefficients. London: Unpublished manuscript.

McManus, I.C., Dewberry, C., Nicholson, S., et al. (2013) Construct-level predictive validity of educational attainment and intellectual aptitude tests in medical student selection: Meta-regression of six UK longitudinal studies. BMC Medicine, 11:243: doi:10.1186/1741-7015-1111-1243.

McManus, I.C., Jonvik, H., Richards, P., et al. (2011) Vocation and avocation: leisure activities correlate with professional engagement, but not burnout, in a cross-sectional study of UK doctors. BMC Medicine, 9: 100 (www.biomedcentral.com/1741-7015/1749/1100).

McManus, I.C., Livingston, G. and Katona, C. (2006) The attractions of medicine: the generic motivations of medical school applicants in relation to demography, personality and achievement. BMC Medical Education, 6: 11.

McManus, I.C. and Ludka, K. (2012) Resitting a high-stakes postgraduate medical examination on multiple occasions: nonlinear multilevel modelling of performance in the MRCP(UK) examinations. BMC Medicine, 10: ((doi:10.1186/1741-7015-10-60)): 60.

McManus, I.C., Ng-Knight, T., Riglin, L., et al. (2015) Doctor, Builder, Soldier, Lawyer, Teacher, Dancer, Shopkeeper, Vet: Exploratory study of which eleven-year olds would like to become a doctor. BMC Psychology, 3:38 (DOI: 10.1186/s40359-015-0094-z), 2015.

McManus, I.C., Richards, P. and Winder, B.C. (1999) Intercalated degrees, learning styles, and career preferences: prospective longitudinal study of UK medical students. British Medical Journal, 319: 542-546.

McManus, I.C., Richards, P., Winder, B.C., et al. (1995) Medical school applicants from ethnic minorities: identifying if and when they are disadvantaged. British Medical Journal, 310: 496-500.

McManus, I.C., Smithers, E., Partridge, P., et al. (2003) A levels and intelligence as predictors of medical careers in UK doctors: 20 year prospective study. British Medical Journal, 327: 139-142.

McManus, I.C. and Sproston, K.A. (2000) Women in hospital medicine: Glass ceiling, preference, prejudice or cohort effect? Journal of Epidemiology and Community Health, 54: 10-16.

McManus, I.C., Thompson, M. and Mollon, J. (2006) Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. BMC Medical Education, 6: 42 (http://www.biomedcentral.com/1472-6920/6/42/abstract).

McManus, I. C. & Vincent, C. A. (1997). Can future poor performance be identified during selection? In P.Lens & G. van der Waal (Eds.), Problem Doctors: A conspiracy of silence (pp. 213-236). Amsterdam: IOS Press.

McManus, I. C. & Wakeford, R. E. (2014). Data linkage comparison of PLAB and UK graduates' performance on MRCP(UK) and MRCGP examinations: Equivalent IMG career progress requires higher PLAB pass-marks. British Medical Journal, 348, g2621.

McManus, I.C., Winder, B.C. and Paice, E. (2002) How consultants, hospitals, trusts and deaneries affect pre-registration house officer posts: a multilevel model. Medical Education, 36: 35-44.

McManus, I.C., Woolf, K., Dacre, J., et al. (2013) The academic backbone: Longitudinal continuities in educational achievement from secondary school and medical school to MRCP(UK) and the Specialist Register in UK medical students and doctors. BMC Medicine, 11:242: doi:10.1186/1741-7015-1111-1242.

Meyer, J.P. (2010) Reliability. New York: Oxford University Press.

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. BMC Medical Research Methodology, 14.

Morrison, J. (2016). Selecting for medical education. Medical Education, 50, 3-4.

Migration Advisory Committee (2015) "Partial review of the Shortage Occupation Lists for the UK and for Scotland".

Musquash, C. and O'Connor, B.P. (2006) SPSS and SAS programs for generalizability theory analyses. Behavior Research Methods, 38: (3): 542-547.

National Recruitment, O. (2010) The scoring and psychometric properties for Stage 2 (Assessment 1) - February 2010. Birmingham: GP National Recruitment Office: Available at http://www.docstoc.com/docs/105084202/Properties-of-2010-Stage-2-assessment.

Newble, D. (2004) Techniques for measuring clinical competence: objective structured clinical examinations. Medical Education, 38: (2): 199-203.

NHS Employers (2014) "Pay and Conditions Circular (M&D) 2/2014". London.

Norcini, J. J., Boulet, J. R., Opalek, A., & Dauphinee, W. D. (2014). The relationship between licensing examination performance and the outcomes of care by international medical school graduates. Academic Medicine, 89, 1-6.

Parkhouse, J. (1991) Doctors' careers: aims and experiences of medical graduates. London: Routledge.

Parkhouse, J. and Howard, M. (1978) A follow-up of the career preferences of Manchester and Sheffield graduates of 1972 and 1973. Medical Education, 12: 377-381

Patterson, F., Archer, V., Kerrin, M., et al. (2011) Design and evaluation of a situational judgement test for selection to the Foundation Programme: Final report. London: Medical Schools Council (Improving Selection to the Foundation Programme); available at http://www.isfp.org.uk/AboutISFP/Documents/Appendix_F_-_Final_Report_of_SJT_pilots.pdf.

Patterson, F., Ashworth, V., & Good, D. (2015). Situational judgment tests: A guide for applicants to the UK Foundation Programme (2nd edition). Cardiff: UK Foundation Programme Office (available at: http://www.foundationprogramme.nhs.uk/pages/medical-students/SJT-EPM).

Patterson, F., Baron, H., Carr, V., et al. (2009) Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. Medical education, 43: (1): 50-57.

Patterson, F., Carr, V., Zibarras, L., et al. (2009) New machine-marked tests for selection into core medical training: evidence from two validation studies. Clinical Medicine, 9: (5): 417-420.

Patterson, F. and Empey, L. (2012a) Psychometric analysis and QA of the GP Stage 3 Selection Process (2010): Results for Rounds 1 and 2: Executive Summary. Derby: Work Psychology Group (available from http://gprecruitment.hee.nhs.uk/Portals/8/Documents/Annual%20Reports/Psychometric%20Analysis%20and%20QA%20of%20the%20GP%20Stage%203%20-%20Executive%20Summary%202010.pdf).

Patterson, F. and Empey, L. (2012b) Psychometric analysis and QA of the GP Stage 3 Selection Process (2011): Results for Rounds 1 and 2: Executive Summary. Derby: Work Psychology Group (available from http://gprecruitment.hee.nhs.uk/Portals/8/Documents/Annual%20Reports/Psychometric%20Analysis%20and%20QA%20of%20the%20GP%20Stage%203%20-%20Executive%20Summary%202011.pdf).

Patterson, F. and Empey, L. (2012c) Psychometric analysis and QA of the GP Stage 3 Selection Process (2012): Results for Rounds 1 and 2: Executive Summary. Derby: Work Psychology Group (available from http://gprecruitment.hee.nhs.uk/Portals/8/Documents/Annual%20Reports/Psychometric%20Analysis%20and%20QA%20of%20the%20GP%20Stage%203%20-%20Executive%20Summary%202012.pdf).

Patterson, F., Ferguson, E., Lane, P., et al. (2000) A competency model for general practice: implications for selection, training, and development. British Journal of General Practice, 50: (452): 188-193.

Patterson, F., Ferguson, E., Norfolk, T., et al. (2005) A new selection system to recruit general practice registrars: Preliminary findings from a validation study. British Medical Journal, 330: 711-714.

Patterson, F., Kerrin, M., & Ashworth, V. (2015). GP Selection Review: Data modelling of Stage 2 scores & impact on selection centre numbers [September 2015]. Derby: Work Psychology Group.

Patterson, F., Kerrin, M., Baron, H., & Lopes, S. (2015). Exploring the relationship between General Practice selection scores and MRCGP examination performance. Final Report to General Medical Council, September 2015. Derby: Work Psychology Group.

Patterson, F. and Lane, P. (2007) Assessment for recruitment. Assessment in medical education and training. Oxford: Radcliffe Publishing, 62-73.

Patterson, F., Lane, P., Ferguson, E., et al. (2001) Competency based selection system for general practitioner registrars. BMJ Career Focus, 323: (7311): S2-7311-27311.

Patterson, F., Lievens, F., Kerrin, M., et al. (2013) The predictive validity of selection for entry into postgraduate training in general practice: Evidence from three longitudinal studies. British Journal of General Practice, DOI:10.3399/bjgp13x674413: e734-e741.

Patterson, F., A. Tavabie, M. Denney, et al. (2013). "A new competency model for general practice: implications for selection, training, and careers." Br J Gen Pract 63(610): e331-338.

Pau, A., K. Jeevaratnam, Y. S. Chen, et al. (2013). "The Multiple Mini-Interview (MMI) for student selection in health professions training - a systematic review." Med Teach 35(12): 1027-1041.

Payne, T., Anderson, N. and Smith, T. (1992) Assessment centres, selection systems and cost-effectiveness: An evaluative case study. Personnel Review, 21: (4): 48-56.

Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI: On the influence of natural selection on the variability and correlation of organs. Phil.Trans.R.Soc.Lond.A, 200, 1-66.

Peile, E. (2013) General practice careers: choices and judgements. British Journal of General Practice, 63: (616): 565-566.

Petrides, K.V. and McManus, I.C. (2004) Mapping medical careers: Questionnaire assessment of career preferences in medical school applicants and final year students. BMC Medical Education, 4: 18.

Pigott, T.D. (2001) A review of methods for missing data. Educational Research and Evaluation, 7: (4): 353-383.

Plint, S. (2014) Securing the Future GP Workforce Delivering the Mandate on GP Expansion GP TASKFORCE FINAL REPORT.

PSSRU (2014) "Unit costs of health and social care 2013-14". Canterbury.

Raghunathan, T. (2016). Missing data analysis in practice. Boca Raton, FL: CRC Press.

Roberts, C., S. Atkins and K. Hawthorne (2014). Performance features in clinical skills assessment: Linguistic and cultural factors in the Membership of the Royal College of General Practitioners examination, King's College London.

Royal Commission. (1968) Royal Commission on Medical Education (The Todd Report), Cmnd 3569. London: HMSO.

Schmidt, F. L. & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Schmidt, F. L. & Oh, I.-S. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. Personnel Psychology, 59, 281-305.

Swanson, D., Norman, G. and Linn, R. (1995) Performance-Based Assessment: Lessons from the Health Professions. Educational Researcher 24: (5): 5 - 11

Tamblyn, R., Abrahamowicz, M., Dauphinee, W.D., et al. (2002) Association between licensure examination scores and practice in primary care. JAMA, 288: (23): 3019-3026.

Thomas, H., Taylor, C., Davison, I., et al. (2010) "National Evaluation of Specialty Selection: Final Report". Birmingham, University of Birmingham.

Thorndike, R. L. (1949). Personnel selection. New York: Wiley.

Tiffin, P.A., Illing, J., Kasim, A.S., et al. (2014) Annual Review of Competence Progression (ARCP) performance of doctors who passed Professional and Linguistic Assessments Board (PLAB) tests compared with UK medical graduates: national data linkage study. British Medical Journal, 348: (g2622).

Tighe, J., McManus, I.C., Dewhurst, N.G., et al. (2010) The Standard Error of Measurement is a more appropriate measure of quality in postgraduate medical assessments than is reliability: An analysis of MRCP(UK) written examinations. BMC Medical Education (www.biomedccentral.com/1472-6920/10/40), 10: 40.

U. K. Foundation Programme Office (2010) Foundation Programme Annual Report 2010: National (UK) Summary. Cardiff: UK Foundation Programme Office (available at http://www.foundationprogramme.nhs.uk/download.asp?file=Foundation_Programme_Annual_Report_Nov10.pdf).

U. K. Foundation Programme Office (2011) Foundation Programme Annual Report 2011: UK Summary. Cardiff: UK Foundation Programme Office (available at http://www.foundationprogramme.nhs.uk/download.asp?file=Foundation_Programme_Annual_Report_Nov11_FINAL.pdf).

U. K. Foundation Programme Office (2012a) Foundation Programme Annual Report 2012: UK Summary. Cardiff: UK Foundation Programme Office (available at http://www.foundationprogramme.nhs.uk/download.asp?file=UK_Foundation_Programme_Annual_Report_2012_FINAL.pdf).

U. K. Foundation Programme Office (2012b) National F2 Career Destination Survey (2012). Cardiff: UKFPO: Available at http://www.foundationprogramme.nhs.uk/download.asp?file=F2_career_destination_report_December_2012.pdf.

U. K. Foundation Programme Office (2013a) F2 Career Destination Report 2013. Cardiff: UKFPO: Available at http://www.foundationprogramme.nhs.uk/download.asp?file=F2_career_destination_report_November_2013.pdf.

U. K. Foundation Programme Office (2013b) Foundation Programme Annual Report 2013: UK Summary. Cardiff: UK Foundation Programme Office (available at http://www.foundationprogramme.nhs.uk/download.asp?file=UK_Foundation_Programme_Annual_Report_2013_FINAL.pdf.

U. K. Foundation Programme Office (2014a) F2 Career Destination Report 2014. Cardiff: UKFPO: Available at http://www.foundationprogramme.nhs.uk/download.asp?file=F2_career_destination_report_2014_-_FINAL_-_App_A_updated.pdf.

U. K. Foundation Programme Office (2014b) Foundation Programme Annual Report 2014: UK Summary. Cardiff: UK Foundation Programme Office (available at http://www.foundationprogramme.nhs.uk/download.asp?file=FP_Annual_Report_2014_-_FINAL_Nov_2014.pdf).

U. K. Foundation Programme Office (2015) Foundation Programme Annual Report 2015: UK Summary. Cardiff: UK Foundation Programme Office (available at http://www.foundationprogramme.nhs.uk/pages/home/keydocs).

van der Vleuten, C.P.M. (1996) The assessment of professional competence: Developments, research & practical implications. Advances in Health Sciences Education, 1: 41-67.

van der Vleuten, C.P.M. and Schuwirth, L.W.T. (2005) Assessing professional competence: from methods to programmes. Medical Education, 39: (3): 309-317.

Wakeford, R. (2014) Predictive validity of selection for entry into postgraduate training in general practice. British Journal of General Practice, DOI: 10.33999/bjgp14X677068: 71-71.

Wakeford, R., Denney, M. L., Ludka-Stempien, K., Dacre, J., & McManus, I. C. (2015). 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, 15.

Wakeford, R., Foulkes, J., McManus, I.C., et al. (1993) MRCGP pass rate by medical school and region of postgraduate training. British Medical Journal, 307: 542-543.

Webb, N.M., Shavelson, R.J. and Haertel, E.H. (2007) "Reliability coefficients and generalizability theory". In Rao, C.R. & Sinharay, S. (Eds.) Handbook of Statistics, Volume 26: Psychometrics. Amsterdam, Elsevier 81-124.

Weekley, J.A. (2004) Scoring situational judgement tests: Does the middle matter? Kenexa HR newsletter.

Wenghofer, E., Klass, D., Abrahamowicz, M., et al. (2009) Doctor scores on national qualifying examinations predict quality of care in future practice. Medical education, 43: (12): 1166-1173.

Wiberg, M. and Sundstrom, A. (2009) A comparison of two approaches to correction of restriction of range in correlation analysis. Practical Assessment, Research & Evaluation, 14: (5): 1-8.

Work Psychology Group (2011) GP Stage 2 Psychometric Report Round 1. Work Psychology Group, Jan 2011.

Work Psychology Group (2012) GP Stage 2 Psychometric Report Round 1. Work Psychology Group, Jan 2012.

Work Psychology Group (2013) GP Stage 2 Psychometric Report Round 1. Work Psychology Group, Jan 2013.

Work Psychology Group (2014) GP Stage 2 Psychometric Report Round 1. Work Psychology Group, Jan 2014.

Work Psychology Group (2015a) GP selection process: Summary of Evaluation Evidence. Unpublished document, Work Psychology Group.

Work Psychology Group (2015b) GP Stage 2 Psychometric Report Round 1. Work Psychology Group, Jan 2015.

List of Abbreviations

List of Abbreviations.

ACCS	Acute Core Common Stem
AKT	Applied Knowledge Test
ANOVA	Analysis of Variance
ARCP	Annual Review of Competence Progression
ASME	Association for the Study of Medical Education
BBT	Broad Based Training
BME	Black and Minority Ethnic
BMJ	British Medical Journal
CCT	Certificate of Completion of Training
CEA	Cost-effectiveness Analysis
CEGPR	Certificate of Eligibility for General Practice Registration
CER	Cost-effectiveness Ratio
CMT	Core Medical Training
CPS	Clinical Problem Solving
CPS(T)	Clinical Problem Solving (Test)
CS	Communication skills
CSA	Clinical Skills Assessment
CST	Core Surgical Training
CT&PS	Conceptual thinking and problem solving

CT1	Core Training Year 1
D-studies	Decision studies
EM algorithm	Expectation Maximization algorithm
EPM	Educational Performance Measure
ES	Empathy and Sensitivity
F1/FY1	Foundation Programme (doctor) Year 1
F2/FY2	Foundation Programme (doctor) Year 2
FPAS	Foundation Programme Application System
FTE	Full time equivalent
FTSTA	Fixed Term Specialty Training Appointment
FtP	Fitness to Practise
GMC	General Medical Council
GPNRO	General Practice National Recruitment Office
GPR	GP Registration
G-studies	Generalisability studies
HEE	Health Education England
ICER	Incremental Cost-effectiveness Ratio
iMRCS	Intercollegiate Membership of the Royal College of Surgeons
IQR	Inter-quartile range
IRT	Item Response Theory
KSS	Kent, Surrey and Sussex (LETB)

LAT	Locum Appointment for Training
LETB	Local Education and Training Board
LRMP	List of Registered Medical Practitioners
LTFT	Less than full time
MAR	Missing at random
MBBS/MBChB	Bachelor of Medicine & Surgery
MCMC	Markov Chain Monte Carlo
MCAR	Missing completely at random
MCQ	Multiple choice question
MCRG	UK Medical Careers Research Group
MDRS	Medical and Dental Recruitment Service
MEE	Medical Education England
MFRM	Many Facet Rasch Modelling
MNAR	Missing not at random
MRCGP	Membership of the Royal College of General Practitioners
MTAS	Medical Training Applications Service
mTiffin	Modified Tiffin Score (see Tiffin et al., 2014)
MVA	Missing Value Analysis (SPSS)
NICE	National Institute for Health and Care Excellence
NTN	National Training Number
O&G	Obstetrics & Gynaecology

OOP	Out of Programme
OSCE	Objective Structured Clinical Examination
PACES	Practical Assessment of Clinical Examination Skills
PI	Professional Integrity
PMM	Predictive mean matching
PMQ	Primary Medical Qualification
PRHO	Pre-Registration House Officer (now the F1 grade)
PSSRU	Personal and Social Services Research Unit
QALY	Quality Adjusted Life Year
RCGP	Royal College of General Practitioners
SD	Standard deviation
SEM	Standard Error of Measurement
SF	Significant figures
SHA	Strategic Health Authority
SJT	Situational Judgement Test; sometimes this is called the Professional Dilemmas Test (PDT)
SRA	Specialty Recruitment Assessment
ST1/2/3/4	Specialty Training Year 1/2/3/4
TOOT	Time out of Training
UCAS	Universities and Colleges Admissions Service
UKFPO	UK Foundation Programme Office
VTS	Vocational Training Scheme

WPBA	Work-place Based Assessment
WPG	Work Psychology Group

~ This page is intentionally left blank ~

SCHOOL OF EDUCATION
UNIVERSITY OF BIRMINGHAM

Edgbaston
Birmingham
B15 2TT