



Improving Selection to the Foundation Programme

Final Report

Medical Schools Council
on behalf of the cross-stakeholder Project Group

August 2011

Acknowledgements

The project described in this report has involved over 1500 people, most of whom have given their time on a voluntary basis. The project team would particularly like to thank:

- The Department of Health, for commissioning the project
- The members of the Project Board and Project Group, for their advice and guidance
- The volunteers from across the UK who have written SJT items
- The foundation doctors and concordance panel members who have reviewed SJT items
- Medical school staff who have helped develop the EPM, and have organised and run the pilot SJTs
- The students who gave up their time to sit the SJT
- The student representatives at medical schools who helped to promote the pilots
- The BMA Medical Students Committee, for their support
- The UK Foundation Programme Office, and the Rules Group, for their advice and support
- The NHS patients, nurses, and doctors who helped with the Job Analysis
- Our specialist advisors and suppliers

As Chair of the ISFP Project Group I should like to record formally Siobhan Fitzpatrick's unstinting dedication to this project and to thank her for her exceptional contribution. Ably assisted by Amy Stringer, and latterly, Oliver Watson, the team has worked tirelessly to deliver this project on time – and within budget – thus permitting the forthcoming Parallel Recruitment Exercise to be undertaken.

Professor Paul O'Neill
Chair, ISFP Project Group

Table of Contents

LIST OF FIGURES	4
LIST OF APPENDICES	4
GLOSSARY	5
1. EXECUTIVE SUMMARY	6
2. INTRODUCTION	7
3. BACKGROUND	7
3.1. Existing arrangements for selection to the Foundation Programme	8
3.2. The case for change	9
4. ISFP PROJECT PLAN	12
4.1. Objectives	12
4.2. Approach and timescales	12
4.3. Management and Governance	13
5. JOB ANALYSIS OF THE ROLE OF THE FY1 DOCTOR	14
5.1. Methods	14
5.2. Professional Attribute Framework	15
6. SJT ITEM WRITING AND QA PROCESS	17
6.1. SJT design	17
6.2. Item development and review	18
6.3. Item writing workshops	19
6.4. Initial item review	20
6.5. Lead item review	21
6.6. Foundation doctor focus groups	21
6.7. Concordance Panel	22
6.8. Test construction	22
6.9. Scoring convention	23
7. SJT PILOTS	24
7.1. Overview	24
7.2. Methods of SJT delivery	24
7.3. Administration	28
7.4. Participation	31
7.5. Feedback from pilots	31
7.6. Initial pilot - key findings	32
7.7. Large-scale pilot - key findings	33
8. EDUCATIONAL PERFORMANCE MEASURE (EPM)	34
8.1. Background	34
8.2. Consultation and piloting	35
8.3. Pilot of draft EPM framework	36
8.4. EPM pilot lessons	38
8.5. EPM Task and Finish Group (EPM Group)	38
8.6. Draft framework	38
8.7. Performance in relation to graduating cohort	39
8.8. Degrees, presentations, prizes and publications	39
8.9. Extra-curricular activities	39
9. ASPECTS OF THE APPLICATION PROCESS USING SJT AND EPM	40
9.1. Aggregating the SJT and EPM scores	40
9.2. Algorithm	40
9.3. Development and testing of SJT item banking software	41

9.4	Defence Medical Services	42
9.5	Academic Foundation Programmes (AFP)	42
10.	COMMUNICATIONS	43
10.1	Development and sign off	43
10.2	Key achievements	43
10.3	Website	43
10.4	Social Media	43
10.5	E-bulletin	43
10.6	Communication documents	44
10.7	MedSoc Presidents	44
10.8	Pilot workshops	44
10.9	Presentations	44
10.10	Press	44
11.	LEGAL ADVICE	45
11.1	SJT results of overseas applicants	45
11.2	Holding the SJT in the UK	46
11.3	Skewing of EPM scores	46
11.4	Failure to provide a decile score as part of the EPM	47
11.5	Weighting SJT and EPM scores	47
12.	MANAGEMENT LESSONS LEARNED	47
12.1	Scheduling and managing a national SJT	47
12.2	The production of SJT items	48
12.3	The performance of the SJT as a selection method	48
12.4	The use of volunteers for SJT development	48
12.5	Paper and electronic delivery of an SJT	49
12.6	The variety of assessments across medical schools	49
12.7	Communication	49
12.8	Ongoing evaluation	50
12.9	Recommendation: Parallel Recruitment Exercise (PRE)	50
13.	COST BENEFIT ANALYSIS (CBA)	50
13.1	Original CBA	50
13.2	Updated CBA	53
13.3	Costs	53
13.4	Feasibility	54
13.5	The case for change revisited	54
14.	THE WAY FORWARD	55
14.1	Final recommendations	55
14.2	Operational Guidance for FP2013	55
14.2.1	SJT Operational Guidance	55
14.2.2	EPM Operational Guidance	56
15.	SUCCESS CRITERIA	57

List of Figures

Figure 1: Project Plan	13
Figure 2: Governance structure for ISFP project	13
Figure 3: Membership of the ISFP Project Board	13
Figure 4: Membership of the ISFP Project Group	14
Figure 5: Professional Attributes Framework	16
Figure 6: Relationship between Professional Attributes and SJT target domains	17
Figure 7: Process of SJT item development and review	19
Figure 8: SJT item writer demographics and job roles	20
Figure 9: Total number of SJT items written by target domain	20
Figure 10: Focus group participant demographics	21
Figure 11: Scoring key for SJT ranking items	23
Figure 12: Methods of SJT test delivery	27
Figure 13: Administrative lessons for SJT delivery	30
Figure 14: SJT pilot participant demographics (initial pilot, Autumn 2010)	32
Figure 15: SJT pilot participant demographics (large-scale pilot, Spring 2011)	33
Figure 16: Participation in three papers of large-scale pilot	33
Figure 17: Alternative selection tools considered during Options Appraisal	51
Figure 18: Relative costs of selection tools in Options Appraisal	51
Figure 19: Criterion for non-financial costs and benefits	52
Figure 20: Non-financial cost benefit analysis	52
Figure 21: The case for change revisited	54

List of Appendices

Appendix A: Legal opinion, July 2011	59
Appendix B: Characteristics of the current system for selection to the Foundation Programme	71
Appendix C: Project Initiation Document, February 2010	77
Appendix D: FY1 Job Analysis	125
Appendix E: Person Specification FP2012	241
Appendix F: Report of the SJT Pilots	245
Appendix G: SJT Pilot Administrators Guidance	375
Appendix H: Report of EPM framework pilot	389
Appendix I: Report of the EPM Task and Finish Group	417
Appendix J: Report of aggregating SJT and EPM scores	435
Appendix K: ISFP Communications Plan	445

Glossary

AFP	Academic Foundation Programme
AoMRC	Academy of Medical Royal Colleges
BMA	British Medical Association
BMA MSC	British Medical Association Medical Students Committee
BME	Black and Minority Ethnic
BMJ	British Medical Journal
CBA	Cost Benefit Analysis
COPMeD	Conference of Postgraduate Medical Deans
CPD	Continuing Professional Development
DH	Department of Health
DPMD	Defence Postgraduate Medical Deanery
EEA	European Economic Area
EMQ	Extended Matching Question
EPM	Educational Performance Measure
FP	Foundation Programme
FPAS	Foundation Programme Application System
FY1/FY2	First/ Second year of an integrated two year Foundation Programme
GMC	General Medical Council
GP	General Practice
ISFP	Improving Selection to the Foundation Programme project
MCQ	Multiple Choice Question
MDAP	Multi-Deanery Application Process
MEE	Medical Education England
Mini-CEX	Mini Clinical Evaluation Exercise
MMI	Multiple Mini Interview
MoD	Ministry of Defence
MSC	Medical Schools Council
MSC-AA	Medical Schools Council Assessment Alliance
MTAS	Medical Training Application System
NACT	National Association of Clinical Tutors
NEAF	National Education Advisors Forum
NEO PI-R	Psychological personality measure
NES	NHS Education for Scotland
NHSE	NHS Employers
NIMDTA	Northern Ireland Medical and Dental Training Agency
OSCE	Objective Structured Clinical Examination
PID	Project Initiation Document
PLAB	Professional and Linguistic Assessments Board
PRE	Parallel Recruitment Exercise
Prince 2	Method of project management
QABME	Quality Assurance of Basic Medical Education
SBA	Single Best Answer
SJT	Situational Judgement Test
SSC	Student Selected Components
UKCDR	UK Collaboration for a Digital Repository
UKFPO	UK Foundation Programme Office

1. Executive Summary

Since 2006 there has been a UK-wide process for selecting graduating doctors to the integrated two year Foundation Programme training in the NHS. Eligible applicants are selected according to an academic quartile score reflecting educational performance in relation to the graduating medical school cohort; evidence of degrees, presentations, prizes and publications; and answers to 'white space' questions on an online application form.

In 2009, in response to concerns about the longevity of the current national application process, the Department of Health commissioned a review of selection to the Foundation Programme, with the aim of recommending a more reliable, robust, valid, feasible and sustainable method for selection which would minimise the risk of successful legal challenge. The ensuing work was a collaborative venture led by the Medical Schools Council (MSC), involving the Academy of Medical Royal Colleges (AoMRC), the British Medical Association (BMA) Medical Students Committee, the Conference of Postgraduate Medical Deans (COPMeD), the General Medical Council (GMC), The National Association of Clinical Tutors (NACT), the Northern Ireland Medical and Dental Training Agency (NIMDTA), NHS Employers, the Scottish Board for Academic Medicine, the Scottish Foundation Board, the UK Foundation Programme Office (UKFPO) and the four UK Health Departments.

The Options Appraisal¹ involved a thorough and detailed appraisal of the methods that might be used for selecting applicants to the Foundation Programme. The Steering Group concluded that the most appropriate selection method would be an invigilated 'Situational Judgement Test' (SJT) to replace 'white space' questions, and a more granular and standardised 'Educational Performance Measure' (EPM) to replace the 'academic quartile' scores.

The DH accepted the recommendations that the SJT and EPM should be developed and piloted for use in the context of selection to the Foundation Programme. In January 2010, the DH commissioned the MSC to run an 18-month project to complete the development and piloting. Representatives from the fifteen stakeholder organisations represented through the initial Options Appraisal convened as the ISFP Project Group, with some small changes in membership, to oversee the project. This document is the final report of the Improving Selection to the Foundation Programme project.

The key achievements of the project are as follows:

- An in-depth literature review on the use of SJTs for selection, particularly in the context of medicine
- An in-depth literature review of the methods of selection into medical education and training up to and including the Foundation Programme
- A formal Job Analysis of the role of the FY1 doctor, which identified nine professional attributes (Commitment to Professionalism, Coping with Pressure, Effective Communication, Learning and Professional Development, Organisation and Planning, Patient Focus, Problem Solving and Decision Making, Self Awareness and Insight and Working Effectively as Part of a Team) and 111 individual behaviours
- The production and evaluation of a design specification (number of items per SJT, duration of the SJT, scoring convention, etc) for a reliable and valid SJT, based on the target professional domains identified in the Job Analysis of the role of the FY1 doctor
- Volunteers from across the UK have been trained to write SJT items against the target professional domains (known as SJT 'items')
- A significant number of SJT items has been developed, quality assured and tested
- The SJT has been piloted by more than 1100 students from 15 UK and 2 non-UK medical schools
- The results of the SJT pilots have been subject to psychometric analysis, which has shown the SJT to be a reliable selection method in this context, capable of differentiating between applicants
- Standards for the delivery and administration of the SJT have been defined

¹ Medical Schools Council (2009) *Selection into the Foundation Programme: An Option Appraisal*

- 'Item banking' software, for the secure storage of the SJT items, has been developed and tested
- Two consultations with all UK medical schools around the number and type of assessments within the undergraduate medical degree programme
- A draft EPM framework has been piloted by 25 UK medical schools and evaluated
- Rules for the production of EPM scores by medical schools have been defined and agreed
- Rules have been developed and agreed for combining SJT and EPM results to give an overall score to applicants for the purpose of selection to the Foundation Programme
- A significant programme of consultation, stakeholder engagement and communication has been completed

The results of the work to develop and pilot new methods of selection demonstrate that the use of the SJT and EPM in combination, to replace the assessment methods currently used, will make selection to the Foundation Programme more reliable, robust and cost effective.

Summary of recommendations

- Selection to the Foundation Programme should reflect the skills, knowledge and professional behaviours of the applicant, reflecting the integrated nature of the Foundation Programme as both education and employment
- Selection to the Foundation Programme from FP2013 onwards should be based upon:
 - o An invigilated Situational Judgement Test (SJT) to assess aptitude for the Foundation Programme (to replace 'white space' questions); and
 - o An Educational Performance Measure (EPM) to reflect educational performance at medical school up to the point of application to the Foundation Programme (to replace quartiles)
- Other aspects of the process of application to the Foundation Programme should remain unchanged, namely a national application process including applicant declaration of foundation school preferences, the use of a matching algorithm and eligibility checking
- There should be a full-scale shadow Parallel Recruitment Exercise (PRE) SJT and EPM run concurrently with the current application process for FP2012
- Using data from tracking, there should be ongoing evaluation of the validity of the SJT and EPM as methods for selection to the Foundation Programme

2. Introduction

This document is the final report of the ISFP Project Group. The aim of the report is:

- To provide the UK health departments with detailed recommendations about whether and how SJTs and EPM scores should be introduced for the selection of applicants to the Foundation Programme; and
- To explain the work and evidence underpinning the recommendations

This report is not designed to be a guide for medical schools, students and others with an interest in selection to the Foundation Programme. Separate guidance will be produced if the recommendations in this report are accepted.

The report begins by describing the background to the project; it then summarises the work done and the findings, before presenting the recommendations. Supporting detail is provided in the appendices.

More information about the project can be found on the ISFP website at www.ISFP.org.uk

3. Background

This chapter summarises the background to the project to allow a general reader to appreciate the context of the report. It starts with an explanation of the existing arrangements for the selection of final year undergraduate medical

students from UK medical schools and eligible non-UK applicants to the Foundation Programme, and then summarises the case for change that has led to the project.

3.1. Existing arrangements for selection to the Foundation Programme

The UK Foundation Programme is a two-year structured integrated training programme for newly graduated doctors, designed to give trainees a range of experience in secondary and primary care environments before beginning specialty training². Successful completion of Foundation Year One (FY1) is a requirement for full GMC registration. The Foundation Programme represents a full-time period of education, training and employment, and selection to the Foundation Programme is thus subject to European employment law. Foundation Programme training is organised and delivered by foundation schools, who liaise with the NHS organisations in their area to ensure the availability of suitable training posts. Confirming the successful completion of the FY1 year is formally the responsibility of the universities that have medical schools under the Medical Act 1983 and its subsequent amendments, although in practice this is delegated to the local postgraduate dean.

Since 2005 there has been a UK-wide application process for selection to the Foundation Programme, consisting of a national, online application open to both UK and non-UK applicants who meet the eligibility criteria. Applicants are awarded an application score, comprised as follows:

- Academic quartiles to reflect educational performance at medical school (40%); and
- Additional academic achievements (degrees, presentations, prizes, publications) (10%); and
- Answers to free text 'white space' questions, which assess the national Person Specification (50).

The graduating medical school provides an academic quartile score (34, 36, 38 or 40 points) to reflect educational performance relative to all of the other students in the applicant's year at that school. Applicants provide evidence of additional academic achievements, which are verified by a national panel against specific criteria. 'White space' questions and their scoring key are developed by the UKFPO Rules Group to assess the elements of the national Person Specification not relating to clinical knowledge and skills, with an extensive Quality Assurance process. The scoring of answers to the 'white space' questions is organised by foundation schools, and is undertaken by trained representatives - mainly clinicians - from the NHS.

The application score, derived by adding the scores for academic quartiles, additional academic achievements and 'white space' questions, are used to determine which applicants should be selected into the national Foundation Programme, and to allocate successful applicants to individual foundation schools. The allocation of applicants to foundation schools is determined by a set of rules (the allocation algorithm), which take into account the applicants' scores and preferences in a particular way. If there are more applicants than posts, then the highest scoring applicants are selected³. If there are more posts than applicants, then all eligible applicants are considered for allocation. Entry into the Foundation Programme is competitive. In 2010, about 7,700 students graduated from UK medical schools, and forecasts from the Medical Schools Council suggest that this number will rise to nearly 8,000 over the next two years. There were also 1,605 applications to the Foundation Programme from graduates of non-UK medical schools; of which 270 were eligible. In total there were 7,253 eligible applicants for 7,073 Foundation Programme places.

Selection to the Foundation Programme comprises three main stages: a) selecting the most suitable eligible applicants to the Foundation Programme, b) allocation to a specific foundation schools (national process) and c) matching to a specific programme within the foundation school (local process). This project to review the methods for national selection to the Foundation Programme relates to (a) and (b) only, but the processes involved with matching to

² See www.foundationprogramme.nhs.uk

³ This is a simplification. Under the prevailing right-to-work legislation, applicants who have an unconditional right to work in the UK will be selected for places first, in order of score. Any unfilled places will then be allocated to the remaining applicants (e.g. those without an unconditional right to work in the UK), again in order of score.

individual programmes remains the responsibility of the relevant foundation schools, and is outside of the scope of the ISFP project. Further detail about the related characteristics of this matching process, including special circumstances and linked applications, are provided in Appendix B.

The annual application process, which is known as the Foundation Programme (FP) recruitment round, starts in the autumn of one year, the results are announced in the spring of the following year, and the applicants take up their posts at the beginning of August in that year. For most applicants, the process runs in parallel with their final year of undergraduate education.

The UK Foundation Programme Office (UKFPO) administers the recruitment rounds under contract to the four UK Health Departments. There is a separate contract with the Department of Health in England (DH) for the provision of the online system, known as the Foundation Programme Application System (FPAS), through which the applications, scoring and allocations are managed. A stakeholder Rules Group maintains the rules and standards for the application process. This includes setting the timetable for each round, agreeing the 'white space' questions and the associated scoring guidelines, quality assuring the scoring, and adjudicating over special cases.

The guiding principles and parameters for the Improving Selection to the Foundation Programme project are that:

- In order to protect patients and to reassure employers, it is necessary to ensure that all foundation doctors are fit for purpose, and, in the event of over subscription to the Foundation Programme, the most appropriate applicants are selected
- Selection to the Foundation Programme should be informed by a measure of clinical skills and knowledge
- Selection to the Foundation Programme should be informed by a measure of meeting the national Person Specification
- Performance in medical school should inform selection to the Foundation Programme
- Allocation to foundation programmes occurs before final year examinations have been taken, and so it is not possible to use performance in Finals as a selection tool

3.2. The case for change

Surveys conducted after each recruitment round have shown a growing level of satisfaction among applicants. However, there are shared anxieties amongst key stakeholders relating to the individual selection tools and the potential for the application process to be subject to legal challenge, as follows:

- The marking of the answers to the 'white space' questions is labour intensive, the cost approaching £2m per year in clinician time
- Applicants are allowed to complete the 'white space' questions online at their convenience. While this is logistically simple, it means that there is no real safeguard that the answers are the genuine sole work of the applicant
- Model answers to the 'white space' questions have been offered for sale over the internet. There are concerns that the impact of such model answers is likely to increase with time, as it will become increasingly difficult to write entirely fresh questions for each successive recruitment round, and the discrimination between applicants will diminish
- While the 'white space' questions appear to offer a practical way to rank large numbers of comparable applicants, and are derived from the national Person Specification, their technical reliability and validity is open to question
- There have been questions about the extent to which the academic quartile scores can be equated across medical schools, particularly since each school derives its scores as it sees fit rather than following a standard approach

- Given these concerns, if there is an increase in the number of eligible non-UK applicants such that the Foundation Programme is over-subscribed, there are concerns that the selection process may be subject to legal challenge from unsuccessful applicants to the Foundation Programme

To date, the impact of these concerns has been softened considerably by the fact that there have been fewer eligible applicants than Foundation Programme places up to the FP2011 recruitment round (meaning that all applicants can get a job) and historically around 90% of applicants have been allocated to their first preference foundation school. Hence, the process has generally been seen as positive by those involved. However, this perspective is somewhat misleading. The percentage of applicants who get their first choice is a product of the matching algorithm that is used and the relative demand for places at different foundation schools - it is not dependent on the application process or the scoring method. Furthermore, there is no guarantee that the number of posts will continue to exceed the number of applicants. The effect of European Directives on the Mutual Recognition of Professional Qualifications⁴ is that applicants from the European Union are entitled to compete equally for Foundation Programme training, provided that they meet the eligibility criteria; and in any event non-EU applicants can apply subject to the prevailing 'right to work' arrangements. In the event that the number of applicants exceeds the number of posts - as is likely in future - then it would be even more important that the application process ensures that the 'best' applicants are selected, and that the process is not open to likely successful legal challenge.

The 2008 Next Stage Review⁵ recognised that the current arrangements were not sufficiently robust, and recommended new work to develop a more reliable and valid selection process. As a result, the Department of Health (DH) commissioned the Medical Schools Council to set up a cross-stakeholder Steering Group to undertake an options appraisal for selection to the Foundation Programme, with the aim of recommending a more reliable, robust, valid, feasible and sustainable method for selection which would minimise the risk of successful legal challenge. The ensuing work was a collaborative venture between fifteen stakeholder organisations⁶, and reviewed evidence for the use of selection tools including personality questionnaires, structured interviews, unstructured interviews, multiple mini interviews, assessment centres, national examinations, portfolios, structured records of achievement, educational performance measurements, situational judgement tests, through:

- Surveys and consultations
- Meetings with stakeholders
- Three independent academic literature reviews
- A report from a panel of international experts in assessment and selection
- An independent cost-benefit analysis

The ISFP Steering Group submitted its final report of the Options Appraisal to the DH in September 2009, in which it made recommendations about the Foundation Programme application process. The relevant findings of the Options Appraisal are summarised as follows:

- The current selection tools, 'white space' questions and academic quartiles, are not sustainable for use in the long term, and the evidence of the Option Appraisal demonstrated a need for change
- As the Foundation Programme is an integrated training programme of education and employment, the selection of applicants should take into account their wider professional attributes for employment as well as their clinical knowledge and skills

⁴ Article 53 of the Treaty on the Functioning of the EU and Directive 2005/36/EC

⁵ DH (2008) *The Next Stage Review: A High Quality Workforce*, Para 32

⁶ The Academy of Medical Royal Colleges (AoMRC), the British Medical Association (BMA) Medical Students Committee, the Conference of Postgraduate Medical Deans (CoPMED), the General Medical Council (GMC), the Medical Schools Council (MSC), the National Association of Clinical Tutors (NACT), the Northern Ireland Medical and Dental Training Agency (NIMDTA), NHS Employers (NHSE), the Scottish Board for Academic Medicine, the Scottish Foundation Board, the UK Foundation Programme Office (UKFPO) and the four UK Health Departments.

- The three selection tools shortlisted for use as a measure of an applicant's wider professional attributes for employment, as defined by the national person specification, were i) structured interviews, ii) Multiple Mini Interviews (MMIs) and iii) a Situational Judgement Test (SJT)
 - Whilst structured interviews are popular with prospective applicants, they would not be a robust basis for selection on the scale of selection to the Foundation Programme (n=8000), which has a common national Person Specification – interviews are more appropriate for selection to specific posts
 - MMIs, in which each applicant undertakes several structured interviews with different interview panels, would be more reliable and more valid, but they would be labour intensive and expensive to implement and sustain
 - The Options Appraisal concluded that an SJT of likely professional behaviours and judgements in hypothetical work situations would be the most appropriate measure of the professional attributes in the national Person Specification. As the SJT would be an invigilated, machine-markable test, it would be a secure, reliable, standardised and relatively inexpensive selection tool
- The two selection tools shortlisted for use as a measure of an applicant's clinical knowledge and skills were i) a special selection test (a 'national exam' for ranking) and ii) a measure of educational performance at medical school based on existing information
 - A national exam has the advantage of being easy to standardise, but would not be fair to all applicants owing to variations in the timing of curricula across medical schools and the timing of the application process (before finals). It would also be prohibitively expensive to develop, and was considered by key stakeholders as likely to undermine educational objectives through distracting applicant focus and distorting the delivery of the medical school curricula
 - The Options Appraisal concluded that the considerable existing information available from medical schools should be used as a measure for applicant educational performance, given that allocation to posts must take place before the results of medical school Finals are known, as an Educational Performance Measure (EPM). In many ways the EPM would evolve from the academic quartiles, but it should be produced in a more robust, complete and standardised way, and possibly be more granular
- Work should begin to develop and pilot the SJT and EPM, with a view to implementation for selection to FP2013 onwards, when the contract for the current on-line FPAS system is due for renewal. Aside from testing the individual effectiveness of the SJT and EPM, the pilots would provide data that could be used to determine the relative weightings to be given to scores for the SJT and EPM components when they are combined to give an overall application score for the purpose of selection to the Foundation Programme
- An initial step in the work should be a formal Job Analysis of the role of the FY1 doctor in order that selection tools might be based on more detailed information about the role
- Other aspects of the process of application to the Foundation Programme should remain unchanged, namely a national application process including applicant declaration of preferences, the use of a matching algorithm and eligibility checking
- Work should continue outside of the scope of the ISFP project to develop common final examination questions between UK medical schools, and to develop an assessment of clinical communication for applicants without English as a first language and who have not interacted with patients in English during the course of their undergraduate medical degree

The DH accepted the recommendations of the ISFP Steering Group, and authorised a short feasibility study to determine how the recommendations might be taken forward. The feasibility study concluded that the recommendations should be implemented through two sequential projects. The first – the current project - would develop, refine and validate the specifications for SJT and EPM through pilots that would take place in parallel with the FP2011 recruitment round. Subject to ministerial approval of the outcome, a second project would then be commissioned to plan and prepare for implementation of the SJT and EPM in a new selection process to be used for live selection for the FP2013 recruitment round onwards

Following the report from the feasibility study, the DH authorised the Improving Selection to the Foundation Programme (ISFP) project, which began in January 2010

4. ISFP Project Plan

The Improving Selection to the Foundation Programme (ISFP) project was set up in January 2010, following a short feasibility study in which a detailed Project Initiation Document (PID) had been produced. Key points from the PID, which is provided as Appendix C, are summarised in this chapter to provide the context against which the outcome of the project is described in subsequent chapters of this report.

4.1. Objectives

The main objective of the project was to design, develop, and pilot the SJT and EPM to replace the existing selection arrangements for use from the FP2013 recruitment round onwards. The scope of the project was to include:

- A detailed Job Analysis of the role of the FY1 doctor to identify the detailed criteria that should be used for the selection of applicants to the Foundation Programme using the SJT
- The development of specifications for the SJT and EPM
- The production of SJT items for use in live recruitment rounds
- The development or purchase of software to store SJT items securely
- The piloting of the SJT and EPM
- Collection of data about costs, resource requirements, timings, etc. for planning a live implementation of the new selection arrangements
- Developing expertise to underpin the success of the live implementation
- Raising awareness of the proposed changes to the selection process, and maintaining the support and cooperation of stakeholders
- Obtaining legal opinion to confirm that the SJT and EPM would be likely to withstand legal challenge

The live implementation of the SJT and EPM would be subject to confirmation of the associated business case and the DH approvals process.

4.2. Approach and timescales

The original plan for the project recognised that the SJT and EPM could be developed independently of each other. Of the two selection tools, the SJT represented by far the larger piece of work, and accounted for almost the entire critical path of the project. Given this, the project was managed in stages that broadly reflected the main steps in the development and piloting of the SJT, which were as follows:

- Stage 1: A Job Analysis and literature review to identify the professional attributes of the FY1 doctor which would be assessed by the SJT; the design specification of an SJT; the production and quality assurance of an initial set of SJT items
- Stage 2: An initial series of micro-pilots to try-out the SJT on a small scale
- Stage 3: The development of further SJT items for a large-scale pilot
- Stage 4: Analysis and reporting of outcomes of the large-scale pilot

The original project plan is illustrated in Figure 1.

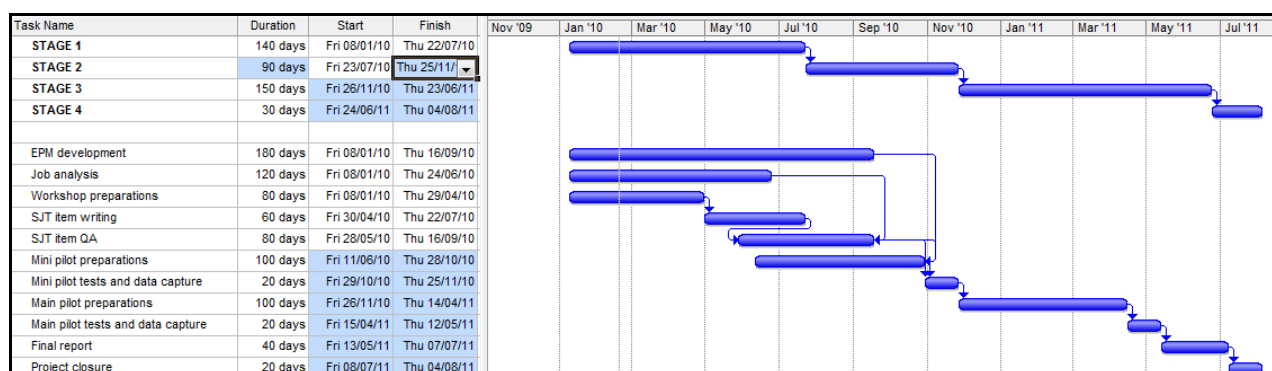


Figure 1: Project Plan

4.3. Management and Governance

The management structure for the ISFP project is illustrated in Figure 2.

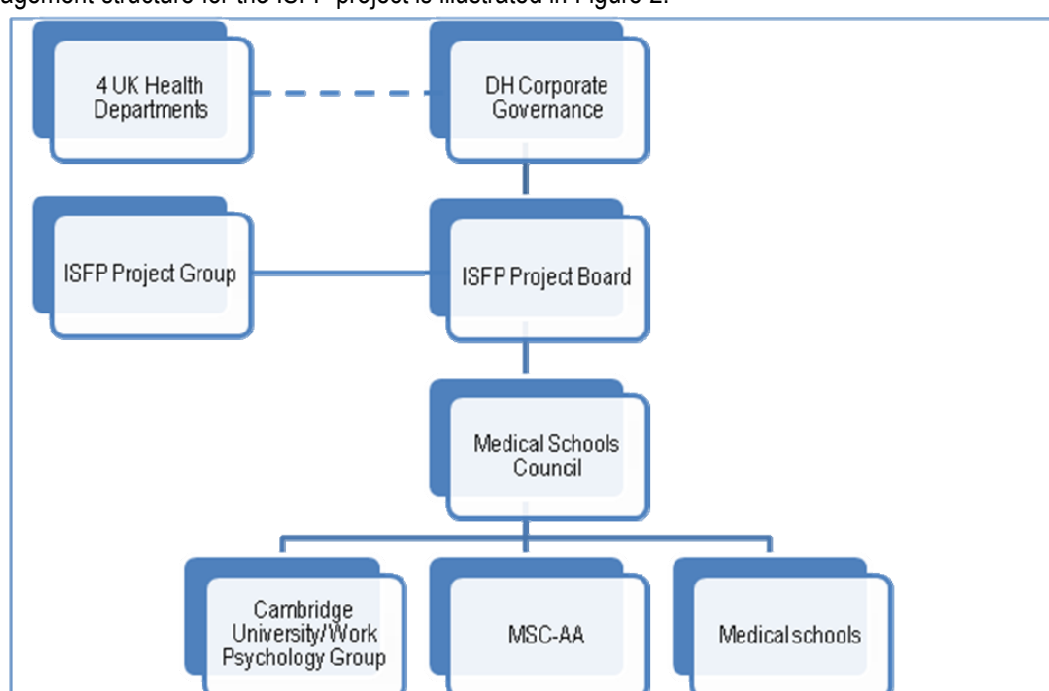


Figure 2: Governance structure for ISFP project

The ISFP Project Group, which was established to continue the work of the ISFP Steering Group during the Options Appraisal, reported to a Project Board established by the Department of Health. Formal highlight reports were produced each month. The membership of the ISFP Project Board is shown in Figure 3.

Project Board Member	Representing
Dr Patricia Hamilton (Chair)	Department of Health
Professor Derek Gallen	UK Foundation Programme Office
Professor Paul O'Neill	Medical Schools Council
Christine Outram	CEO, MEE
Dr Katie Petty-Saphon	Medical Schools Council
Professor Anthony Weetman	Medical Schools Council

Figure 3: Membership of the ISFP Project Board

Fifteen stakeholder organisations were represented through the ISFP Project Group, which reviewed all key products and the Project Risk Register, and provided advice and guidance for the project, on behalf of the ISFP Project Board. The membership of the ISFP Project Group is shown in Figure 4:

Project Group Member	Representing
Professor Paul O'Neill (Chair)	Medical Schools Council
Professor Dinesh Bhugra	Academy of Medical Royal Colleges
Dr Ian Curran	Medical Educators
Dr Ashley Fraser	NHS Employers
Professor Derek Gallen	UK Foundation Programme Office and COPMeD
Martin Hart	General Medical Council
Duncan Henderson	NHS Education for Scotland
Matthew Langdon	Department of Health England (on behalf of the 4 UK Health Departments)
Andrew Matthewman	Department of Health England (on behalf of the 4 UK Health Departments)
Professor Simon Maxwell	Prescribing Skills Assessment
Terry McMurray	NIMDTA
Professor Jim McKillop	Scottish Board for Academic Medicine
Dr Katie Petty Saphon	Medical Schools Council
Nick Deakin/ Karin Purshouse	British Medical Association Medical Students Committee
Dr Kim Walker	UK Foundation Programme Office (Foundation School Directors & Managers)
Professor Anthony Weetman	Medical Schools Council

Figure 4: Membership of the ISFP Project Group

The project was managed by a core team in MSC: Denis Shaughnessy, Project Manager, Siobhan Fitzpatrick, Senior Policy Officer, Amy Stringer, Communications Officer, and Oliver Watson, Policy Officer. The core team produced many of the deliverables from the project, as well as coordinating the activities of the other participants and stakeholder groups, such as the individual medical schools and MSC-AA.

Funding awarded by the DH was held in a restricted account controlled by Jessica Pugh, Finance Manager of Universities UK.

5. Job Analysis of the role of the FY1 doctor

A systematic, multi-method job analysis, which incorporated a literature review, semi-structured critical incident interviews and observations, was undertaken to define the professional attributes required to be effective as an FY1 doctor. This triangulation approach was used to increase the credibility and validity of the results of the Job Analysis. A total of 294 individuals working as, or with, FY1s were involved in the Job Analysis and supporting consultation. The outcomes of this research are described in behavioural terms in the form of a Professional Attributes Framework (Figure 5), which sets out the key dimensions and example descriptors of effective performance in the role. These findings were used to determine the domains to be targeted for selection to the FY1 role, and to define the test specification for the SJT. The full report of the Job Analysis of the role of the FY1 doctor is available as Appendix D.

5.1. Methods

The literature review, which covered all existing policy documentation and relevant publications, was conducted in two parts. The first part focused on existing policy documentation and relevant publications in the literature to assist in identifying the key skills, abilities and professional attributes of an FY1 doctor. The second part of the literature review focused on SJTs and their use in high stakes selection. The results provided context for the Job Analysis of the role of

the FY1 doctor, and developing the understanding of the methods that will be undertaken when developing the SJT. The detailed search strategy can be found in the report of the Job Analysis (Appendix D).

A total of 51 semi-structured interviews were carried out with foundation doctors, clinical/ educational supervisors, Directors/ Deans/ Heads of Medical Education at medical schools, Trust Medical Directors, foundation school directors, Foundation Programme Directors and patient representatives. The interviews were conducted face to face (4) or by telephone (47) and lasted between 30-45 minutes. Interviewees were asked for examples of specific incidents from the FY1 job that were characteristic of effective or non-effective performance, and thus to identify the key skills and attributes needed to perform as a FY1 doctor. The interview transcripts were analysed using template analysis to identify the key themes within the data. All 51 transcripts were coded according to an initial template. The template underwent a number of iterations, and quality checks were carried out.

Observations (shadowing) of 13 FY1 doctors were carried out at four locations throughout the UK. A total of 29 hours of observations was undertaken by two researchers. The purpose of observing FY1 doctors was to gain a greater insight and understanding into the tasks and responsibilities of a FY1 doctor, as well as the professional attributes required to perform the role successfully. The researchers recorded the tasks, responsibilities, behaviours and attributes of the FY1 doctor. A record was kept of the timings of actions and behaviours to gain an understanding of the temporal nature of the role. The observations were analysed and the behaviours classified according to the template created following the interviews. Where behaviours could not be associated with existing classifications in the template, new classifications were introduced.

5.2. Professional Attribute Framework

Analysis of the data from the interviews and observations resulted in nine professional attributes and 111 behavioural descriptors being defined. These nine attributes are set out in a 'Professional Attribute Framework', in Figure 5, which details the behavioural descriptors that are representative of each attribute specific to the FY1 role.

Professional Attribute	Professional Attribute - definition
1. Commitment to Professionalism	Displays honesty, integrity and awareness of confidentiality & ethical issues. Is trustworthy and reliable. Demonstrates commitment and enthusiasm for role. Willing to challenge unacceptable behaviour or behaviour that threatens patient safety, when appropriate. Takes responsibility for own actions.
2. Coping with Pressure	Capability to work under pressure and remain resilient. Demonstrates ability to adapt to changing circumstances and manage uncertainty. Remains calm when faced with confrontation. Develops and employs appropriate coping strategies and demonstrates judgement under pressure.
3. Effective Communication	Actively and clearly engages patients and colleagues in equal/open dialogue. Demonstrates active listening. Communicates verbal and written information concisely and with clarity. Adapts style of communication according to individual needs and context. Able to negotiate with colleagues & patients effectively.
4. Learning and Professional Development	Demonstrates desire and enthusiasm for continued learning, takes responsibility for own development. Willing to learn from others and from experience. Is open and accepting of feedback. Demonstrates a desire and willingness to teach others.
5. Organisation and Planning	Manages and plans workload effectively, displaying efficient time management and delivering tasks on time. Able to prioritise effectively and re-prioritise where appropriate. Is conscientious and maintains accurate records.

Professional Attribute	Professional Attribute - definition
6.Patient Focus	Ensures patient is the focus of care. Demonstrates understanding and appreciation of the needs of all patients, showing respect at all times. Takes time to build relationships with patients, demonstrating courtesy, empathy and compassion. Works in partnership with patients about their care.
7.Problem Solving and Decision Making	Demonstrates an ability to assimilate a range of information and identify key issues. Engages with the wider issues and thinks creatively to solve problems and reach appropriate decisions. Is proactive and demonstrates initiative. Is able to attend to detail.
8.Self Awareness and Insight	Demonstrates awareness of the boundaries of their own competence and willing to seek help when required, recognising that this is not a weakness. Exhibits appropriate level of confidence and accepts challenges to own knowledge.
9. Working Effectively as Part of a Team	Capability & willingness to work effectively in partnership with others and in multi-disciplinary teams. Demonstrates a facilitative, collaborative approach, respecting others' views. Offers support and advice, sharing tasks appropriately. Demonstrates an understanding of own and others' roles within the team and consults with others where appropriate.

Figure 5: Professional Attributes Framework

The Professional Attributes Framework was checked against the findings from the first part of the literature review. This confirmed that all of the professional attributes identified through the review had already been identified through the other two Job Analysis methods (interviews and observations), thus confirming the validity of the Professional Attributes Framework.

An online survey to validate the outcomes of the Job Analysis with a larger number of individuals gained 230 respondents. The survey was sent to a number of different organisations (with a request for it to be cascaded to their membership) and to individuals working as or with FY1s, including the BMA Junior Doctors Committee, the UKFPO Foundation Doctors Advisory Board and the UKFPO Foundation School Directors and Managers networks. The results from the validation survey indicated that all nine Professional Attributes were considered by key stakeholders to be important to be effective as an FY1 doctor and important to evaluate at selection.

The Professional Attributes Framework was mapped against existing data, including the national Person Specification and the FY1 curriculum. The mapping demonstrated good concordance between the Professional Attributes Framework and the FY1 curriculum. The mapping demonstrated that the Job Analysis and the resultant Professional Attributes Framework showed good concordance with the national Person Specification, although the Professional Attributes Framework identified one additional attribute - Learning and Professional Development - which was not specified in the FP2011 national Person Specification. Subsequently, the UKFPO Rules Group accepted the recommendation of the ISFP Project Group to include an additional statement in the FP2012 national Person Specification requiring commitment to learning and professional development (Appendix E).

The next step of the Job Analysis was to define the professional attributes that could be measured in an operational SJT, as not all can be directly measured in this format. Aspects of Organisation and Planning, and Problem Solving and Decision Making, for example prioritisation or attention to detail, are implicit in the format and preparation for an SJT and thus embedded in the test. Following the outcome of the Job Analysis, a panel (n=7) of researchers and occupational psychologists expert in SJT development and design, identified that five composite professional domains would be the most appropriate to target in an operational SJT and thus provide a psychometrically robust and valid test. Figure 6 illustrates the relationships between the nine professional attributes and the five SJT target domains.

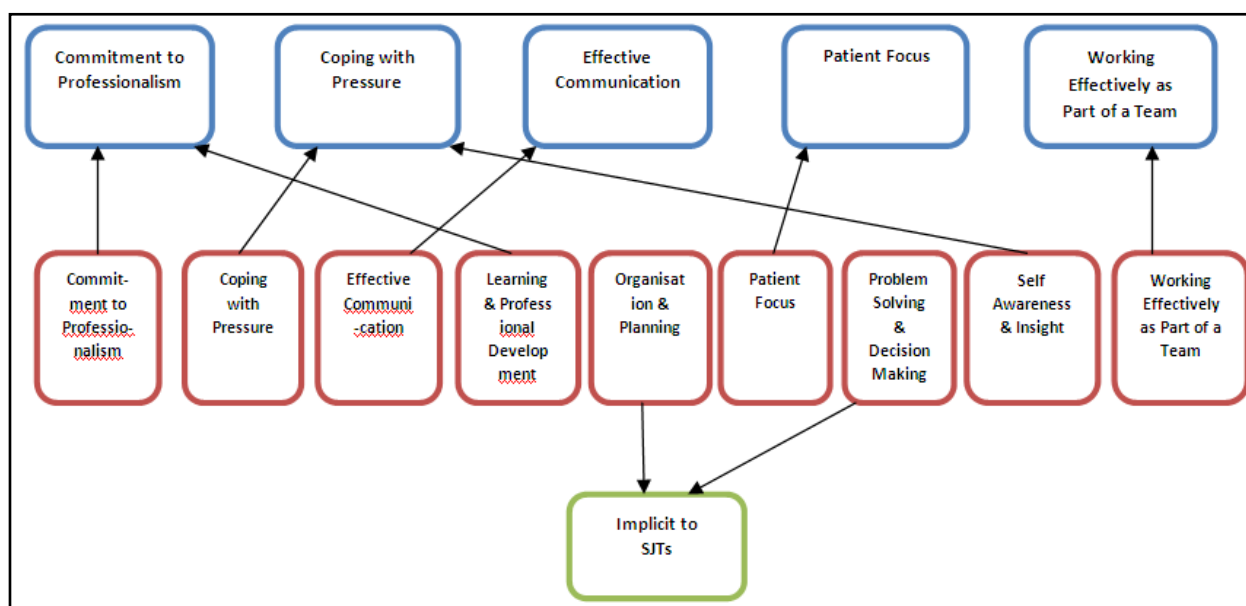


Figure 6: Relationship between Professional Attributes and SJT target domains

6. SJT Item Writing and QA Process

The University of Cambridge and the Work Psychology Group were contracted to design, develop and evaluate the SJT for selection to the Foundation Programme, as described in this chapter. Full information is available in Appendix F, the report of the SJT pilots.

6.1. SJT design

The Professional Attributes Framework (Figure 5) identified through the findings of the detailed Job Analysis of the role of the FY1, and the lessons from existing use of SJTs in high stakes selection as identified through the extensive literature review, informed an initial test specification for the SJT, including the length of the test, appropriateness of content, response format and response instructions.

Ten micropilots involving (n=32) final year medical students at four medical schools (12 male, 20 female; 15 white, 17 non-white) sought feedback on the use of SJTs for selection to the Foundation Programme, the appropriateness and relevance of the content, and the test specification for response format and response instruction. The micropilots involved a short (30 minute) pilot test in exam conditions and small group discussions on 16 SJTs of four response types. All attendees were sent briefing materials in advance of the micropilot, and were asked to sign a confidentiality disclaimer on the day.

The feedback from the ten micropilot sessions was used alongside information from the literature reviews, and existing good practice from the SJT used for selection into GP training – an existing successful use of SJTs for selection into postgraduate medical training - to inform the SJT design specification used for item development and piloting. Further details can be found in Appendix F. One of the purposes of the pilots was to evaluate whether the specification was appropriate. Analysis and review of the design specification occurred throughout all stages of the development and review process.

The SJT is designed to be staged in the context of the Foundation Programme, with applicants assuming the role of an FY1 doctor in answering the items. Knowledge based instructions ('what should you do') are used instead of behavioural based response instructions ('what would you do'). The evidence from the literature reviews showed that knowledge based instructions are more appropriate for high stakes selection contexts where faking and coaching are potential threats, and the GMC places a high premium on probity.

Two types of response format were used; ranking and multiple choice. The rationale for this was informed by experience in designing successful high stakes SJTs in other contexts. Using both response formats enables a fuller range of scenarios to be used, for example some scenarios require an applicant to differentiate between singular actions in response to a scenario that vary in appropriateness, whilst in other scenarios it is necessary to do more than one thing. Different response formats do not enable analysis of separate traits or skills. Whilst the ranking items allows for greater granularity in the data due to the more complex response/scoring format, they are typically more time-consuming to complete and are not appropriate for all scenarios, therefore a balance with multiple response items is needed. Approximately two thirds of items used in the test were of the ranking answer format (rank the five options in order of appropriateness), and one third were of the multiple choice format (select three from eight options). This structure also follows the validated methodology from the SJT for use in short-listing to GP training posts.

The SJT items are mapped to the five SJT target domains (Figure 6) for item development and test construction purposes, but each item does not exclusively measure a target domain in isolation. As such the SJT should not be viewed as assessing five individual domains or as a 'blueprint'. The content of the scenario and the correct response keys are likely to relate to the targeted domain, but a number of the other response options are likely to tap into one or more of the other SJT target attribute domains. This increases the complexity of items, and avoids the possibility of a 'template answer' being provided by applicants.

The SJT test content should cover the range of target domains in roughly equal proportions. Similarly within and across the target domains, the test should cover a range of relevant topics e.g. dealing with a difficult colleague; being asked to undertake an unfamiliar procedure.

The SJT is designed to last for two hours and contain 60 items. In the initial pilot, 65 items were piloted in two hours; and in the large-scale pilot, 60 items were piloted in two hours. Sixty items were shown to be sufficient to cover the five target attribute domains in a sufficiently reliable way without overloading applicants. Sixty items also enables a greater distribution of scores and thus differentiation between applicants. Allowing two minutes per item is consistent with previous evidence and experience, which has demonstrated good reliability and validity.

6.2. Item development and review

In line with good practice in design and development of assessment, an individual test item should not be used until its psychometric properties are known, and the quality, relevance and validity of the question and its responses are established. The process of development and review for the SJT items is both detailed and necessary; ensuring that only those items meeting all the quality criteria are piloted. Once the results of the pilot are known, the performance of items (consistency of answers between applicants, level of difficulty etc) can be discerned. Some items may, at this stage, need further refinement and review before they can be piloted again; other items can be deemed to be the final version of the item, ready for inclusion in an item bank and for use in live selection to the Foundation Programme. It is vital that the psychometric properties of live test content is known, in order to ensure fairness of assessment, and in order to facilitate standard setting or test equating between multiple papers.

There were two phases of SJT item development and review. Phase One ran between April and October 2010 and Phase Two ran between November and March 2011. Figure 7 outlines the process of the SJT item development and review process followed for both Phases One and Two. The individual stages of item development and review are described in more detail in sections 6.3 – 6.7.

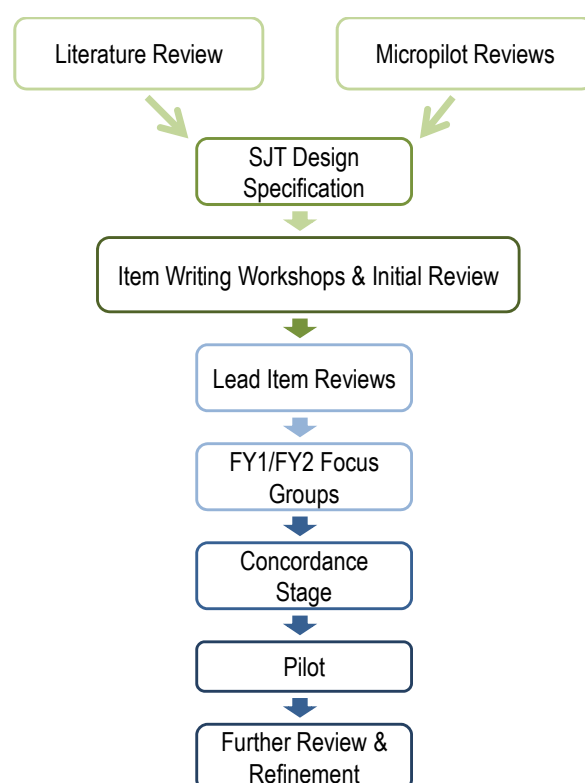


Figure 7: Process of SJT item development and review

6.3. Item writing workshops

Five item writing workshops were held between April and May 2010 (Stage 1 of the ISFP project) and two item writing workshops were held between November and December 2010 (Stage 3 of the ISFP project). A total of n=89 individuals from across England, Wales, Scotland and Northern Ireland were trained in SJT item-writing principles for selection to the Foundation Programme. Of these, nine were involved in item-writing during both phases of item development. The one-day item-writing workshops were accredited by the Royal College of Physicians, and attendees were awarded 6 CPD points.

Volunteer item-writers were recruited via the UKFPO Foundation School Directors), the Society for Academic Primary Care, and via a Psychiatry contact, in order to ensure that the breadth of expertise in developing items reflected a range of medical specialties (n=59 in 23 acute specialties, n=12 from 2 community specialties, 17 = not-declared) and UK regions. The national Person Specification for the role required that volunteers be familiar with the roles and responsibilities of an FY1, have worked with FY1 doctors within the previous two years, be willing to commit to attending a one day training workshop and be willing to engage in follow-up work. The demographic data of the item-writers and their job role are provided in Figure 8.

Gender	Male	53
	Female	34
	Not Stated	2
Age group	25 and under	0
	26-35	7
	36-45	38
	46-55	26
	56-65	15
	66 and over	0
	Not stated	3

Ethnicity	White	57
	Black	0
	Asian	20
	Mixed	0
	Chinese	3
	Other	5
	Not Stated	4
Role	Clinical Tutor	6
	Clinical/Educational Supervisor	44
	Foundation School Director	15
	Lay Representative	2
	Medical School Director	4
	Other	16
	Not stated	2

Figure 8: SJT item writer demographics and job roles

All item writers were sent briefing materials in advance of the workshop, and were asked to sign a confidentiality and code of conduct form on the day. Each workshop was facilitated by a minimum of two facilitators trained in SJT item-writing principles. A member of the ISFP Project Group attended each workshop to introduce the context of the ISFP project and the role of SJTs in future selection to the Foundation Programme. The facilitators provided an overview of the test specification for selection to the Foundation Programme and provided training in SJT item-writing principles. During the workshop, item writers worked in pairs and small groups to create and review scenarios and plausible responses for the SJT items, receiving feedback from the two trainers.

In total, over the seven item writing workshops, 453 items were written, by 82 item writers. Seven item writers did not submit any items either at or following the workshops. This is a return rate of 92% and an average of 5.5 items per item writer. In the first phase of workshops, item writers were asked to aim to write between six and eight items, on the day and as follow-up work. However only two thirds of item writers actively contributed to follow-up work (for n=262 items). In the second phase of workshops, item writers were asked to write between three and five items on the day, to encourage item writers to produce better quality items and spend more time trying to refine and develop each item. In the second phase, 90% of item-writers actively contributed to follow-up work. Overall, the number of items written by each item writer ranged from one to 19. In total 265 ranking items and 198 multiple choice items were developed. This split was in line with the design specification of the SJT, which is for two thirds ranking and one third multiple choice. The number of items relevant to each of the target domains is illustrated in Figure 9.

Target domain	Number of items written
Commitment to Professionalism	180
Coping with Pressure	91
Effective Communication	68
Patient Focus	46
Working Effectively as Part of a Team	68

Figure 9: Total number of SJT items written by target domain

6.4. Initial item review

All SJT items written at or following the item-writing workshops were logged onto a spreadsheet, which was also used for version control as items were refined and reviewed. Initially, SJT items were reviewed by a core team of

psychologists trained in SJT item writing principles. Where necessary, item writers were asked to provide clarification on the content of their items, and resubmit them with amendments. In some cases, there were several iterations between the item-writer and the psychologist; in most cases one or two amendments were sufficient. One of the goals of this method is to maintain buy-in from trained item writers in that they maintain ownership of the SJT items they have contributed to the programme.

Following the initial review in the first phase of workshops, 290 items were deemed to meet the quality criteria for the SJT items and were ready for the next stage of review. 115 items were deemed not suitable to be taken any further; an attrition rate of 31%. Item writers were informed of the status of the items they had written. During the second phase of workshops, 70 of the 77 items were ready for the next stage of review. Seven items were deemed not suitable and were not taken any further in the process; an attrition rate of 9%. This was a great improvement from the first set of workshops, and can be attributed at least in part to amendments made to the structure of the training.

6.5. Lead item review

The 360 items which passed the quality controls following initial review by psychologists were then reviewed by a lead item reviewer. Twelve lead item reviewers were clinicians identified during and following the SJT item writing workshops as those individuals who appeared to have fully understood the item writing principles and were willing to be involved further in the process. Whilst the SJT items written at workshops were written in pairs and reviewed in small groups, review by a separate clinician trained in SJT item-writing would ensure validity and realism to the content of the SJT item scenario and response key.

Lead item reviewers were asked to consider whether the items reflected the target domains, if they were clear and realistic, if terminology was commonly used across the UK and if the answer key was appropriate. This review stage took place electronically and remotely. Lead item reviewers had already signed a confidentiality and code of conduct form at the item-writing workshops, and they were sent no more than 25 items for review in order to minimise the security risk. Of the 360 items reviewed by lead item writers, 54 items were deemed not suitable to go any further, leaving 306 items to go forward to the next stage.

6.6. Foundation doctor focus groups

The purpose of the foundation doctor focus groups was to gain feedback from job incumbents about the relevance and fairness of the SJT items, to ensure content validity and face validity of the SJT. Twenty focus groups of between three and five foundation doctors were held in five regions of England and Northern Ireland, with a total of 63 individuals taking part. Participants were asked to sign a confidentiality and code of conduct form. The demographic data of the foundation doctors involved in focus groups are provided in Figure 10.

Gender	Male	16
	Female	46
	Not Stated	1
Age	Mean age	26.3
	Range	23-39
Ethnicity	White	32
	Black	2
	Asian	19
	Mixed	3
	Chinese	3
	Other	2
	Not Stated	2

Figure 10: Focus group participant demographics

Participants were asked to review the test instructions for clarity and ease of understanding. They were invited to discuss in small groups the scenario content of individual SJT items and the response options. They were also asked to provide a possible answer key and about the timing of the test (60 items in 120 minutes). For security reasons no individual had sight of more than 20 items.

All participants in the focus groups agreed that the test instructions were clear and easy to understand. Following feedback from the focus groups, 20 of the 306 items reviewed were deemed not suitable to go forward to the next stage of review, with queries on a further 11 items. This feedback was considered alongside the concordance - comments and results. In light of feedback from the foundation doctor focus groups, small revisions were made to the remaining 275 items before these were subject to concordance analysis.

6.7. Concordance Panel

Following best practice in SJT development, the aim of a concordance stage was to identify a high level of consensus between Subject Matter Experts on the answer keys to determine whether items were ready to be piloted. Where items did not achieve a high level of concordance, they were re-examined and reviewed; some were for further workshop and lead item review.

Subject Matter Experts were volunteers who met a similar person specification as for SJT item writers in that they were required to be familiar with the roles and responsibilities, skills and attributes to be successful in the role of FY1 and to have worked with FY1 doctors within the previous two years. Subject Matter Experts were recruited via the UKFPO and Medical Schools Council networks to complete a trial of SJT items in exam conditions. They were also invited to give written and verbal feedback on the content of the items. All Concordance Panel members signed a confidentiality and code of conduct form.

Four Concordance Panels of between eight and 19 Subject Matter Experts (total n=60) were conducted. Four papers containing a total of 286 items were reviewed. At this stage, the tests were not constructed as final tests (e.g. spread of domains) as the aim of the concordance panels was to analyse individual items. Using established criteria of acceptance levels, items were deemed either to have acceptable levels of concordance (259) or unacceptable levels of concordance (27). Items with above satisfactory levels of concordance were deemed suitable for piloting. Consideration was also given to verbal and written feedback from the concordance panel when selecting items. Twenty one items (the 21 items which demonstrated satisfactory levels of concordance, rather than good or excellent) were removed from the test paper in light of feedback from the Concordance Panel. Items with lower, but still satisfactory, levels of concordance were also considered in light of the feedback from the item writer and foundation doctor focus group key. If the concordance key differed from the item writer and foundation doctor focus group key, it was not seen as appropriate for the item to be used in the pilot. Following the Concordance Panel review, 200 items were selected for inclusion in the initial and large-scale pilots.

6.8. Test construction

Test construction for the piloting phase was based on the following parameters and recommendations:

- There should be roughly two thirds ranking and one third multiple choice items within a test paper, to allow granularity of performance data and a range of scenarios
- There should be, as far as possible, an equal split between the domains being targeted
- Each item, as well as being mapped against a domain, were mapped by topic area. Scenarios were assigned to three broad topics of a 'colleague' topic, a 'patient' topic or a 'personal' topic. Within these three broad topic areas, scenarios were further assigned to different topic areas. A good range of topic areas should be used within each paper
- Items should be spread throughout the paper to ensure they are not grouped together by 'topic' or 'domain'

- Items that include gravely serious scenarios (e.g. those involving deaths or serious mistakes by the FY1) should not be used at the beginning of the paper, as applicants may find these scenarios stressful initially; placing these items later on in the test allows applicants time to familiarise themselves with the context of the test and what is expected of them

6.9. Scoring convention

In contrast with clinical knowledge items, with SJT items there is often no definitive correct answer. Two scoring methods were piloted: rational scoring which uses a predetermined answer key from concordance data; and empirical scoring answer key determined through use of live data. Following best practice, the SJT scoring key is based on rational scoring determined through:

- consensus at the item review stages of item writers and lead item review
- expert judgement in a Concordance Panel review
- review and analysis of the pilot data

When determining how to score each item type, consideration needs to be given to the amount of valid information which can be derived from each test item. For example, a ranking item – which takes longer for the applicant to assess and respond – should receive points for answers which are partially correct, as this provides the maximal amount of information about that applicant's performance.

For ranking items, 20 marks were available for each item. For each of the five response options up to four marks were available. Marks were awarded for near misses. If a participant tied two options, they received 0 marks for each of the options they tied, as they have not answered the question. There was no negative marking. Although 20 marks were available for ranking items, if a participant offered a complete answer to the item, the lowest possible available score is 8 and the highest possible score is 20. The scoring key for ranking items is illustrated in Figure 11.

Keyed Rank	Applicant rank				
	1	2	3	4	5
1	4	3	2	1	0
2	3	4	3	2	1
3	2	3	4	3	2
4	1	2	3	4	3
5	0	1	2	3	4

Figure 11: Scoring key for SJT ranking items

For multiple response items, four marks were available for each option answered correctly, with a total of 12 marks available for each item. If an applicant answered more than three options, they scored 0 for the whole item, as it was incorrectly answered. There was no negative marking employed. This represents a lower total number of marks for multiple response items available than for ranking items, to reflect that ranking items are able to provide more detailed data than multiple response items and require more effort from participants.

Participants who did not complete an item received 0 for that item. Participants were not credited with random guesses on the items that they did not complete (identified by sporadic or unusual patterns of answers).

Scoring SJTs is a complex process and the choice of methodology for how to score the items is not straightforward. Alternatives have been considered, including not providing points for near misses, or systematically subtracting eight from each ranking item.

7. SJT Pilots

7.1. Overview

There were two phases of SJT piloting; an initial pilot in autumn 2010 and a large-scale pilot in spring 2011. The initial pilot took place in four UK medical schools, involving 455 participants. The large-scale pilot took place in 13 UK medical schools, involving 639 participants. Two non-UK pilots took place during spring 2011 involving 43 participants; however, as the sample size was so small, these data have been removed from analysis so as to avoid any potential misinterpretation. Feedback from the non-UK participants is still considered.

The purpose of the initial pilot was to ensure sufficient confidence in the psychometric characteristics of the SJT to justify a larger pilot, to inform the design and development of an SJT for selection to the Foundation Programme, and to expose any unforeseen issues with the administrative and logistical arrangements. The purpose of the large-scale pilot was to undertake further psychometric analysis to inform the recommendations as to the design and operational (live) test specification, to continue to pilot administrative and logistical arrangements, and to pilot a larger number of SJT items and therefore develop the item bank.

The pilots have demonstrated that the SJT is an appropriate method for selection to the Foundation Programme, as the pilots demonstrate construct validity and criterion related validity, and that an SJT of 60 items with the right psychometric properties is both reliable and able to differentiate sufficiently between applicants. Over 95% of participants in the pilots evaluated the SJT, and there was good consensus that it is relevant, fair and better than the current white paper questions, demonstrating stakeholder acceptability.

This chapter summarises the practical findings from the pilots and the key findings from the analysis of pilot data. Further analysis of data from the pilots is included within Appendix F.

7.2. Methods of SJT delivery

Over the two phases of piloting, the method for the delivery of the SJT was trialled on paper, online and using a hybrid method of computerised handsets, as described in detail in Figure 12. Security of SJT item content was paramount throughout the pilots.

Competitive tenders were invited for the paper-based delivery. Stephen Austin was commissioned for the printing, delivery, collection and scanning for the initial SJT pilot, and Cambridge Assessment was commissioned for the same services for the large-scale SJT pilot.

Five medical schools in the large-scale pilot volunteered to run the SJT on computers; the other schools involved in the pilots did not have the facilities to offer this to the whole final year cohort. Cambridge Assessment provided the use of, and technical support for, online software to deliver the SJT for the large-scale pilot, at zero cost to the project, as this provided an opportunity for the organisation to pilot their newly developed online delivery software.

The use of Banxia handsets, owned by the host university of one medical school involved in the initial pilot, were provided at zero cost to the ISFP project.

	Paper	Computer	Handsets
Description	• 1 x 36pp question booklet • 1 x 2pp OMR answersheet • 1 x 2pp glossary • Applicants completed the OMR answersheet; the papers were then scanned and answers converted to a .csv file	• Secure login with timed access (centrally controlled) • Answers were recorded in real-time • Cambridge Assessment staff attended all pilot sites to provide technical support • Papers were provided as a back-up, but were not used	• Battery operated individual handsets were used, and remotely connected to a 'hub' • Participants completed the paper-based pilot and then entered their answers onto the handsets (extra time to do so) • Answers were entered in real-time and then uploaded in one go
Number of participants	• 455 participants in the initial pilot from four UK schools; 421 participants from eight UK schools and 43 non-UK participants from two non-UK schools in the large-scale pilot completed the SJT on paper	• 218 participants from five UK schools in the large-scale pilot completed the SJT on computers	• 53 participants from one UK school in the initial pilot completed the SJT using handsets
Applicants with special requirements	• Coloured papers and enlarged papers were provided by request • Extra time was arranged and facilitated by the pilot sites	• No provision was made during the pilots, but reassured that the software could allow for – with prior notice – e.g. font size adaptation • Extra time could be provided – but in reality participants with extra time completed the test on paper	• No adaptations were made. The screen size is fixed, and it would be difficult to adapt. The only feasible option would be for applicants with special needs to record their answers in another format, for these to then be manually entered • Separate hubs could be used for sessions of different test lengths
Security (summary)	• Papers were securely stored in tamper-evident bags by the printer. One working day in advance of the pilot, boxes of papers were couriered to a named contact at the medical school, and signed for. Boxes of papers were stored overnight according to local medical school protocol for secure storage. Papers were counted out and in to ensure that all copies of all papers were returned	• Secure network and secure storage of content (items) assured throughout • It was not possible to take a screen shot, press 'escape' or the back button • The administrator had to approve each participant to join the session (only recognised users) • Unique login details were provided on desks – to prevent anyone accessing the server from outside of the room	• Data were stored and downloaded in an encrypted file • The data was stored in format 'ABCDE' separately from the SJT items – it would be difficult to match the two sets of information

	Paper	Computer	Handsets
Administration (preparation)	Arranging secure overnight storage of boxes of papers	- Arranging secure overnight storage of boxes of papers - The IT lead for the pilot site spent circa 2 days' time pre-loading computers with access to the software via the network.	- Arranging secure overnight storage of boxes of papers - The length of test and question formats had to be pre-loaded onto a 'session' – this took circa 2 days' admin time
Administration (on the day)	- Papers were provided on the desks before participants entered the room - Example items provided	- Computers were logged in to the secure server when participants entered the room - Short briefing on how to use the software, plus opportunity to practise use of software - At one school the computer server was re-booted as the test was about to start; all participants were re-logged back onto the system within 15 minutes - Incompatibility with screen size at one school – participants had to scroll	- Short briefing on how to use the handsets, plus opportunity to practise use of handsets - Handsets were distributed after questions papers were collected, but with answer sheets still on desks
Venues	Typical exam halls were used – desks sufficiently spaced to allow invigilators to walk between desks	- In some venues, computers were widely spaced, or participants sat at alternate desks. All desks faced the same direction - Difficult to administer extra time – whilst access could be provided to applicants within the same session, they needed to end the session together to minimise disruption	As for paper-based pilots
Participant feedback	- There was no negative participant feedback - Some participants commented that they were not aware the glossary was available (it was indicated in the briefing and included as a separate paper) - UK participants were familiar with the format of the OMR forms and there were rarely comments about the mode of delivery; not all	- Overall, participants found the software easy to use, navigate, clear where the glossary was, and clear which answers had been attempted and/ or completed - There was a mixed reaction to the ease of use of the drag & drop answers - Most negative feedback related to the pilot site which had experienced the server	- Overall, quite negative feedback - Applicants were unfamiliar with the use of handsets, and it took some time to get used to entering the data - Participants commented that it was difficult to scroll between answers, and there was no indication if an answer had been completed or not

	Paper	Computer	Handsets
	non-UK participants were familiar with the concept of the OMR form, however the sample questions helped with this • A small number of participants suggested that the SJT might work well on computers	reboot • It should be noted that the participants of the computer tests were those participants who already routinely use computers for medical school assessments	
	• Analysis of the mean scores for participants taking the SJT as paper-based and computer-based showed no statistical difference, indicating that the mode of delivery did not affect the answers. • Analysis of the raw answers entered by handsets and the answers entered by the same participants on paper did not identify any differences – however it would be difficult to defend which set of answers should count in the event of a discrepancy.		
Pros	• Participant familiarity with OMR sheets • All UK schools routinely run paper-based assessments for large numbers; with sufficient notice, there is no difficulty in arranging for a large, flat, well-lit hall with sufficient capacity	• Auto-generated answers • No issue about being unable to read handwriting for RA numbers etc • Prompts for incorrect number of answers given or if partial answer	• Auto-generated answers • No issue about being unable to read handwriting for RA numbers etc
Cons	• There is an extra stage of papers being sent for scanning. This adds 1 day to the timeline per 1,000 papers • Some 'messy' papers were difficult to scan	• Fewer than one third schools have facility to deliver assessments via computers for whole cohort in only one or two sittings – even when allowing for booking several computer rooms for use in parallel • Issues of compatibility of system with multiple operating systems and screen sizes • Issues of providing individual adaptations (extra time etc) • Papers would need to be printed and provided to schools as a back-up – this would add to cost	• Participant and invigilator unfamiliarity • Difficulties in navigating through the questions – other handsets are available • Recommendation that applicants completed both paper-based and entered into handsets is both time-consuming and additional resource • Issues of providing individual adaptations (extra time etc) • Papers would need to be printed and provided to schools as a back-up – this would add to cost

Figure 12: Methods of SJT test delivery

7.3. Administration

Guidance for invigilators was developed by the ISFP project team, informed by existing protocols for universities and national examination bodies, to support standardised administration at the pilot sites. The guidance document included recommendations around the number of invigilators, the venue to be used, the applicant briefing to be read to participants, as well as policies to address e.g. a security breach. The Administrators guidance is provided as Appendix G.

A tailored administration document which included a background to the project, instructions for administration on the day, invigilator checklists, a briefing for participants and a feedback sheet was sent to all named leads for the pilot sites. Tailored versions of these materials were developed for those schools delivering the SJT on paper, computer and handsets, and for the one school completing a NEO PI-R personality measure. A reference guide was also produced by Cambridge Assessment for administrators who would be using the online software.

Staff and students involved in the pilots were asked to treat the SJT pilots as a live selection test, under examination conditions. The majority of participants in the initial and large-scale pilots indicated on their evaluation forms that the SJT was well run and well invigilated. Feedback from SJT administrators informed the running of subsequent pilots. There were a small number of administrative issues, the lessons of which are summarised in Figure 13.

Pilot findings		Lesson
Venues	Initial pilot: Two pilots took place in lecture theatres – this made it impossible for invigilators to walk between desks, and for students to leave to go to the bathroom without disturbing others. In at least one venue there was no clock in the venue	SJTs must be taken in exam conditions in a typical exam venue, with sufficient space to walk between desks. Venues to provide details in advance
	Large-scale pilot: Participating schools were informed of the detailed requirements in advance. Most venues were suitable. However, the computer room at one venue had high background noise and two venues had no clock visible	Clocks must be functioning and visible
Leadership	Initial pilot: At one pilot there appeared to be no lead invigilator – the MSC ‘observer’ assumed this role	Nominated responsible officer at each pilot site
	Large-scale pilot: Lead invigilators were in place at each participating school. The involvement of academics was beneficial for student engagement	Nominate academic ‘champion’ for SJT
Student behaviour and invigilation	Initial pilot: At some sites a small number of students did not treat the pilot as seriously as a real application – chatting when finding a seat, leaving the test early. At others, students were allowed to leave the test venue early before the end of the session. In at least one case, the instructions for applicants were not read aloud at the start of the test	Invigilators at several sites were briefed by the lead administrator prior to the pilot. Briefing to students should be revised to make sure that it is an appropriate length
	Large-scale pilot: Student behaviour was good at most sites though a small number were talking as they took their seats at some pilot sites. Invigilators at many sites reported that the briefing to applicants was too long; this led to some invigilators abbreviating the information	
Instruction - ‘should’ or ‘would’	Initial pilot: A number of students were confused around the importance of the ‘should’ instructions	Ensure the briefing and written instructions are clear about the significance of the ‘should’ instruction
	Large-scale pilot: Adjustments to the applicant briefing and test materials appear to have been successful	
Concern about regional variance	Initial pilot: A number of students were concerned that the most appropriate ranking of the SJT answers might vary between	Ensure that concordance and review of items is undertaken by individuals in a range of specialties and a range of geographic

Pilot findings		Lesson
	clinical settings and between different regions	regions
	Large-scale pilot: This concern did not arise in feedback sessions	
Access arrangements	Initial pilot: No requests to adjust for access arrangements apart from enlarged paper size at one pilot site	Establish special requirements of applicants at an early stage and build in to delivery specifications
	Large-scale pilot: A variety of adjustments were required including extra time, different coloured paper, coloured overlays, enlarged paper, enlarged font sizes (min size 14pt), different font types and rest breaks	Financial implications of the provision for multiple access arrangements to be considered
RA numbers	Initial pilot: Some students mis-entered or left blank their RA numbers	Ensure that candidates understand anonymity is ensured. Invigilators to check papers against master list of RA numbers
	Large-scale pilot: There was some mis-entering and missing RA numbers, however all students made aware of anonymity	Reinforce importance of accuracy to applicants
Guidance for administrators	Initial pilot: Some administrators wanted more modular guidance based on different roles on the day	Have separate checklists for i) lead invigilator and ii) other invigilators. Seek to shorten the guidance for administrators where possible
	Large-scale pilot: Some felt that documentation to be overly detailed	
Test paper delivery (courier)	Initial pilot: The test papers were not delivered by the couriers to responsible officer; instead they were handed over to some other person at the medical school	Ensure that the contractual arrangements with the couriers contain clear instructions for delivery
	Large-scale pilot: All papers delivered to responsible officer	
Computer delivery of the SJT	Large-scale pilot: At one school the computer server was re-booted as the test was about to start (all participants were logged back onto the system within 15 minutes). Incompatibility with screen size at one school – participants had to scroll	Ensure technical support is on-hand
		Continue to provide papers for SJT use as a back-up
Scoring key	Initial pilot: Students wanted more of a guide around how marks are allocated and how to split their time effectively	Share key elements of marking scheme
	Large-scale pilot: Scoring key shared with students in feedback	

Figure 13: Administrative lessons for SJT delivery

7.4. Participation

Participation in the SJT pilots was voluntary. Responses and feedback was anonymous; individual feedback on performance was not provided, as the test was still in development. A minimum participation of 1,000 students was sought in order to provide statistical confidence in the findings of the SJT pilots. A minimum sample size for the psychometric analyses for each paper was $n=100$, which was achieved for every SJT paper, as illustrated in Figure 14 overleaf. Given this, there can be a reasonable degree of confidence in the findings of the pilots.

The initial pilot in autumn 2010 highlighted a lower than hoped participation amongst students. The timing of the pilots was the most frequent explanation for this, with many participants being about to take or having just taken final exams – or away on a ‘reflection week’. Additional schools were asked to run pilots, and other measures were taken to maximise student awareness of the role of the pilot SJTs, and incentives provided to ensure participation. Workshops (briefing sessions) on the ISFP pilots were delivered in ten of the fifteen UK medical schools, and this noticeably increased participation in the pilots. Other incentives to participate in the pilots included:

- The experience of taking part in an SJT as preparation for future selection processes
- Feedback on the SJT and answers to a subset of items
- A random cash prize draw (one per school)
- £1,000 provided to student organisations for 50% turnout (large-scale pilot only)

7.5. Feedback from pilots

Feedback was sought from participants involved in the pilots through an evaluation questionnaire. A member of the ISFP project team or Work Psychology Group attended the majority of SJT pilot sessions to provide a feedback session to participants on the SJT and answers to a subset of questions, using a standardised slide-set. The feedback session was an opportunity to explain the role of the SJT pilot in developing future selection methods, to dispel any concerns raised, and to hear the views of participants. Overall feedback was positive, particularly when comparing the SJT with ‘white space’ questions.

During the initial pilot, evaluation forms were completed by 428 participants. 78.7% of participants agreed or strongly agreed that the content of the test seemed relevant to the FP. 67.7% agreed or strongly agreed that the content of the test appeared to be fair. Only 32.1% of participants agreed or strongly agreed that the results of the test should help selectors to differentiate between weaker and stronger participants, however 31.5% neither agreed nor disagreed with this statement.

During the large-scale pilot, evaluation forms were completed by 652 participants; this includes some participants from the non-UK sample who took the initial pilot paper. 89% of the participants felt that the test instructions were clear and easy to understand, while 85% thought that the information given about the pilot was clear and helpful. 85% of participants also agreed or strongly agreed that the content of the SJT seemed relevant to the FP. 80% felt that the scenario content was appropriate for their level of training and 75% considered that the difficulty level was appropriate. Overall, 73.3% of participants agreed or strongly agreed that the content of the test was fair, but when considering whether the results of the test would help differentiate between the strong and weak participants, 42% neither agreed nor disagreed.

The majority of participants commented that the SJT is a better method of selection than ‘white space’ questions. Only a very small number of participants felt that the SJT should not be adopted in its current form, and these participants mostly believed that sufficient improvements could be made to make the SJT a useful selection tool. Feedback suggested that participants felt the test was about the right length in terms of time and content, though they often felt that there was a large amount of information to read in the time. It has been suggested that the stage that students are at in their course could affect the answers they give. Several feedback sessions have raised a lack of awareness amongst participants about the different responsibilities of clinical supervisors and educational supervisors.

7.6. Initial pilot - key findings

Figure 14 outlines the gender, ethnicity and nationality of the 455 participants at four UK schools who took part in the initial pilot. The mean age of the sample was 24.7 years, with a range of 22-43 years.

		No. of participants	% of participants
Gender	Male	159	34.9%
	Female	275	60.4%
	Not declared	21	4.6%
Ethnicity	White	319	70.1%
	Asian	67	14.8%
	Black	10	2.2%
	Chinese	10	2.2%
	Mixed	12	2.6%
	Other	7	1.5%
	Not declared	30	6.6%
Nationality	British	378	83.1%
	Non-British	30	6.6%
	Not declared	47	10.3%

Figure 14: SJT pilot participant demographics (initial pilot, Autumn 2010)

The pilot paper consisted of 65 items (42 ranking and 23 multiple choice) to be completed in two hours. Five of the items were used to give feedback to participants. The total available score for the paper was 1116. Analysis was adjusted to exclude erratic answer patterns and large amounts of missing data.

The majority of participants (94.9%) completed the test within two hours, supporting the intention that the SJT is a 'power test' not a 'speed test'. However it should be noted that attitude to a pilot which does not count towards their live application is likely to differ from a live application process, where applicants could be expected to spend longer reflecting and reviewing their answers. Statistical analysis demonstrated a high level of internal reliability ($\alpha=0.74 - 0.77$), only slightly below the level of internal reliability required for a high stakes assessment ($\alpha=0.80$). Psychometrically, over half of the items worked well. The items that did not perform well psychometrically are undergoing further review and refinement.

Two scoring conventions were considered – rational scoring, whereby the scores are determined using a pre-determined answer key, and empirical scoring, where items are administered to a large-scale pilot sample. For items where the empirical key differs from the rational key, agreement was sought with Subject Matter Experts as to whether the empirical key is appropriate. There was a wide range of scores in the initial pilot (minimum 682 and maximum 952 for rational scoring; minimum 710 and maximum of 992 for empirical scoring). This range of scores demonstrates that although there were a number of tied scores, overall the test is able to differentiate between applicants.

Female participants scored slightly higher than male participants (by approximately 0.3 standard deviations). White participants scored slightly higher than Black and Minority Ethnic (BME) participants. This is a typical and comparable finding with many selection methods and examinations. Caution should be employed when interpreting these results due to the small sample sizes.

All participants in the initial pilot were required to complete a work style evaluation questionnaire, which provides insight into social desirability. Work using the Social Desirability scale suggests that applicants 'faking' their responses to reflect their perception of desirable rather than what they would do in a situation was not a serious issue. In theory

applicants cannot prepare for the test, although there is an inevitable risk that private companies set up 'revision' courses and guides.

The findings from the initial pilot indicated that the majority of items worked well psychometrically, the reliability of the test was good and it was able to differentiate between participants. The pilots confirmed the design specification for the SJT, and provided lessons for the administrative and logistical delivery of the SJT.

7.7. Large-scale pilot - key findings

Figure 15 outlines the gender, ethnicity and nationality for the 639 participants in the large-scale pilot. The mean age of the sample was 24.8 years, with a range of 21-48 years. Fourteen participants (2.2%) self declared that they had a disability, as defined by the Equality Act 2010.

		No. of participants	% of participants
Gender	Male	228	35.7%
	Female	401	62.8%
	Not declared	10	1.6%
Ethnicity	White	436	68.2%
	Asian	89	13.9%
	Black	14	2.2%
	Chinese	31	4.9%
	Mixed	23	3.6%
	Other	24	3.8%
	Not declared	22	3.2%
Nationality	British	553	86.5%
	Non-British	69	10.8%
	Not declared	17	2.7%

Figure 15: SJT pilot participant demographics (large-scale pilot, Spring 2011)

The large-scale pilot paper consisted of three papers, in order to maximise the number of new SJT items piloted. Each paper consisted of 60 items (40 ranking and 20 multiple choice) to be completed within two hours. Within each of the papers, 15 of the items were anchor items common to each paper to assist with analysis, plus the same five items used in the initial pilot for feedback. At this stage, it was not possible to use these to equate the tests as this was the first time the items had been piloted. A total of 1,040 marks were available for each version of the SJT paper.

As outlined in Figure 16, there was an unequal split of participants between the three papers. Efforts were made to ensure an even number of participants across the three papers, although the number of participants at each pilot site could not be predicted.

	No. of participants	% of overall sample
Paper One	138	21.6%
Paper Two	344	53.8%
Paper Three	157	24.6%

Figure 16: Participation in three papers of large-scale pilot

In total, 614 participants (96.0%) completed all items, confirming that the SJT is a power test, and that two hours is an appropriate amount of time to complete 60 SJT items.

Analysis was adjusted to remove participants who had an erratic answer pattern or a large amount of missing data. For all three papers, test statistics showed a relatively high level of internal reliability ($\alpha=0.71 - 0.76$), before being adjusted to exclude poorly performing items. When only those items that have good psychometric properties were analysed and corrected for test length using the Spearman Brown formula, all three papers show good internal reliability ($\alpha=0.79 - 0.85$), sufficient for that required for a SJT being used in a high stakes assessment ($\alpha=0.80$).

The range of scores (minimum 658 and maximum 922) is similar to that achieved in the initial pilot. This is a good spread and shows that the test is able to differentiate between participants. Analysis showed a high level of agreement between the rational keying and empirical keying ($r = .96, p < .001$).

Female participants scored slightly higher than male participants (by approximately 0.5 standard deviations) on Paper One; there were no significant differences by gender on Paper Two and Three. White participants scored higher than BME participants on all three tests. This is a typical and comparable finding with many selection methods and examinations. It should be noted that there was variation between the papers as to how gender and ethnicity affected results, suggesting that a strong pattern of group differences cannot be identified. Further research could help to explain observed group differences.

Around 100 of the 135 items in the large-scale pilot were of sufficient quality to be used in future tests. The remaining items require further review and refinement, as they did not achieve the psychometric properties for a live SJT item in the large-scale SJT pilot. The number of items that were of sufficient quality for future use is typical for SJT development and further review will help to identify the reasons for poorly performing items.

In order to evaluate the construct and criterion related validity, participant performance (tested on two small separate cohorts) on a NEO PI-R personality measure, and comparison with OSCE performance and written assessments, were evaluated. The pilots demonstrated positive correlation with high levels of extraversion, openness to values and achievement. From this, it can be concluded that the SJT broadly measures attributes relevant to the skills needed for work in the Foundation programme.

8. Educational Performance Measure (EPM)

Following extensive consultation and piloting, all stakeholders have reached agreement on a standardised EPM framework of performance at medical school in relation to the graduating cohort up to the point of application to the Foundation Programme, which meets the agreed quality criteria. This next chapter details the process of consultation and piloting, the lessons learned, and decisions taken in developing the EPM framework. The final EPM framework is provided in Appendix I.

8.1. Background

Applicants to the Foundation Programme currently receive a score based on performance at medical school in relation to their cohort, as ranked into four quartiles worth 40, 38, 36 and 34 points. This score is then combined with their score from an online application form, 'white space' questions (total of 60 points), which includes 10 points for degrees, presentations, prizes and publications.

Concerns about the use of academic quartiles, raised before and during the detailed Option Appraisal, include the comparability of applicants from different medical schools and discrimination of applicants at the margins between quartiles. Given issues around the lack of transparency and lack of consistency in the allocation of points for quartiles across medical schools, one of the main drivers to review the quartile system was to ensure defensibility in the event of legal challenge. Stakeholder feedback showed strong support for the use of some measure of academic performance as well as non-academic and possibly extra-curricular activities in the selection process. The International Expert Panel, whose advice was sought during the 2009 Options Appraisal, supported the principle of

making greater use of information accumulated during medical school, to recognise that the Foundation Programme is an extension of education and training, both to value applicants' academic achievements, and using the evidence of past performance as the greatest predictor of future performance.

The ISFP Project Group (previously the Steering Group) recommended that a standardised measure of applicant educational performance up to the point of application to the Foundation Programme be piloted and developed. It was envisaged that the EPM would be derived using a standardised and transparent framework of existing performance measures, agreed with medical schools in consultation with applicants and other stakeholders, and that this measure would address some of the concerns around comparability of applicants from different schools, and that it would enable greater granularity when combined with the SJT. All UK and non-UK medical schools would be required to provide an EPM score to the UKFPO.

Under the current system of applications to the Foundation Programme, applicants receive up to 10 points for previous degrees, presentations, prizes and publications which meet the criteria specified by the UKFPO. These points are awarded within the 'white space' questions. There was full stakeholder support for recognising the most exceptional applicants and their additional educational achievements as evidenced by degrees, presentations, prizes and publications, and incorporating this into the EPM.

8.2. Consultation and piloting

There were two in-depth consultations with all 31 UK medical schools to identify the type of information on student performance available at the point of application to the Foundation Programme, and the practicalities of using this within the EPM framework. At this stage, wider stakeholder consultation was not sought.

The first consultation, in September and October 2009, gathered evidence around the information currently used to inform quartile rankings, and the assessment information on applicant performance currently collected and utilised by UK medical schools. A summary of responses is provided in Appendix H, and can be summarised as follows:

- All schools record data on student performance relating to clinical skills and curriculum knowledge. Measures of professionalism and records of extra-curricular activities are rarely collected systematically, and are impossible to quantify
- There is a range of assessments covering clinical skills, curriculum knowledge, Student Selected Components (SSCs) and measures of professionalism, as well as summative, formative and progress-testing, used by medical schools. The timing at which the subject knowledge is assessed varies between schools, as different curricula approach the Tomorrow's Doctors outcomes framework in different ways
- Assessment formats include Single Best Answer (SBA), Multiple Choice Questions (MCQs), Extended Matching Questions (EMQs), Observed Structured Clinical Examinations (OSCEs), case presentations and essays. The number of assessments of each type, and the weightings between the different curriculum elements, varies considerably between medical schools
- The GMC framework Tomorrow's Doctors specifies the minimum requirements of a competent doctor, and as such, the majority of assessments devised by medical schools are designed to assess whether a student's performance is satisfactory or not, rather than to differentiate between the most exceptional students and those closer to the borderline. Given this, a full ranking within a cohort is neither desirable nor feasible. Some assessments, for example Student Selected Components, are typically pass/ fail. For other assessment types, a grade is awarded - often in line with university policy -, and for others, a raw score or percentage
- As part of the undergraduate MBChB or MBBS programme, students have the opportunity to undertake additional study towards an intercalated degree, usually taking an extra year of study to complete. Intercalation is usually by competitive entry, based on performance at medical school, and is not available to all students. Between 0% (Warwick – all graduate entry) and 100% (Cambridge, Imperial, Oxford – excluding graduate entrants) of students at a given school intercalate, with a median of 25% across the country

Analysis of the initial consultation led to the proposal that in place of overall quartiles to reflect performance in relation to the cohort, a more fair and granular method would be to record two quartile measures of performance – one relating to clinical skills or practical assessments, and a second relating to the underpinning scientific knowledge or written exams. The two scores would then be combined to provide an overall score, with additional granularity to the current quartile points.

A second consultation, in November and December 2009, sought educationalist opinion around the principles of an EPM framework on the basis of two quartile measures, including the relative weighting between the two components, the weighting between the different years of study, as well as more detailed questions around how additional points for additional academic achievements (degrees, presentations, prizes and publications) might be awarded. The responses are summarised as follows:

- Given that medical schools have diverse approaches to delivering the undergraduate curriculum and a range of assessments of student performance on the required outcomes, it should be the responsibility of individual schools to determine how many, which, and the relative weightings of, assessments to be representative of educational performance at the point of application to the Foundation Programme
- The weighting between measures of clinical 'practical' skills and 'written' curriculum knowledge should be 50:50. Some schools did raise concerns about the practicalities of providing separate scores for these two elements
- Performance across all years of study should be recognised, with a greater emphasis on performance in the later years of study. Differences in the length of course – for example between fast-track graduate entry (4 years) and standard entry (5/6 years) - should be taken into account
- The majority (25) of schools believed that the EPM should recognise additional academic achievements (presentations, prizes, and publications) in addition to performance in relation to the cohort. There were also proposals to recognise the most exceptional students within the cohort
- There was mixed opinion around whether the EPM should award points for possession of degrees in addition to the MBChB or MBBS, with 13 schools (7 schools with Graduate Entry (GE)) commending additional points for previous degrees; 7 (4 GE) schools commending points only for degrees gained during undergraduate medical degree; and 6 (3 GE) schools commending no points for additional degrees

8.3. Pilot of draft EPM framework

A draft EPM framework for piloting, informed by the two consultations, was agreed by the ISFP Project Group. The pilot EPM framework called for applicants to be given two scores - one relating to written assessments and the other relating to practical assessments – according to a specific prescription of weightings between earlier (40%) and later years (60%) of the course. The pilot EPM framework also specified rules that:

- Only data from summative assessments, not formative ones, would count
- Where re-sits have been taken, the marks would be capped at the pass mark except in the case of extenuating circumstances
- Schools would devise appropriate weightings for modules and assessments
- Schools whose students transfer to another school partway through the course would be required to provide the school of graduation with the necessary data to allow early years' data to be included
- The calculations would be made using standard-set raw marks, not grade points

Of the 25 UK medical schools involved in the EPM pilot in May 2010, 22 schools provided performance data for 5,373 applicants to FP2010 in the form of separate scores for written and practical assessments. EPM data were provided as normalised scores, to allow comparisons to be made within a single school cohort. Three schools provided EPM scores for a further 442 students in different cohorts, although the scores could not be correlated with other information. The remaining six medical schools indicated that they would in principle be able to adhere to the draft

framework, but had not been able to do so within the timeframe. As the pilot used retrospective data, the EPM scores could then be analysed with the original quartile scores supplied by the medical school, and with the application form scores achieved by applicants to the FP2010. The report of the pilot of the draft EPM framework is provided in Appendix H, and the findings can be summarised as follows:

- For some schools, the pilot EPM framework required the use of additional assessment data (for example during the early years of study) not usually included in the calculation of quartiles; other schools reported that the framework limited the number of assessments that could be used (for example by requiring a split between 'written' and 'practical' EPM scores). Typically, between three and fourteen assessment scores were used across the two measures of performance; some of these were raw scores but others were on the basis of grade points
- To illustrate the diversity of scores provided by participating medical schools:
 - One school provided written EPM scores consisting of 13 possible scores, for 228 applicants, between 0.5 and 2 (2 being the lowest performing applicant and 0.5 the best performing applicant) whereas another school provided written EPM scores as 210 possible scores (223 applicants) within a range of 139.6 and 397.6
 - Eight schools were able to provide a full discrete ranking for the cohort, although there was feedback that this level of differentiation was artificial
 - Twelve schools provided written EPM scores that enabled division into four equal sized quartiles. Owing to tied scores, with one exception, for the remaining schools there was a difference in group size of between four and ten applicants
 - Three schools provided practical EPM scores which enabled a full rank (no tied scores). For 16 of 25 schools, applicants could be ranked into four equal quartiles for practical EPM scores. There was a significant variation in the number of tied scores at the other nine schools, for example at one school, four times as many applicants were placed in the third quartile (126) as in the fourth (32)
- The Pearson's correlation co-efficient was calculated to correlate the written EPM scores and practical EPM scores with applicants' original quartile points score. Across 22 schools, the written EPM scores had a strong correlation of 0.86, but across 20 schools there was only a moderately strong correlation of 0.65 for the practical EPM scores with the original quartile score. This indicates that the EPM pilot framework places a different emphasis on the elements currently used to inform quartile points
- The Pearson's correlation co-efficient for written EPM scores and practical EPM scores was 0.51, with just 2,306 of the 5,373 applicants placed in the same quartile for both measures, and 899 applicants (17%) ranked more than two or more quartiles differently (e.g. 1st in one measure and 3rd or 4th in the other)
- For comparison of quartile scores with written EPM scores 67% of students were in the same quartile; 30% of students moved to one quartile either side of their original rank; and 3% of students moved by more than 2 quartile ranks. For comparison of quartile ranks with practical EPM scores 46% of students were in the same quartile; 44% of students moved to one quartile either side of their original rank; and 10% of students moved by more than 2 quartile ranks

Different methods of combining scores for the written EPM scores and practical EPM scores were modelled, including addition, multiplication, division and standard deviation. Whilst distance from the mean is the statistically more robust measure, the unknown variables (difficulty of assessments, comparability between schools) cast doubt on the reliability of any measure other than combining two scores by addition.

By adding together the points for written EPM quartiles and practical EPM quartiles, the range of points achieved ranged from four possible overall bandings to seven possible bandings. Further analysis of the underlying data indicated that it would be possible to further increase the differentiation between applicants (granularity) by extending the different number of bandings of points available for performance at medical school from four to seven, and again from seven to ten. Increasing the number of bandings can help to differentiate applicants by i) providing a greater

spread of scores and ii) reducing the number of applicants with tied scores, without greatly affecting the values of scores achieved in terms of mean, median or mode.

8.4. EPM pilot lessons

The lessons of the EPM pilot can be summarised as follows:

- The EPM pilot demonstrated the importance and the benefits of a nationally agreed framework in terms of transparency and consistency
- Averaging an applicant's performance over a 'basket' of assessments should be more representative of the applicant's capability than taking any one of the assessments individually
- The EPM pilot highlighted that a framework which specifies the proportion of assessments from different stages in medical school is inappropriate, since it does not recognise the variation across schools in the timing of courses and assessments. It is not practical to specify the exact composition of the EPM, and the different approaches to assessment and learning should be encouraged
- An EPM framework that calls for a clear split between assessments of curriculum knowledge and clinical skills is not feasible, as one competency underpins the other
- Not all schools were able to adhere to the EPM pilot framework in full, as the university systems did not record raw scores only grades in some instances, and it was not possible to identify re-sit marks for three or four years previously. An EPM framework would need to be sufficiently flexible to allow for a transitional period if these rules are to be required
- The pilots indicated that in principle, there is scope to make the EPM more granular than quartiles, provided that the underlying data enable sufficient differentiation between applicants. However many education leads within the medical schools raised concerns that a high level of granularity might be artificial, if assessments are designed to assess minimum competence e.g. pass/fail, rather than to differentiate between the most exceptional and the more satisfactory students. Statistically, increasing the number of bandings can help to differentiate applicants by i) providing a greater spread of scores and ii) reducing the number of applicants with tied scores, without greatly affecting the mean, median or mode. Issues of feasibility and desirability would need to be addressed
- Whilst many schools consult on and publish the composition of quartiles, there is wide variation between medical schools in the way they produce quartiles. The pilot revealed the different compositions of current measures of performance at medical school – which include assessments in all years in some schools to a single assessment in a single year in others. The pilot also raised issues around the comparability of students within a single cohort who had taken a different combination of assessments, for example comparing those on graduate entry with standard entry, as well as the minority of students who transfer between institutions at the end of third year

8.5. EPM Task and Finish Group (EPM Group)

An EPM Task and Finish Group (EPM Group) was convened to recommend specific and workable standards for the EPM framework, in consultation with stakeholders. The EPM Group was Chaired by Professor Mike Greaves, University of Aberdeen, and its membership of twelve included representation from medical school staff, medical students, employers, and the Department of Health. The EPM Group met face to face once in November 2010, with subsequent work carried out by individuals on the EPM Group, and decisions taken electronically. The following chapter outlines the conclusions of the EPM Group, and the process by which these conclusions were reached in consultation with stakeholders.

8.6. Draft framework

A draft EPM Framework was agreed by the EPM Group in February 2011, and proposed to the relevant stakeholder groups (medical school staff, students, employers, ISFP Project Group, ISFP Project Board), and revised to reflect feedback and further recommendations. The final EPM framework was approved by the EPM Task and Finish Group

(April 2011), the Medical Schools Council (May 2011), the ISFP Project Group (May 2011) and the ISFP Project Board (May 2011). Legal opinion was sought (June 2011) and the outcome of this advice is detailed in Appendix A.

8.7. Performance in relation to graduating cohort

The EPM Group recommended that the most meaningful comparable measure of performance is of the applicant relative to the graduating cohort at their medical school. There can be no legal requirement for any non-UK medical school to change its curriculum in order to provide a score for application to employment; furthermore, although all UK medical degree courses teach to the GMC's curriculum framework, Tomorrow's Doctors, the timing and format of assessments varies considerably between medical schools. Given this, it is neither desirable nor possible to define an EPM framework that stipulates the types of assessment taken in particular years of the degree course.

The EPM Group recommended that the EPM framework should require each school to devise a representative 'basket of assessments' covering the curriculum, in consultation with its student body. The standards for the production of EPM scores need to allow for the variety of ways in which assessments are conducted, and must reflect the outcomes to be achieved rather than the means by which they are achieved. In consultation with staff, students and assessment experts, the EPM Group also recommended rules for the consistency of the type of assessments to be included, and the treatment of special cases, for example re-sit assessments and transfer students.

The EPM Group recommended that performance in relation to the graduating cohort be measured in as close to equal sized deciles as possible, allowing for tied scores. The EPM group recognised that the degree of granularity cannot be increased beyond the resolving power of the underlying assessments, particularly given that a number of medical schools record grades rather than raw scores. However, a basket of assessments would provide multi-source performance data, and permit a greater degree of differentiation between applicants. This would be fairer for applicants at the margins of quartiles, whilst not increasing the level of granularity artificially.

8.8. Degrees, presentations, prizes and publications

The EPM Group recommended that a small number of points are available to reward academic excellence, as demonstrated by the possession of a degree, or presentation, prize or publication. The EPM Group accepted the BMA Medical Students Committee's (BMA MSC) recommendation that these points should be divided into two categories – degrees and then separately presentations, prizes and publications.

There was an extensive consultation with stakeholders around the eligibility of additional points for degrees (whether all previous degrees, only degrees taken within the length of the programme, or only degrees that extended the length of the programme), and the criteria for presentations, prizes and publications. The final recommendations were agreed by all stakeholders, and accepted by the UKFPO Rules Group.

8.9. Extra-curricular activities

The EPM Group recommended that there are no points for extra-curricular activities. The BMA MSC had proposed three possible methods of recording and scoring extra-curricular activities: credits, a standardised portfolio and a free text space to declare additional achievements. However, the EPM Group assessed that the three options were neither cost-effective nor valid, particularly given that there is no evidence that an applicant who has undertaken extra-curricular activities would perform better as an FY1 doctor than any other applicant. Furthermore, the added value of undertaking additional activities and responsibilities (for example through sports, volunteering, committee representation, caring for a relative etc) would already be reflected through performance in the SJT, which assesses the professional attributes of a foundation doctor.

9. Aspects of the application process using SJT and EPM

9.1 Aggregating the SJT and EPM scores

Expert advice was received about how the points for the EPM and the SJT should be aggregated, to provide a total application score. The advice was informed through modelling of score combination techniques, informed by the data of the SJT spring pilots as a basis to estimate the means and standard deviations of key parameters. It was agreed that the aggregation method should provide an appropriate and defensible outcome but also that it should look intuitive, thus gaining a high level of acceptability. This advice concluded that the SJT results should be recalibrated to one decimal point to a 50 point scale and added or averaged with the EPM on a 50 point scale without further data modification. The report of this analysis is provided as Appendix J.

9.2 Algorithm

The report of the ISFP Options Appraisal recommended that there should be further review of the matching algorithm used to match successful and eligible applicants to the Foundation Programme to foundation schools.

Feedback from applicants indicates that the reasons for choosing particular foundation schools are as much a factor of location and personal preference, as they are of the programmes available. Furthermore, patterns of preferences suggest that strategic choices are made by applicants, particularly amongst the London foundation schools which are historically usually oversubscribed.

Applicant preferences, and application scores, are loaded onto a central computer system, and a macro is used to match applicants to foundation schools according to the pre-defined rules. In the event of oversubscription, the highest scoring n applicants are selected onto the Foundation Programme, and the remaining applicants placed on a reserve list until such time as a place becomes available.

The matching algorithm used up to FP2011 was 'First Preference First', giving priority to applicants for whom a foundation school is their first choice. Applicants are placed into rank order according to their application score. The rank ordered list is then processed by allocating each applicant in turn to her first choice school if a place at that school remains, and if not, deferring further consideration of the applicant to Phase 2. In Phase 2 each applicant in turn is assigned to the foundation school highest on her preference list that still has at least one place remaining.

The alternative matching algorithm considered for selection to the Foundation Programme is known as the 'Serial Dictator', which gives priority to applicants with the highest scores first. Applicants are placed into rank order according to their application scores, and each applicant is allocated in turn to the foundation school highest on her preference list that still has at least one place remaining.

Whilst the 'First Preference First' algorithm maximises the number of applicants placed in their first choice foundation school, and thus provides greatest applicant satisfaction, it is possible and likely that an applicant A with a low score will be allocated to a given foundation school, even though applicant B may have a higher score, by virtue of their choices. The effect of this is that applicants may select their preferences strategically, and secondly, that applicants may be allocated to a foundation school very low on their list of preferences (once they have missed their first preference) if other schools higher on their list are full. The 'Serial Dictator' algorithm achieves a lower allocation to first choice school, but in no case is applicant A, with a lower score, allocated to a foundation school in place of higher scoring applicant B.

Each time a matching algorithm is run, there is random tie breaking for applicants with the same score vying for places in an over-subscribed foundation school. It is worth noting that, because of the rather coarse grained scoring system of the current system, the number of ties that have to be broken is very large, containing several hundred applicants. The

greater granularity provided by SJTs will improve this situation. There is currently a judgement call as to how these tied ranks should be split, as follows:

- The 'greedy' rule, which maximises the number of 1st choices, then subject to that, maximises the number of 2nd choices, then the number of 3rd choices, and so on
- The 'generous' rule, which minimises the number of last choices, then subject to that, minimises the number of (last-1)th choices, and so on
- The 'amended-generous' rule, which maximises the number of 1st choices, then subject to that, minimises the number of last choices, then the number of (last-1)th choices, and so on
- The 'amended-greedy' rule which seeks a profile that is the same length as the best possible generous profile, but subject to that, maximises the number of 1st choices, then maximises the number of 2nd choices, and so on

The FP2011 preferences were re-run using the First Preference First and Serial Dictator matching algorithms. The data analysis revealed that with both matching algorithms, the same 5,979 of the 7,073 applicants (85%) would have been allocated to their first choice foundation school. In the First Choice First algorithm, an additional 412 applicants were allocated to their first choice school (90%). Of the 400 applicants allocated to their second choice foundation school through the Serial Dictator algorithm, 134 would have been allocated to their first choice foundation school using the First Choice First algorithm, but 187 applicants would have been worse off, placed between their 3rd and 18th choice foundation school. With both matching algorithms, more applicants in total are placed in the foundation schools of first, second and third choice (92.5% with 'Serial Dictator' and 92.3% with 'First Choice First'). Overall, 412 applicants are better off under the 'First Choice First' algorithm, and 458 applicants are better off under the 'Serial Dictator' algorithm. However, there is a question of fairness as to which applicants should be allocated to foundation schools lower on their preferences. It should also be considered that the FP2011 data may contain an element of strategic preferences.

The ISFP Project Group reviewed the data, and recommended to the UKFPO Rules Group that the 'Serial Dictator' matching algorithm be used to match applicants to foundation schools, on the basis that it was the fairest method by which to do so. At its meeting in May 2011, the UKFPO Rules Group agreed to implement the new matching algorithm, 'Serial Dictator', from applications to FP2012, one year before any changes to the selection tools used are implemented.

9.3 Development and testing of SJT item banking software

Since February 2010, The Medical Schools Council Assessment Alliance (MSC-AA) has been developing a new online item bank to store, tag, quality assure and facilitate the delivery of Situation Judgement Test items. Development work has been shared between the MSC-AA Information Systems Team and Epigenesys Ltd., a software development company based in Sheffield. The functional specifications of the system are derived from the work performed by the MSC-AA Information Systems Team on the UKCDR (UK Collaboration for a Digital Repository) specification.

The development methodology adopted was test-driven and feature-driven. The development team generally works on the basis of short iterations of 2 or 3 weeks within which features are designed, implemented and tested in an iterative manner.

A beta version of the system was released in May 2011, and is now extensively used to author and quality assure items. The last large piece of functionality, allowing users to download items directly from the bank, is scheduled to be delivered in October 2011. Since its beta release, the system has been extensively tested in terms of functionality, usability and security. It has now been used in the framework of more than 5 authoring events as well as 20 review events across the UK, and user feedback has been largely positive.

As at 1 August 2011, the system holds a total of 91 SJT items. As additional authoring and quality assurance events are planned in the near future, and as existing content is transferred to the software, this number is expected to rise significantly over the course of the next few months.

Technology

The new MSC-AA item bank is a web-based application designed to support the complete quality assurance cycle of the SJT items, from authoring to exam readiness. It also features a multi-level taxonomy tagging system, allowing the tagging of items against an unlimited number of taxonomies.

The application is written in Ruby and makes use of MySQL (relational database management system) and MongoDB (a document-based database) for item storage. JQuery (Javascript library) is also used on the front-end to allow for a rich user experience.

The application is stored on Epigenesys's servers in Sheffield, and scheduled backups of the database are made on a regular basis. Connection to the server is made through a secure, encrypted, https certificate. Each author and reviewer has unique login credentials to the system, which can be revoked at any time by the system administrators.

Security

Penetration testing was carried out on the system by in April 2011 by ProNetExpert Ltd., a security expert company based in Edinburgh. Although security of the system was deemed satisfactory, additional security measures have been implemented to increase the security of the system. Another penetration testing session is scheduled to take place before the end of year 2011, to assess the security level of the application after the implementation of the item delivery feature set.

9.4 Defence Medical Services

Each year approximately 60 military cadets, who have been supported by the MoD through medical school, are allocated by the Defense Deanery to supernumerary Foundation Programmes. A memorandum of understanding between the Defence Postgraduate Medical Deanery (DPMD) and the regional deaneries ensures that military foundation doctors are cost neutral, such that the Trust receives the service component as a 'free good' and does not charge for the educational component. These students currently do not apply through a competitive process for a Foundation Programme and this process is unlikely to change.

The ISFP Project Group recommends that all applicants to the Foundation Programme should complete the national application process using the SJT and EPM, in addition to any separate processes to recruit to the Defence Medical Services.

9.5 Academic Foundation Programmes (AFP)

Around 5% of Foundation Programmes are Academic Foundation Programmes (AFP), which provide foundation doctors with the opportunity to develop research, teaching and leadership/ management skills in addition to the competences outlined in the Foundation Programme curriculum. Recruitment to the AFP is currently a regionally coordinated process, usually comprising a standard application form and interview (telephone, video conference or panel), as well as evidence of academic performance at medical school. Applicants are nominated for an AFP by their graduating medical school, and eligibility is checked nationally. Currently, recruitment to an AFP takes place up to six months before selection to the general Foundation Programme. If AFPs are not filled, the posts may be allocated to the general Foundation Programme. Similarly if applicants are unsuccessful, they are still eligible to apply through the national application process. For FP2011, 1,328 applicants applied to 472 AFPs through a two-stage process. There was a single national offers date, followed by a cascade system.

The ISFP Project Group recommends that all applicants to the Foundation Programme should complete the national application process using the SJT and EPM, in addition to any separate processes to recruit to AFPs, before their appointment can be confirmed.

10. Communications

The communications strategy was produced by the Communications Officer for the Medical Schools Council. A number of key areas were addressed in the strategy including the communication objectives and key messages, different stakeholder groups grouped via importance, key activities to help raise awareness of the project, risks, resources and budget. The ISFP Communications Plan is provided as Appendix K.

10.1 Development and sign off

The strategy was developed in April 2010 and circulated to a variety of stakeholders, including the UKFPO, for comments. This was then discussed and approved at a Project Group meeting on 15 April 2010, after which it was presented to Project Board members for final approval and sign off.

10.2 Key achievements

Stakeholder groups were involved in a variety of ways, with information tailored to their needs. All comments from these stakeholders were collated and fed back to the ISFP Project Group throughout the duration of the project. The key achievements are described in sections 10.3 – 10.10.

10.3 Website

An ISFP website was developed to host all relevant information about the project and to help gather feedback from users via an online discussion forum. The website was launched on 5 May 2010 and as at 20 July 2011 there had been 11,675 visits (10,623 from the UK) and 33,351 page loads, with visitors spending an average of 2 minutes 7 seconds on the website. Website visitor numbers varied each month, and the higher levels of visitors coincided with the various communications that were sent out and key project milestones e.g. pilots. During the 2010 summer months and over the Christmas period there were very few visits, suggesting that most users were students. Google Analytic data demonstrate that 72% of all visitors came directly to the site, which means they had clicked on the links included in the communications or that they knew the website address. An additional 22% were referred by other websites demonstrating that the weblink which stakeholders put on their website were worthwhile. The most viewed pages were the home page, the SJT main page, and the example SJT question and answers page. The website will continue to be developed during the Parallel Recruitment Exercise (PRE) as the information about the proposed changes develops.

10.4 Social Media

A Facebook Group was set up in May 2010. As at 20 July 2011 there were 1,115 members and 21 comments had been posted on the discussion forum and main page. This has proved an effective way of reaching over 1,000 medical students about the project by sending messages to them at key stages of the project.

A video podcast involving Professor Paul O'Neill and Nick Deakin – who was the BMA MSC representative on the ISFP Project Group at the time – was produced and hosted on YouTube. As at 20 July 2011 there were 1,993 unique views and 5 comments had been posted.

There were 44 comments posted on the ISFP discussion forum between 16 April 2010 and 26 May 2011. The ISFP Project Group responded to some of these as they were either questions or areas that needed to be clarified.

10.5 E-bulletin

An e-bulletin sign up function was added to the ISFP website to allow interested stakeholders to register for project updates. E-bulletins were sent every other month between October 2010 and July 2011 to the 244 registered users.

Updates were also sent via other organisation's e-bulletins including: UKFPO via its Medical Student Board, Foundation School Directors and Foundation Schools Managers networks, BMA MSC and NHS Employers.

10.6 Communication documents

A number of key documents were sent to medical students, either about the project as a whole or specifically about the SJT pilots. These included: a key facts document to all medical students, a leaflet for all students which was circulated with the Student BMJ and Student BMA News to 20,000 subscribers, a Project Handbook, FAQ documents, letters from Professor Paul O'Neill to all students outlining the project developments in October 2010 and February 2011, and updates for school newsletters. Tailored information was also sent to students at pilot schools including emails about the pilots and workshops with follow up emails two weeks later, and pilot letters to all students outlining what to expect on the day of the pilot and what they needed to bring with them.

10.7 MedSoc Presidents

To ensure that information was reaching as many medical students as possible, the MedSoc Presidents at each medical school were contacted on a number of occasions with information about the project and asked to circulate to all students. Twenty-two MedSoc Presidents replied to confirm they had circulated this information.

10.8 Pilot workshops

Workshops were held at eleven medical schools involved in the pilots between November 2010 and April 2011. The purpose of these was to inform students about the project and to encourage participation in the SJT pilots. Unfortunately, only one of the four schools in the initial pilot was able to host a workshop. This was due to the timing of the pilots and the fact that some students were not back at medical school after the summer period with enough time for this to be delivered. The proportion of medical students who attended the workshop and then attended the pilot was extremely high. As a result, it was agreed that it would be beneficial to hold these workshops at schools involved in the large-scale pilot. Ten schools were able to host a workshop, with the remaining five schools being sent some personalised slides to circulate to all final year students instead. Attendance at the workshops varied due to the timing of these, although feedback was extremely positive. The workshops provided an excellent opportunity to talk through student concerns and gather their feedback on the proposed changes. All schools covered the cost of room hire.

10.9 Presentations

Updates, both oral and written, were provided at a number of key meetings including: UKFPO Medical Student Board, UKFPO Rules Group, UKFPO Careers Annual Conference, BMA Medical Students Committee, Medical Schools Council, MCAN and NEAF (undergraduate and postgraduate careers advisors), GMC Undergraduate Board, Medical Programme Board, UK Scrutiny Group, Medical School Secretaries, Royal Society of Medicine Intercalation Conference and Universities UK staff meetings. NHS Medical Workforce Directors and Medical Directors were also sent email updates.

10.10 Press

Medical press were contacted during each milestone of the project – mainly with information about the pilots. The Student BMA News featured a number of articles, the Student BMJ included two small updates at each pilot stage on their news bite page, the BMJ Careers wrote an article about pre-employment competency checks which mentioned the ISFP project and small number of other medical websites included information which was circulated in press releases.

11. Legal Advice

The arrangements for selecting applicants to the Foundation Programme must comply with the relevant UK and European legislation, notably that relating to employment, equalities, human rights and data protection. Given this, the ISFP Project Group has taken legal advice about the recommendations in this report. The legal opinion is provided as Appendix A.

A briefing for counsel was drafted by the ISFP project team, and agreed with DH. The brief outlined all of the key recommendations relating to the proposed selection methods that had been piloted, summarised as follows:

- Selection will be based on the results of an invigilated SJT and the EPM
- The SJT will be run on a small number of dates in the UK
- Applicants will bear their own travel costs for sitting the SJT
- The EPM will have two components, one reflecting each student's academic ranking in their cohort, expressed in deciles (e.g. the top 10% of the students in the cohort will be in the first decile, the next 10% in the second decile and so on), the other component being awarded for exceptional academic achievements (including additional degrees)
- Each medical school will be required to base its decile framework on a representative basket of academic assessments undertaken by all of the students in a cohort, and to publish the composition of the basket of assessments
- Where applicants cannot provide a decile score, they will be treated as if they were in the lowest decile
- Applicants who re-sit their final year will be assessed on the basis of their original decile (e.g. the decile achieved before the re-sit year), plus any points for additional academic achievements at the time of application
- The professional attributes to be tested by the SJT will be included in the national Person Specification

The legal opinion was sought from a firm of solicitors familiar with the existing arrangements for selection to the Foundation Programme. The opinion was received on 1 July 2011, and reviewed by the ISFP Project Group. In summary, the legal opinion is that the proposed use of the SJT and EPM is consistent with the relevant legislation. Specifically, the solicitors had no concerns about the legality of:

- The inclusion in the national Person Specification of the professional domains identified in the Job Analysis
- The use of an invigilated SJT
- The use of the EPM
- Selecting applicants on the basis of a combined SJT/ EPM score
- The award of additional points for degrees, presentations, prizes and publications, provided that care was taken to ensure that comparable achievements were equally recognised
- Requiring medical schools to publish the basis upon which their EPM scores were determined
- Judging re-sit applicants on the basis of their original decile (e.g. the decile they were awarded in relation to the original cohort before they re-sat a year)
- The possibility of charging applicants to take the SJT
- The possibility of setting a threshold SJT score for the purpose of excluding very low-scoring applicants from selection on the basis of not meeting the national Person Specification

The review has suggested that there remains some small risk of a successful legal challenge in relation to certain equality and diversity considerations, but this appears to be an improvement relative to the risks that are present with the current selection arrangements.

The specific issues raised by the legal opinion are considered in more detail in sections 11.1 – 11.5.

11.1 SJT results of overseas applicants

There was no feasible way to pilot the SJT with a representative sample of overseas applicants during the initial or large-scale pilots. Of the 112 overseas applicants to FP2011 who completed clinical assessments in the UK, only 7 volunteered to participate in the pilot SJT. Members of the ISFP project team travelled to two non-UK medical schools to deliver the SJT pilot, however there were only 43 non-UK participants in total. These data were removed from analysis, so as to not distort interpretation of the results based on a small sample. It will not be possible to compare the performance of UK and non-UK applicants on the SJT until the SJT is used for live selection in FP2013, which will be the first opportunity to obtain data from a representative population of overseas students.

The SJT should at least reduce the risk of a successful challenge compared with the current 'white space' questions. In terms of language, the current 'white space' questions are a composition test, whereas the SJT is only a comprehension test, which is less onerous.

The ISFP Project Group recommends that the SJT performance of applicants from non-UK medical schools should be analysed after each recruitment round.

11.2 Holding the SJT in the UK

The legal review has suggested that holding the SJT in the UK alone could present a small risk of challenge on the grounds that applicants having to travel to the UK to take the test might be disadvantaged relative to those who live here. The solicitors have recommended that, in order to minimise the risk, the rationale for holding the test in the UK should be properly scrutinised and documented. We agree with this recommendation, and propose that the option of delivering the SJT overseas should be evaluated in more detail. Our current expectation is that delivering paper-based SJTs overseas will not be feasible; however, it may become a viable prospect if the SJT is delivered as an electronic test in the future.

The solicitors have also suggested an investigation of the possibility of setting up a grant scheme to contribute towards the travel expenses of applicants who can be shown suffer financial hardship as a consequence of having to travel to the UK to take the SJT. The ISFP Project Group suggests that this should be taken forward in the run-up to the FP2013 recruitment round. In the meantime it should be noted that the requirement for non-UK applicants to travel to the UK to sit the SJT is no more onerous than the current arrangements requiring eligible non-EU applicants to travel to the UK to sit the clinical assessment as part of eligibility checking for selection to the Foundation Programme.

11.3 Skewing of EPM scores

The current proposals for the EPM require each individual medical school to use a representative 'basket' of assessments to rank all of their students in a given cohort. Since the nature and number of assessments varies so much from school to school, the EPM framework cannot specify which individual assessments should be taken into account by any given school. Given this, the solicitors have questioned whether it might be possible for a medical school to base its EPM scores on a deliberately skewed combination of assessments in order to favour its students. Fortunately the nature of the EPM deciles prevents schools from being able to gain an advantage for their cohort. Each applicant is given a score relative to all of the other applicants from the same school, so any change to the 'basket' of assessments that favoured one applicant from the school would necessarily disadvantage another from the same school- there is no way to skew the results for the cohort as a whole.

This argument does not apply in the case of a medical school from which only a single student is applying, since in principle it would be possible for the school to skew the selection of assessments to favour that one student while ignoring the effect on all their others. This is an issue that has already been encountered by the UKFPO in relation to the existing academic quartile scores. In practice the scope for such skewing may be reduced by requiring all schools to publish the composition of their EPM basket of assessments and by providing evidence of where the student sits in relation to the whole cohort. It should be noted that this can only currently be required of UK schools.

11.4 Failure to provide a decile score as part of the EPM

Currently, where applicants are unable to provide an academic quartile score they are treated as if they were in the lowest quartile. It is proposed that this arrangement should continue for the decile component of the EPM, so that applicants who do not provide a decile score will be treated as if they were in the lowest decile. The legal advice is that this is a sensible arrangement; however the solicitors have suggested that where applicants can show that they are unable to provide a score through no fault of their own then it would be fairer to provide some other means of judging their academic performance. In practice, however, there appears no workable way of achieving this. Aside from the EPM itself, there is no separate standard test that can be used reliably to equate the academic performance of one applicant compared with a significant sample of others. In principle the grade of an applicant's final degree, such as 1st, 2.1, 2.2 etc, might be mapped in some way to deciles, but there are two problems with this approach: most medical degrees are not graded as other university degrees are, and in any case most applicants do not sit their finals exams until after they have applied to the Foundation Programme.

11.5 Weighting SJT and EPM scores

It is proposed that the SJT and EPM scores of an applicant should be combined to provide a single overall score for the purpose of selection to the Foundation Programme. This raises the question of the relative weighting to be given to each component to arrive at the overall total. The legal advice is that the weighting given to each component should reflect its relative value as a predictor of the qualities of the applicants; and that whatever weightings are used, they should be published.

12. Management Lessons Learned

One of the purposes of a pilot project is to learn lessons to become better prepared for a live implementation. In this respect, the current project has been very insightful. There has been important learning in the following areas, all of which inform the recommendations in this report to underpin a smooth implementation of the new selection methods:

- Scheduling and managing the delivery of a national SJT in multiple venues
- The practicalities of producing and quality assuring SJT items
- The performance of the SJT as a selection method
- The involvement of volunteers for SJT item development and review
- Paper and electronic delivery options for the SJT
- The variety of assessments across medical schools
- Communications
- Ongoing evaluation
- Recommendation: Parallel Recruitment Exercise (PRE)

These topics are considered in more detail in sections 12.1 – 12.9.

12.1 Scheduling and managing a national SJT

There are two main competing factors to be taken into account in relation to the scheduling of the SJT. On the one hand, many students at UK medical schools spend some part of their final year undertaking an elective period of practical learning overseas; the timing of these 'electives' varies from school to school, so that there is no single day in the academic calendar upon which all medical students are in the UK. Requiring large numbers of students to travel back to the UK to take the test during their elective would be costly and disruptive, so this argues in favour of allowing the SJT to be run on a range of dates to suit the timing of electives. On the other hand, there is a significant cost associated with increasing the number of dates on which the SJT may be run, in terms of the provision of venues and also the number of items that would be needed in a single recruitment round. There is a possible risk that students at an earlier sitting of the SJT might leak information about the test to students attending a later sitting. To prevent this, different versions of the SJT test paper need to be produced for each sitting, and the different papers need to be 'test equated' to ensure that the overall results take into account any variances in the relative difficulty of the papers.

Aside from the extra cost of producing extra test items, there are other complexities that arise if the SJT is to be scheduled to run on more than one sitting. For example, in order to analyse the outcome of the tests in psychometric terms, it is important that every version of the paper is taken by a sufficiently large and representative sample of applicants, which means that care must be taken to ensure that an adequate number of applicants attends each sitting. While these issues were apparent at the outset of the project, the pilot has provided first-hand experience of their practical implications, which are taken into account in the recommendations.

Other points of learning that have arisen in this area include: the practicalities of distributing and collecting test papers across the UK in a secure and reliable way; the pitfalls associated with the need for each applicant to have a unique identifier that allows their SJT answers to be matched in an unambiguous way to their national Foundation Programme application form; the variation in the nature and capacity of the venues available at medical schools for running assessments; and the time and effort required to produce, quality assure, and mark SJT papers.

12.2 The production of SJT items

The project has adopted a model for the development and quality assurance of SJT items that combines features of two other similar processes, namely the development of academic test items by the MSC-AA and the development of SJT items used by the National Recruitment Office for GP speciality training. The pilot has provided an opportunity to test the process, and gather metrics that will be needed to plan development of SJT items for the longer term. More specifically, the pilot has provided insights including the productivity of 'item writing workshops'; the proportion of draft SJT items that are suitable for piloting and for live use; the time and effort required to generate and quality assure items; the challenges of maintaining the security of the SJT items.

12.3 The performance of the SJT as a selection method

In all around 1,100 students have taken SJT papers produced during the pilot. This has generated a significant amount of data that have been subjected to psychometric analysis. As a consequence the pilot has yielded important information about the various aspects of the performance of the SJT as a selection method, including:

- The reliability of the SJT
- The time required for applicants to complete the SJT
- The number and type of items to be included within the SJT
- The granularity of the SJT results
- The validity of the SJT

These topics are reported in detail in the report of the SJT pilots (Appendix F).

12.4 The use of volunteers for SJT development

The project has trialled a particular approach to the development and quality assurance of SJT items that is reliant on the participation of volunteers at every stage of the process. More specifically, the volunteers include:

- Clinicians familiar with the role of an FY1 doctor, who draft the SJT items
- Foundation doctors, who review the drafted items to confirm that: the items present scenarios that FY1 doctors might realistically encounter in practice
- 'Concordance Panel' members, who sit SJT tests to act as proxies for good applicants

The decision to use volunteers was in part influenced by the MSC-AA model, in which academic questions for use in medical school finals are produced and quality assured by volunteers. This contrasts with the GP model, in which clinicians are paid to draft the SJT items. The experience gained in the pilot has suggested that it is possible to encourage a sufficient amount of voluntary participation overall, but the use of volunteers introduces constraints and leads to unpredictable outcomes.

12.5 Paper and electronic delivery of an SJT

The project has gained experience of three options for delivery of the SJT, namely:

- A paper-based SJT
- An electronically delivered SJT
- A hybrid approach, in which the SJT was completed on paper but applicants used special handsets to record their answers electronically

In each case, the SJTs were delivered under invigilated conditions at venues provided by medical schools. Details about each of the options are presented in section 7.2. The overall learning suggests that the paper-based test is currently the most feasible of the three options trialled. The main reason for this has been, in some medical schools, the lack of suitable facilities for delivering the tests electronically. That said, the creation, distribution, and collection of different versions of paper tests, across a range of venues and dates, is a considerable administrative burden, and presents some challenges in terms of maintaining the physical security of the papers.

Two further options for delivery that could be considered are the electronic delivery of tests either through i) commercially available test venues, or through ii) the use of centrally owned laptops or i-pads. These two methods would avoid the constraints imposed by a lack of infrastructure to deliver assessments to the whole cohort at some medical schools and would overcome the issues of adapting software and access to different hardware and software systems. In any case, if multiple versions of the test are required – rather than for use on common national dates – a larger number of items would be needed and would add to the costs.

12.6 The variety of assessments across medical schools

At the outset of the project there was an expectation that a framework for the EPM could define the type and numbers of assessments that should be taken into account when evaluating student educational performance in relation to cohort. Indeed just such a framework was originally drafted and piloted in the first half of the project. However, this exercise served to highlight the diversity that existed across medical schools in the following areas:

- **Optional v mandatory assessments.** Some schools offered their students a wider choice of optional programmes and assessments
- **Practical v written assessments.** Some schools maintained a sharp distinction between practical and written assessments, while others ran tests that combined both elements
- **Granularity of outcomes.** Some assessments resulted in very detailed scores, while others might result only in a binary pass/fail outcome. Moreover, some schools retained the detailed 'raw' scores achieved by their students, while others (as a matter of university policy) retained records in the form of grades (e.g. A, B, C...) and did not retain the raw scores
- **Timing and number of assessments.** Some schools implemented a small number of detailed assessments, while others ran a larger number of smaller assessments. There are also differences in the timing of assessments; some schools tended to hold practical assessments at an earlier stage than other schools

Given the above, it became clear that a prescriptive formula for the construction of an EPM score, of the sort that might dictate a certain mix of practical and written assessments taken from specific years of the curriculum, could not be defined in a way that would provide comparable results across all schools. Instead, it would be necessary for each school to determine which 'basket of assessments' would yield the most representative outcome locally.

12.7 Communication

The importance of communications, via a variety of media, must not be underestimated. The key messages must be conveyed to final year medical students taking part in the Parallel Recruitment Exercise (PRE), but also importantly to future cohorts of applicants, both UK and non-UK. It is important that the stakeholder groups aware of the forthcoming changes include employers, deaneries, foundation schools, and clinicians currently involved in marking the white-

space questions – who would in future be involved in SJT item-writing and quality assurance. It will be important to provide the evidence and rationale for the changes, and to ensure familiarity and awareness before the changes are used for live selection to the Foundation Programme.

12.8 Ongoing evaluation

The SJT pilot has demonstrated that the SJT is an appropriate method for selection to the Foundation Programme, as it is able to differentiate sufficiently between applicants, the majority of items performed well and the reliability of the measurement was good.

The commissioning of a new online application process facility provides the opportunity to include a mechanism to facilitate a separate research project to correlate anonymous data collected through selection to the Foundation Programme with past and future performance data, with permissions.

12.9 Recommendation: Parallel Recruitment Exercise (PRE)

Running the pilots has demonstrated the importance of all elements of a live application process being fully piloted before live implementation. Whilst 15 UK medical schools have been involved in the SJT pilots, it is recommended that all 30 UK medical schools with graduating students be involved in a full-scale Parallel Recruitment Exercise of both the SJT and EPM. Legal opinion indicated that participation in the SJT could not be mandated, as it would not 'count' towards live application. However, the opportunity to participate in the SJT should be made available to all applicants to FP2012.

A full-scale pilot SJT and EPM provides a unique opportunity:

- To pilot the SJT administrative arrangements in all venues, ahead of live recruitment
- To pilot new SJT items for inclusion in the SJT item bank
- For UK medical schools to consult and publish the 'basket of the assessments' to inform the decile component of the EPM score. In the event that an applicant reappplies to FP2013, their original decile score will count – so this will facilitate smooth transition to the new points system in FP2013
- To obtain additional data to evaluate further the use of the SJT and EPM for selection to the Foundation Programme
- To address any administrative or other issues ahead of live implementation

A communications strategy for the PRE has been produced by the Medical Schools Council's Communications Officer and the ISFP Communications Consultant, and should be followed. This outlines the ways in which students and other stakeholders will be kept informed.

13. Cost Benefit Analysis (CBA)

The decision to develop and pilot the SJT and EPM was based on the outcome of a Cost Benefit Analysis (CBA) which weighed the relative advantages and disadvantages of the SJT and EPM in combination with other alternatives.

13.1 Original CBA

The original CBA was undertaken in the summer of 2009, and weighed the relative advantages and disadvantages of the SJT and EPM compared with other competing options that had been identified by an expert panel. The other options considered during the Options Appraisal are summarised in Figure 17.

Option Name	Summary
Do Nothing	Continuing with the current approach to selection, retaining the 'white space' questions and academic quartiles
Structured Interviews/ EPM	Each applicant would meet with a panel of interviewers and be asked a defined set of questions with an agreed, structured scoring system. The scored interview would replace the use of 'white space' questions. There would continue to be a need for medical schools to provide something similar to the academic quartile score to take into account the applicant's academic performance
Multiple Mini Interviews (MMI)/ EPM	MMIs are an extension of structured interviews, in which the applicant rotates around a series of short interviews with each interview being designed to explore a particular element of the personal specification for foundation training. Each interview is structured with a calibrated scoring system. Again the scored MMIs would replaced the use of 'white space' questions, and medical schools would continue to provide something similar to the academic quartile score to take into account the applicant's academic performance
National Assessment	Every applicant would undertake a standard assessment (which is separate from their Finals) of their clinical knowledge and skills. The assessment would replace the use of the 'white space' questions and would obviate the need for a separate academic score from medical schools

Figure 17: Alternative selection tools considered during Options Appraisal

The CBA followed the recommendations of the Treasury's 'Green Book' which provides a framework for the formal evaluation of options. Estimates were made of the costs of each option over five years based on rates current in 2009, with a 3.5% annual discount factor. The relative costs of the options were estimated as in Figure 18.

Option	Five year cost
Do Nothing	£21m
Structured Interview/ EPM	£23m
MMI/ EPM	£26m
National Assessment	£32m
SJT/ EPM	£19m

Figure 18: Relative costs of selection tools in Options Appraisal

To evaluate the non-financial pros and cons, the options were scored against a suitable set of evaluation criteria, shown in Figure 19.

Criterion Name	Definition
Reliability	The technical reliability of the selection technique associated with the option. Broadly this means the likelihood that applying the technique in the right way will give consistent results.
Validity	The technical validity of the selection technique associated with the option. Broadly this means the extent to which the technique is actually measuring the characteristics of a 'good' doctor.
Granularity	The extent to which the selection method provides a fine-grained set of scores for the applicants.
Consistency	The extent to which the selection techniques can be expected to be consistently applied across the UK.
Longevity	The extent to which the performance of the option can be maintained over successive recruitment rounds.

Criterion Name	Definition
Educational Impact	The extent to which the option supports or undermines educational objectives.
Fairness	The extent to which the option offers a level playing field for applicants.
Compliance	The extent to which the option discourages, prevents, or otherwise guards against the effects of cheating or malpractice.
Transparency	The ease with which applicants can understand what is expected of them and why they achieved their ranking.
Applicant Burden	The extent to which the option minimises the costs and effort for applicants.
Medical Time	The extent to which the option minimises the amount of medical staff time required for selection.
Feasibility	The ease with which the option could be brought into successful live use.
Public Opinion	The extent to which the general public, when presented with a sound-bite description of the process, are likely to agree with it.

Figure 19: Criterion for non-financial costs and benefits

The options were scored against each of the criteria on a scale of 0 to 10, where 0 represented the worst option in terms of a given criterion, and 10 the best. The criteria were then each given a weighting to take into account their relative importance, and the significance of the spread represented by the scores from 0 to 10. The results of this are summarised in Figure 20.

		Reliability	Validity	Granularity	Consistency	Longevity	Educational Impact	Fairness	Compliance	Transparency	Applicant Burden	Clinical Time	Short Term Feasibility	Long term Feasibility	Public Opinion
Do Nothing		0	3	5	0	0	4	7	0	0	5	2	10	3	0
MMI/ EPM		10	8	5	2	8	4	4	7	5	3	0	4	3	10
Structured Interview/ EPM		8	7	0	2	8	4	4	7	6	3	1	5	3	10
National Exam		10	0	10	8	9	0	0	10	7	0	0	0	0	10
SJT/ EPM		10	10	10	10	9	6	10	10	3	3	10	2	7	5
Weighting (Totals)	100	10	10	1	7	12	15	5	7	2	4	8	10	10	1
Do Nothing	290	0	29	5	0	0	59	34	0	0	20	16	98	29	0
MMI/ EPM	516	98	78	5	14	94	59	20	48	10	12	0	39	29	10
Structured Interview/ EPM	501	78	69	0	14	94	59	20	48	12	12	8	49	29	10
National Assessment	361	98	0	10	55	106	0	0	69	14	0	0	0	0	10
SJT/ EPM	775	98	98	10	69	106	88	49	69	6	12	78	20	69	5

Figure 20: Non-financial cost benefit analysis

A sensitivity analysis showed that the combination of SJT/ EPM had the lowest cost and highest score regardless of any realistic changes to the underlying assumptions, and so should be the preferred option on both counts.

13.2 Updated CBA

The CBA has been revisited in the light of the experience gathered during the project. Most of the assumptions made in the original cost benefit analysis have been shown to remain valid. The two main factors needing revision are costs and feasibility, considered in section 13.3 and 13.4.

13.3 Costs

A key assumption underlying the original cost estimates was that the all-in cost of developing, maintaining and delivering the SJT would be the same as the costs incurred by the GMC in developing, maintaining and delivering Part A of the PLAB (Professional and Linguistics Assessment Board) test, which is used to assess doctors who have qualified in medicine outside of the EEA. The PLAB test appeared to be a reasonable comparator on the grounds that, like the SJT, it was an invigilated, machine-markable, multiple-choice test, for which the items were developed by clinicians. The PLAB Part A fees at the time were £145 per applicant, which equated to £1.13m per year for an SJT to be taken by 8,000 applicants.

The deciles component of the EPM will be maintained by individual medical schools, which is at no cost to the national delivery of the recruitment process; although as with the current system of quartiles, does require significant staff resource. There will need to be a method of verification for points for additional academic achievements.

Current estimates suggest that the costs for the use of the SJT and EPM in combination for selection to the Foundation Programme will be lower than first expected, being nearer to £0.9m per year, rather than £1.13m per year. This sum excludes the running costs of the UKFPO and FPAS. The costs cover:

- The secure storage of SJT items
- Maintaining a bank of quality-assured SJT items to support the year-on-year requirements for selection to the Foundation Programme. This is to include:
 - Analysing the performance of items
 - Removing or improving items with inadequate performance characteristics
 - Updating items that may have become dated as a result of changes to terminology, standards etc
 - Forecasting the demand for new items
 - Creating and quality assuring new items to meet the projected demand, including the piloting of new items
- Defining the standards for the arrangement, conduct and invigilation of SJTs by medical schools
- Providing suitable guidance about the standards, and ensuring that the standards are understood, and adhered to, by medical schools
- Creating, printing, distributing, delivering, collecting and marking SJT papers (assuming the test is delivered on paper), and providing the results of the test to UKFPO in an agreed format
- Arranging, hosting, invigilating and administering SJTs at venues throughout the UK
- The maintenance and evaluation of the EPM framework and standards
- Providing suitable guidance about the EPM framework, and ensuring that the standards are understood, and adhered to, by UK medical schools
- Promoting awareness and understanding of the SJT and EPM
- Providing decile points for UK applicants, as part of the EPM
- Verification of decile points for non-UK applicants, as part of the EPM
- Verification of points for degrees, presentations, prizes and presentations within the EPM

13.4 Feasibility

Given that the SJT/ EPM option has been piloted, and the other options have not, this is now a more feasible proposition than it was when the original CBA was performed. Therefore the score for the feasibility of SJT/ EPM needs to be increased relative to the other options, which strengthens the case for SJT/ EPM in terms of the financial and the non-financial factors.

13.5 The case for change revisited

Figure 21 summarises the case for change; it lists the main concerns about the current selection process, and indicates the corresponding improvements that can be expected to be brought about by the proposed changes.

Concern about current system	SJT and EPM
The use of 'white space' questions is not sustainable as a selection tool and will become steadily less discriminatory between eligible applicants given that i) there is a limited range of new questions that can be generated	SJTs will draw upon a bank of items to be available for each application round. Given that the situations experienced in the Foundation Programme are varied and complex, new items can be built incrementally and continuously against the detailed Job Analysis
The marking of 'white space' questions is labour intensive, the cost in clinician time approaching £2 m per year	SJTs are machine markable. The clinician time involved in the development and quality assurance of SJT items is expected to be between £50,000 and £100,000 annually
The use of 'white space' questions in non-invigilated conditions – and the availability of model answers on the internet – raises concerns about the risk of plagiarism and coaching	SJTs will be undertaken in invigilated conditions in the UK on no more than three national dates It is not possible to revise for the SJT as the scenarios are complex and answers relate to judgement rather than knowledge
While the 'white space' questions offer a practical way to rank large numbers of comparable applicants, their technical reliability and validity is open to question	There is evidence for the validity of SJTs dating back more than 30 years. The SJT pilots demonstrate the technical reliability, internal reliability, and validity for use for selection to the Foundation Programme
The academic quartile system makes it difficult to compare fairly between applicants from different medical schools, as it is not standardised or subject to quality assurance across medical schools	The EPM includes a standardised framework for deciles, within which medical schools have the flexibility to define in consultation with their student body, the contents of the 'basket of assessments'. Schools will be required to publish their locally agreed deciles framework, which will facilitate transparency and quality assurance from the wider community. The move to deciles will also be fairer to applicants at the margins
If there is an increase in the number of eligible non-UK applicants such that the Foundation Programme is over-subscribed, the selection process may be more likely to be subject to legal challenge from unsuccessful applicants to the Foundation Programme	The use of a valid SJT in invigilated conditions, and EPM in combination are a more reliable, robust, fair and valid method of selection to the Foundation Programme, and thus likely to withstand legal challenge

Figure 21: The case for change revisited

14. The Way Forward

14.1 Final recommendations

- Selection to the Foundation Programme should reflect the skills, knowledge and professional behaviours of the applicant, reflecting the integrated nature of the Foundation Programme as both education and employment
- Selection to the Foundation Programme from FP2013 onwards should be based upon:
 - An invigilated Situational Judgement Test (SJT) to assess aptitude for the Foundation Programme (to replace 'white space' questions); and
 - An Educational Performance Measure (EPM) to reflect educational performance at medical school up to the point of application to the Foundation Programme (to replace quartiles)
- Other aspects of the process of application to the Foundation Programme should remain unchanged, namely a national application process including applicant declaration of foundation school preferences, the use of a matching algorithm and eligibility checking
- There should be a full-scale shadow Parallel Recruitment Exercise (PRE) SJT and EPM run concurrently with the current application process for FP2012
- Using data from tracking, there should be ongoing evaluation of the validity of the SJT and EPM as methods for selection to the Foundation Programme

Assuming the final recommendations of the ISFP project are accepted, more detailed arrangements to operationalise the implementation for FP2013 are set out below. These will also be subject to ongoing evaluation and can be adapted as appropriate for future recruitment rounds.

14.2 Operational Guidance for FP2013

- Scores for performance on the SJT and EPM should be equally weighted
- A matching algorithm, as agreed by the UKFPO Rules Group, should be used to match applicants to foundation schools according to their preferences and their application scores
- Matching of applicants to individual programmes should remain a local process, managed by the foundation school to which the applicant is allocated
- All applicants to the Foundation Programme, including applicants to academic and to the defence medical services, should complete an SJT and EPM. Separate methods of recruitment to these programmes should continue to run in parallel
- In the event of oversubscription, the highest scoring applicants should be selected to the Foundation Programme. The remaining eligible applicants should be allocated to Foundation Programme posts as vacancies arise, according to their application score
- On application, applicants should provide consent that their subsequent performance will be tracked anonymously
- The communications strategies for both the PRE and the live application process should be implemented

14.2.1 SJT Operational Guidance

- The SJT should consist of 60 live questions. Around two thirds of the questions should be in the format 'rank five options' and one third of the questions in the format 'select three from eight options'
- The content of the SJT should be mapped against the professional domains identified in the Job Analysis
- The SJT should be taken under invigilated examination conditions in the UK, on no more than two nationally agreed dates between November and January. An additional date should be offered to those applicants with extenuating circumstances. The SJT should also be offered at the same venue as the clinical assessments undertaken as part of eligibility checking

- A single SJT question paper should be used on any one date and different versions of the test used for test equating purposes
- Applicants may only take the SJT once during a single recruitment round
- There should be national guidance for the standards of delivery of the SJT by test centres, and a Service Level Agreement in place to require that these standards are adhered to
- There should be agreed national rules for e.g. eligibility of extenuating circumstances, provision for special requirements
- UK medical schools should be responsible for delivering the SJT to their graduating cohort, in line with national guidelines
- A separate venue (central location) should be booked for the delivery of the SJT to non-UK applicants, in line with national standards
- The SJT should be paper-based and machine-markable in the first instance, with further consideration of computerised delivery of the SJT in future years
- There should be centralised printing, scanning of answer sheets and marking
- The Quality Assurance processes recommended by the review of the SJT pilots should be followed.
- The scoring convention used for the pilots should be used for FP2013
- Only SJT items with confirmed acceptable psychometric properties through post-test analysis will count towards the applicant's SJT score
- There should be a programme for developing new SJT items. Alternative models of item-writing should be trialled and reviewed for effectiveness
- New items should be piloted in addition to the SJT live test content within the SJT papers, in order to develop the SJT item bank for use in live selection
- Secure SJT item-banking arrangements should be in place
- Further work should be undertaken to explore whether candidates who perform extremely poorly in the SJT do or do not meet the national Person Specification.

14.2.2 EPM Operational Guidance

- The EPM score should be derived from educational achievements in three parts:
 - Performance at medical school in relation to the final year cohort, banded by deciles
 - Additional degrees (Bachelors, Masters and Doctorates)
 - Academic achievements including presentations, prizes and publications
- All UK medical schools should publish the composition of the framework by which the decile points are determined, informed by local consultation, according to the principles of the EPM framework
- UK medical schools should provide a decile score to the UKFPO in an agreed format
- Non-UK applicants should submit a Dean's Statement confirming i) the size of the graduating cohort, ii) their class rank within the graduating cohort and iii) the decile points awarded
- Where an applicant is unable to provide a valid decile point score, they will receive the same number of points as the lowest decile
- Applicants should provide evidence of degrees, presentations, prizes and publications, to be eligible for these points. A mechanism for verification of the evidence should be put in place

15. Success Criteria

The success criteria for the ISFP Project are as follows:

- **Completeness.** All of the project's products have been completed and accepted
- **Legality.** Legal advice has confirmed that the proposed approach is robust
- **Finances.** The projected costs of implementing the new approach are less than the costs of the current process
The actual costs of the project are within the approved budget
- **Reliability and Validity.** The tests have confirmed that SJT and EPM provide a reliable and valid selection method
- **Practicality.** The project has confirmed that: it is possible to schedule and run the SJT for the expected numbers of applicants, taking into account the need to avoid electives; medical schools are able to provide EPM scores to the required standard, and there is a satisfactory means of dealing with applicants that cannot provide EPM scores
- **Buy-in.** The key stakeholders have confirmed their support for the approach, including:
 - The 4 UK Health Departments
 - All foundation schools
 - All medical schools
 - BMA Medical Students Committee
 - UKFPO
 - COPMED
 - General Medical Council
 - NHS Employers
 - Medical Education England
 - Medical Royal Colleges
- **Robustness.** Piloting has shown that it is possible for medical schools to implement the SJT securely, taking into account the need to keep questions secret, to ensure the identity of the applicants taking the tests, and to mark the scripts correctly
- **Timeliness.** The key products have been delivered in good time for the start of the application process for FP2013
- **Sustainability.** It is possible to develop a sufficiently large number of SJT items over many years

Improving Selection to the Foundation Programme

Appendix A

Legal opinion, July 2011



**ADVICE TO THE UK FOUNDATION PROGRAMME ON ARRANGEMENTS FOR
RECRUITMENT TO THE FOUNDATION PROGRAMME IN THE UNITED KINGDOM**

I set out below the legal questions that have been raised, together with my advice.

1. Are the new arrangements compliant with all relevant legislation?

We have considered in particular the requirements of employment and equalities law, human rights, data protection and public law. There are some potential areas of challenge in relation to the area of equality and diversity. Otherwise, we consider that the proposals are consistent with relevant legislation.

2. If not, what aspects might be challenged and on what grounds?

The particular areas of challenge are dealt with in response to the specific queries below.

Recognising that there are potentially issues around the different aspects of the process, further questions related to the different aspects are:

Inclusion of professional domains in the FP Person Specification

3. Is it reasonable for the FP Person Specification to include the professional domains identified through the Job Analysis of the F1 post undertaken as part of this project?

We have no concerns about the inclusion of the professional domains in the person specification.

Use of an invigilated SJT

4. It is proposed that selection should in part be determined by an assessment of likely professional judgements and professionalism as these traits strongly relate to job performance. Is this a reasonable measure of the qualities of the applicant in



relation to the UKFPO Person Specification, in particular, professional judgements in likely work based contexts?

It is right that the selection of successful candidates should be based on traits which are proven to relate to a strong performance in the role. Clearly, those traits should also be reflected in the person specification. Therefore, provided that the matters being assessed fall within one of the categories set out in the person specification, we consider that this is an acceptable and reasonable measure of an applicant's qualities.

Having reviewed the paper prepared following the pilot SJTs, we note that there was evidence that non-British participants scored lower than British participants on all tests and that the difference in scores, although small, was statistically significant. This does raise the prospect of potential claims of discrimination on the basis of race.

Section 19 of the Equality Act provides that a person (A) discriminates against another (B) if A applies to B a provision, criterion or practice ('PCP') which is discriminatory in relation to a relevant protected characteristic of B's. Subsection (2) of section 19 goes on to explain that a PCP is discriminatory in relation to a relevant protected characteristic of B's if—

- (a) A applies, or would apply it to persons with whom B does not share the characteristic,
- (b) it puts or would put, persons with whom B shares the characteristic at a particular disadvantage when compared with persons with whom B does not share it,
- (c) it puts, or would put, B at that disadvantage, and
- (d) A cannot show it to be a proportionate means of achieving a legitimate aim.

The protected characteristics which are covered under this provision are age, disability, gender reassignment, marriage and civil partnership, race, religion or belief, sex and sexual orientation.

Applying these provisions to the proposal that applicants undertake the SJT, we consider that applicants from overseas could seek to make out grounds (a) to (c) on the basis of the data



which currently exists. We are mindful of the fact that the numbers in the non-British group are small, which renders the data unreliable as being representative of the group as a whole. However, on the face of it the pilot scheme does demonstrate that as a group non-UK candidates are at a disadvantage.

The question that will need to be addressed by the Foundation Programme is whether this disadvantage can be justified as a proportionate means of achieving a legitimate aim. In this circumstance, it is our view that the goal of recruiting the most suitable candidates onto the foundation programme is a legitimate aim. We note that the research literature indicates that SJTs have significant validity in predicting job performance and can offer incremental validity over methods such as ability tests and personality questionnaires. Furthermore, they may be fairer than other methods in contexts where candidates have little or no experience in the target job, and tend to show smaller differences between candidate groups (e.g. based on race) than cognitive ability tests. These factors are all relevant in considering whether the use of the SJT and the need for it to be undertaken in the UK could be justified.

The question of whether the means of achieving this aim are proportionate is more difficult to assess on the basis of the information currently before us, because we do not have any information about what element of the test is creating the disadvantage. If it is the content of the test itself, we consider that it is likely that it would be considered to be proportionate given the very thorough analysis that has been undertaken into the attributes to be tested and their correlation with strong performance in the FY1 role. However, it is necessary to consider whether there are other means of achieving the same end that do not have the same potential to discriminate against certain protected groups. Clearly, to date the Foundation Programme has managed applications without using the SJT. Is there evidence that in doing so it has not always selected the most appropriate candidates or that the current selection process has resulted in anomalies or hardship for applicants? Can some of the benefits of the SJT in terms of predicting job performance and general fairness be achieved by other selection methods that are less likely to have a detrimental impact on some applicants?

Secondly, if the disadvantage is created by a factor other than the content of the test itself, such as language or cultural misunderstanding, then there may be an expectation that more could be done to address the issue. For example, whilst it is reasonable and lawful to expect that a doctor coming to the UK will have a knowledge of the English language which is



commensurate with an ability to undertake safe and effective clinical practice, there may still need to be some recognition of the fact that English is not the first language by offering (say) additional time for non-English speakers to complete the test.

Based on the issues set out above we consider that this is an area which might be open to challenge and where there is a small risk that such challenge could be successful. We recommend that in order to reduce this risk, some research could be undertaken into the reasons for the difference in score between UK and non-UK participants and that the results of this research, together with the outcomes of it, be recorded. This would be consistent with the general approach to assessing the equality impact of new measures that a public body proposes to take. We are, however, mindful that the difference in scores appear to actually be very small and, as mentioned, the sample used in the pilot test was very small so that it could be argued that it was not representative. It is therefore questionable whether, in reality, the difference is really sufficient to support an argument of discrimination.

5. Is it reasonable to run the SJT only on a small number of days (c. 3)?

Yes, in our view, bearing in mind the costs and practical difficulties involved in running an SJT, we consider that it is reasonable to limit the available days to a relatively small number. However, the limitation may be relevant when considered alongside the requirements to undertake the SJT in the UK, and to pay any travel expenses for attending the SJT, which are considered below.

6. To simplify administration it is assumed that UK applicants would sit the SJT at a venue provided by their medical school, while overseas applicants would be invited to sit the test at a UK venue selected to be convenient for overseas travel (e.g. London). Is it reasonable to require all applicants to undertake the SJT in the UK?

As with the matters discussed at 4 above, there is a risk of indirect discrimination claims from applicants that will be placed at a disadvantage by the requirement to undertake the SJT in the UK. Most obviously, those applicants from overseas will be required to travel to the UK in order to undertake the SJT. The question for the Foundation Programme is whether it could



show that the requirement to undertake the SJT in the UK is a proportionate means of achieving a legitimate aim.

The question of whether or not the requirement for candidates to sit the SJT in the UK is justified will depend upon the reasoning behind the decision that the test should be UK based. The key question which will be asked is whether it is possible to offer an arrangement whereby the SJT can be undertaken by overseas applicants without the need to travel, in order to reduce the potential adverse impact for applicants of attending test centres in the UK. For example, could the test be undertaken electronically and invigilated by someone local, or could it be taken at their own medical school? We note the assertion that the decision to hold the tests in the UK is to 'simplify administration'. If a decision were taken to hold tests at overseas medical schools, what would the consequences and barriers be? Would the administration and cost associated with it be so significant as to be potentially prohibitive? Has this been considered in any detail?

As above, we consider that this is an area which might be open to challenge and where there is a small risk that such a challenge could be successful. We recommend that in order to reduce this risk, detailed consideration be given to the justification for only holding tests in the UK, and/or to the possibility of tests being undertaken outside the UK; and that the decision taken as a result of that consideration, and the reasons for it, be recorded.

7. Is it reasonable to require applicants to pay their own travel expenses for attending the SJT?

This requirement will put applicants from overseas at a further disadvantage given the much higher costs that will be associated with taking the test. UK based participants need only pay the cost of attending their usual medical school whereas overseas applicants will have to fund significant travel costs and, potentially, accommodation as well.

As with the other examples of indirect discrimination, the Foundation Programme will need to demonstrate that the requirement for applicants to pay their own travel costs is a proportionate means of achieving a legitimate aim.



The briefing note does not provide an explanation of the reasons for the decision that applicants should bear their own costs. However, I anticipate that it is cost driven (if this is not the case please let me know) given the enormous expenditure which would be associated with paying travel costs of all applicants.

The legal position in relation to cost as a legitimate aim is currently unsettled. Historically, it has been the case that discriminatory practices cannot be justified on the basis of cost alone, although cost may be taken into account alongside other factors. However, two recent Employment Appeal Tribunal decisions criticised this approach and suggested that where the costs involved in avoiding indirectly discriminatory practice would far outweigh the impact of the discrimination, then cost might properly be considered to be a legitimate aim. There is likely to be further case law on this issue and it is an area which we expect to see develop significantly over the coming months.

In view of the cumulative impact of this requirement; the requirement to undertake SJT in the UK and to do so on a limited number of days, we consider this area is vulnerable to challenge. It may therefore be worth considering what steps could be taken to mitigate against the potentially unfair/discriminatory consequences of these requirements, such as introducing hardship grants for those overseas applicants who can demonstrate that they will suffer significant financial hardship if they are required to fund their own travel for the SJT.

- 8. The funding arrangements for SJT have yet to be determined. One option would be to require applicants to pay a fee to take the test, the fee being set to recover the cost of the SJT. Would such an option be reasonable? Do Whitley Council rules permit it? If that option were adopted, could the fees due from UK applicants be included in, and taken out, of their medical school fees? (e.g. the medical school fees for membership of the Medical Schools Council Assessment Alliance - who would be hosting the SJT item banking arrangements and responsible for development of the SJT for future years - would be increased to include an element to cover a fee to take the SJT).**

We consider that it is lawful and reasonable to charge a fee in connection with the SJT, provided that the fee is payable by all applicants equally. Further, it is our view that the fees



due from UK applicants could be included within medical school fees if that were the preferred approach, and medical schools agree to it.

- 9. One possibility being considered is that a threshold SJT score or ranking might be set, below which applicants might be subject to further scrutiny as to whether they meet the UKFPO Person Specification for employment. For example, the answers to the SJT items of the 5% lowest scoring applicants might be subject to further investigation. Would such an approach be reasonable and provide sufficient evidence to exclude the candidate from the application process?**

Yes, provided that the process is set out in advance, is applied equally and fairly to all candidates and that each applicant is considered against the same, specific criteria, this approach would seem to be reasonable and appropriate.

Use of the EPM

- 10. Is it reasonable to use the EPM as a measure of the applicant's overall academic achievement, including clinical skills and knowledge, up to the point of application to the FP?**

Yes, we have no concerns about the use of the EPM in principle as a measure of academic achievement.

- 11. Is it reasonable to award points for the achievement of additional degrees, prizes, publications and presentations up to the point of application to the FP?**

Yes, this would appear to be a reasonable and sensible approach. However, it must be borne in mind that the allocation of points for such additional merits may need to take into account non-standard achievements which are of a similar level, including those gained outside the UK. It must also take account of academic achievements which may now be outdated but which have comparable status in order to avoid any possible taint of discrimination on grounds of age.



This may, in practice, be a very difficult process to manage and assess fairly and may leave the scheme open to criticism and challenge where disputes arise about the relative merits of a particular achievement.

12. Is it reasonable to require each applicant (whether from a UK or overseas medical school) to have a decile score provided in a verifiable form by their medical school?

There is a risk that this will put overseas applicants at a disadvantage if their medical schools are unwilling or unable to provide decile scores. If the UKFPO has evidence of the willingness of medical schools, including those outside the UK, to provide decile scores, this will be highly relevant in assessing the impact of this proposal. Provided the impact assessment does not suggest that those applicants from overseas will be placed at a significant disadvantage as a result of this requirement, and that the scores produced are properly and fairly comparable to one another, this is a reasonable requirement.

As above, there is the potential for dispute here which may ultimately lead to criticism and challenge. The process will therefore need to be scrutinised closely and must be absolutely transparent.

13. If applicants cannot provide evidence of their decile score, is it reasonable to treat them as if they were in the lowest decile? (We believe that any other approach, such as treating them as an average applicant, would be open to abuse, in that an applicant with a low EPM score might claim to be unable to provide it, and thus gain an advantage.)

Yes, this would be reasonable if the applicant is unable to produce evidence of their decile score where they are unable to show that the failure is caused by any external factor outside their control. However, as a matter of fairness if, for example, there is clear evidence the failure to produce the score is due to an error or failure on the part of the medical school, there may need to be an alternative mechanism in place for assessing academic achievement on an individual basis.



14. Is it reasonable to require UK medical schools to publish the framework upon which they have derived the EPM rankings for their applicants?

Yes, we consider that it is appropriate and reasonable that the frameworks used by medical schools be published. This will help to ensure that fair and comparable results are produced across medical schools and should help to eradicate potential criticisms or challenges.

You should, however, be aware that having those frameworks in the public domain will leave them open to scrutiny from disgruntled applicants and, if there are elements of unfairness arising from the frameworks, it may aid a challenge.

15. Is it reasonable to require non-UK medical schools to do likewise?

Yes. However, in order to ensure fairness and comparability, detailed guidance must be issued in order to aid the medical schools in assessing the level of achievement in comparison with standard UK results. Further, there must be sufficient flexibility in the process to allow for recognition of non-standard achievements which are of comparable weight to the standard merits which are awarded points.

16. Where an applicant re-sits the final year of their medical school degree course, is it reasonable to use their original EPM (e.g. their ranking in their original cohort at medical school, rather than their ranking compared with the cohort among which they spend their re-sit year. The basket of assessments up to the point of application would be unchanged, as this would reflect only time at medical school prior to the final year)?

Yes, we consider that this would be a reasonable approach.

17. Is it reasonable to allow each medical school to determine which assessments should count towards the ranking?

It is our view that to allow too much flexibility may lead to an increased risk of criticism and challenge. There could be the potential for medical schools to skew their data to ensure that their candidates score highly, thereby improving their results and their own reputation.



However, there clearly needs to be some degree of flexibility in place in recognition of the fact that medical schools outside the UK will not apply the same assessments.

I consider that the best approach would be to allow some discretion to the medical schools within a central framework and subject to detailed guidance.

SJT/EPM weighting

18. It is proposed that each applicant will be given an overall score derived from the SJT and EPM for the purpose of selection. In deriving this overall score. Is it reasonable to use a combination of the two measures?

Yes, we consider that this approach is reasonable.

19. Is it reasonable to give an equal weighting to the SJT and EPM?

It is our view that the weighting awarded to each score should be determined by the relative value of the information obtained. For example, if the studies show that performance on the SJT has a significantly higher correlation with subsequent performance in the FY1 role, it would be sensible to apportion slightly more weight to that element of the process than to the EPM.

Whatever weighting is determined to be appropriate, it must be applied equally and fairly and it should be made known to the applicants in advance of the process.

Peter Edwards

Capsticks Solicitors LLP

1st July 2011

Improving Selection to the Foundation Programme

Appendix B

Characteristics of the current system for selection to the Foundation Programme

Characteristics of the current system for selection to the Foundation Programme

UKFPO Foundation Programme Rules Group

The Foundation Programme Rules Group is responsible for defining the rules and processes for recruitment to the Foundation Programme. The Rules Group comprises representatives from all major stakeholders, including COPMed, foundation schools, the GMC, medical schools, BMA Medical Students Committee, NHS Employers, the four UK Health Departments and the UKFPO. All major changes to the application process proposed by the Rules Group are subject to legal scrutiny and equality impact assessment. Written proposals are then submitted to the four UK health departments and to the Foundation Programme Standing Committee (which acts as the governing body for the UKFPO) for approval.

Certification of competence

In the UK, the General Medical Council (GMC) certifies competence to practise as a doctor, and specifies the legally required minimum standards. UK medical schools base their curricula on the framework set out by the GMC and determine their procedures for assessing whether students should graduate and become licensed to practise with provisional registration with the GMC. Schools do so in accordance with the standards and outcomes set by the GMC's Undergraduate Board in *Tomorrow's Doctors*, as verified both by external examiners and by the Committee's procedures for Quality Assurance of Basic Medical Education (QABME). The system of continuous external examiner scrutiny, and formal QABME visits twice every ten years, provides the GMC with the evidence it requires to confirm that the education delivered by UK medical schools prepares graduates for the knowledge, skills, attitudes and behaviours to assure patient safety as a practising Foundation doctor.

In 2005, the GMC undertook an extensive study of the need for a common final examination¹, and later published a *Strategic Proposal for Assessment*² in which it recommended consideration of the implications of either shared questions within UK medical school finals or a single national examination. This was considered as part of the initial Options Appraisal (MSC, 2009), and the ISFP Project Group have recommended that work should continue outside of the scope of the ISFP project to develop common final examination questions between UK medical schools, and to develop an assessment of clinical communication for applicants without English as a first language and who have not interacted with patients in English during the course of their undergraduate medical degree.

History of recruitment to medical training posts

A national recruitment process for the Foundation Programme was introduced in 2005 for programmes beginning in 2006, with seven of the 17 Deaneries recruiting on line through the Multi-Deanery Application Process (MDAP)³. The process was developed by a UK Strategy Group, which included the four Chief Medical Officers. MDAP introduced common, nationally agreed person specifications, application forms and scoring guidelines, and facilitated applications into foundation schools geographically remote from the graduating medical school. Applications included personal statements which were considered difficult to score objectively. The proposal to interview, for example, the lowest scoring 5% of applicants was rejected by consultants and students.

A fully internet-based application process was introduced for FP2007 and which included all 17 deaneries across the UK. The Medical Training Application System (MTAS) included an online application form and an academic ranking provided by the applicant's graduating medical school to determine the overall score. Eligibility criteria were also established. Applicants were allocated to a foundation school based first on their preference list, then their score. In

¹ http://www.gmc-uk.org/education/documents/strategic_outcomes_final_report_jun_2006.pdf

² http://www.gmc-uk.org/education/documents/strategic_proposal_for_assessment.pdf

³ Developed by the West Midlands Deanery in 2000, the MDAP was later used by four deaneries in 2004. The MDAP removed several features of the original system, including interviews, the introduction of anonymised scoring of applications by panels, and appointment based on applicants with the highest scores receiving their first preference.

2007, the UK Foundation Programme Office (UKFPO) was established to take responsibility for coordinating the online recruitment process across the UK.

Current UK recruitment process

Selection into the Foundation Programme comprises three main stages: a) selecting the most suitable eligible applicants to the Foundation Programme, b) allocation to a specific foundation schools (national process) and c) matching to a specific programme within the foundation school (local process). This project to review the methods for national selection into the Foundation Programme relates to (a) and (b) only, but the processes involved with matching to individual programmes remains the responsibility of the relevant foundation schools, and is outside of the scope of the ISFP project.

There were 7,253 eligible applications to FP2011 for 7,073 available places. Around 57% of UK applicants applied to the foundation school associated with their graduating medical school. Where the number of local graduates is not matched by the number of available Foundation Programme places, some regions such as London and Scotland are net exporters and some students are obliged to move.

Eligibility checking

All applicants to the Foundation Programme must confirm that they meet the eligibility criteria. For applicants who have graduated from a UK medical school within the previous 24 months, this is confirmed by the graduating medical school. All applicants who have graduated from non-UK medical schools are required to complete Eligibility applications. As part of the eligibility criteria, applicants are required to demonstrate they meet the requirements for provisional registration with the GMC by the start of the Foundation Programme, including demonstrable English language ability, criminal records clearance and declaration of Fitness to Practise. All applicants are subject to pre-employment checks by their employing Trust. No post is guaranteed until all eligibility and pre-employment checks have been undertaken and approved. An academic transcript and a letter of nomination from the medical school Dean are also required.

The UKFPO Eligibility Office requires all UK and non-UK applicants who qualified from medical school more than two years previously, to pass an assessment of their clinical skills. Medical schools are invited to tender to host the clinical skills assessments. The University of Manchester was awarded the contract for FP2009 and a Service Level Agreement has been extended each year since. For FP2011, 112 applicants of the 225 who were required to undertake clinical skills assessments completed them.

Non-EU nationals and UK nationals without EC rights who graduate from outside of the EU are also required to prove their Professional and Linguistic Assessment Board (PLAB) status. The PLAB examination, a GMC requirement, is an English language test, a written knowledge examination, and a clinical assessment. Non-EU nationals are required to provide evidence of their right to work as a doctor in training in the UK. For FP2011, the UKFPO's Eligibility Office processed 1,605 applications from non-UK applicants. Of those, 1,037 were eligible to apply, but only 880 completed application forms. Of those, 267 were eligible for a Foundation Programme.

The UKFPO makes reasonable allowances for the small number of applicants to the Foundation Programme with refugee status, on an individual basis, in recognition that it may be difficult for these applicants to produce the necessary paperwork to prove their eligibility.

Online application form and white space questions

The UKFPO website provides information on programmes available in all foundation schools, the application form and the online guidance two weeks in advance of the application period. During this time, applicants register and their details are checked against eligibility requirements. Applicants have two weeks to complete and submit their application form, to include their rank of all 26 foundation schools in order of preference. For FP2011, the online

application form included five competency based questions mapped against the person specification. Anonymised application forms are scored by the applicant's first choice foundation school. Each form is scored horizontally by a panel of two trained people, at least one of whom is a clinician, using nationally agreed scoring criteria. A number of applications are verified where the scoring panel doubt the truthfulness of an answer or have serious concerns about the reported behavior of the applicant.

Academic quartile ranking

The medical school Dean issues a statement to the UKFPO providing a verifiable academic quartile ranking for each applicant, based on a local ranking of academic performance in comparison with other students at the same medical school. For FP2011, 34 points were awarded for the fourth quartile, 36 points for the third quartile, 38 points for the second quartile and 40 points for the first quartile. The means of obtaining this rank is defined locally by each medical school.

Allocation to foundation school (national level)

Applicants are allocated to Foundation Schools through a matching algorithm which takes account of their foundation school choices, then their application score. In the event of oversubscription to a particular foundation school, applicants with the same score at the cut-off for allocation are then randomly allocated. Once all first choices have been allocated, applicants are allocated according to their score to the next choice foundation school on their list with vacancies.

Matching to Foundation Programme (local level)

Applicants rank the programmes available in their allocated foundation school. The method for matching students and programmes varies locally, and may include an interview. The score of the UKFPO process is used in this matching process by most schools. Applicants are informed of their programme match via the online application system, FPAS. The employing Trusts run pre-employment checks such as Criminal Records Bureau clearance and reference checks.

Special circumstances

Medical schools may recommend individual students to be allocated to a particular foundation school on the grounds of special circumstances (caring responsibilities, health or educational reasons). Applicants apply to their medical school for pre-allocation, and requests are considered by a local panel. Approval results in applicants being pre-allocated to a foundation school providing that the applicant also meets the minimum criteria of the UKFPO application process. There were 241 approved Special Circumstances applications to FP2011.

Linked applications

Linked application forms enable two applicants to be allocated to the same foundation school (with spaces available at the level of the lower score of the two applicants). This was first introduced for FP2009. Applicants with special circumstances may not apply for a linked application.

Improving Selection to the Foundation Programme

Appendix C

Project Initiation Document February 2010

Improving Selection to the Foundation Programme

Project Initiation Document

February 2010

Contents

Contents	80
1. Introduction	83
2. Background	84
Existing Arrangements	84
The Case for Change	85
The Way Ahead	86
3. Objectives	87
4. Blueprint for the New Selection Process	88
Overview	88
Organisation and Responsibilities	89
Online Application	90
EPM	91
SJT Definition and Set-up	91
SJT Delivery	91
Selection	92
Allocation to Foundation School	92
Matching to Posts	92
Pre-Employment Checks	93
Reserve List Allocation	93
5. Scope and Exclusions	94
Job Analysis	94
Person Specification	94
SJT	94
EPM	94
Models of Score Weightings	95
Algorithm	95
Common Content of Assessment (CCA)	95
Online Application System	95
Trials and Pilots	95
Communications	95

Evaluations, Legal Checks and Approvals	96
6. Constraints	97
Timescales	97
Legal Compliance	97
Standards	97
Funding and Resources	97
Approvals	97
7. Approach.....	98
Overview	98
Job Analysis	98
SJT	99
EPM	101
Score Modelling	102
8. Products.....	103
9. Assumptions.....	106
Timescales	106
Test delivery	106
UK-Wide Approach	106
Test Purpose	106
ApplicationMMC	106
The Selection Process	106
Board Costs	106
EPM	107
Procurement.....	107
Scope	107
10. Dependencies and Interfaces	108
External Dependencies	108
11. Success Criteria and Evaluation	109
Success Criteria	109
Evaluation	110
12. Organisational Structure.....	111

13. Activity Plan.....	113
The Critical Path and End Date	113
Staging	113
14. Resources and Funding	116
Budget and Funding Summary.....	116
SJT Specialists and MSC-AA Manpower and Software Development	117
Legal Advice.....	117
T&S	117
Medical Schools	118
Communications	118
MSC Core Team	118
15. Project Controls.....	119
Business Case	119
Stages	119
Products	119
Quality Management	119
Project plans	119
Communications Plan	119
Project reporting.....	119
Meetings	119
Actions and Issues	120
Risk Management	120
Change Control	120
Lessons Learned.....	120
16. Quality.....	121
Responsibilities	121
Overarching Quality Expectations	121
Quality Methods	121
17. Project Tolerances	123
18. Communications	123
19. Risks	123

1. Introduction

- 1.1. This is the Project Initiation Document (PID) for a project to develop and pilot new assessment methods for the selection and allocation of applicants to the Foundation Programme in the UK. The project has been commissioned by the Department of Health in England (DH) on behalf of the four UK Health Departments, and is being run by the Medical Schools Council (MSC).
- 1.2. The current version is a draft for review by the Project Board.

2. Background

- 2.1. This section summarises the background to the project to allow a general reader to appreciate the context of the PID. It starts with an explanation of the existing arrangements for the recruitment to the Foundation Programme of final year undergraduate medical students in the UK, and then summarises the case for change that has led to the project.

Existing Arrangements

- 2.2. In the UK the first two years of training for newly graduated doctors are known as the 'Foundation Programme'¹. The training is organised and delivered by 'Foundation Schools', who liaise with the NHS organisations in their area to ensure the availability of suitable training posts. Assuring the quality of F1 posts and confirming the successful completion of the F1 year is however the responsibility of the universities that have Medical Schools under the Medical Act 1983 and its subsequent amendments.
- 2.3. Since 2006 a UK-wide process has been followed to recruit medical students into posts within the Foundation Programme. The annual process, which is known as the Foundation Programme (FP) recruitment round, starts in the autumn of one year, when prospective candidates submit an online application form; the results are announced in the spring of the next year, and the applicants take up their posts in late summer. For most applicants, the process runs in parallel with their final year of undergraduate education.
- 2.4. Entry into the Foundation Programme is competitive. Each applicant is awarded a score built of two components. One component is determined by the quality of the applicant's answers to the so-called 'white space' questions on the online application form. The other component is an 'academic quartile' score given to the applicant by their Medical School, based on their educational performance relative to all of the other students in their year at the school.
- 2.5. The scores are used to determine which applicants should be selected into the Foundation Programme, and to allocate successful applicants to Foundation Schools. If there are more applicants than posts, then the highest scoring applicants are selected². If there are more posts than applicants, then all eligible applicants are considered for allocation.
- 2.6. The allocation of applicants to posts takes place in two stages. In the first stage, each applicant is allocated to a specific Foundation School; in the second, the applicant is allocated to one of the posts associated with that school. The allocation of applicants to Foundation Schools is determined by a set of rules (the allocation algorithm), which take into account the applicants' preferences and scores in a particular way.
- 2.7. The United Kingdom Foundation Programme Office (UKFPO) administers the recruitment rounds under contract to the four UK Health Departments. There is a separate contract with the Department of Health in England (DH) for the provision of the online system, known as ApplicationMMC, through which the applications, scoring and allocations are managed. The ApplicationMMC contract comes to an end in June 2011, and may be extended for one year only to June 2012.
- 2.8. A 'Recruitment Rules Group', consisting of stakeholder representatives, maintains the rules and standards for the process. This includes setting the timetable for each round, defining the 'white space' questions and associated scoring guidelines, quality assuring the scoring, and adjudicating over special cases. The scoring of answers to the 'white space' questions is organised by the Foundation Schools, and is performed by trained representatives- mainly clinicians- from the NHS.

¹ See www.foundationprogramme.nhs.uk

² This is a simplification. Under the prevailing right-to-work legislation, applicants who have an unconditional right to work in the UK will be selected for places first, in order of score. Any unfilled places will then be allocated to the remaining applicants (i.e. those without an unconditional right to work in the UK), again in order of score.

The Case for Change

- 2.9. Surveys conducted after each recruitment round have shown a growing level of satisfaction among applicants. However, the current process has some actual and potential shortcomings of a serious nature, as follows:
- The marking of the answers to the 'white space' questions is labour intensive, the cost approaching £2m per year.
 - Applicants are allowed to complete the 'white space' questions online at their convenience. While this is logistically simple, it means that there is no real safeguard that the answers are the genuine work of the applicant.
 - Model answers to the 'white space' questions have been offered for sale over the Internet. There have been concerns that the impact of such model answers is likely to increase with time, as it will become increasingly difficult to invent entirely fresh questions for each successive recruitment round.
 - There have been questions about the extent to which the academic quartile scores can be equated across Medical Schools, particularly since each school derives its scores as it sees fit rather than following a standard approach.
 - While the 'white space' questions appear to offer a practical way to rank large numbers of comparable applicants, and are derived from the Person Specification, their technical validity is open to question.
- 2.10. To date the impact of these concerns has been softened considerably by the fact that there have been fewer applicants than posts (meaning that all applicants can get a job) and around 90% of applicants are allocated to their first preference Foundation School. From this perspective the process has generally been seen as positive by those involved. However, the perspective is somewhat misleading. The percentage of applicants who get their first choice is a product of the algorithm that is used and the relative demand for places at schools- it is not dependent on the selection or scoring method. Furthermore, there is no guarantee that the number of posts will continue to exceed the number of applicants. The effect of European employment law is that the Foundation Programme jobs must be offered on an equal footing to applicants from the European Union; and in any event applicants from elsewhere can apply subject to the prevailing 'right to work' arrangements. In the event that the number of applicants did exceed the number of posts, then it would be more important that the process for selection did indeed pick the 'best' applicants from the pool available, and was not open to legal challenge.
- 2.11. The Next Stage Review *A High Quality Workforce* recognised that the current arrangements were not sufficiently robust, and recommended new work to develop a more reliable and valid selection process. As a result, the DH commissioned the Medical Schools Council (MSC) to set up the Foundation Programme Steering Group (FPSG) to conduct an options appraisal of selection into Foundation training, with the aim of recommending a reliable, valid, feasible and sustainable method for selection, which minimises the risk of successful legal challenge. The ensuing work was a collaborative venture between all the major stakeholders, and included:
- Surveys and consultations
 - Meetings with stakeholders.
 - Three independent literature reviews,
 - A report from a panel of international experts in assessment and selection.
 - An independent cost-benefit analysis.

2.12. Having considered the evidence, the FPSG submitted its report to DH in September 2009, in which it made recommendations about the future of the FP recruitment process. The relevant findings of the report may be summarised as follows:

- While there is much that is right about the current selection process, the 'white space' questions and academic quartiles are not sufficiently robust, and need to be replaced.
- The selection of applicants should take into account their wider professional attributes as well as their clinical skills and knowledge.
- Simple interviewing would not be a robust basis for selection, as the available evidence shows that interviewing is neither a reliable nor a valid selection technique. So called 'Multi Mini Interviews' (MMIs) in which each applicant undertakes several structured interviews in succession, would be more reliable and more valid, but they would be labour intensive and expensive to implement.
- An invigilated, machine-markable test of the applicant's professional attributes (a so called 'Situational Judgement Test' or SJT) would be the best replacement for the 'white space' questions, as it should be highly secure, reliable, and standardised. It also had the advantage of being relatively inexpensive, and of being used successfully for the shortlisting of entrants into specialty training for General Practice.
- The two main options for gauging an applicant's clinical skills and knowledge were to hold a special selection test (the 'national exam for ranking'), or to consider the evidence available from the applicant's performance at Medical School. The former had the advantage of being easy to standardise; however, it was very expensive, and was seen by key stakeholders as undermining educational objectives. It would also not be possible to hold such an exam before Finals (because of variations in the timing of the curricula across Medical Schools), which would be too late for the allocation of posts. The FPSG therefore recommended that the considerable amount of information available from Medical Schools about the performance of the applicant- the so called 'Educational Performance Measure' or EPM- should be used for the purpose of selection. The EPM would be similar in many ways to the academic quartile score, but would be produced in a more robust and standardised way, and possibly be more granular.
- Work should begin to develop and pilot SJT and EPM, with a view to implementation after 2012 when the contract for the current application service comes to an end. Aside from testing the individual effectiveness of SJT and EPM, the pilots would provide data that could be used to determine the relative weightings to be given to SJT scores and EPMs when they are combined to give an overall score to applicants for the purpose of selection.
- An initial step in the work should be a formal job analysis for the Foundation Programme in order that selection tools might be based on more detailed information about the role.

The Way Ahead

2.13. A feasibility study proposed that the recommendations should be implemented through two projects running in series. The first- the current project- would develop, refine and validate the specifications for SJT and EPM through pilots that would take place in parallel with the 2010/11 recruitment round. Subject to ministerial approval of the outcome, a second project would then be commissioned to implement SJT and EPM in a new recruitment process to run from the 2012/13 round onwards.

2.14. DH has given conditional approval for the first of these projects, which is the subject of this PID.

3. Objectives

- 3.1. The principal objective of the project is to design, develop, and pilot the new assessment techniques to ensure that they are suitable to replace the existing selection arrangements for the 2012/13 recruitment round and beyond. This is to include:
- A detailed job analysis of the role of the FP doctor to identify the detailed criteria that should be used for the assessment of FP applicants.
 - The development of specifications for the SJT and EPM.
 - The production of SJT items.
 - The development or purchase of software to store SJT items securely.
 - The piloting of SJTs and EPMs.
 - Collecting data about costs, resource requirements, timings, etc, for planning a live implementation of the new selection arrangements.
 - Developing expertise to underpin the success of the live implementation.
 - Spreading awareness of the proposed changes to the selection process, and maintaining the support and cooperation of stakeholders.
 - Obtaining legal opinion to confirm that the assessment methods are suitable in all legal respects.
- 3.2. The implementation of these changes is subject to confirmation of the associated business case, which will be subject to review at the end of the project, and the associated DH approvals process.

4. Blueprint for the New Selection Process

- 4.1. This section presents a description of the intended arrangements for FP recruitment for placements from 2013 onwards.
- 4.2. This description does not cover recruitment to academic and military posts, which is subject to special arrangements.

Overview

- 4.3. There will be an annual UK-wide process (the FP round) for selection into the Foundation Programme.
- 4.4. Rounds will be managed by UKFPO on behalf of the Health Departments of the UK countries.
- 4.5. Rounds will run to a UK-wide timetable.
- 4.6. Rounds will be facilitated by an online system administered by UKFPO.
- 4.7. There will be a standard person specification for Foundation Programme posts.
- 4.8. There will be a preliminary 'pre-enrolment' stage for a round during which prospective applicants declare their interest and are subjected to eligibility checks.
- 4.9. One of the eligibility conditions is that the applicant's Medical School must agree to provide for the applicant a verifiable score- the 'Educational Performance Measure' (EPM) - conforming with published standards.
- 4.10. The EPMs of all eligible applicants will be entered onto the online system following the pre-enrolment stage where each applicant's eligibility status is checked.
- 4.11. Eligible applicants will be invited to complete a single application form via the online system.
- 4.12. In their application, the applicant will specify their preferences for Foundation Schools. To support this, information about all of the Foundations Schools and their associated training programmes will be available to the applicant via the online system.
- 4.13. All applicants will be invited to take an invigilated selection test- the Situational Judgement Test (SJT)- designed to evaluate their professional attributes relevant to the person specification.
- 4.14. The SJT will be defined and maintained by Medical Schools Council Assessment Alliance (MSC-AA). The test items for SJTs will be held in a secure item bank managed by MSC-AA.
- 4.15. The SJT will be machine-marked or computer based- in either case it will not need to be marked manually.
- 4.16. UK Medical Schools will be responsible for SJTs for their students. UKFPO will be responsible for the tests for other applicants. The tests will be held in the UK on a set number of dates, expected to be fewer than 10, selected to avoid electives.
- 4.17. The tests will result in an SJT score being awarded to each applicant. These scores will be loaded onto the online application system.
- 4.18. The online system will calculate for each eligible applicant an overall score which will be a weighted sum of their EPM and SJT score. The rules for the weightings will be the responsibility of UKFPO.

- 4.19. Applicants will be selected and allocated to Foundation Schools by the online system according to a set of rules that will take into account the number of places available, the overall scores, applicant preferences, and the right-to-work status of applicants.
- 4.20. Once applicants are allocated to Foundation Schools, each School will be responsible for implementing local arrangements for matching their allocated applicants to individual training posts.
- 4.21. Once applicants are matched to individual posts, their details will be passed to the relevant employer, who will be responsible for pre-employment checks and employment contracts.
- 4.22. The SJT scores and answers will be retained centrally for quality assurance purposes.

Organisation and Responsibilities

- 4.23. This section provides a summary of the responsibilities of the organisations that participate in the delivery of the FP rounds, namely:
 - UKFPO and its associated Foundation Programme Rules Group (FRG)
 - Medical Schools Council
 - Medical Schools
 - Foundation Schools
 - Employers.
- 4.24. The responsibilities of these organisations will be as described below.
- 4.25. UKFPO and the FRG will be responsible for:
 - Defining the timetable, eligibility criteria, person specification, and the rules and process for selection and allocation to Foundation School.
 - Developing the application form.
 - Obtaining final legal sign-off.
 - Providing guidance about the round for applicants, Foundation Schools and Medical Schools.
 - Checking the eligibility of applicants from Medical Schools outside of the UK.
 - Providing and managing the online application system.
 - Managing the SJT for applicants from Medical Schools outside of the UK.
 - QA of EPMs from Medical Schools.
 - Managing the round overall.
 - Handling queries from stakeholders.
 - Reporting outcomes to the Health Departments.

4.26. Medical Schools Council will be responsible for:

- Developing and maintaining the standards for the production of EPMs.
- Providing standardised guidance to ensure consistency across Medical Schools.
- Designing, creating and maintaining the SJT bank*.
- Making SJTs available to Medical Schools and to UKFPO*.
- Obtaining agreement for common dates on which to hold SJTs.

*this will be through the MSC-AA.

4.27. UK Medical Schools will be responsible for:

- Identifying their eligible applicants.
- Preparing EPMs for each of their applicants.
- Managing the delivery of SJTs for their applicants.
- Providing all relevant applicant information; including EPMs and SJT scores to the Foundation Schools for entering onto the online system. [The project will need to consider how this might work in practice; one option recently considered is for the SJT answer sheets to be centrally scanned by MSC-AA.]

4.28. Overseas Medical Schools will be responsible for providing verifiable EPMs for their applicants.

4.29. Foundation Schools will be responsible for:

- Publishing information about their school, and its programmes and posts, on the online system to help applicants decide their preferences.
- Matching their allocated applicants to individual programmes and posts, and ensuring the results of the matching are entered on the online system.
- Liaising with employers to handover applicant details.

4.30. Employers will be responsible for completing pre-employment checks.

Online Application

4.31. There will be an agreed process through which all eligible applicants are assigned a user name on the online system. UK Medical Schools will be responsible for collating details of their applicants for loading onto the online system. The UKFPO eligibility office will perform this function for other applicants.

4.32. The timetable will include an opening date and a closing date for the submission of applications.

4.33. Eligible applicants will be responsible for registering on the online system and completing and submitting an online application form before the closing date. Late applications will not be accepted.

- 4.34. The application form is likely to be broadly similar in scope to the current FP application form, except that it will exclude the 'white space' questions.

EPM

- 4.35. The EPM will be a score provided for the applicant by their Medical School according to an agreed standard framework. The framework will be defined and maintained by the Medical Schools Council (MSC) working with its constituent Medical Schools and in consultation with medical student representatives.
- 4.36. As well as providing the EPMs for applicants, each participating Medical School must provide a formal declaration confirming that they have complied in full with the standard framework. Declarations are to be received by the UKFPO before the opening date for applications.
- 4.37. The treatment of applicants that are unable to provide an acceptable EPM needs to be determined within the project. At present, where an applicant is unable to provide an acceptable quartile score, the applicant is not deemed ineligible but is assumed to have the lowest possible quartile score. [This issue is more likely to arise in relation to overseas applicants.]

SJT Definition and Set-up

- 4.38. The SJT will be a paper-based machine-markable test sat under invigilated conditions. (An option that could be investigated during the project is for the tests to be delivered online.)
- 4.39. The items for a given test will be drawn from a UK-wide bank of items created, quality assured and administered by the MSC-AA.
- 4.40. The test items will be produced and quality assured by trained volunteers to standards defined and maintained by MSC-AA.
- 4.41. MSC-AA will be responsible for ensuring the quality and availability of test items.

SJT Delivery

- 4.42. UK Medical Schools will be responsible for scheduling and managing the delivery of the test for their students according to national guidance. The guidance, which will be produced by MSC-AA, will specify:
- The instructions for compiling tests from the item bank.
 - The time that should be allowed for the test.
 - The notice that should be given to applicants invited to the test.
 - The steps that should be taken to verify the identities of participating students.
 - The invigilation arrangements, and standards for test facilities.
 - The guidance and rubric to be given to applicants.
 - The process for submitting results for processing on the online system.
 - The arrangements for notifying applicants of the results.
 - Recommendations for dealing with applicants who fail to attend, interruptions (eg fire alarms), appeals, etc.
 - Security arrangements and retention periods to apply to test papers and answer sheets.

- Handling of special needs applicants.
- 4.43. SJTs will be run on nationally agreed dates, chosen so that at least one of the dates will not clash with the timing of the electives at any given Medical School. The questions for each sitting will be drawn at random from the question bank.
- 4.44. UKFPO will be responsible for arranging and delivering SJTs for applicants not from UK Medical Schools. These SJTs will be held in the UK.
- 4.45. There will be a UK-wide deadline by which all SJT scores must be loaded on the online system. It is currently assumed that Medical Schools will send copies of all answer sheets to MSC-AA, and that MSC-AA will be responsible for scanning the sheets and loading the results. However, other options will be considered during the project.

Selection

- 4.46. Once all SJT scores have been entered onto the online system, the existing selection rules will be applied to determine which eligible applicants are to be allocated a place on the Foundation Programme:
- 4.46.1. Eligible applicants with an unconditional right to work in the UK will be selected ahead of any others.
 - 4.46.2. Subject to the rule above, higher scoring applicants will be selected ahead of lower scoring applicants.
 - 4.46.3. Where a number of equally scoring applicants are competing for the last available places, the selection from among those applicants will be random. [To illustrate this, suppose there are 7,000 places and 7,050 applicants. Suppose that 6,990 applicants have scores of more than a certain value X, say. All of those applicants have been selected, leaving 10 places available for the remaining lower-scoring applicants. If 15 of these applicants all have the same score X, and 10 of them need to be picked to fill the remaining places, then the 10 will be picked at random.]
- 4.47. These rules will be implemented by the online system, which will identify those applicants who have been selected and those who have not.

Allocation to Foundation School

- 4.48. Each selected applicant will be allocated to a Foundation School according to a predefined set of rules (the allocation algorithm) which will take into account the scores and preferences of applicants, and the number of training places available at each school.
- 4.49. The allocation algorithm will have been defined, agreed with stakeholders, and tested as part of the usual work of UKFPO.
- 4.50. The allocation will be performed by the online system, following which each Foundation School will have online access to the application forms of its allocated applicants.
- 4.51. [In practice, allocation and selection are likely to be combined as a single process of the online system, but they are presented separately here for clarity.]

Matching to Posts

- 4.52. Once applicants have been allocated to a Foundation School, the Schools will be responsible for matching their allocated applicants to individual posts, and for communicating the outcome to applicants and employers.

- 4.53. Foundation Schools may implement their own local processes for matching; however, the results of the matching process must be entered onto the online system by the deadline specified in the timetable.
- 4.54. The arrangements for providing employers with access to the application forms will continue as is - namely, the Foundation Schools will be responsible for downloading the application details and transferring them securely to the relevant employing organisation.

Pre-Employment Checks

- 4.55. Employers will be responsible for liaising with applicants to complete pre-employment checks, making offers of employment and finalising employment contracts.

Reserve List Allocation

- 4.56. It is possible that issues will arise after matching that prevent individual vacancies being filled (eg applicants withdraw, or fail their finals or pre-employment checks). In such cases the UKFPO would expect to run two small allocation rounds through which any remaining applicants would be allocated to the unfilled posts using the scores and rules from the main round. If there are still unfilled posts after this, then employers will be responsible for recruiting to them directly without UKFPO involvement.

5. Scope and Exclusions

5.1. The detailed scope of the project is considered below under the following headings:

- Job Analysis
- Person Specification
- SJT
- EPM
- Modelling Weightings of SJT and EPM scores
- Algorithm
- CCA
- Online Application System
- Trials and Pilots
- Communications
- Evaluations, Legal Checks and Approvals.

Job Analysis

5.2. The project includes the commissioning and delivery of a structured analysis of the job of the Foundation doctor. The analysis will include a literature search, and interviews with employers, Foundation doctors, Foundation School Directors, educational supervisors and patient representatives. The outcome of the analysis will be justified recommendations about whether and how the Foundation Programme person specification should be updated, and corresponding recommendations for the factors to be taken into account by the design of SJT items.

Person Specification

5.3. The project may make recommendations for the update of the person specification in line with the findings of the job analysis. However the responsibility for implementing any changes to the person specification lies outside the project (with UKFPO).

SJT

5.4. The project includes:

- The design, development and verification of a suitable framework for the delivery of SJT for the purpose of selection. This includes: analysis of the requirements; the development of a test specification and sample tests, the development or purchase of software to permit item banking, distribution and analysis; the verification of the specification through piloting; training and guidance for test developers; scoring rules; procedures and facilities for storing, and distributing test items, and for delivering, marking, calibrating, and quality assuring tests including psychometric performance.
- Associated guidance for applicants for use in the pilots.
- The commissioning of expertise and services as required to support the above.

5.5. The project excludes the production of the live SJT and scoring rules for the 2012/13 round; this will be covered by a subsequent project to be run by UKFPO, based on the outcome of the pilots.

EPM

5.6. The project includes:

- The definition and verification of requirements and standards for the production of EPM for use in recruitment to the Foundation Programme.
- Liaison with all UK Medical Schools to ensure that the standards are workable and implemented.
- Liaison with some Medical Schools outside the UK to inform the development of the standards.
- The production of guidance for UK and non-UK Medical Schools.

5.7. The project excludes:

- The production of guidance for Medical Schools in languages other than English.
- Liaison with Medical Schools to ensure the availability of the EPMs for applicants for the 2012/13 round- this will be managed by UKFPO outside of the programme as part of their operational management of the round.

Models of Score Weightings

5.8. The project includes work to determine the respective weightings to be given to SJT scores and EPMs for the purpose of selection.

Algorithm

5.9. The project excludes any changes to the allocation algorithm.

Common Content of Assessment (CCA)

5.10. The project excludes the work associated with the extension of CCA to all Medical Schools, and its transfer from UMAP to MSC-AA; this will be separately managed by MSC.

5.11. The project does include the work required to ensure that the MSC-AA item bank (or an equivalent) can be used to create, store and distribute SJT test items and to support the analysis of item performance.

Online Application System

5.12. The project excludes the procurement of an online application service to support recruitment rounds from 2012/13; that will be the responsibility of a subsequent project run by UKFPO.

Trials and Pilots

5.13. The project includes:

- The piloting and assessment of the use of SJT and EPMs.

5.14. The project excludes:

- The running of live recruitment rounds, which will remain the responsibility of UKFPO.

Communications

5.15. The programme includes:

- Scheduled and ad-hoc reporting to the appropriate governance bodies (eg the four UK health departments).

- Consultation with stakeholders to define and confirm requirements.
 - Liaison with participating Foundation Schools, Medical Schools, doctors and medical students to arrange, run and evaluate trials and pilots.
 - Publication of information about the project, its objectives, its deliverables, and its progress, for stakeholders.
- 5.16. The project excludes liaison with individuals, employers, Foundation Schools, Medical Schools or applicants in connection with the running of live recruitment rounds, which will remain the responsibility of UKFPO.

Evaluations, Legal Checks and Approvals

- 5.17. The project includes:
- The production of evaluation reports.
 - The production of a business case for implementation.
 - Obtaining legal advice to confirm that the detailed arrangements are legally robust.
- 5.18. The production of any associated impact assessments will remain the responsibility of DH staff.

6. Constraints

- 6.1. The approach to the project is subject to the following constraints.

Timescales

- 6.2. The piloting of any changes must be scheduled to fit in with the timetables for the annual recruitment rounds, which are set each year by UKFPO.
- 6.3. The current contract for the ApplicationMMC service cannot be extended beyond the 2011/12 recruitment round; allowing time for approvals and reprocurement, this means that the Assessments project must deliver its final report by July 2011.
- 6.4. The timing of SJT tests must be scheduled to suit the pattern of electives across Medical Schools.
- 6.5. At least eight weeks notice will need to be given for project activities that require the time of clinical staff.

Legal Compliance

- 6.6. Any changes to the recruitment process must be in accordance with relevant UK and EU employment and right-to-work legislation, and with UK public law standards of reasonableness and fairness.

Standards

- 6.7. Any personal information collected through the programme must be held and processed in accordance with the Data Protection Act and guidance provided by the guidance provided by the Information Commissioner's Office.

Funding and Resources

- 6.8. Some of the development and quality assurance work will be performed by volunteers, with their time being provided to the programme free of charge by their employer. As a result the programme will have limited access to their time, and will not be able to call upon it at short notice. Typically, it can be expected that such working groups will be able to meet no more than once a month.

Approvals

- 6.9. The work is subject to the DH approvals processes.

7. Approach

7.1. This section defines the approach to project. It begins with an overview, then considers the main elements of work individually in the following order:

- Job Analysis
- SJT
- EPM
- Score Modelling

Overview

7.2. DH has commissioned MSC to implement the project. MSC in turn will commission some elements of the work from third parties.

7.3. A core team within MSC will be responsible for:

- All project deliverables, bar those to be outsourced (see below).
- Obtaining the necessary approvals for the project.
- Defining and maintaining the project plan.
- Managing risks, issues and progress.
- Reporting to the Project Board.
- Commissioning and managing sub-contracts.
- Managing project-level communications and relationships with stakeholders.
- Ensuring a smooth handover to UKFPO of the outputs of the project.

7.4. The following elements of work will be commissioned externally by the core team:

- The Foundation Programme job analysis.
- The development of the SJT, and the associated item bank arrangements.
- Advice about the relative weightings of SJTs and EPMS.
- Legal opinions and other specialised QA activities.

Job Analysis

7.5. The purpose of the job analysis is to identify:

- The professional attributes that should be considered in ranking applicants for the purpose of selection and allocation.
- The indicators of the attributes, and how these are to be tested.

- 7.6. The job analysis will be commissioned from Cambridge University. The core team will draw-up the requirements for the analysis for QA by representatives of the Project Group.
- 7.7. The job analysis is expected to include a literature search, and structured interviews with employers, clinical supervisors and Foundation Programme doctors. The output of the work is to be a report which will provide the information needed to inform the design of the SJT.

SJT

- 7.8. It is proposed that the arrangements for creating and managing SJTs should build upon the substantial work that has already been completed by the Universities Medical Assessment Partnership (UMAP) in building infrastructure and processes for the development and management of assessment tests. UMAP has a network of trained writers who draft assessment items at specially facilitated workshops, and a thorough quality assurance process for refining the drafts for incorporation in a national bank, from which test items are distributed to Medical Schools for use in finals. At present 16 Medical Schools participate in this scheme, but it has been agreed that all Medical Schools will do so in future, and the scheme will become owned by the Medical Schools Council Assessment Alliance (MSC-AA). It is proposed that this scheme should be used for the operational production of SJT, subject to the following developments:
- New training and guidance for item writers specifically for the SJT items, based on a specification for SJT.
 - An increase in the number of trained writers to provide the additional capacity required. These item writers will have experience of working with FP doctors, and are likely to include, for example, Clinical Tutors and Medical School Directors of Education.
 - Additional workshop, specially blueprinted to suit the SJT requirements.
 - Developing QA arrangements to cover SJT.
 - Changes to the software of the item bank to manage the SJT items separately from those for other assessments, and to allow test answers to be uploaded for analysis.
 - New guidelines for the administration of SJT tests by Medical Schools.
- 7.9. The MSC will be responsible for managing the introduction of these changes, but will commission specialist expertise (expected to include Cambridge University and MSC-AA³) to deliver the following outputs and activities:
- The SJT test specification, including the test content, the criteria for setting/modifying test content, the item types and response formats, test length, scoring conventions, standards for rubric.
 - The guidance and training materials for item writers.
 - The QA of individual test items.
 - Facilitation of the item writing workshops.
 - The design of pilots and trials of the effectiveness of the SJT, and the associated analysis, reporting, and any subsequent revisions of the specification and guidance (arranging and managing the trials will remain the responsibility of MSC).

³ From here in this document MSC-AA will be used to refer to MSC-AA or UMAP

- 7.10. The project team will draw-up the requirements for the specialist work, and negotiate contracts.
- 7.11. The sequence of the work will be as described below.
- 7.12. The test specification will be produced in draft, and subject to QA review.
- 7.13. An initial set of test items will then be produced by volunteers at workshops facilitated by the specialists. The workshops will be arranged by MSC. In preparation for the workshops, the specialists will produce draft guidance materials. The outputs from the workshop will include: comments from the participants about the quality of the guidance; draft test items; productivity metrics that can be used to plan operational item production.
- 7.14. The draft items will then be subject to a rigorous quality assurance process designed by the specialists. If necessary, the guidance materials will be updated to reflect any systematic issues found at this stage.
- 7.15. Further item-writing workshops will be held as required to produce sufficient items for an initial mini-pilot, the purpose of which is to assess the characteristics of the test are broadly suitable before investing effort in developing further items for a large-scale evaluation. In particular the mini pilot will be designed to assess:
- The time taken to complete the test.
 - Whether the tests provide a good spread of results.
 - Whether the rubric is sufficiently clear, and the format can be readily understood.
 - Whether the test items are sufficiently unambiguous and appear to have face validity and fairness from the candidate perspective.
- 7.16. The mini pilot will be organised by MSC with the support of one or more Medical Schools, the final-year students from which will act as the candidates for the test. This pilot will also be extended to include applicants undertaking the clinical assessment as part of eligibility checks, thus including a number of non-UK applicants. [The detailed planning of the pilots would also consider the option of involving F1 doctors- this might be particularly relevant if there is a specific requirement for participants not from UK Medical Schools.] The venues for the test will be provided by the participating Medical Schools.
- 7.17. Subject to successful completion of the mini-pilot, further item writing workshops will be held as required to produce the materials for the main pilot. This will be on a larger scale, involving at least 4-5 Medical Schools. The purpose of the main pilot is to provide sufficient data to underpin a rigorous psychometric assessment of the SJT, specifically to confirm:
- The reliability of the test.
 - The validity of the test in relation to the requirements of the Person Specification and the other findings of the job analysis.
 - The fairness of the test in relation to the relevant legal requirements (eg the test should not systematically favour candidates by gender or ethnicity).
- 7.18. The pilot will also be used to test the administration procedures and guidance.
- 7.19. One output of the main pilot will be a detailed assessment report describing the manner in which the SJT provides a robust basis for selection to the Foundation Programme. This must provide sufficient evidence to convince stakeholders and must be completed by July 2011. The other main output will be a set of recommendations for changes to the banked test items, the test specification, or the guidelines.

EPM

- 7.20. The purpose of the EPM is to provide a standard framework by which Medical Schools can provide reliable scores for the clinical knowledge and skills of their students. The key outputs of the EPM work are:
- A definition of the standards to be satisfied by EPMs.
 - Guidance for Medical Schools on how to implement the standards.
 - The processes and templates to be used by Medical Schools to pass EPMs to the online application system.
 - A standard declaration to be signed by heads of Medical Schools to conform that the EPMs are valid and have been produced in accordance with the standards.
 - Guidance for applicants.
- 7.21. The development and introduction of the EPM standards will be managed by MSC, and will build upon the information gathered through a questionnaire sent to all UK Medical Schools to query the availability of quantitative information about the educational performance of applicants.
- 7.22. The responses to the questionnaire will be collated and analysed to identify a common subset of quantitative information that might be provided by all schools and be relevant to selection to the Foundation Programme. Based on the outcome of the analysis, MSC will produce an options paper for review, which will describe and assess the options that are consistent with the common subset. These might include, for example:
- The indicators from which EPMs should be derived- eg whether prizes should be taken into account.
 - The use of a single score or multiple measures.
 - The type of scores (eg quartiles, Z-scores, etc).
- 7.23. The draft will be issued to Medical Schools for comment, and legal advice sought.
- 7.24. In parallel with the review, the MSC will begin to draft the processes, templates, and associated guidance for Medical Schools. These will be subject to initial QA by members of the Project Group and issued in draft to the Medical Schools for comment.
- 7.25. The feedback from the review, and the comments received from the Medical Schools, will be collated by the MSC core team who will liaise with the Medical Schools to agree the resolution of issues associated with the comments, and for updating the draft products accordingly. Where the issues cannot be resolved they will be escalated first to the Chair of MSC, and then if necessary to the Project Board.
- 7.26. The new standards, templates and procedures are to be ready for initial trialling in the 2010/11 recruitment round, when they are to be used by Medical Schools to produce EPMs in parallel with the academic quartiles.
- 7.27. The MSC will obtain written agreement from all UK Medical Schools to the final version of the standard to be used from 2012/13 onwards.

Score Modelling

- 7.28. The purpose of the modelling is to determine how SJT scores and EPMS should be combined for the purpose of selection and allocation.
- 7.29. The modelling will be conducted by an appropriate specialist commissioned by UKFPO. The main inputs to the work are expected to be the Foundation Programme person specification, the SJT and EPM design criteria, and scores and outcomes from the earlier trials and previous live rounds. The output of the work will be a report showing the impact of different options for combining the EPM and SJT scores, and a recommended option. The paper will be subject to QA by members of the Project Group, and approval by the Process Project Board.

8. Products

8.1. This section lists the high-level deliverables from the project, broken down under the following headings:

- Project Management and Approvals
- Job Analysis and SJT
- EPM
- Scoring Rules
- Communications

8.2. More detailed product breakdowns will be produced on a stage by stage basis.

Product Category	Product Name	Notes
Project Management and Approvals		
	PID	
	Project Filing System	
	Risk Register	
	Action and Issues Log	
	Quality Log and product Checklist	
	Product Descriptions	Formal product descriptions will be produced for key products only.
	Communications Plan	Comms products relating to specific categories are itemised separately below
	Stage Plans	
	End Stage Reports	
	Highlight Reports	
	Checkpoint Reports	
	Project Board Meetings and Minutes	
	Project Closure Report	
	Business Case	A high-level business case will be produced for review at the end of each project stage. A more detailed business case will need to be produced to support the case for the procurement of a replacement online system.
	Evaluation Report	The report could be commissioned from an independent organisation or produced as a project assurance activity by user representatives.
	Legal Approvals	
SJT and Job Analysis		
	Requirements	To define the services and deliverables expected from Cambridge.
	Contract with Cambridge	
	Detailed plan for the work	This will be produced by the core team with input from the supplier.
	Literature study	
	Job analysis interviews and questionnaires	

Product Category	Product Name	Notes
	Job analysis report and recommendations	
	Recruited volunteers for item writing	Need to get a group of question setters.
	Draft specification of SJT	What skills are covered? How many questions? Marking scheme? Format of the answers, etc. Pass criteria. Realistic scenarios, unambiguous, relevant to NHS/F1, takes account of E&D agenda, test attitudes to E&D. What are the delivery arrangements- on paper or electronic? Number of questions, number of possible answers, duration of test. Is the test flat or an 'IRT' tree structure? Need to take account of differences in the structure of curricula across the Medical Schools
	Item writing workshops	
	Quality Assurance process	
	Specification for item bank	To define how the MSC-AA item bank will be used to store and distribute SJT
	Item banking arrangements	May need interim arrangements for the pilots pending the set-up of MSC-AA
	Training material for item writers	
	Initial pilot test	
	Arrangements for initial pilot	Who is involved, when will it happen, what are the deliverables, who is doing what, etc
	Initial evaluation report and recommendations	
	Updated materials and specifications	Assume the workshop materials and SJT specifications may need to be updated in the light of the learning from the pilot.
	Large pilot test questions	
	Large pilot evaluation report and recommendations	
	Sample tests and explanatory info for applicants	
	Live questions in bank for 2012/13	
	Common administration procedures	Security and safe storage, security escalation arrangements, how far apart the desks, Illness DNAs, special circs, spoiled papers, uploading scores, retention policy, appeals, fire alarms etc. Equipment needed, ID checks, invigilation standards, data prep (how do we ID students uniquely- nb MS student numbers unique if school prefix added, but overseas students might be an issue??) NB the procedures for uploading SJT results onto the online application system will be developed by the processes project. However, there needs to be an equivalent process for uploading results to the item bank to support future analysis. These two processes could be combined.

Product Category	Product Name	Notes
	Analysis and reporting requirements	A definition of the steps that will be taken to analyse the outcomes of the SJT.
	Analysis and report	The product of the requirements above.
EPM		
	Requirements	A formal statement of the requirements that must be satisfied by EPMs in order for them to provide a robust basis for selection
	Options paper	If there is more than one option
	Specification	A formal definition of the EPM standard.
	EPM template	The template for the delivery of EPMs. Assume this is a spreadsheet which contains EPMs linked to some ID of the applicant.
	Declaration	Form of words to be signed by head of Medical School to confirm that EPMs are valid.
	Delivery process	A written description of the process to be followed for providing EPMs for piloting. NB needs to cater for UK and overseas applicants.
	Guidance for Medical Schools	Explanatory information to ensure that Med Schools understand the requirements and processes for generating EPMs.
	Information for Applicants	Applicants will need to know how the EPMs are formed, who is responsible for providing them, etc
	Medical Schools workshop	Medical Schools will be invited to a workshop at which the EPM standards, requirements and processes will be explained. NB may need more than one workshop across the UK.
Scoring Rules		Modelling the EPM/SJT weightings
	Assessment Criteria	A written definition of the criteria that will be used to determine the optimum way to combine SJT and EPM scores.
	Analysis and recommendations	A report showing the effect of modelling different options, comparing these with the assessment criteria, and recommending the preferred option.
	Published rules	The formal definition of the overall score.
	Rationale	An explanation of the rationale for the scoring rules, which can be published to pre-empt queries and criticisms.
Communications		
	See Section 20	

9. Assumptions

- 9.1. This section lists the assumptions upon which the plan for the project is based.

Timescales

- 9.2. The new arrangements for selection and recruitment must all be introduced in time for the 2012/13 round. Allowing time for approvals and for the procurement and implementation of a new online system, the current project must be completed by the end of June 2011.

Test delivery

- 9.3. Applicants from UK Medical Schools will attend SJT tests arranged by their Medical School.
- 9.4. The SJT will be taken under invigilated conditions in the UK.
- 9.5. The UKFPO will be responsible for arranging the SJT for overseas applicants.
- 9.6. There will be a small number of nationally agreed dates, selected to avoid clashes with electives, upon which Medical Schools may schedule SJTs. The test items used in a given SJT sitting will be drawn randomly from a large bank of test items, to minimise the risk that questions leaked from the earlier sitting could give advantages to applicants taking the later sitting.
- 9.7. The duration of the SJT will be of the order of 2-3 hours.
- 9.8. The SJT items will ultimately be stored in the item bank provided by MSC-AA. Until the MSC-AA bank is available for storing SJT items, a temporary store will be built or purchased by MSC.
- 9.9. SJT items will be produced by trained volunteers, whose time will be provided gratis by their employer.

UK-Wide Approach

- 9.10. The new selection arrangements will continue to be provided on a UK-wide basis, with no requirement to support local variations over and above the degree of flexibility provided in the current arrangements.

Test Purpose

- 9.11. The purpose of the SJT is to support only selection and allocation: the tests are not required to provide diagnostic outputs to identify the development needs of applicants.

ApplicationMMC

- 9.12. There will not be a requirement for applicants, Medical Schools or Foundation Schools to use ApplicationMCC for the trialling of SJT and EPM.
- 9.13. There may be a requirement for ad-hoc data downloads from ApplicationMMC to support analysis. Such downloads will be obtained on a 'business as usual' basis by UKFPO.

The Selection Process

- 9.14. The selection process will be as described in Section 4.

Board Costs

- 9.15. The members of the Project Board and Project Group (excluding core team members) will contribute their time free of charge to the project. The project will pay for reasonable travelling expenses.

EPM

- 9.16. All UK Medical Schools will agree to provide the EPMs to the required standard and at no cost to the project.

Procurement

- 9.17. The procurement of the specialist services to develop the SJT will be conducted according to the MSCs normal practice for procurement, and will not be subject to EU procurement rules.

Scope

- 9.18. The scope is as defined in Section 5.

10. Dependencies and Interfaces

External Dependencies

- 10.1. The table below lists the activities or deliverables outside the direct control of the programme, upon which the success of the programme depends. It also identifies the escalation route to be used where issues arise in relation to the dependencies.

Owner	Dependency	Escalation
Medical Schools	Agreeing and implementing the EPM arrangements.	MSC Chair
Medical Schools	Participation in pilots.	MSC Chair
Medical Schools	SJT administration.	Head of School
Methods Consulting Ltd	Cooperation in providing data (eg white space scores) from ApplicationMMC to be used in the analysis of SJT and EPM.	DH Contract Manager
DH	Funding, and timely sign-off of stages etc	SRO
DH	Obtaining approvals from other UK countries.	SRO
NHS employers and clinicians	Participation in SJT item writing at no cost to the project.	SRO
BMA and other professional organisations	Agreement to changes to the national recruitment process if recommended by results of pilot	SRO and MEE
Students and FP doctors	Participation in pilots	Heads of Medical Schools and Foundation Schools

11. Success Criteria and Evaluation

Success Criteria

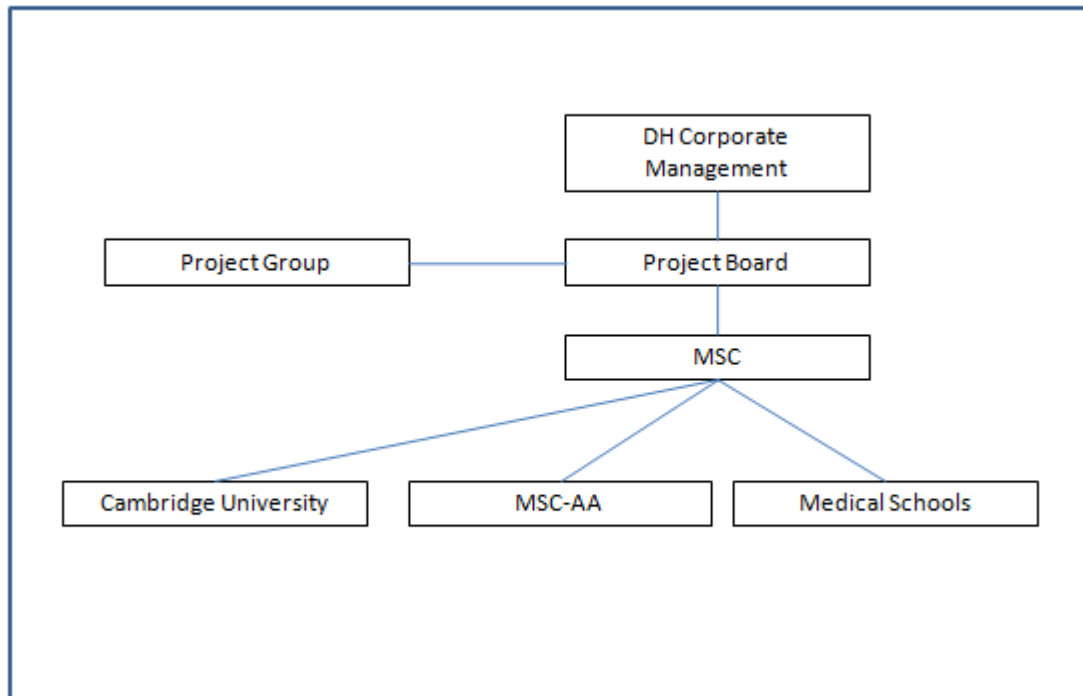
- 11.1. The project success criteria are as follows:
- 11.2. **Completeness.** All of the project's products have been completed and accepted.
- 11.3. **Legality.** Legal advice has confirmed that the proposed approach is robust.
- 11.4. **Finances.**
 - 11.4.1. The projected costs of implementing the new approach are affordable, and less than the costs of the current process.
 - 11.4.2. The actual costs of the project are within the approved budget.
- 11.5. **Reliability and Validity.** The tests have confirmed that SJT and EPM provide a reliable and valid selection method.
- 11.6. **Practicality.** The project has confirmed that: it is possible to schedule and run the SJT for the expected numbers of applicants, taking into account the need to avoid electives; Medical Schools are able to provide EPMs to the required standard, and there is a satisfactory means of dealing with applicants that cannot provide EPMs.
- 11.7. **Buy-in.** The key stakeholders have confirmed their support for the approach, including:
 - All Health Departments
 - All Foundation Schools
 - All Medical Schools
 - BMA Medical Students Committee
 - UKFPO
 - COPMED
 - NHS Employers
 - General Medical Council
 - Medical Royal Colleges
 - Medical Education England
 - General Medical Council.
- 11.8. **Robustness.** Piloting has shown that it is possible for Medical Schools to implement the SJT securely, taking into account the need to keep questions secret, to ensure the identity of the applicants taking the tests, and to mark the scripts correctly.
- 11.9. **Timeliness.** The key products have been delivered in good time for the start of the 2012/13 recruitment round.
- 11.10. **Sustainability.** It is possible to develop a sufficiently large number of SJT items to support selection over many years.

Evaluation

- 11.11. By July 2011 the proposed approach needs to be piloted and evaluated to give a firm basis for starting work on the requirements and approvals to procure a replacement online system. There will need to be a formal evaluation, starting around June 2011, to provide the necessary confidence in the way forward, specifically considering the following subset of the evaluation criteria:
- 11.12. The evaluation must confirm that the combination of SJT and EPM provides a valid and reliable method that can be practically deployed in a robust way for selection purposes,
- 11.13. The evaluation will include a legal review of the following:
- The proposed selection process.
 - The nature or the SJTs and the arrangements for delivering them.
 - The standards and process for the provision of EPMs.
- 11.14. The Net Present Value of the proposed approach will be re-assessed and compared with the 'Do Nothing' option to ensure that the economic case for the project is still sound.
- 11.15. The evaluation must confirm that the programme is still on-track to deliver in good time for the start of the 2012/13 round.

12. Organisational Structure

12.1. The organisational structure of the project is illustrated in the diagram below.



12.2. The Project Board will be chaired by Patricia Hamilton, the Director of Medical Education in DH, who is the Senior Responsible Officer (SRO) accountable for the success of the work. The other members of the Project Board are listed below:

Wendy Russell	DH Policy Lead	
Lindsey Proctor	DH Project Manager	
Matthew Langham	DH Analytical Support	
Derek Gallen	Chair UKFPO	(Senior User)
Christine Outram	MEE Managing Director	(Senior User)
Paul O'Neill	Chair Project Group	(Senior Supplier)
Katie Petty-Saphon	MSC Executive Director	(Senior Supplier)

12.3. Project assurance and stakeholder representation will be provided by the Project Group, chaired by Professor Paul O'Neill, which includes representatives of the following organisations:

- The four health departments
- Medical Schools
- Foundation Schools
- NHS Employers
- British Medical Association
- General Medical Council

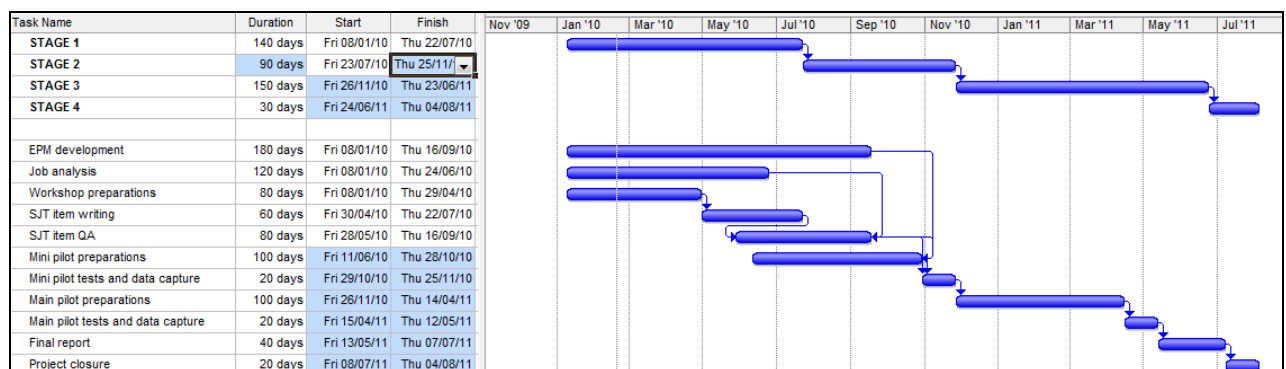
- The UKFPO
- 12.4. A core team within the Medical Schools Council (MSC) will be responsible for the performance of the work, and for managing the input from Medical Schools and the two main sub-contractors Cambridge University and MSC-AA.

13. Activity Plan

- 13.1. This section presents a high-level activity plan for the project, beginning with an assessment of the critical path.

The Critical Path and End Date

- 13.2. The end date for the project is determined by the need to validate the proposed selection process to a sufficient degree to allow legal advice to be sought and approval to be given to procure, develop and test a new online application system in time for the 2012/13 round (note that the contract for the current system can only be extended until June 2012). Given this, the final evaluation report from the project needs to be submitted around July 2011.
- 13.3. The critical path is mainly determined by the SJT development, and can be identified by working back from the final report as follows:
- 13.4. The final report cannot be completed until the main pilot has been run and the outcomes analysed.
- 13.5. The main pilot cannot be run unless the SJT is ready to be piloted.
- 13.6. The SJT needs trialled on a small scale basis to ensure it is ready for piloting.
- 13.7. The small scale trial cannot be run until an initial set of SJT items has been drafted and quality assured.
- 13.8. The SJT items cannot be drafted until the item-writers have been trained.
- 13.9. The item writers cannot be trained until they have been recruited, the training sessions scheduled, and the training materials produced.
- 13.10. The overall timeline is illustrated below.



Staging

- 13.11. In line with good practice the project will be run in stages. Each stage, bar the final one, will include the production of a detailed plan for the next stage in line, so that more detailed plans will be signed-off on a stage by stage basis as each project progresses.
- 13.12. Each stage will include fortnightly checkpoint meetings and bi-monthly project board meetings.

13.13. To ensure continuity and minimise unnecessary delays, project work will continue in parallel with end-stage approvals, rather than being suspended while approvals are sought. For the same reason, minor tasks associated with one stage may be scheduled to start near the end of the preceding stage, in effect anticipating approval. However, no major commitments will be made to continue work beyond the end of a stage. In particular, the contracts for externally commissioned work should contain appropriate breakpoints to allow for the possible premature closure of the project at each stage boundary.

13.14. The stages will be as follows:

- Stage 1- Set-up, Analysis and Design. January 2010 to July 2010.
- Stage 2- Initial Pilot. August 2010- December 2010.
- Stage 3- Main Pilot and Report January 2011 to July 2011.
- Stage 4- Closure July 2011 to August 2011.

13.15. These stages are described in turn below.

Stage 1- Set-up, Analysis and Design. January 2010 to July 2010.

13.16. In this stage the project will be set-up, the Foundation Programme job analysis will be completed, an initial version of the SJT specification will be prepared, and draft EPM standards and guidance will be developed.

13.17. **Set-up.** The key project controls- PID, risk register, meeting schedules, action log, communications plan, etc- will be set-up by the project team.

13.18. **Job analysis.** The requirements for the work will be drafted by MSC and subject to review by representatives of the Project Group. Cambridge will be responsible for planning and managing the work, which will be subject to fortnightly reviews with the core team. Four months have been allowed for the performance of the work and the production of a report, with a subsequent month for the report to be considered and accepted by the Project Group and Project Board.

13.19. **SJT.** Again the MSC will be responsible for the development of requirements for review by representatives of the Project Group. Five months have then been allowed for the initial design work and the production of a first set of SJT items for initial pilots. The SJT work can start in parallel with the job analysis, but some time must be allowed after the job analysis has been completed for further changes and refinements to the SJT design.

13.20. **EPM.** The responses to a questionnaire issued to Medical Schools will be used to determine what information they hold that could be used for constructing an EPM. The responses to the questionnaire will be analysed to scope the possibilities for the design of an EPM, and an options paper will be published for consultation. In parallel with that consultation, work will begin on drafting the guidance and processes that will be followed by Med Schools when they produce EPMs. It is assumed that most of the products will be developed in an iterative way, with successive drafts being subject to review (rather as the UKFPO Rules Group manages its work presently).

Stage 2- Initial Pilot. August 2010 to December 2010.

13.21. In this stage: some Medical Schools will produce EPMs in parallel with quartiles, as the basis for assessing the validity of the standard and associated guidance; the SJT will be trialled on a small scale at one or two Medical Schools.

Stage 3- Main Pilot and Report. January 2011 to July 2011.

- 13.22. In this stage: the SJT will be piloted on a larger scale, with around 10 Medical Schools; legal advice will be sort in relation to the detailed arrangements for SJT and EPM; the evaluation report will be produced.

Stage 4- Closure.

July 2011 to August 2011.

- 13.23. In this stage the project will be closed down, and work handed over to the UKFPO.

14. Resources and Funding

Budget and Funding Summary

14.1. The table below summarises the expected costs for the project, broken down by financial year:

Project Costs (£ex VAT)

Category	2009/10	2010/11	2011/12	Totals
SJT specialists and job analysis	£100,000	£325,000	£150,000	£575,000
MSC-AA manpower and software development	£10,000	£400,000	£200,000	£610,000
Working groups T&S	£8,550	£49,200	£14,250	£72,000
Med School pilot expenses	£0	£232,500	£0	£232,500
Legal Advice	£0	£20,000	£30,000	£50,000
Independent Review	£0	£20,000	£20,000	£40,000
Communications	£0	£75,000	£25,000	£100,000
Project team	£131,200	£393,600	£164,000	£688,800
Contingency	£100,000	£0	£0	£100,000
Project totals	£349,750	£1,515,300	£603,250	£2,468,300

14.2. An indicative breakdown by Stage is as follows:

Project Costs (£ex VAT)

Category	Stage 1	Stage 2	Stage 3	Stage 4
SJT specialists and job analysis	£250,000	£150,000	£150,000	£25,000
MSC-AA manpower and software development	£135,000	£300,000	£150,000	£25,000
Working groups T&S	£32,100	£17,100	£19,950	£2,850
Med School pilot expenses	£0	£22,500	£210,000	£0
Legal Advice	£0	£10,000	£40,000	£0
Independent Review	£0	£10,000	£30,000	£0
Communications	£25,000	£25,000	£25,000	£25,000
Project team	£300,000	£150,000	£206,000	£32,800
Contingency	£100,000	£0	£0	£0
Project totals	£842,100	£684,600	£830,950	£110,650

14.3. The items in the summary tables are explained in more detail below.

SJT Specialists and MSC-AA Manpower and Software Development

- 14.4. These headings cover the cost of the specialist resource responsible for devising, producing and validating the SJT. The costs are based on a quotation provided by Cambridge University, and indicative figures provided by UMAP. Specifically, the costs are intended to cover:
- The completion of the job analysis.
 - The initial definition of the SJT specification.
 - The training of item writers.
 - Facilitation of workshops for item writing.
 - Devising and managing the process for the QA of SJT items.
 - The production of test items.
 - Any necessary technical modifications to the item bank to support SJT.
 - The development of the pilot strategy, and of the guidance for pilot participants.
 - Provide the item banking service for the pilots.
 - The evaluation of the pilots, and subsequent refinements of the test definition.
- 14.5. There is considerable uncertainty around the cost and timing of the work required to ensure that the item bank is suitable for use for SJT. The figures provided should therefore be considered as a prudent allowance- they are not based on detailed bottom-up estimates.

Legal Advice

- 14.6. The costs for legal advice are those to be incurred by MSC in relation to verifying the legality of the proposed changes to the selection process, once these have been defined in more detail. The costs have been estimated by extrapolating the actual costs incurred for earlier advice relating to the recommendations of the FPSG.

T&S

- 14.7. The travel and subsistence costs have been estimated from the assumptions shown below.
- The Project Board will meet every other month.
 - The Project Group will meet every month.
 - The checkpoint meetings will not be face-to-face.
 - The project will pay T&S for volunteers attending item writing workshops, there being a total of 100 attendances.
 - An allowance of 10 attendances a month has been made for ad-hoc meetings (eg formal reviews for quality assurance).
 - The individual T&S costs will vary between £300, say for travel from Manchester to London, to £0 for locally-based participants, with the average being £150 per participant.
- 14.8. The effect of these assumptions is shown in the table below:

T&S Costs	2009/10	2010/11	2011/12
Project board attendances per month	4	4	4
Project group attendances per month	10	10	10
QA meeting attendances per month	5	5	5
Active months in the year	3	12	5
Total attendances for routine meetings	57	228	95
SJT workshop attendees		100	
Total attendees	57	328	95
Average cost per attendee	£150	£150	£150
Total cost	£8,550	£49,200	£14,250

Medical Schools

- 14.9. The costs payable to Medical Schools for their participation in the pilots have been estimated by comparison with the actual costs incurred by Foundation Schools in administering the marking of white space questions. It is assumed that:
- 14.10. Each of the 31 Medical Schools will hold one substantial SJT test as part of the piloting.
- 14.11. The effort required will amount to 10 days of administrative time and 5 days of management time per SJT test per Medical School.
- 14.12. The cost of the venue for the test will average £3,000 per Medical School.

Communications

- 14.13. The costs for communications are based on detailed estimates presented separately in the project communications plan.

MSC Core Team

- 14.14. Estimates for the resource requirements for the MSC core project team have been produced on a 'bottom-up basis' from estimates for individual products listed in Section 8.

15. Project Controls

- 15.1. The controls to be applied for the management of the project will be a sub-set of PRINCE 2 as defined below.

Business Case

- 15.2. A high level business case will be produced by the project team for approval by the Project Board. The business case will be revisited at the end of every stage to ensure that there is a valid basis for proceeding to the next stage.

Stages

- 15.3. The project will be run in stages, with each stage subject to Project Board approval to proceed.

Products

- 15.4. The planning and controls will be product-based. There will be a master product list, maintained by the project team, showing the status of all planned and delivered projects. Product descriptions will be produced for the most important deliverables, as agreed by the Project Board.

Quality Management

- 15.5. The quality management approach will be as defined in Section 16.

Project plans

- 15.6. There will be a high-level plan identifying the timeline for the project overall, and for the stages.
- 15.7. There will be separate detailed plans for each of the stages within the project.
- 15.8. The plans will include formal definition of tolerances within which the project team has delegated authority- any planned or expected changes that breach the tolerances will be the subject of exception reports requiring Project Board approval.

Communications Plan

- 15.9. The project team will produce a detailed communications plan for approval by the Project Board. The plan will be reviewed at least at the end of each stage.

Project reporting

- 15.10. Project highlight reports will be produced monthly, to an agreed standard format, by the project manager and distributed to members of the Project Board for information.
- 15.11. Exception reporting will be used to raise and manage events that cannot be managed within the tolerances of the Project.
- 15.12. End stage reports will be produced at the end of each project stage for Project Board approval.
- 15.13. An End of Project report will be submitted to the members of the Project Board during the final month of the project. The report will cover the achievement of the project's objectives, performance against planned target time and cost, and recommendations for dealing with any outstanding issues.

Meetings

- 15.14. **Project Boards.** The Project Board (which will contain business, user, and supplier representatives) will meet at least quarterly to assess the progress of the project and provide

any decisions or direction required by the project manager. Board meetings will also consider any exception reports.

- 15.15. **Checkpoint meetings.** The core project team will meet fortnightly to review progress against plan, agree forward actions, deal with risks and issues, and prepare reports for the Project Boards. The checkpoint meetings might be held by conference call to reduce travel costs.

Actions and Issues

- 15.16. The project manager will maintain a joint log covering all issues and actions agreed between the core team and the workstream leads (note it is not intended that the log should include actions or issues that arise within the externally commissioned workstreams).

Risk Management

- 15.17. There will be a risk register for the project.
- 15.18. The risk register will be updated monthly and reviewed at the project board meetings.
- 15.19. Each risk will have a business owner at project board level accountable for ensuring that the proposed mitigation/management actions are appropriate and progressed.
- 15.20. The key risks are considered in Section 19.

Change Control

- 15.21. Requests for change relating to externally commissioned work will be managed as follows:
- 15.22. Once a new requirement becomes apparent, at the request of the project manager the relevant workstream leads will produce recommendations for how to deal with the new requirement (where this is not already clear), with an assessment of the likely impact on costs, risks and timescales. The project manager will determine, through reference to the Project Board as required, whether the change should be implemented, and if required will ask the workstream lead to produce a contract change note (CCN) for approval by the Project Board.
- 15.23. The CCN will include:
- A brief description of the change.
 - Any additional charges.
 - The detailed impact of the change, in terms of requirements, deliverables, timescales, documentation, staffing and dependencies.
- 15.24. The CCN will become effective when signed by the project manager (following Project Board approval) and the workstream lead. Where a significant amount of work is required to produce recommendations and assess the impact of a change, then by agreement this will be charged through a separate CCN.

Lessons Learned

- 15.25. The End Stage Reports will consider lessons learned from each stage, and make recommendations as required for building upon them in subsequent stages.

16. Quality

- 16.1. This section defines the responsibilities for quality and quality assurance, the project-level quality expectations, and the management methods to be followed.

Responsibilities

- 16.2. The Project Board is responsible for quality. The Project Board will delegate quality assurance for certain products to specialists or user representatives who will be independent of the individuals or organisations responsible for the delivery of those products.

Overarching Quality Expectations

- 16.3. The changes to be made to the selection process must:
- Be robust, transparent, valid, reliable, fair and effective.
 - Be properly trialled, so that their effectiveness can be judged on evidence acquired through practical use.
 - Comply with all applicable legislation and standards.
 - Be subject to independent evaluation and scrutiny, including a review of legal compliance, and of the data security arrangements.
 - Be sufficiently scaleable to cope with likely fluctuations in numbers of applicants and posts.
 - Be capable of supporting recruitment rounds from 2012/13 until at least 2016/17.
 - Satisfy the expectations established in the cost benefit analysis that formed part of the options appraisal undertaken by MSC on behalf of DH.

Quality Methods

- 16.4. The applicable quality methods will be specified on a product by product basis within the relevant stage plans. Typically these methods will be some combination of the following options:
- Informal review
 - Formal review
 - Scripted acceptance testing
 - Trialling
 - Legal check.
- 16.5. These methods are described in turn below.
- 16.6. **Informal Review.** In this method, the draft product will be distributed to reviewers, typically by email, for them to review individually against the relevant quality criteria. The product owner will collate and respond to the comments from the individual reviewers, liaising with the reviewers as necessary to clarify concerns and agree their resolution. The response to the comments will be in writing. Once the reviewers have agreed to the written responses, a further version of the product will be produced with the necessary changes incorporated. Where agreement cannot be reached, or the concerns raised appear to conflict with the quality criteria, a project issue will

be raised. The review comments and responses will be filed. This will be the default method for reviewing documents.

- 16.7. **Formal Review.** This will be as per the informal review except the reviewers will go through their comments and concerns together in a chaired meeting at which the resolution of issues will be agreed. This method will be used for the review of documents in cases where the issues to be addressed are likely to be complex or the subject of conflicting views.
- 16.8. **Scripted Testing.** This method will be used to quality assure software and systems for user acceptance purposes (i.e. in addition to earlier testing undertaken by the producer). Test scripts will be produced to cover the expected behaviour of the product. Testers following the scripts will compare the actual characteristics of the product with the expected characteristics and record any variances, which will be graded by severity according to an agreed specification. The testing will be subject to a test plan which will define:
- The timing and location of the tests
 - The selection and briefing of testers
 - The arrangements for supporting the testers
 - The requirements for test datasets and environments
 - The arrangements for recording, and reviewing test results, for correcting defects and re-testing, and for reporting outcomes for final sign-off.
- 16.9. **Trialling.** This method will be used for quality assuring systems, processes, and the tests to be undertaken by applicants. As with the scripted acceptance testing, volunteers will conduct trials following written instructions, in accordance with a defined test plan, the outcome of the trials being recorded and compared with the expected results or other quality criteria.
- 16.10. **Legal Check.** Certain end products of the programme will be subject to professional legal scrutiny to ensure compliance with law and to assess the risk of legal challenge.

17. Project Tolerances

- 17.1. All changes or developments that may lead to the inability of the project to meet the key milestones shown below must be escalated without un-necessary delay to the Chair of the Project Board.
- 17.2. The project team has delegated authority to make changes that will not:
- Increase overall costs.
 - Materially increase the overall risk of completion.
 - Delay the achievement of key milestones.
- 17.3. Any such changes should be notified to the Project Board through the highlight reports.
- 17.4. The key milestones are as follows:

Milestone	Date
Completion of job analysis	August 2010
Initial pilot	November 2010
Main pilot	April 2011
Final report	July 2011

18. Communications

- 18.1. There is a separate communications plan for the project.

19. Risks

- 19.1. The project risk register is separately documented in the highlight reports provided to the Project Board.

Improving Selection to the Foundation Programme

Appendix D

FY1 Job Analysis



FY1 Job Analysis Report

Improving Selection to the Foundation Programme

Work Psychology Group and the University of Cambridge

Professor Fiona Patterson

Vicki Archer

Dr Maire Kerrin

Victoria Carr

Louise Faulkes

Phillipa Coan

Dr David Good

Final Report

July 2010

Acknowledgements

This project was conducted by the Work Psychology Group in collaboration with the University of Cambridge on behalf of the *Improving Selection to the Foundation Programme* Project Group. The ISFP Project Group, Chaired by Professor Paul O'Neill, includes representatives from fifteen organisations: the Academy of Medical Royal Colleges (AoMRC), the British Medical Association (BMA) Medical Students Committee, the Conference of Postgraduate Medical Deans (CoPMED), the four UK Departments of Health, the General Medical Council (GMC), the Medical Schools Council, the Northern Ireland Medical and Dental Training Agency (NIMDTA), NHS Education for Scotland (NES), NHS Employers (NHSE), the Scottish Board for Academic Medicine, the UK Foundation Programme Office (UKFPO) and clinical tutors. The ISFP Project is funded by the Department of Health.

The Co-directors of the project to develop a Situational Judgement Test for ISFP are Professor Fiona Patterson and Dr David Good. The core project team consisted of Professor Fiona Patterson, Vicki Archer, Dr Maire Kerrin, Victoria Carr, Louise Faulkes and Phillipa Coan in conjunction with the Medical Schools Council project team including Professor Paul O'Neill, Professor Tony Weetman, Dr Katie Petty-Saphon, Dr Denis Shaughnessy, Siobhan Fitzpatrick and Amy Stringer.

The project teams would like to thank the Cambridge University Hospitals NHS Foundation Trust, University Hospital of South Manchester NHS Foundation Trust and NHS Highland in Scotland for enabling insightful shadowing of Foundation doctors. We are especially grateful to staff at the UKFPO, AoMRC and UK medical schools for facilitating the research, and to the individual Foundation School Directors and Managers, Foundation Programme Directors, Foundation doctors, Medical School Education Deans, Consultant clinical tutors and patient representatives who shared their time and expertise. We would like to add a personal thank you to Dr Alys Burns, East of England Deanery; Dr Ian Curran, London Deanery; Lynne Hobbs, Methods; Professor Stephen Powis, Royal Free Hampstead NHS Trust; and Dr Kim Walker, NHS Education for Scotland, for their contributions to this research.

Contents

1.	Executive Summary.....	1
2	Introduction	3
2.1	Purpose of the job analysis	3
2.2	Structure of this report	3
3	Job Analysis	4
3.1	Introduction	4
3.2	Job analysis approach	4
3.3	Literature review.....	6
3.4	Semi-structured interviews.....	8
3.5	Observations (Shadowing)	11
3.6	Validation questionnaire.....	12
4	Analysis of the Data	14
4.1	Analysis strategy for the interviews.....	14
4.2	Creation of the template	15
4.3	Analysis strategy for the observations.....	15
4.4	Validation of Professional Attributes Framework.....	15
5.	Results of Job Analysis	18
5.1	Professional Attributes Framework	18
5.2	Findings from the literature review	18
5.3	Findings from the interviews	29
5.4	Findings from observations.....	32
5.5	Findings from validation questionnaire	33
5.6	Summary of findings	36
5.7	Mapping on to existing material.....	37
6.	Situational Judgement Test Design Specification.....	41
6.1	SJT literature review	41
6.2	Purpose of the test.....	42
6.3	Target attribute domains for the SJT	43
6.4	Context.....	47
6.5	Item selection methodology	47
6.6	Response formats	48
6.7	Test length.....	49
6.8	Test structure	49
6.9	Scoring key	49

FY1 Job Analysis Report In confidence to Medical Schools Council

6.10 Multiple forms	50
7. Summary and Conclusions	51
8. References	52
Appendix 1: Literature Review Specification & Search Strategy	55
Appendix 2: Job Analysis Participant Briefing.....	59
Appendix 3: FY1 Doctor Job Analysis Questions	61
Appendix 4: Patient Representative Briefing.....	63
Appendix 5: Observation Observe and Record Sheet.....	65
Appendix 6: Foundation Year One Doctor Professional Attributes Questionnaire	66
Appendix 7: Foundation Programme Commencing August 2010	69
Appendix 8: FY1 Literature Review Table	71
Appendix 9: Situational Judgement Test - Literature Review Table	83
Appendix 10: Project Team Biographies.....	104

1. Executive Summary

- 1.1 The purpose of the job analysis was to define the professional attributes (as opposed to clinical knowledge and skills) that are expected in the role of a Foundation Year One (FY1) doctor. The results will be used to determine the target domains (and test specification) for the SJT which will be used alongside a measure of educational attainment (Educational Performance Measurement) for selection to the Foundation Programme.
- 1.2 A systematic, multi-method job analysis, which incorporated a literature review, semi-structured critical incident interviews and observations, was carried out. This triangulation approach was used to increase the credibility and validity of the results.
- 1.3 The literature review, which covered all existing policy documentation and relevant publications, was conducted in two parts. The first part aimed to identify the key skills and professional attributes of a FY1 doctor.
- 1.4 Part two of the literature review focussed on Situational Judgement Tests (SJTs) and their use in high stakes selection. The results help to put the purpose of the job analysis in context and aid an understanding of the methodology that will be undertaken when developing the SJT.
- 1.5 In all, 51 semi-structured interviews were carried out with Foundation doctors, clinical/educational supervisors, Directors/Deans/Heads of Medical Education at Medical Schools, Trust Medical Directors, Foundation School and Programme Directors and patient representatives. The interviews were conducted face to face or by telephone and lasted between 30-45 minutes. Interviewees were asked for examples of specific incidents from the FY1 job that were characteristic of effective or non-effective performance, and thus to identify the key skills and attributes needed to perform as a FY1 doctor.
- 1.6 Observations (shadowing) of FY1 doctors were carried out at four locations throughout the UK. 13 FY1 doctors were observed. A total of 29 hours of observations were undertaken by two researchers. The purpose of observing FY1 doctors was to gain a greater insight and understanding into the tasks and responsibilities of a FY1 doctor, as well as the professional attributes required to perform the role successfully. The researchers recorded the tasks, responsibilities, behaviours and attributes of the FY1 doctor. A record was kept of the timings of actions and behaviours to gain an understanding of the temporal nature of the role.
- 1.7 A survey was conducted to validate the outcomes of the job analysis with a larger number of individuals. The validation survey was sent to a number of different organisations (with a request for it to be cascaded) and individuals, including; the BMA Junior Doctors Committee, the UKFPO Foundation Doctors Advisory Board and the UKFPO Foundation School Directors and Managers. 230 individuals responded.

- 1.8 The interview transcripts were analysed using template analysis¹¹ to identify the key themes within the data. All 51 transcripts were coded according to an initial template. The template underwent a number of iterations, and quality checks were carried out at after half of the transcripts had been coded.
- 1.9 The observations were analysed and the behaviours classified according to the template. Where behaviours could not be associated with existing classifications in the template, new classifications were introduced.
- 1.10 Analysis of the data from the interviews and observations resulted in nine professional attributes being defined, namely: (1) Commitment to Professionalism, (2) Coping with Pressure, (3) Effective Communication, (4) Learning and Professional Development, (5) Organisation and Planning, (6) Patient Focus, (7) Problem Solving and Decision Making, (8) Self Awareness and Insight and (9) Working Effectively as Part of a Team. These were set out in a 'Professional Attribute Framework' which lists the behavioural descriptors that are representative of each attribute.
- 1.11 The contents of the Professional Attributes Framework were checked against the findings from the first part of the literature review. This confirmed that all of the professional attributes identified through the review had already been identified through the other two job analysis methods (i.e. interviews and observations), thus confirming the validity of the Framework.
- 1.12 The results from the validation survey indicated that all nine Professional Attributes were rated as important for a FY1 doctor and important to evaluate at selection.
- 1.13 The Professional Attributes Framework was mapped against existing data, including the FY1 Person Specification and the FY1 Curriculum. The mapping demonstrated that the job analysis results and the resultant Professional Attributes Framework showed good concordance with the UKFPO FY1 Person Specification, although the Professional Attributes Framework identified one attribute; Learning and Professional Development that is not present in the FY1 Person Specification. The mapping also demonstrated good concordance between the Professional Attributes Framework and the FY1 Curriculum.
- 1.14 This report outlines the design specification for the SJT. Section 6 details the SJT target attribute domains, the length of the test, response formats and the test structure.
- 1.15 In conclusion, the job analysis has resulted in a Professional Attributes Framework which identifies the professional attributes that are important for a FY1 doctor, and provides target attribute domains for the SJT. It is recommended that the detail provided within the Professional Attributes Framework is used going forward to assist in the development of the existing FY1 Person Specification.

2 Introduction

2.1 Purpose of the job analysis

- 2.1.1 This report outlines the methodology and results of a job analysis carried out to support the development of a Situational Judgement Test (SJT) for the selection of trainee doctors onto the Foundation Programme. The purpose of the job analysis was to define the professional attributes (as opposed to clinical knowledge and skills) that are required by the role of the Foundation Year One (FY1) doctor. The results are used to determine the target domains for the SJT and will be used to inform the test specification.
- 2.1.2 The outcomes of this analysis are described in behavioural terms in the form of a 'Professional Attributes Framework', which sets out the key dimensions and example descriptors of effective performance in the role.
- 2.1.3 The SJT will be used alongside a measure of educational attainment (Educational Performance Measurement, or EPM) which will measure clinical skills and knowledge. The EPM is not within the remit of this report.

2.2 Structure of this report

This report has four main sections:

- Section 3 describes the job analysis methodology and research conducted to identify the professional attributes that are judged as important by a range of stakeholders for a FY1 doctor.
- Section 4 provides an outline of the data analysis methodology employed.
- Section 5 describes the findings from the job analysis, and provides a framework of the key professional attributes (with associated behavioural descriptors) for the FY1 doctor (Professional Attributes Framework). The content validity of this framework is illustrated by showing how these attributes map onto existing policy documentation, including the Foundation Programme curriculum.
- Section 6 presents in detail the design specification for the Situational Judgement Test and provides a matrix of the target domains and example scenarios.
- A summary and conclusions are provided in Section 7. Further materials relevant to sections of the report are provided in the Appendices.

3 Job Analysis

3.1 Introduction

- 3.1.1 The cornerstone of an effective selection system is a thorough analysis of the relevant knowledge, skills, attitudes and other personal qualities associated with successful performance in a given role, as this enables accurate identification of domains to be targeted at selection.
- 3.1.2 Job analysis is a systematic process for collecting and analysing information about jobs. When properly conducted, a job analysis provides objective evidence of the skills and abilities needed for effective job performance and thus provides support for the use of selection procedures measuring those skills and abilities. As such, a comprehensive job analysis is typically regarded as best practice as a first step in designing any selection system. Research evidence also shows that conducting a job analysis is important for an organization to defend its human resource management practices against legal challenge¹.
- 3.1.3 As the name suggests, a job analysis should be analytical, meaning that the job analysis process should break down the job into its main component parts rather than describing the job as a whole. The end result of the job analysis should be a documented, structured, and thorough report which outlines the essential requirements of the job².
- 3.1.4 For purposes of inclusion, diversity, and buy-in from stakeholders, it is important to gather information from a range of individuals. By gathering a diversity of opinions and views, an accurate and representative description of the job is more likely to be reached².
- 3.1.5 There is no given rule for the number of individuals who should participate in a job analysis, however guidance is provided on the acceptable number of participants needed to obtain reliable and accurate information about a job³. For a job that has a population of around 8000 (as does the F1 role in the UK), 40 individuals should be involved in the process to obtain 80% confidence in the results. To obtain 90% confidence in the results, 100 individuals should be involved in the entire job analysis process, including interviews, observations and respondents to surveys.
- 3.1.6 Job analysis is a process and not a single methodology. There are multiple ways to gather and analyse job information. Best practice advises a multi method approach to job analysis, which gathers information from multiple sources using multiple means, as this allows identification of aspects of a job that may not be accessible through a single methodology⁴. Data collected from multiple sources also prevents potential biases from a single source⁵ and using multiple methods allows for convergence of results to make the job analysis more comprehensive⁶.

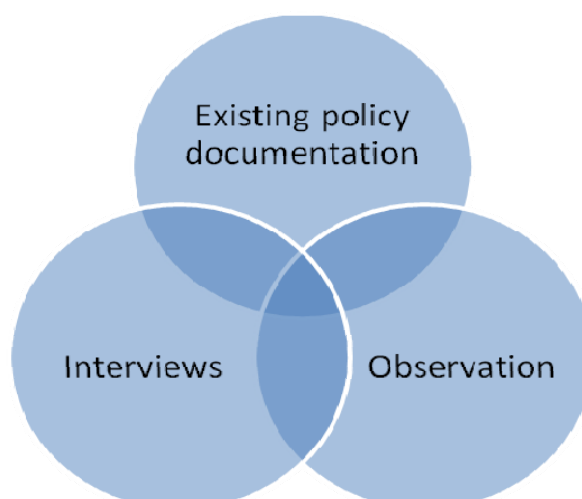
3.2 Job analysis approach

- 3.2.1 In this study, a previously validated multi-method job analysis approach was followed^{7,8}. The job analysis methods used included:

- a full literature review of existing research and relevant policy documentation
- Semi-structured critical incident interviews with individuals familiar with the role of a FY1 doctor, including Foundation School Directors and Educational Supervisors, as well as Foundation Doctors and patient representatives
- Direct observations (shadowing) of FY1 doctors in the workplace, using structured checklists
- A validation questionnaire-based survey that asked respondents (a range of stakeholders) to rate the relative importance of each of the professional attributes and the importance of assessing/evaluating each attribute at the point of selection

3.2.2 Figure 1 below depicts the multi-method approach used and demonstrates how the results from the individual methods are to be triangulated. Triangulation is a technique that refers to the application and combination of several research methodologies. The purpose of triangulation in qualitative research is to increase the credibility and validity of the results.

Figure 1: Triangulation Approach to Job Analysis



3.2.3 When deciding which method to employ as part of the multi method approach to job analysis, the project team took several considerations into account. The overriding consideration was the requirement for a thorough and in depth approach that followed best practice by gathering information from multiple sources using multiple means. Other considerations included; availability of individuals from whom to collect data, utility, the availability of information, and the need to for minimal interference.

3.2.4 When undertaking the job analysis, evaluation criteria for the outcomes of the job analysis were agreed at the outset. The evaluation criteria for this job analysis are as follows:

Face validity – this is when a model or procedure ‘looks right’ to the applicant/incumbent or others e.g. does the Professional Attributes Framework appear relevant and realistic and reflect the professional attributes required for successful performance in the FY1 role?

This was established through ensuring job analysis interviews were conducted with FY1 doctors and other relevant parties that were familiar with the professional attributes required for successful performance. The validation questionnaire also assisted in establishing face validity by asking whether the professional attributes identified were important to the FY1 role.

Content validity – this is when a model or procedure appears to cover a representative sample of the dimensions being targeted e.g. does the framework accurately reflect the professional attributes required for successful performance within the role?

This was established through conducting interviews with individuals familiar with the role, observing FY1 doctors in the workplace and the use of a validation questionnaire to ensure that the professional attributes that are required for the FY1 role were identified and agreed by a cross-section of individuals who had specific knowledge and experience of the FY1 role. The outcomes of the job analysis are then mapped onto the existing FY1 Person Specification and the FY1 curriculum to ensure that the outcomes of the job analysis relate to what is already known about the role.

Criterion-related validity – this is whether outcomes of, or scores on, a selection test or procedure predict future job performance and/or predict measures of performance that are already accessible e.g. do the outcomes of the job analysis relate to successful performance? This will be established as part of the implementation of the Situational Judgement Test (more details are provided in the SJT design section 6) and will continue to be addressed as part of the on-going process.

Political validity⁹ - selection models, procedures or decisions must account for a wide variety of perspectives *between* different stakeholder groups (beyond the candidate and employing organization) who hold potentially diverse views. It is important that a model or procedure for selection has buy-in and acceptance from all key stakeholders. Political validity was established through the use of the validation questionnaire and will continue to be addressed throughout the development and evaluation process.

Construct validity – this is establishing that the method or procedure is sampling distinct constructs (or attributes/skills/proficiencies) that reflect what they claim to sample e.g. does the Professional Attributes Framework identify discrete and independent attributes that reflect successful performance in the FY1 role?

Construct validity was established through the evaluation of the theoretical framework by occupational psychologist and researchers experienced in job analysis and competency design (see section 4.4). Further validation of the framework will be carried out following the autumn 2010 pilot of the Situational Judgement Test.

3.3 Literature review

- 3.3.1 The literature review was conducted in two parts, both of which are presented within this report.

- 3.3.2 Part one of the literature review focuses on existing policy documentation and relevant publications in the literature to assist in identifying the key skills, abilities and professional attributes of a FY1 doctor.
- 3.3.3 This aspect of the literature review forms part of the triangulation approach to the job analysis of the FY1 role and allows for the identification of specific skills, abilities and professional attributes that are documented in the literature which may be incorporated into the proposed framework.
- 3.3.4 Part two of the literature review focuses on Situational Judgement Tests and their use in high stakes selection. The results assist in placing the purpose of the job analysis in context and understanding the methodology that will be undertaken when developing the Situational Judgement Test.
- 3.3.5 The following is a brief summary of the search strategy used for the literature reviews. This was approved in advance by representative members of the ISFP Project Group. The full search strategy can be found in the appendices (Appendix 1).
- 3.3.6 A comprehensive literature search was conducted across the most relevant databases including: Medline, ERIC, CINAHL, Psycinfo, PubMed and BioMed Central. In order to identify the key skills, abilities and professional attributes of the Foundation Programme doctor the following search terms were used:

Junior doctor	AND	Skills
Preregistration doctor		Knowledge
PGY1		Abilities
Preregistration housing officer		Attitudes
PRHO		Competencies
House Officer		Capabilities
Foundation doctor		Interpersonal
Foundation year		Professional attributes
Medical student		Non cognitive attributes
Medical undergraduate		Professional competence
Intern		
F1		

- 3.3.7 The search strategies created were designed for maximum sensitivity (recall) to ensure that no papers of significance were overlooked. The searches covered the last twenty years of research (1990-present day). The search methodology involved three main stages:
- The above search strategy was conducted along with reviews of relevant reports including Good Medical Practice (2006), the New Doctor (2009), Tomorrow's Doctors (2009), The Foundation Programme Curriculum (2010), Rough Guide to the Foundation Programme (2010), and the Foundation Programme Person Specification (2010).

- ii) All documents were scanned for relevance on the basis of their title and abstract. Non-relevant papers were rejected at this stage on the basis of the following inclusion criteria:
 - Relevant to identified topic
 - Adequate methodology
 - Informed opinion which adds to the picture of the topic
 - Peer reviewed (research publications only)
 - English language
- iii) A detailed review of the relevant papers was undertaken including checking all reference lists for additional articles of relevance.

3.4 Semi-structured interviews

- 3.4.1 Job incumbent and supervisor/stakeholder semi-structured interviews are advantageous as they allow the input from a relatively large number of individuals to be included as part of the analysis and also enable a future focused/orientated approach to be employed. It is possible that job incumbents may present an exaggerated representation of their job, but they can provide insights into the job that are unavailable through other methods.
- 3.4.2 Semi-structured interviews were carried out with individuals (representing various stakeholder groups) familiar with the role of a FY1 doctor. Stakeholders were identified after discussion with the Project Board and included Foundation School Directors, Heads of Undergraduate Medical Education, Foundation Doctors, nursing staff and patient representatives. The job roles of the interviewees are described in Table 1 and the primary specialities of the interviewees' outlined in Table 2.
- 3.4.3 Invitations were cascaded through the UKFPO Foundation Doctor Advisory Board, Foundation School Directors and Managers, and the Medical Schools Council Assessment Alliance. Patient representatives were identified through AoMRC, Royal College of Anaesthetists (RCoA), and Royal College of General Practitioners (RCGP) patient/lay groups.
- 3.4.4 All interviewees were sent a briefing document (Appendix 2) which provided the interviewee with some background information about the project, and an outline of what the interview would involve. This information was important to prepare individuals for the interview and gain consent. Following all the interviews, thank-you letters were sent to all interviewees.
- 3.4.5 A total of 51 participants were interviewed (of whom 8 were patient representatives). The relatively large number of interviews conducted was to ensure that the information gained was representative and to ensure that all the relevant areas of the FY1 job role were covered. Information about the gender, age and ethnicity of the interviewees is presented in Table 3.
- 3.4.6 The semi-structured interviews lasted approximately 30 to 45 minutes and were conducted either face to face or by telephone. The interviews incorporated the Critical Incident Technique methodology¹⁰. Using this previously validated methodology⁶, interviewees were asked for examples or specific incidents from the FY1 job that were characteristic of effective

or non-effective performance, and thus to identify the key skills and attributes needed to perform as a FY1 doctor. This also provided contextual information that was used to assist with the development of some of the scenarios for the SJT (full details of interview questions can be found in Appendix 3). The interviews were audio recorded and then transcribed prior to analysis.

- 3.4.7 Interviewing a range of individuals with different job roles and different experiences/knowledge of FY1 doctors helped to ensure that the information gathered covered the breadth of the FY1 job role and enabled different perspectives to be considered. The involvement of the patient representatives was important to ensure a patient-focused perspective. However, a limitation of using patient representatives is their possible unfamiliarity with the role of the FY1 and difficulty with differentiating the FY1 role with the role of other doctors that they may have encountered. With this in mind, the patient representatives were provided with an outline of the FY1 job role and examples of where and how they may have encountered a FY1 doctor (Appendix 4). In the absence of familiarity with FY1s specifically, interviewees were asked to consider the role of a doctor in general.

Table 1: Interviewees' Roles

Job Role	N
Foundation Year One/Two Doctor	6
Clinical Tutor/Clinical/Educational Supervisor	16
Foundation Programme/School Director	10
Director/Head/Dean of Undergraduate Medical Education	10
Patient representative	8
Nursing staff	1

Table 2: Interviewees' Primary Speciality

Primary Speciality	N
Endocrinology	1
Gastroenterology	2
Ear Nose and Throat	2
Geriatrics	2
Clinical Oncology	1
Paediatrics	1
Medicine	1
Physiology	1
Anaesthetics	2
General Practice	1
General Surgery	1
Haematology	1
Obstetrics and Gynaecology	1
Mixed (FY1/2)	6
Not Stated	28

Table 3: Demographics of Job Analysis Participants

		Interviews (including patient reps)	Observations (FY1s)	Validation Questionnaire (range of stakeholders)
Gender	Male	27	3	105
	Female	24	10	121
	Not Stated	0	0	4
Age range	25 and under	2	9	16
	26-35	3	4	31
	36-45	12	0	75
	46-55	12	0	74
	56-65	7	0	31
	66 and over	0	0	0
	Not stated	15	0	3
Ethnicity	White	35	10	191
	Black	0	0	3
	Asian	3	2	23
	Mixed	0	0	5
	Chinese	0	1	2
	Not Stated	13	0	6

3.5 Observations (Shadowing)

- 3.5.1 Direct observation can provide insights that other methods cannot. However, care has to be taken that the behaviour of the individual being observed is not altered by the presence of the researcher.
- 3.5.2 The purpose of observing FY1 doctors was to gain a greater insight and understanding into the tasks and responsibilities of a FY1 doctor, as well as the professional attributes required to perform the role successfully. Additionally, a direct behavioural observation of doctors in the workplace gives a greater understanding of the temporal dynamics of the role and how the job is actually performed throughout the day. In addition, the observation study provided an opportunity to observe interactions between the FY1 doctors and other healthcare professionals (nurses, pharmacists).
- 3.5.3 Observations were carried out at four locations throughout the UK. Shadowing FY1 doctors in four different locations provided a greater understanding of the breadth of the role, and how the role may differ in different rotations and localities. For example, observations in the Western Isles provided useful information with regards to the role of a FY1 doctor in a more rural location, compared to a city hospital, such as in Manchester.
- 3.5.4 Four locations were identified by the project team as suitable for shadowing; Wythenshawe Hospital, University Hospital of South Manchester NHS Foundation Trust; Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust; Raigmore Hospital, Inverness; and Western Isle General Hospital, Stornaway.
- 3.5.5 The hospitals were asked if the researchers could sample the three core groupings of specialities; *acute*, *family/community-based* and *interventionist* specialties; recognising that family based specialties are less frequently undertaken by FY1s. This was to try and ensure that shadowing covered the breadth of the FY1 role. The majority of shadowing took place in the acute and interventionist specialties; however the family/community-based specialties were sampled in the semi-structured interviews.
- 3.5.6 In total, 13 FY1 doctors were shadowed for periods between one hour and five and a half hours, with a total of 29 hours of observation taking place. Demographic details of the doctors shadowed are outlined in Table 3. Shadowing took place during day and night shifts. Details of the specialties where the doctors worked and the hospitals in which they were posted are detailed in Table 4.
- 3.5.7 Two experienced occupational psychologist researchers, who had received the appropriate clearance to shadow the FY1s, were involved in the observations. As part of the shadowing, the researchers provided an introduction and a short briefing about the background of the project and the aims for the shadowing to the necessary individuals, including the FY1 doctor(s) to be shadowed. It was important that the purpose of the shadowing and how the information would be used was explained to ensure the doctors did not fear evaluation of them personally. It was made clear that the study was designed to evaluate the FY1 role in

general as part of a broader job analysis study. The researchers then shadowed the specified doctor, ensuring that they did not interfere with their role or care of the patients.

- 3.5.8 The researcher recorded the tasks, responsibilities, behaviours and attributes of the FY1 doctor, using the 'observe' and 'record' elements of Observe, Record, Classify and Evaluate (ORCE), a previously validated methodology⁴. The researchers observed the FY1 doctors and recorded all behaviour and interactions with other individuals. Where conversations took place, verbatim comments were recorded. A record was kept of the timings of actions and behaviours to gain an understanding of the temporal nature of the role. Appendix 5 shows the documentation used to observe and record the behaviour.

Table 4: Details of Observations

Hospital	Speciality
Wythenshawe Hospital, Manchester	Gastroenterology
	Trauma and Orthopaedic Surgery
Addensbrooke Hospital, Cambridge	HPB Surgery
	Cardiology
	Respiratory Medicine
	Infectious Diseases
Raigmore Hospital, Inverness	Trauma and Orthopaedic Surgery
	Respiratory Medicine
	General Medicine
Western Isle General Hospital, Stornoway	General Medicine
	General Surgery

3.6 Validation questionnaire

- 3.6.1 The purpose of the validation questionnaire was to validate the outcomes of the job analysis with a larger number of individuals. The aim was not to rank the professional attributes in order of importance but rather to rate their importance both in terms of importance to the role of a FY1 and importance for evaluating at selection.
- 3.6.2 The validation questionnaire - which was produced as an online survey - asked the respondent two questions; to rate the importance of each professional attribute (as identified through the job analysis) for a FY1 doctor on a scale of 1-6 (1=little importance, 6=extremely important), and to rate the importance of evaluating each professional attribute at selection, again on a scale of 1-6 (1=not at all, 6 = essential). The questionnaire is shown in Appendix 6.
- 3.6.3 An invitation to participate in the survey (accessible through a weblink) was sent out to a number of different individuals and organisations (with a request for it to be cascaded with the aim that it would be seen by a large number of people), including; BMA Junior Doctors Committee, the UKFPO Foundation Doctors Advisory Board, UKFPO Foundation School

Directors and Managers, Association of UK University Hospitals (AUKUH) Medical Directors, National Association of Clinical Tutors (NACT), London Deanery, AoMRC Patient Liaison Group, Medical Schools Council Assessment Alliance, plus all job analysis interview participants and item writing workshop¹ attendees, including those who expressed an interest but were unable to take part. The link was also posted on the Medical Schools Council website and included in the UKFPO bulletin. Individuals were given three weeks to complete the survey.

- 3.6.4 The total number of respondents to the survey was 230. Due to the way the survey was cascaded, it is not possible to calculate a response rate; however following best practice regarding acceptable number of responses to surveys⁷, the numbers are seen as sufficient. Table 3 outlines the demographic details of the individuals who responded to the survey. Table 5 outlines the job roles of those who responded to the survey.

- 3.6.5 The data was then collated and analysed.

Table 5: Job roles of survey respondents

Job Role	N
Foundation Year One/Two Doctor	46
Clinical Tutor/Clinical/Educational Supervisor	78
Foundation Programme/School Director/Manager	52
Director/Head/Dean of Undergraduate Medical Education	10
Patient representative	6
Other	38

¹ Five SJT item writing workshops were carried out between April and May 2010. Individuals familiar with the FY1 role (excluding Foundation Doctors) and those who work closely with FY1's were invited to attend. The purpose of these workshops was to generate SJT items for the SJT pilot in autumn 2010.

4 Analysis of the Data

4.1 Analysis strategy for the interviews

- 4.1.1 51 interview transcripts were available for analysis. Template analysis was used, as it is a systematic and well-structured approach to handling textual data¹¹.
- 4.1.2 Template analysis is a well-established technique in job analysis research, and allows the researcher to thematically analyse relatively large amounts of qualitative data¹². The 'template' is created as a list of codes that represent themes (or in this case behavioural descriptors which represent the professional attributes required for a FY1 to perform their role) within textual data (in this case, interview transcripts). The codes are typically organised hierarchically allowing a clear (and transparent) representation of the associations between themes. For example there are broad themes (or domains that represent professional attributes) within which subsidiary themes (or behavioural descriptors) will fall. It should be noted that codes are specified not only for themes found in most of the transcripts but also for those that appear in only a minority of transcripts in line with established template analysis methodology¹³.
- 4.1.3 In template analysis, an initial template is created and used to code the textual data. When some relevant text is found that does not fit logically with the existing codes, a change to the template is required. Where the required change in the coding structure is significant, the researcher may need to adjust the earlier coding of transcripts to fit the new version of the template. The work may require iterations of such changes to the template¹⁴.
- 4.1.4 One of the main advantages of template analysis is that it allows the reader to gain a clear overview of the themes identified in the analysis and therefore lends itself appropriate for job analysis research. It also enables the researcher to reduce large amounts of unstructured text into a structured format which is relevant and manageable for the evaluation¹⁵.
- 4.1.5 One of the disadvantages with template analysis is the time taken to review and code all data¹⁶. Another possible limitation of template analysis is the researcher's involvement in the research process, and how this may shape its outcomes. Reflexivity² is therefore required to ensure the researcher is aware of how their own assumptions about the phenomenon under investigation may influence the formulation of the model¹². Involvement of other researchers throughout the process to quality check progress is recommended¹³.

² Reflexivity refers to an awareness of the researcher's contribution to the construction of meanings throughout the research process, and an acknowledgment of the impossibility of remaining 'outside of' one's subject matter while conducting research

4.2 Creation of the template

- 4.2.1 The interviews were coded according to the following definition '*behaviours identified to explain effective FY1 performance that represent professional attributes*' (as opposed to clinical knowledge and skills).
- 4.2.2 An initial template was devised based on themes emerging from analysis of a subset of the interview transcripts. Three experienced occupational psychologist researchers independently analysed three transcripts and produced their own initial template. A one hour meeting was then held between the three researchers to agree an initial template. During the meeting, the three independently created templates were compared, contrasted and discussed until a consensus was reached as to an initial template.
- 4.2.3 The initial template was modified after five interviews had been coded, and then again after 16 interviews and a fourth version of the template was created after 28 interviews. At each modification, previous transcripts were reviewed and updates made based on the new coding.
- 4.2.4 At this stage, after 28 transcripts had been coded, the two researchers who had previously been involved in creating the initial template were provided with two transcripts and asked to code them independently using the most recent (fourth) template. This was done to provide a quality check of the analysis to ensure it was not being systematically distorted in some manner by the researchers' own preconceptions and assumptions. The similarities and differences were discussed to agree revisions to the themes.
- 4.2.5 The fifth iteration of the template was created by analysing the final 23 interviews, reviewing the first three transcripts used for creating the initial template and through successive readings of all the transcripts.

4.3 Analysis strategy for the observations

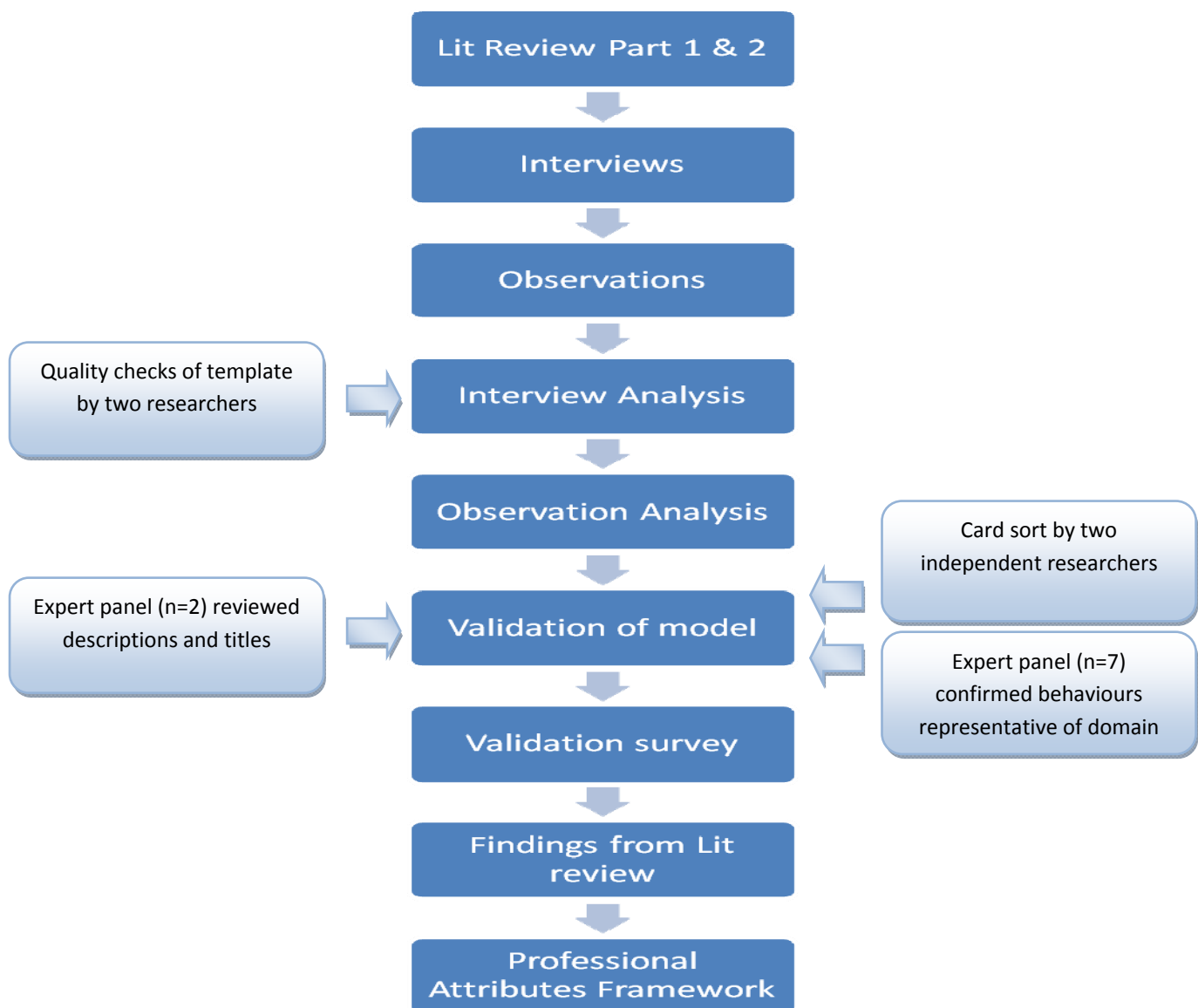
- 4.3.1 From the 29 hours of observations a total of 470 behaviours were observed. Of these 470 behaviours, analysis showed that there were 61 unique behaviours observed and the remaining behaviours were replicated.
- 4.3.2 The behaviours observed and recorded from the observations were classified against the fifth iteration of the template. It was found that the majority of the behaviours were already represented in template, thus confirming its validity. A sixth version of the template was created to take into account the few remaining behaviours that were not present in the fifth version.

4.4 Validation of Professional Attributes Framework

- 4.4.1 The sixth template grouped the behavioural descriptors into nine broad themes. To ensure that the behavioural descriptors were grouped under the correct broad theme, and were representative of that theme, further analysis was carried out. Each behavioural descriptor from the template analysis was recorded on a separate card with the aim of grouping the

descriptors into similar themes. This technique is known as card sorting, where a group of subject experts are guided to generate a category tree or groupings¹⁷. To perform the card sort, two experienced occupational psychology researchers (who had previously not been involved in the job analysis interviews/observations or subsequent analysis) worked together for two hours to group the behaviours recorded on the cards into similar themes. This resulted in eight broad themes.

- 4.4.2 The two researchers who had carried out the card sort and the lead researcher held a one hour meeting to discuss the outcomes of the card sort and finalise the broad themes and inclusion of the behavioural descriptors within these themes. The eight broad themes elicited from the card sort were compared with the nine broad themes elicited from the template analysis. The themes and the behavioural descriptors were discussed and justification for inclusion considered until consensus was reached regarding the final number of broad themes and grouping of the behavioural descriptors within each of these broad themes.
- 4.4.3 The final template, or framework, consists of nine broad themes. These broad themes are from here on in referred to as 'professional attributes'. The framework is termed the Professional Attributes Framework and provides a full description of the data and outlines the behavioural descriptors grouped under the nine professional attributes that are expected in the role of a FY1 doctor.
- 4.4.4 Each professional attribute can be defined as '*a set of specific behaviour patterns, including knowledge, skills and abilities, a person is expected to have to perform effectively as a FY1*'. Professional attribute headings were then defined and short descriptions of the attributes were developed, based on the elements of which they consisted, thus using a post-hoc approach to labelling them⁴.
- 4.4.5 In order to further validate the framework, the headings and descriptions were independently reviewed by two senior occupational psychologists who had not previously had access to the data.
- 4.4.6 An expert panel of researchers and occupational psychologists experienced in job analysis and competency design (N=7) then reviewed the content and the constructs within the framework for two hours. This process was intended to confirm the results and that the professional attributes are representative of the attribute defined, thus helping to establish the content validity (and to some extent construct validity) of the framework.
- 4.4.7 Figure 2 provides an illustration of the job analysis methodology and analysis strategy.

Figure 2: Job Analysis Methodology and Analysis Strategy

5. Results of Job Analysis

5.1 Professional Attributes Framework

- 5.1.1 The Professional Attributes Framework (Table 6) consists of nine professional attributes (that represent the professional attributes important for the role of a FY1 doctor) and provides a definition of each of the professional attributes and the behavioural descriptors that represent the professional attribute.
- 5.1.2 The nine professional attributes are: (1) Commitment to Professionalism, (2) Coping with Pressure, (3) Effective Communication, (4) Learning and Professional Development, (5) Organisation and Planning, (6) Patient Focus, (7) Problem Solving and Decision Making, (8) Self Awareness and Insight and (9) Working Effectively as Part of a Team.
- 5.1.3 Each of these professional attributes is represented by a range of behavioural descriptors. The behavioural descriptors describe the detail of the professional attribute and are therefore core to what the professional attribute represents.
- 5.1.4 Within the professional attribute Effective Communication, further categories emerged. These are:
- General communication
 - Communication with patients
 - Communication with colleagues
 - Written communication
- 5.1.5 The content validity of this framework is illustrated by showing how these attributes map onto existing policy documentation, including the Foundation Programme curriculum. This is described in more detail in section 5.7.

5.2 Findings from the literature review

- 5.2.1 The main findings from the literature review were compared with the Professional Attributes Framework to identify if there were any professional attributes that had been identified within the literature review that had not been identified during the interviews or observations. Table 7 outlines the main findings of the literature review and how the findings map onto the framework.
- 5.2.2 The review indicated that all findings from the literature review had been identified through the other two job analysis methods thus confirming the validity of the Professional Attributes Framework.
- 5.2.3 The full FY1 literature review is presented in detail in Appendix 8.
- 5.2.4 A summary of the Situational Judgement Test literature review findings can be found in section 6.

Table 6: Professional Attributes Framework and Behavioural Descriptors

Professional Attribute	Professional Attribute Definition		Behavioural Descriptors *I = behaviour from Interview O = behaviour from Observation
1.Commitment to Professionalism	<i>Displays honesty, integrity and awareness of confidentiality & ethical issues. Is trustworthy and reliable. Demonstrates commitment and enthusiasm for role. Willing to challenge unacceptable behaviour or behaviour that threatens patient safety, when appropriate. Takes responsibility for own actions.</i>		1. Is punctual (I) 2. Takes responsibility for own actions/work (I &O) 3. Owns up to mistakes (I) 4. Takes responsibility for own health and well-being (I) 5. Demonstrates commitment to and enthusiasm/motivation for role (I) 6. Understands/is aware of the responsibility of the role of being a doctor (I) 7. Is reliable (I) 8. Displays honesty towards others (colleagues and patients) (I) 9. Trustworthy (I) 10. Identifies/challenges unacceptable/unsafe behaviour/situations when appropriate(colleague/organisational issues) (I) 11. Challenges others knowledge where appropriate (I) 12. Understands/demonstrates awareness of ethical issues, including confidentiality (I)
2.Coping with Pressure	<i>Capability to work under pressure and remain resilient. Demonstrates ability to adapt to changing circumstances and manage uncertainty. Remains calm when faced with confrontation. Develops and employs appropriate coping strategies and demonstrates judgement under pressure.</i>		1. Remains calm and in control of situations (I&O) 2. Manages uncertainty/ambiguity (I) 3. Effectively deals with outcomes of mistakes/decisions (I) 4. Exhibits flexibility/adaptability when dealing with changing circumstance or issues(I&O) 5. Employs effective coping mechanisms for dealing with stress/emotions (I&O) 6. Demonstrates good judgement under pressure (I)

			7. Does not give up easily (I) 8. Deals appropriately with confrontational/difficult situations (colleagues/patients) (I&O) 9. Seeks support when faced with stress/pressure (I&O)
3.Effective Communication	<i>Actively and clearly engages patients and colleagues in equal/open dialogue. Demonstrates active listening. Communicates verbal and written information concisely and with clarity. Adapts style of communication according to individual needs and context. Able to negotiate with colleagues & patients effectively.</i>	General	1. Listens effectively (I&O) 2. Ensures surroundings are appropriate when communicating (I) 3. Understands/responds to non verbal cues (I) 4. Uses non-verbal communication effectively (I)
		With Patients	1. Uses language that is understood by patients/relatives and free from medical jargon (I&O) 2. Demonstrates sensitive use of language (I&O) 3. Communicates information to patients clearly and concisely (I&O) 4. Adjusts style of communication according to patient's/relative's needs (I&O) 5. Adjusts how much information to provide according to patient's/relative's needs (I) 6. Provides information to patients and keeps them updated (I&O) 7. Readily answers patient's and relative's questions (I&O) 8. Ensures has all the relevant information before communicating to patients/colleagues (I&O) 9. Asks questions/seeks clarification to gain more information /understanding about the patient (I&O) 10. Finds out patient's/relative's level of knowledge/understanding (I) 11. Allows patients/relatives to ask questions/uses silence effectively (I&O) 12. Checks patient's/relative's understanding (I&O) 13. Summarises information/reflects back to patients to clarify their own understanding (I&O)

		With colleagues	1. Asks questions of colleagues to gain more information (I&O) 2. Provides/summarises information accurately and concisely to colleagues (I&O) 3. Provides only relevant information to colleagues (I&O) 4. Keeps colleagues informed/updated (<i>about patients and about where they will be</i>)(I&O) 5. Is able to negotiate/use diplomacy (I&O) 6. Knows exactly what are asking for and why (I) 7. Is assertive where necessary (I) 8. Adapts style of communication according to need and situation (I) 9. Clarifies information to check their own understanding (O)
		Written	1. Displays high standards of written communication (I) 2. Uses concise and clear written communication (I) 3. Has legible handwriting (I)
4.Learning and Professional Development	<i>Demonstrates desire and enthusiasm for continued learning, takes responsibility for own development. Willing to learn from others and from experience. Is open and accepting of feedback. Demonstrates a desire and willingness to teach others.</i>		1. Demonstrates a desire and enthusiasm to learn and improve (I) 2. Is willing to learn from others (I&O) 3. Keeps up to date with knowledge (I) 4. Is willing to learn outside of work through reading around subject/access to internet (I) 5. Learns from previous experiences (I&O) 6. Learns from mistakes (I) 7. Takes responsibility for own learning (I&O) 8. Is open to and accepts of feedback (I&O) 9. Is willing to teach others (I&O)

5.Organisation and Planning	<i>Manages and plans workload effectively, displaying efficient time management and delivering tasks on time. Able to prioritise effectively and re-prioritise where appropriate. Is conscientious and maintains accurate records.</i>		1. Demonstrates effective time management (I&O) 2. Prioritises tasks effectively (I&O) 3. Manages and plans workload/tasks (I&O) 4. Checks tasks are complete/follows up tasks (I&O) 5. Maintains accurate records (I&O)
6.Patient Focus	<i>Ensures patient is the focus of care. Demonstrates understanding and appreciation of the needs of all patients, showing respect at all times. Takes time to build relationships with patients, demonstrating courtesy, empathy and compassion. Works in partnership with patients about their care.</i>		1. Able to gain trust from patients/instil confidence (I) 2. Is empathetic towards patients (I&O) 3. Is able to maintain appropriate distance from patients/relatives (I) 4. Is willing to make self available to patients (for support/provide continuity/get to know them) (I) 5. Understands/considers the differing needs of patients (I&O) 6. Recognises that patients have different backgrounds/values/beliefs (I) 7. Shows respect for patients (I&O) 8. Works jointly with the patient about their care (I) 9. Recognises the patient as the centre of care (I&O) 10. Shows a genuine interest in patients (I&O) 11. Shows compassion towards patients/relatives (I&O) 12. Is willing to spend time with relatives (I) 13. Builds rapport with patients (I&O) 14. Is polite, courteous and presents an open manner when dealing with patients (I&O) 15. Provides reassurance to patients (I&O) 16. Considers patient's safety at all times (I)

7.Problem Solving and Decision Making	<i>Demonstrates an ability to assimilate a range of information and identify key issues. Engages with the wider issues and thinks creatively to solve problems and reach appropriate decisions. Is proactive and demonstrates initiative. Is able to attend to detail.</i>		1. Makes appropriate decisions based on all the relevant information (I) 2. Demonstrates ability to use own initiative (I&O) 3. Able to assimilate and integrate information effectively (I) 4. Understands problems from a wider perspective/takes into account whole picture (I) 5. Demonstrates good attention to detail (I) 6. Is proactive; takes charge of a situation and makes things happen (I&O)
8.Self Awareness and Insight	<i>Demonstrates awareness of the boundaries of their own competence and willing to seek help when required, recognising that this is not a weakness. Exhibits appropriate level of confidence and accepts challenges to own knowledge.</i>		1. Is willing to seek help when needed (I) 2. Able to say no when faced with a situation/task they are not comfortable with/admit that they don't know (I) 3. Accepts that asking for help is not a weakness (I) 4. Recognises boundaries/limitations of own competence (I) 5. Displays appropriate level of confidence in self/own ability (I&O) 6. Accepts challenges to own knowledge (I) 7. Accepts mistakes will be made (I)
9. Working Effectively as Part of a Team	<i>Capability & willingness to work effectively in partnership with others and in multi-disciplinary teams. Demonstrates a facilitative, collaborative approach, respecting others' views. Offers support and advice, sharing tasks appropriately. Demonstrates an understanding of own and others' roles within the team and consults with others where appropriate.</i>		1. Demonstrates an understanding of others' roles within the team , including a recognition of others' knowledge/skills/abilities/roles/responsibilities (I&O) 2. Able to identify and utilise the most appropriate person for a task/situation (I&O) 3. Aware of own role/responsibilities within team (I&O) 4. Builds rapport and establishes relationships with other team members (I&O) 5. Supports others (in a non-task related way) (I) 6. Is willing to offer assistance with tasks/workload (I&O) 7. Is willing to offer/provide advice (I&O) 8. Values other people's opinions and contributions (I&O)

			9. Shows respect for colleagues (I&O) 10. Delegates and shares tasks effectively(I&O) 11. Maintains harmony within the team (I&O) 12. Identifies when others are in difficulty/are struggling (I) 13. Able to take on board team members' perspectives (I&O) 14. Able to adapt role within team where necessary i.e. take on role of leader (I) 15. Shares knowledge and expertise with colleagues (I&O) 16. Is willing to take direction from others (O) 17. Consults with others and asks for advice (I&O) 18. Makes others aware of own workload (I&O)

Table 7: Professional Attributes Mapped against Findings from FY1 Literature Review

Professional Attribute	Definition	Findings from Literature Review
Commitment to Professionalism	<i>Displays honesty, integrity and awareness of confidentiality & ethical issues. Is trustworthy and reliable. Demonstrates commitment and enthusiasm for role. Willing to challenge unacceptable behaviour or behaviour that threatens patient safety, when appropriate. Takes responsibility for own actions.</i>	• Reports³ - F1 doctors must: Be honest in their relationships with patients, colleagues and employers; Act with integrity (e.g. act without delay if they have good reason to believe they or a colleague may be putting a patient at risk; never abuse a patient's trust); Demonstrate that they are able to take appropriate action if their own health, performance or conduct or that of a colleague puts patients, colleagues or the public at risk; Know how errors can happen and how to manage risk; Accept the moral and ethical responsibilities involved in providing care (e.g. demonstrate sound knowledge of confidentiality); know the principles of equality and diversity; Be able to adapt to changing circumstances throughout their professional life. • Wright & Carrese (2001) looked at what values and attitudes physicians tried to pass on to medical trainees to encourage their professional development and integrity was a key theme encompassing trust and honesty. • Hesketh et al (2003) interviewed 57 PRHOs throughout Scotland regarding learning outcomes of the PRHO year and found a key theme surrounded developing appropriate attitudes to ethics.
Coping with Pressure	<i>Capability to work under pressure and remain resilient. Demonstrates ability to adapt to changing circumstances and manage uncertainty. Remains calm when faced with confrontation. Develops and employs appropriate coping strategies and demonstrates judgement under pressure.</i>	• Reports - F1 doctors must: Show initiative and deal effectively with pressure and/or challenge; Offer assistance in an emergency. • Schon (1983) argues that professional competence is the ability to manage ambiguous problems, tolerate uncertainty, and make decisions with limited information. • Lempp et al (2003) asked 16 Pre Registration House Officers (PRHOs) to comment on their specific experiences in the first few months as a PRHO. Many commented on the adoption of coping strategies to deal with heavy demands and clinical tasks e.g. prioritising work and asking for help when needed. • Brennan et al (2010) aimed to ascertain how 31 medical graduates, across five hospitals, found the transition from medical school to foundation year using semi-structured interviews and an audio diary approach. Key experiences included having to manage uncertainty and cope with feelings of stress and anxiety concerning their level of competence (particularly regarding prescribing).
Effective Communication	<i>Actively and clearly engages patients and colleagues in equal/open dialogue. Demonstrates active listening. Communicates verbal and written information concisely and</i>	• Reports - F1 doctors must: Be able to communicate clearly, sensitively and effectively with patients and their relatives (e.g. listen to patients, ask for and respect their views about their health; keep them informed about the progress of their care; make sure patients are informed about how information is shared within teams); Know how to develop and maintain successful relationships with patients in a range of situations (e.g. within a consultation, when breaking bad news, when dealing with complaints, when communicating with patients who have mental illness or mental and physical disabilities); be able to communicate in different ways (including spoken, written and electronic

³ Reports refer to the following documents: Good Medical Practice (2006), the New Doctor (2009), Tomorrow's Doctors (2009), The Foundation Programme Curriculum (2010), Rough Guide to the Foundation Programme (2010), and the Foundation Programme Person Specification (2011).

	<i>with clarity. Adapts style of communication according to individual needs and context. Able to negotiate with colleagues & patients effectively.</i>	methods); Be able to effectively interface with different specialties and with other professionals. • Hoffman et al (2009) - students who were rated as 'outstanding' on a house officer potential scale had excellent communication and presentation skills. • Wright & Carrese (2001) looked at what values and attitudes physicians tried to pass on to medical trainees to encourage their professional development and communication was a key theme encompassing body language, openness, sitting down when speaking with patients, silent observation, avoidance of confrontation and effective listening. • Stewart et al (1999) identified tasks of pre-registration doctors by asking a panel of experts who were clinical tutors who had at least 2 years experience of supervising PRHOs. The final tasks included those requiring communication, e.g. establishing and maintaining good working relationships with other staff; giving advice on individual patient care, communicating information between hospital and community, talking to patients and relatives and handling difficult patient interactions. • Hesketh et al (2003) interviewed 57 PRHOs throughout Scotland regarding learning outcomes of the PRHO year and found a key theme surrounded communication with patients, relatives and the healthcare team. • Katinka et al (2005) interviewed 1150 junior doctors on the key competencies required of professional practice. All doctors rated communication skills as highly important in their current pre-registration year.
Learning and Professional Development	<i>Demonstrates desire and enthusiasm for continued learning, takes responsibility for own development. Willing to learn from others and from experience. Is open and accepting of feedback. Demonstrates a desire and willingness to teach others.</i>	• Reports - F1 doctors must: Internalise the importance of CPD and self-directed learning (demonstrated through effective assessments); Keep up to date with the developments in their field; Be willing to respond constructively to the outcome of appraisal, performance review and assessment; Teach their peers and other students (if required to do so); Contribute to appraisal, assessment or review of students and other colleagues they work with; Develop the skills, attitudes and practices of a competent teacher. • Hesketh et al (2003) interviewed 57 PRHOs throughout Scotland regarding learning outcomes of the PRHO year and found a key theme surrounded personal development. This included recognising self education/personal development as a lifelong process, accepting/coping with increased responsibility, developing confidence, developing further competence from clinical experience on the job, and developing appropriate attitudes to personal health and well-being.
Organisation and Planning	<i>Manages and plans workload effectively, displaying efficient time management and delivering tasks on time. Able to prioritise effectively and re-prioritise where appropriate. Is conscientious and maintains accurate records.</i>	• Reports - F1 doctors must: Manage their own time under supervision; demonstrate good time management; maintain accurate record-keeping (e.g. medical notes, discharge letter/summaries, outpatient letters, prescriptions). • Hoffman et al (2009) students who were rated as 'outstanding' on a house officer potential scale had excellent preparation skills which included being accurate, having good attention to detail, good time management, being well organised, confident, resourceful and prepared. • Hesketh et al (2003) interviewed 57 PRHOs throughout Scotland regarding learning outcomes of the PRHO year and found a key theme surrounded personal development which included managing time well. • Bogg et al (2001) 90 PRHOs from three teaching hospitals in the Northwest of England completed a diary of work-based activities and 18 were then interviewed as a follow-up. Time management was found to encompass 43% of their day to day activities. • Stewart et al (1999) identified tasks of pre-registration doctors by asking a panel of experts who were clinical tutors

		who had at least 2 years experience supervising PRHOs. The final list of tasks included self management skills which encompassed time management, efficient use of resources, identifying priorities, for example when on call knowing who to attend to first and what to do first.
Patient Focus	Ensures patient is the focus of care. Demonstrates understanding and appreciation of the needs of all patients, showing respect at all times. Takes time to build relationships with patients, demonstrating courtesy, empathy and compassion. Works in partnership with patients about their care.	• Reports - F1 doctors must: Demonstrate that they are taking increasing responsibility for patient care (under supervision and with appropriate discussions with colleagues); Ensure effective relationships with patients are established and maintained (e.g. creating an environment where the doctor can encourage and support the patient to share all information relevant to the consultation); Introduce themselves to patients and colleagues with appropriate confidence and authority; Recognise that patients are knowledgeable about themselves and the effect their health has on their daily life (and they should be aware that relatives and carers often have to be included in the information given to patients); Respect patients regardless of their lifestyle, culture, beliefs etc; Work in partnership with patients – give patients the information they want in a way they can understand and support patients in caring for themselves. • Hoffman et al (2009) – students who were rated as ‘outstanding’ on a house officer potential scale had excellent interpersonal skills which included being committed to patient care, good at building rapport, friendly, works well with patients and families, is compassionate, caring and has a good bedside manner. • Wright & Carrese (2001) looked at what values and attitudes physicians tried to pass on to medical trainees to encourage their professional development and both respect (encompassing patient dignity, patient preferences, being non-judgemental, respect, tolerance and confidentiality) and caring (encompassing empathy, compassion, caring, concern, understanding and consideration) were key themes.
Problem Solving and Decision Making	Demonstrates an ability to assimilate a range of information and identify key issues. Engages with the wider issues and thinks creatively to solve problems and reach appropriate decisions. Is proactive and demonstrates initiative. Is able to attend to detail.	• Reports - F1 doctors must: Prioritise tasks and information effectively; Solve problems, analyse and use numerical data; Have knowledge around diagnoses and clinical decision-making. • Hoffman et al (2009) – students who were rated as ‘outstanding’ on a house officer potential scale had excellent clinical skills which included problem solving skills • Hesketh et al (2003) interviewed 57 PRHOs throughout Scotland regarding learning outcomes of the PRHO year and found a key theme surrounded appropriate decision-making skills, clinical judgement and reasoning.
Self Awareness and Insight	Demonstrates awareness of the boundaries of their own competence and willing to seek help when required, recognising that this is not a weakness. Exhibits appropriate level of confidence and accepts challenges to own knowledge.	• Reports - F1 doctors must: Recognise personal and professional limitations and be willing to ask for help when necessary; Reflect on practice and be self critical; Be committed to self-directed life-long learning. • Hoffman et al (2009) – students who were rated as ‘outstanding’ on a house officer potential scale had excellent self directed learning skills which included accepting feedback, being a good listener, having intellectual curiosity, being an avid learner, bright and a thinker. • Stewart et al (1999) identified tasks of pre-registration doctors by asking a panel of experts who were clinical tutors who had at least 2 years experience supervising PRHOs. The final list of tasks included self management skills which encompassed knowing when to contact a more senior doctor for help, clear knowledge of one’s own competencies and understanding his or her own role and responsibilities as doctor and employee. • Hesketh et al (2003) interviewed 57 PRHOs throughout Scotland regarding learning outcomes of the PRHO year

		and found a key theme surrounded personal development. This included being aware of limitations and seeking help. • Brennan et al (2010) aimed to ascertain how 31 medical graduates, across five hospitals, found the transition from medical school to foundation year using semi-structured interviews and an audio diary approach. Key experiences included knowing when to ask for help and learning from prior experiences.
Working Effectively as Part of a Team	<i>Capability & willingness to work effectively in partnership with others and in multi-disciplinary teams. Demonstrates a facilitative, collaborative approach, respecting others' views. Offers support and advice, sharing tasks appropriately. Demonstrates an understanding of own and others' roles within the team and consults with others where appropriate.</i>	• Reports - F1 doctors must: Work effectively as a member of a team (supporting others and efficient handovers); Respect everyone they work with; Share information where necessary; Listen and take account of others health professionals and agencies; Be able to demonstrate effective teamworking and leadership skills; Be willing to lead when faced with uncertainty and change. • Hoffman et al (2009) - students who were rated as 'outstanding' on a house officer potential scale had excellent teamwork skills which included being hardworking, willing to help, having initiative, independence, leadership skills, positive to work with, dependable, reliable, mature and a key team player. • Stewart et al (1999) identified tasks of pre-registration doctors by asking a panel of experts who were clinical tutors who had at least 2 years experience supervising PRHOs. The final list of tasks included self management skills which encompassed understanding the roles of others and working as part of a multidisciplinary ward team. • Hesketh et al (2003) interviewed 57 PRHOs throughout Scotland regarding learning outcomes of the PRHO year and found a key theme surrounded their role as doctor within the health service. This included working as a team, adapting to different levels of support, accepting principles of collective responsibility and adapting to different working protocols. • Katinka et al (2005) 1159 junior doctors responded to a survey on the key competencies required of professional practice. All doctors rated teamwork skills as highly important in their current pre-registration year. Brennan et al (2010) aimed to ascertain how 31 medical graduates, across five hospitals, found the transition from medical school to foundation year using semi-structured interviews and an audio diary approach. Key experiences included working with others – particularly with senior doctors and the multidisciplinary team.

5.3 Findings from the interviews

5.3.1 Example quotes from the interviews are provided to illustrate the professional attribute and the behavioural descriptors.

5.3.2 Commitment to Professionalism

“What you don’t want is someone who has said they have done something when they haven’t done it. Honesty is crucial”

“Honesty about their own health” “They need to be able to make a judgement call to see if they feel they can still see patients appropriately”

“Driven and energetic” “It’s making sure you do your job to the best of your ability”

“They will be a role model for society” “You want people who don’t just assume that they are still students and haven’t taken that step forward”

“Junior doctors often see the bad side of doctors and they are too scared to admit it...a doctor should feel empowered to report on others if needs be”

“Don’t leave it for someone else to do your work...it’s your job to do as much as you can, taking things on for yourself”

5.3.3 Coping with Pressure

“Dealing with stressful situations and dealing with personal anxiety...then dealing with the outcomes of decisions that you have made”

“Need to function in a rapidly changing environment”

“The good doctor remains calm rather than getting panicked. They are able to make sound clinical judgement”

“Dealing with confrontational areas – difficult colleagues and problems”

“Someone who does it well partially removes themselves from the situation and almost leaves them outside and doesn’t get caught up in the emotion from the pressure”

“The students who do well are not easily put off or are not easily frightened by the situations”

5.3.4 Effective Communication

"Asking open questions...reflecting back...check back all the time that people are understanding what you are saying and this may mean tailoring what you say – maybe you don't use technical terms"

"Body language is very important" "they need to pick up on social cues, body language"

"When requesting a CT scan...you need to know exactly what you want and why you want it and you need to have all the information they are going to ask you" "The capacity for persuasion and negotiation are quite important skills"

"Handovers need to be done well, so communication is crucial – important information needs to be given out about the patients."

"You need to know what the appropriate information is to transfer, making sure that it is transferred accurately and that it is received by the correct people" "express yourself clearly and concisely"

"Important that keep nurses informed...at the end of a ward round sit down with a nurse and go through everything"

"It is important that you ask questions so you can make an informed decision about what to do next"

"Being able to transfer verbal through to written and to a high standard"

5.3.5 Learning and Professional Development

"Able to...learn from mistakes"

"It's important to reflect, thinking about what they are doing so they can learn from it"

"They can learn a huge amount from nursing staff"

"They need a degree of curiosity and an acceptance that not all learning will take place at work"

"A successful doctor has a desire to go on learning...and the motivation to do that"

"They need to be responsible for their own learning"

5.3.6 Organisation and Planning

"It's vital for doctors to be able to plan"

"Need to have good record keeping; everything you say and do needs to be documented"

"Need to know what can wait or making sure that things get done"

"Being able to prioritise, especially when there are competing priorities"

"Need good time management skills"

5.3.7 Patient Focus

"Portraying there is time to talk and having a relationship with the patient"

"Being patient-centred and understanding what is more important for the patient"

"Recognising where the patient is coming from, age, emotion etc"

"A friendly and competent approach, an open manner"

"To engage with compassion"

"Important to maintain the clinical distance while being empathetic"

"Establish a relationship of trust"

5.3.8 Problem Solving and Decision Making

"To know as soon as possible exactly what's going on, what's happening...get up to speed quickly and efficiently"

"They can integrate different bits of information from different directions about the same patient and come up with a sensible diagnosis and plan"

"It's about asking the bigger questions, thinking outside the box"

"They are being proactive, getting involved with different things"

5.3.9 Self Awareness and Insight

"If the FY1 is not sure, they need to be able to say no...to admit that they are unsure"

"Having an appropriate level of being able to ask people for help and understanding that this is a professional attribute and not a weakness"

"Being able to appear confident without arrogance"

"Having the humility to accept that others will know better than you"

5.3.10 Working Effectively as Part of a Team

"You have to be inclusive, you have to treat them as an equal and value their contribution"

"How you deal with people who are struggling and how you support them"

"Work as part of a team, recognising their own role and boundaries within the team...and good at recognising boundaries of different people's roles"

"Need to appreciate the pressure nurses are under e.g. number of beds, management issues"

"They need to be able to make friends, they work together so if they are unfriendly...the work of the team suffers"

"They need to be able to share information and workload"

"There may be a point when you need to help out with someone else's job"

5.4 Findings from observations

- 5.4.1 The majority of behaviours observed during the observations were classified within Effective Communication (18) and Working Effectively as part of a Team (15). Behaviours were also observed in Organisation and Planning (5), Patient Focus (9), Learning and Professional Development (5), Coping with Pressure (5), Problem Solving and Decision Making (2), Self Awareness and Insight (1) and Commitment to Professionalism (1). Table 6 details the particular behaviours observed in more detail.
- 5.4.2 The majority of behaviours from the Professional Attributes Framework were observed during the shadowing of FY1 doctors, thus demonstrating that the attributes identified in the interviews covered the breadth of the FY1 role. Those behaviours that were not observed during the shadowing are those behaviours that are less easily identified, for example behaviours within Commitment to Professionalism (takes responsibility for own health and well being, trustworthy) and Self Awareness and Insight (acceptance that asking for help is not a weakness, accepts mistakes will be made.)

5.4.3 There was no observed difference in the professional attributes demonstrated by the FY1's between the four locations where the observations were carried out.

5.4.4 Table 8 below provides examples of the behaviours observed.

Table 8: Illustrative behaviours from FY1 observations

Professional Attribute	Example observations
Effective Communication	<i>"Introduces self to patient, asks if it is ok to take blood from them. Explains what he will be doing. Apologises to patient and to nurse that he could not get blood. Checks with nurse if there is anything else he needs to do"</i>
Working Effectively as Part of a Team	<i>"Speaks to a nurse about some blood tests requests. Asks a nurse if it is ok if she does one as he is quite busy. Nurse apologises about something that happened day before. Tells her not to worry and jokes about it".</i>
Organisation and Planning	<i>"Checks blood test on computer. Documents in patient's notes their blood results that are on a computer that should have been done by someone else the day before".</i>
Patient Focus	<i>"Goes to see patient and asks what happened. Examines patients and listens to concerns. Says she understands how difficult it is and how it must be affecting his life and that they will try and do everything they can to resolve the situation".</i>
Learning & Professional Development	<i>"After examination, asks the registrar/consultant what she missed and what she didn't do when she examined the patient".</i>
Coping with Pressure	<i>"Speaks to the registrar where he felt that they had not supported them and had gone behind their back. Asks in future that they come directly to him if they have a problem. Remained calm when discussing the situation".</i>
Self Awareness and Insight	<i>"Nurse unsure whether should give aspirin to the patient. Reassures the nurse that it is ok to give aspirin".</i>

5.5 Findings from validation questionnaire

5.5.1 The findings from the survey show that all nine professional attributes were seen as important for a FY1 doctor. Table 9 outlines the minimum and maximum ratings, the mean rating of importance (1=little importance, 6=extremely important), the standard deviation and the percentage scores.

5.5.2 Commitment to Professionalism, Coping with Pressure, Effective Communication, Organisation and Planning, Patient Focus, Self Awareness and Insight and Working Effectively as Part of a Team were all seen as very important for a FY1 doctor (mean rating >5). Learning and Professional Development and Problem Solving and Decision Making were

seen as the least important on the nine professional attributes, but were still viewed as important.

Table 9: Mean Rating of Importance of the Professional Attributes (n=230)

Professional Attribute (in alphabetical order)	Min	Max	Mean	SD	Rating of importance (percentage) ⁴					
					1	2	3	4	5	6
Commitment to Professionalism	2	6	5.4	0.73	0	0	0	10%	36%	53%
Coping with Pressure	3	6	5.2	0.78	0	0	2%	16%	39%	43%
Effective Communication	2	6	5.5	0.67	0	0	1%	4%	39%	56%
Learning & Professional Development	1	6	4.9	0.89	1%	0	3%	26%	46%	23%
Organisation and Planning	3	6	5.2	0.70	0	0	1%	13%	49%	37%
Patient Focus	1	6	5.3	0.82	1%	0	2%	12%	36%	49%
Problem Solving and Decision Making	2	6	4.9	0.92	0	0	7%	28%	37%	28%
Self Awareness and Insight	3	6	5.2	0.78	0	0	3%	13%	43%	40%
Working Effectively as Part of a Team	3	6	5.4	0.69	0	0	2%	6%	42%	50%

5.5.3 Table 10 outlines the results from the question which asked respondents to rate the importance of evaluating the professional attributes at selection. The table details the minimum and maximum ratings, the mean rating of importance (1=not at all, 6 = essential), the standard deviation and the percentage scores.

5.5.4 The findings show that Commitment to Professionalism, Effective Communication, Patient Focus, Self Awareness and Insight and Working Effectively as Part of a Team were all seen as important to evaluate at selection (mean rating >5). Coping with Pressure, Learning and Professional Development, Organisation and Planning and Problem Solving and Decision Making were seen as less important to evaluate at selection; however these were still seen as desirable.

⁴ (1=little importance, 6=extremely important)

Table 10: Mean Rating of Importance to Evaluate at Selection (n=230)

Professional Attribute (in alphabetical order)	Min	Max	Mean	SD	Rating of importance (percentage) ⁵					
					1	2	3	4	5	6
Commitment to Professionalism	1	6	5.2	0.98	0	2%	3%	18%	31%	46%
Coping with Pressure	1	6	4.9	0.97	0	0	8%	22%	38%	31%
Effective Communication	1	6	5.3	0.82	1%	0	1%	11%	37%	50%
Learning & Professional Development	1	6	4.6	1.03	1%	1%	9%	32%	37%	20%
Organisation and Planning	1	6	4.9	0.92	1%	0	5%	27%	42%	25%
Patient Focus	1	6	5.1	1.10	1%	1%	4%	19%	30%	44%
Problem Solving and Decision Making	1	6	4.8	0.97	0	0	8%	28%	37%	26%
Self Awareness and Insight	1	6	5.1	0.90	0	0	6%	16%	41%	37%
Working Effectively as Part of a Team	1	6	5.1	0.90	1%	1%	2%	16%	44%	36%

5.5.5 Statistical analysis was carried out to identify whether some attributes were rated significantly more important to be evaluated at selection than others. Although the intention of this survey was not to rank the professional attributes, the results from this could be used to inform the professional attributes to be targeted in the SJT.

5.5.6 A repeated measures ANOVA was carried out to test for significant differences between the mean ratings of the professional attributes. As a large number of statistical tests were being carried out simultaneously the Bonferroni correction was applied⁶. The results showed that there were 21 significant differences in the mean ratings of the professional attributes.

5.5.7 Specifically, the results show that Effective Communication was rated as the most important attribute to evaluate at selection as its mean rating was significantly higher than all of the professional attributes with the exception of Commitment to Professionalism. Learning and Professional Development was significantly rated the least important to be evaluated at

⁵ (1=not at all, 6 = essential)

⁶ The Bonferroni correction is a multiple-comparison correction used when several dependent or independent statistical tests are being performed simultaneously. In order to avoid a lot of spurious positives (as high as 1 in 20), the significance level is lowered to account for the number of comparisons being performed. This conservative analysis strategy was adopted so as not to over-interpret the results.

selection as its mean rating was significantly lower than all the professional attributes with the exception of Problem Solving and Decision Making.

- 5.5.8 The fact that some attributes have been rated by respondents as less important to be evaluated at selection may indicate that these attributes are less able to be assessed at selection and/or are able to be learned in training and will develop as FY1s progress through training. A number of the comments provided on the survey supported this interpretation (see table 11).
- 5.5.9 Participants in the survey were offered the opportunity to comment on the survey and add any further attributes that they felt were important for a FY1 doctor. A total of 96 comments were received that were relevant to the role of a FY1 doctor. The researcher content-analysed the comments; a resultant summary of the core themes is provided below in Table 11.
- 5.5.10 56 of the comments discussed possible further professional attributes. Review and analysis by the lead researcher of the additional attributes proposed revealed that they were already covered as part of the Professional Attributes Framework, so no new themes or attributes were identified. This suggests that the Professional Attributes Framework is a comprehensive account of the professional attributes that are expected in the FY1 job role.

Table 11: Summary of Validation Questionnaire Comments

Comment Theme	N
Highlighting attributes already covered in the Professional Attributes Framework	56
Consideration that the Foundation Programme is a learning process and some of the attributes are more likely to be learned and develop during the programme	19
Importance of clinical knowledge and competence	6
Drawbacks to the current process for selection into the Foundation Programme and suggestions for alternatives	15

5.6 Summary of findings

- 5.6.1 Broadly the same sets of behaviours were found from each of the three research methods- interviews, observations and the literature review. This provides a clear indication that the professional attributes outlined in the Professional Attributes Framework are representative of the attributes that would be expected from a FY1. This finding is replicated in the validation survey.

5.7 Mapping on to existing material

- 5.7.1 As part of the job analysis, it is important to ensure that the information obtained and any resulting frameworks map onto existing relevant material about the role, such as person specifications and curriculums. This helps assist with ensuring that the information obtained as a result of the job analysis is content valid i.e. that it is relevant and appears to be representative of the job role.
- 5.7.2 Figure 3 below demonstrates how the Professional Attributes Framework maps to the existing FY1 Person Specification. The current Person Specification is not intended to provide detailed information and only provides a high level definition of each of the attributes. Mapping between the Professional Attributes Framework and the FY1 Person Specification is therefore unlikely to show a very high level of concordance as the two models are not directly comparable. The high level definition of the attributes in the FY1 Person Specification also means that mapping is based on a high level interpretation of the Person Specification attributes.
- 5.7.3 The results of this mapping showed good concordance with the existing FY1 Person Specification. Four of the Person Specification attributes map directly onto one of the job analysis professional attributes. Two of the Person Specification attributes (understanding of principles of equality and diversity and initiative and ability to cope effectively with pressure) each map onto three of the job analysis professional attributes. Learning and Professional Development is the only professional attribute identified by the job analysis that does not clearly map onto the FY1 Person Specification.
- 5.7.4 It is therefore demonstrated that the current FY1 Person Specification does not contain any professional attributes that were not identified in the job analysis and the resultant Professional Attributes Framework. However, the Professional Attributes Framework did identify professional attributes that are not present in the Person Specification. Therefore the Professional Attributes Framework appears to show a good level of concordance with the UKFPO FY1 Person Specification.
- 5.7.5 Figure 4 displays how the Professional Attributes Framework maps onto the FY1 curriculum. More information about the syllabus can be found in the Foundation Programme Curriculum 2007.
- 5.7.6 The area of Recognition and Management of Acutely Ill that is defined in the FY1 Curriculum is related to clinical skills, rather than professional attributes and therefore does not map directly onto any of the job analysis professional attributes. It is therefore not presented in Figure 4.
- 5.7.7 The majority of the professional attributes map onto more than one of the syllabus areas outlined in the FY1 Curriculum. For example, Commitment to Professionalism, Patient Focus, Working Effectively as Part of a Team, Learning and Professional Development and Self Awareness and Insight each map onto three of the syllabus areas in the FY1 Curriculum and Effective Communication and Organisation and Planning map onto two of the syllabus areas

in the curriculum. Problem Solving and Decision Making and Coping with Pressure map onto one of the syllabus areas in the FY1 Curriculum. Therefore, the job analysis results and the resultant Professional Attributes Framework show good concordance with the Foundation Year One Curriculum.

Figure 3: Professional Attributes Framework and FY1 Person Specification

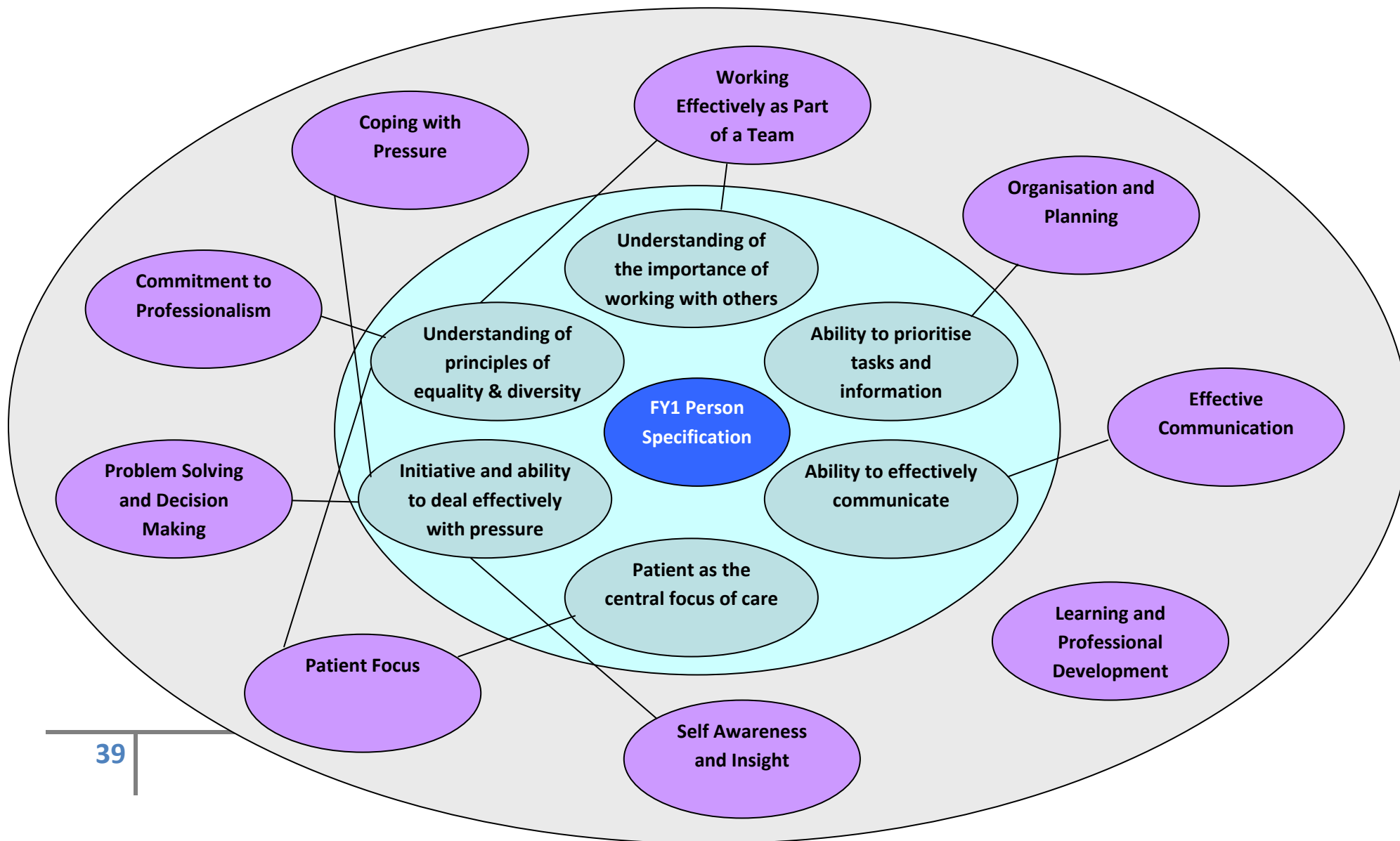
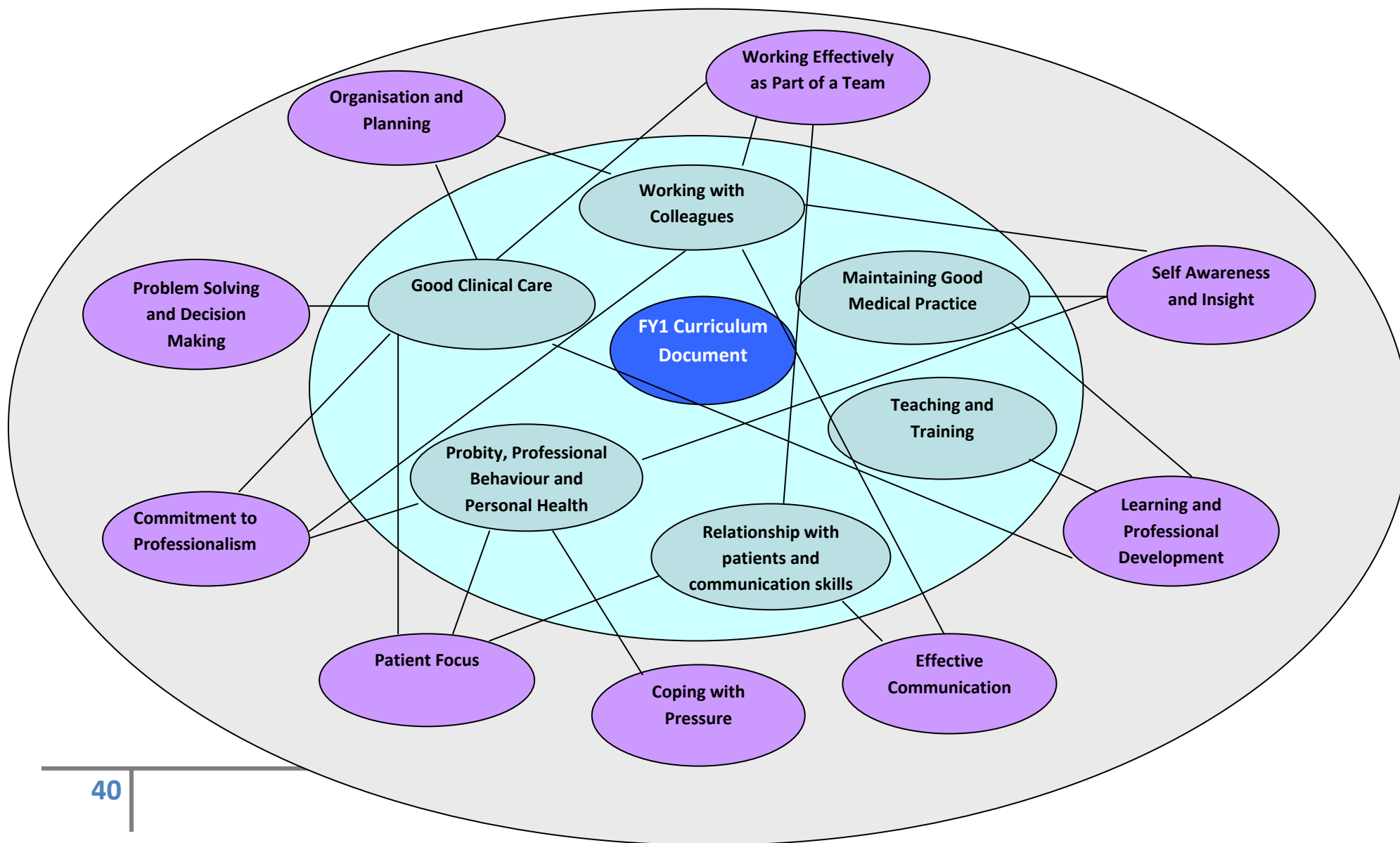


Figure 4: Professional Attributes Framework and FY1 Curriculum



6. Situational Judgement Test Design Specification

6.1 SJT literature review

- 6.1.1 The findings from the SJT literature review can be summarised under 13 headings. These are; Definition, Reliability, Criterion related validity, Construct validity, Fairness and adverse impact, Candidate reactions, Test format, Response instructions, Susceptibility to faking, Coaching and practice effects, Developing alternate forms, Scoring SJTs and General issues. The full SJT literature review can be found in Appendix 9.
- 6.1.2 Situational judgement tests (SJTs) describe job-related problems requiring the application of relevant knowledge, skills, abilities or other characteristics (KSAOs) to solve¹⁸. The scenarios are created on the basis of job analysis findings and critical incidents assessment¹⁹. SJTs can contain a variety of answer formats, such as choosing the best option or worst option, or rating the effectiveness of several alternatives; and responses are typically scored by comparing them to the judgements of experts in the respective field¹⁸.
- 6.1.3 Over the last 15 years SJTs have become increasingly popular in large-scale selection, typically at the shortlisting stage²⁰. Key benefits include being cost effective to develop, administer and score whilst being versatile in their presentation (written, web or video-based formats)²⁰. Despite becoming popular for practical reasons, their construct validity remains elusive and there is increasing uncertainty over how SJTs work and what they actually measure²¹. Whilst some argue that SJTs measure a single construct (e.g. practical intelligence²²), there is a growing consensus that SJTs are essentially measurement methods that can be designed to measure a variety of constructs, both cognitive and non cognitive²⁰. Subsequently, their predictive validity is likely to depend on what criterion the SJT is targeting. For example, an SJT that is designed to target interpersonal skills will more likely predict interpersonally oriented performance²³.
- 6.1.4 Generally, SJTs have been shown to have moderate to good reliability and predictive validity^{24, 25} and incremental validity over cognitive ability and personality measures^{26, 27}. SJTs have also been shown to have reduced adverse impact compared to cognitive ability tests²³, are typically met with favourable applicant reactions²⁸, and are considered to have high perceptions of face validity²⁹. Nevertheless, the literature highlights a range of moderators that influence these conclusions, including the format of the SJT (e.g. video-based vs. Written³⁰) as well as the types of response instructions used (e.g. knowledge based: 'what should you do?' vs. behavioural tendency: 'what would you do?'³¹). Research has also explored susceptibility to faking³², and the effectiveness of different scoring methods¹⁹, and has started to explore the coachability of SJTs³³.

6.2 Purpose of the test

- 6.2.1 The purpose of the Situational Judgement Test (SJT) is to improve selection into foundation training, by replacing the existing 'white space' questions⁷. It will be used in the final year of medical school to assess medical school students applying for the UK Foundation Programme. It is a test of aptitude, not attainment and therefore different from an examination. It will be used alongside an educational performance measurement (EPM) which is defined elsewhere.
- 6.2.2 The aim of the design and development of the SJT is to create a valid and efficient selection method. A robustly designed SJT will enhance the predictive validity of the process and improve standardization and consistency nationally, while reducing the potential for plagiarism. As a machine-marked test, the SJT will increase the efficiency of the process by avoiding the cost of the manual marking of 'white space' answers.
- 6.2.3 Designing a robust selection process based on a thorough job analysis of the role, that assesses applicants against core professional attributes, will also assist in enhancing patient safety.
- 6.2.4 The evaluation criteria used to judge the effectiveness of the FY1 SJT are set out below:
- **Reliability**– this will be established through psychometric analysis of the data following the autumn 2010 pilot and future pilots.
 - **Discriminatory** – it is essential that the test is able to discriminate between candidates. This is being addressed by making sure that the response formats that are used ensure the greatest amount of granularity and by ascertaining the level of difficulty of the items. Perceived difficulty is being established through ratings by the concordance panel (section 6.5.2) and FY2 focus groups and through statistical analysis (6.5.4). The test will need to consist of items of differing levels of difficulty, with enough 'difficult' items to ensure the test is discriminatory.
 - **Face validity** – this is a measure of whether the selection test or procedure 'looks right' to the applicant or others i.e. the link between the test and performance on the job is clearly evident. Face validity will be established through the use of candidate evaluation (FY2 focus groups and an evaluation questionnaire following completion of the SJT pilot) to establish whether the candidates feel that the SJT items are realistic and an accurate reflection of what FY1 doctors experience in their role. This will also address issues of perceived fairness on the part of the candidates.
 - **Content validity** - this is the extent to which the selection test or procedure covers a representative sample of the domains being assessed, as judged by subject-matter-experts e.g. the test items are relate to the professional attributes required for successful performance within a role. This will be established by the SJT items being

⁷ DH (2008) *A High Quality Workforce: NHS Next Stage Review*, available from http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_085840

reviewed by the researchers and stakeholders as part of the design and development process to ensure they are content specific.

- **Criterion-related validity** – this can be evaluated by examining the predictive and/or concurrent validity of the test.
 - *Predictive validity* - this is the extent to which the outcomes of a selection test or procedure predict future job performance. This can be established through longitudinal studies where individuals' scores on the SJT are compared to a performance measurement in the FY1 doctor role.
 - *Concurrent validity* - this is the extent to which the outcomes of, or scores from, a selection test or procedure correlate with other measures of performance that are already accessible. This can be established by comparing individuals' scores on the SJT to performance on other assessment measures, such as medical school exams or bespoke measures of supervisors.
- **Construct validity** – this is the extent to which the selection test or procedure is measuring the constructs (or attributes/skills/proficiencies) that it claims to be measuring. For example, that the SJT is actually measuring the professional attributes that underpin successful performance in the job. This can be evaluated in various ways such as through the use of measures of relevant constructs administered at the same time as the SJT, and comparison between these scores and the scores on the SJT.

Construct validity is notoriously difficult to establish within the SJT measurement methodology. It is suggested that *“although the research on SJTs so far indicates that they hold promise as valid predictors of job performance, the construct validity of SJTs remains elusive”*³⁴. Unlike *“purer”* measures of cognitive ability or personality, SJTs reflect complex, contextualised situations and events. The complex nature of the scenario and the use of numerous responses within an SJT item make it difficult to unpick how each item is measuring a particular construct or dimension.

- **Political validity**⁹ – this is whether the test or procedure is acceptable to stakeholders and will be established through stakeholder consultation at all stages of the design and development of the SJT.

6.3 Target attribute domains for the SJT

- 6.3.1 One of the aims of the job analysis was to define the professional attributes that could be measured in an SJT. Not all professional attributes are suitable to be measured directly in an SJT and therefore the professional attributes that are to be used in the operational test need to be identified. These suitable professional attributes will be termed 'SJT target attribute domains'.
- 6.3.2 The content of the scenario of a single SJT item should aim to target one of the SJT target attribute domains. This ensures that the scenario is relevant to the professional attributes

that are expected of an FY1 doctor. However, the nature of SJT items (consist of a scenario and a number of response options) means that a single SJT item will not exclusively measure one target domain. The content of the scenario and the correct response keys are likely to relate to the targeted domain, but a number of the other response options are likely to tap into one or more of the other SJT target attribute domains (see issues regarding construct validity above). The benefits of this are that this not only does the complexity of the item increase, but the possibility of a template answer being provided by candidates is avoided.

6.3.3 Following the outcome of the job analysis, a panel (N=7) of researchers and occupational psychologists expert in SJT development and design, considered which of the nine professional attributes would be the most appropriate to target in an operational SJT and thus provide a psychometrically robust and valid test.

6.3.4 The criteria by which the SJT target attribute domains were derived are outlined below.

- *Relevant to the outcomes of the job analysis.* The SJT target attribute domains should be derived from the job analysis and should be rated as 'important' and 'important to assess at selection' through the validation questionnaire. Although the validation questionnaire indicated relative importance, this in itself should not be taken as the sole indication of the most appropriate attribute domains to be sampled in the SJT. This will assist in demonstrating that the attribute domains have content validity.
- *Relative independence and discreteness.* The SJT target attribute domains should be relatively independent from each other. This will assist in instructing item writers and will inform the design of validation studies.
- *Bound in appropriate and relevant language.* The language used to define the SJT target attribute domains should be appropriate and tailored to the medical context and relevant to that of the role of a FY1 doctor. It should mirror the language used in relevant documentation e.g. *Tomorrow's Doctors*. This will assist in demonstrating face and content validity.
- *Map to the FY1 Person Specification.* The SJT target attribute domains should map to the professional attributes that are outlined as requirements in the FY1 Person Specification.

6.3.5 This resulted in the expansion of the original four domains, to five domains. These five SJT target attribute domains are:

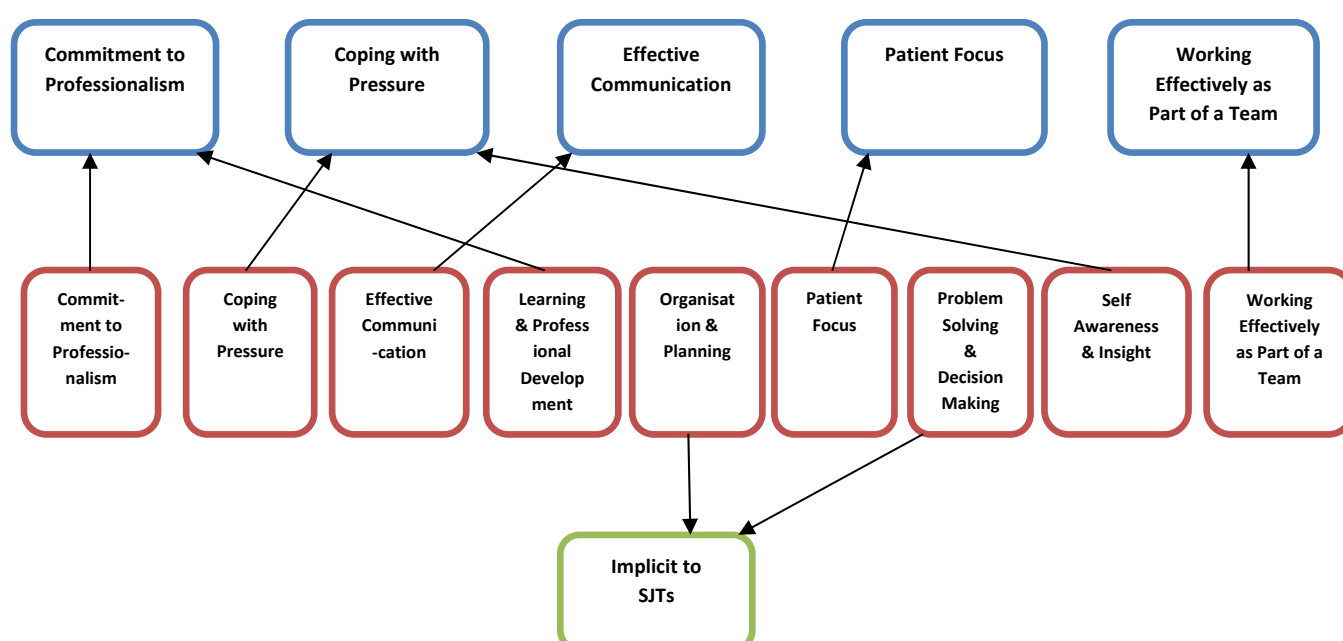
- Commitment to Professionalism
- Coping with Pressure
- Effective Communication
- Patient Focus

- Working Effectively as Part of a Team

6.3.6 Figure 5 illustrates the relationships between the nine professional attributes and the five SJT target attribute domains.

6.3.7 Organisation and Planning and Problem Solving and Decision Making are the two of the nine domains that are not explicitly sampled in the final five SJT target attribute domains. As outlined above, aspects of Organisation and Planning are implicit in the format and methodology of the SJT and thus embedded in the test. For example, response styles may reflect prioritisation e.g. rank order the importance of the tasks. Aspects of Problem Solving and Decision Making can also be seen to be implicit to undertaking an SJT (e.g. making appropriate decisions based on all the relevant information, attention to detail).

Figure 5: Relationship between Professional Attributes and SJT target domains



6.3.8 When all SJT items have been submitted and finalised, those SJT items that have been deemed as useable, will be reviewed by the SJT workshop facilitators and item reviewers (N=3) and re-categorised in accordance with the five SJT target attribute domains.

6.3.9 A matrix which outlines the SJT target attribute domains and possible SJT scenarios (as identified from the job analysis interviews) is provided in Table 12.

Table 12: Matrix of SJT Target Attribute Domains and Sample SJT Scenarios

Commitment to Professionalism	• Issues of confidentiality e.g. hearing a colleague talking about a patient outside of work/leaving confidential paperwork around • Challenging inappropriate behaviour e.g. Consultant speaking to a colleague/patient in an inappropriate way • Commitment to learning e.g. need to go to teaching but also needed on the ward
Coping with Pressure	• How respond when you make a mistake e.g. provide wrong medication to patient • Dealing with confrontation e.g. angry relative • Seeking help when not sure of the correct procedure/best way of doing things
Effective Communication	• Gathering information and communicating your intentions to nursing staff or other colleagues • Negotiating for a scan from radiology • Listening and effectively communicating with an unhappy patient or relative
Patient Focus	• Identifying that that patient's views and concerns are important and they should have input into their care • Considering that a patient may have different needs from others around them • Spending time trying to understand a patient's concerns and empathising with them
Working Effectively as Part of a Team	• Recognising and valuing the skills and knowledge of nursing staff, when faced with a disagreement about a patient's care • Consulting with colleagues about how to share workload fairly • Offering assistance and support to colleagues when they are not able to handle their workload

6.4 Context

- 6.4.1 The Situational Judgement Test has been designed to be set in the context of the Foundation Programme, with the applicants assuming the role of a FY1 doctor. Applicants' (medical school students) views were canvassed to ascertain whether the content of sample SJT items, set in the context of the Foundation Programme, was felt to be appropriate to their level and whether they had sufficient experience to be able to respond to the items. Responses indicated that the students felt this context was fair and appropriate.

6.5 Item selection methodology

- 6.5.1 The selection criteria for the items to be used in an operational test are outlined below:
- Items must be content valid (test appears to assess the attributes/skills required for successful performance within a role) and face valid (test 'looks right' to the candidate). This will be established through the use of FY2 focus groups and through item review.
 - Items must be judged as unambiguous. During development, any items containing potential ambiguity that cannot be resolved, or for which consensus cannot be reached on the scoring key, will be discarded.
 - The collection of items forming the test must come from a good range of the five SJT target domains, with roughly an equal number of items from each of the domains.
 - Two thirds of the items used in a test should have the ranking answer format, and one third should have the multiple choice format. Using more items with the ranking format will produce more granularity, as more information is able to be gained from each item. This also follows the validated methodology from the GP SJT²⁶.
- 6.5.2 A concordance stage will be carried out prior to the pilot. Following a previously validated methodology²⁶, the concordance panel involves a panel of Subject Matter Experts, in this case clinicians (N= ≥10), who have not been involved in item development, taking a SJT that consists of around 80 trial items. The aim of the concordance stage is to identify a high level of consensus between experts on the item keys.
- 6.5.3 A concordance analysis will be undertaken to ensure the experts are in agreement over the keyed response for each trial item. The selection criteria for items for the pilot will be based on the psychometric analysis of the results from a concordance panel i.e. those items that show high concordance will be selected for the pilot.
- 6.5.4 All of the test items used in the pilot will be subject to extensive psychometric analysis. Analysis will include: consistency in responses between concordance panel, item writers and candidates; pattern of scores by demographic groups, and subgroup differences (e.g. age, gender, ethnicity, place of medical school); reliability of items within domains and reliability

of test overall; and psychometric analysis at an item level (differential item functioning⁸, facility values⁹ and item partials¹⁰). This analysis may result in some being removed from the bank.

- 6.5.6 Candidate reactions will be evaluated at piloting stages, specifically to address perceptions of fairness and related fairness and job relevance issues. A bespoke questionnaire will be designed based on a previously validated questionnaire²⁶ so that reactions can be closely evaluated. It is important to assess candidate reactions, as negative reactions could compromise the validity and utility of the SJT.

6.6 Response formats

- 6.6.1 Research shows there are many different response formats that could be used in an SJT. These include:

- Choose best and/or worst response
- Rate effectiveness of all options (e.g. on Likert scale or effective vs not effective)
- Rank effectiveness of all options
- Choose multiple options

- 6.6.2 Ranking options can be better than rating or choose best and/or worst, as it provides more data from each item and more variance in responses. A ranking format also replicates one of the key skills required by FY1s; that of prioritisation, so may be seen as most suitable for the SJT. Choosing multiple options works well for items/scenarios where ranking is not appropriate.

- 6.6.3 Applicants (medical school students) views were canvassed regarding the SJT response formats. All four types of response format in order to gauge candidate reactions. Evaluation of the information from the micropilots indicated that the ranking and the multiple choice response instructions were viewed favourably by most participants.

- 6.6.4 The mini pilot is piloting both ranking and multiple choice types of response formats. The final response format(s) will not be proposed until after the mini pilot in autumn 2010.

⁸ DIF is a procedure used to determine if test items are fair and appropriate for assessing the ability of various demographic groups. It is based on the assumption that test takers who have similar ability (based on total test scores) should perform in similar ways on individual test items regardless of their sex, race, or ethnicity.

⁹ The facility value is the mean score achieved for the item. If the facility value is very low, then the item may be too difficult and may not yield useful information. If the facility value is very high, then the item may be too easy and may not provide useful information or discriminate between candidates.

¹⁰ Item partials examine the correlation between item score and overall test score, not including the item itself. This shows how well an item differentiates between high and low scorers.

6.7 Test length

- 6.7.1 Test length considerations include reliability, logistics, acceptability, coverage of target domains and item format and length.
- 6.7.2 The number of items in, and the duration of, the SJT will be determined through piloting. It is proposed that the test in the first pilot should last for two hours and contain 60 items. This number of items is sufficient to cover the five target attribute in a sufficiently reliable way without overloading the candidates, while two minutes per item is consistent with GP SJT which has demonstrated good reliability and validity¹⁸. This will also ensure that the test is a 'power test' rather than a 'speed' test i.e. a test of an individual's aptitude in the target attribute domains rather than a test of how fast they can read.
- 6.7.3 Applicants (medical school students) views were canvassed as to whether they felt 60 items in two hours was appropriate and enough time to respond to the items. All students canvassed felt this would be sufficient time to complete the test.

6.8 Test structure

- 6.8.1 The test structure will be established through piloting. As far as possible, (taking in to consideration the other considerations outlined below) the items will be presented in ascending difficulty (assuming that some items are more difficult than others). The difficulty of items will be rated initially by the FY2 focus groups and the concordance panel.
- 6.8.2 The five SJT target attribute domains will be spread throughout the paper, as grouping items together by domain could promote cueing. Likewise, items that cover similar topics (e.g. dealing with colleagues) will also not be grouped together. Items that include gravely serious scenarios (e.g. those involving deaths or serious mistakes by the candidate) should not be used at the beginning of the paper, as candidates may find these scenarios distressing initially if they are not familiar with the format of the SJT; placing these items later on in the test allows the candidate time to familiarise themselves with the context of the test and what is expected of them.
- 6.8.3 Items with the same response format will be grouped together to avoid confusion on the part of the candidate when completing the items.

6.9 Scoring key

- 6.9.1 In contrast with clinical knowledge items, with SJT items there is often no definitive correct answer. The scoring key is determined through:
 1. Consensus at the item review stage of item writers and initial subject matter experts.
 2. Expert judgement in a concordance study.
 3. Review and analysis of the pilot data. Consideration will be given to how the candidates keyed the items and the consensus between this and the key derived from the first two stages.

- 6.9.2 There are a number of scoring conventions that can be used, examples of which are provided below:
- a. Set number of points for identifying correct options.
 - b. Several methods if rating using numbered rating scales or 'best/worst' responses
 - c. Ranking items – scores based on how close responses are to expert key.
- 6.9.3 When determining how to score each item, consideration needs to be given to the amount of valid information which can be derived from each test item. For example, for a ranking item, a candidate could receive points for getting items partially correct, or only receive points for providing the fully correct answer. It is suggested that the former option is employed as this provides the maximal amount of information about the candidate's performance.
- 6.9.4 The scoring conventions to be employed will be finalised after the pilot through psychometric analysis.

6.10 Multiple forms

- 6.10.1 In a national recruitment round there will be a need for an SJT to be held on at least two, and possibly three, separate days. Three 'parallel forms' of the test will therefore need to be created each year to reduce the impact of applicants passing on information to those taking the test on later days. These parallel forms of the test will contain some items in common to equate the level of difficulty.
- 6.10.2 Multiple forms of the test will not need to be created until the spring 2011 pilot.
- 6.10.3 Good practice would also suggest that there should be enough items available to create a 'reserve' paper that could be called into action if a breach of security comes to light before the testing date.

7. Summary and Conclusions

- 7.1 A thorough multi-method job analysis was carried out to identify the professional attributes that are required for the effective performance of a FY1 doctor.
- 7.2 The job analysis included interviews with those familiar with the FY1 role, observations of FY1 doctors and a validation survey that asked participants to rate the importance of the professional attributes identified. A total of 294¹¹ individuals participated in the job analysis, providing a wide range of perspectives.
- 7.3 Through analysis of the data nine professional attributes were identified, each represented by a number of behavioural descriptors as outlined in the Professional Attributes Framework (Table 6). These professional attributes were mapped onto the FY1 Person Specification and FY1 Curriculum.
- 7.4 The Professional Attributes Framework lists the behavioural descriptors that represent each of the professional attributes. The value of the framework is the level of detail that is provided. This level of granularity is essential when selecting an individual for a role; therefore the Professional Attributes Framework could be used to extend and improve the person specification going forward.
- 7.5 The number of behavioural descriptors derived from the job analysis will provide a rich source of information for selection processes in the future, including the development of SJT items.
- 7.6 Five SJT target domains were identified to be sampled in the Situational Judgement Test. Analysis following the pilots of the SJT will provide further evidence of the validity of the Professional Attributes Framework.

¹¹ A number of individuals may have contributed to more than one of the job analysis methods e.g. an interview and responded to the survey. 294 is the total number of respondents, but total number of participants may be lower

8. References

1. Gutman, A. (2000). Recent Supreme Court ADA rulings: Mixed messages from the Court. *The Industrial-Organizational Psychologist*, 37, 31–41.
2. Thompson E and Thompson T.A. (1982). Court standards for job analysis in test validation. *Personnel Psychology*, 35, 865-874.
3. Biddle, D.A (2009) Guidelines Orientated Job Analysis: A job analysis process for selection procedure, development and validation. Biddle Consulting Group Inc.
4. Arnold J, Silvester J, Patterson F, Cooper C, Robertson I, Burnes B. (2010). *Work Psychology* (5th Edition). Pearson: FT Prentice Hall.
5. Gatewood R.D & Field, H.S. (2001). *Human Resource Selection*. Harcourt, Inc. New York.
6. Brumback, G.B, Romashko, T, Hahn, C.P & Fleishman, E.A. (1974). *Model Procedures for Job Analysis, Test Development, and Validation*. Washington DC: American Institutes for Research.
7. Patterson F, Ferguson E, Lane P, Farrell K, Martlew J, Wells A. (2000). A competency model for general practice: implications for selection, training, and development. *British Journal of General Practice*, 2000, 50, 188-193.
8. Patterson F, Ferguson E & Thomas S. (2008). Using job analyses to identify core and specific competencies for three secondary care specialties: Implications for selection and recruitment. *Medical Education*, 42, 1195-204
9. Patterson, F & Zibarras, L. (in press). Exploring the construct of perceived job discrimination and a model of applicant propensity for case initiation in selection. *International Journal of Selection & Assessment*.
10. Flanagan J, C. (1954). The Critical Incident Technique. *Psychological Bulletin*, 51.4, 327-359
11. Miles M, Huberman M (1994). *Qualitative data analysis*. 2nd ed. London: Sage.
12. King, N (1998). Template Analysis, in Symon, G. & Cassell, C. (Eds.) *Qualitative Methods and Analysis in Organisational Research: A Practical Guide*. London: Sage.
13. King, N (2004). Using templates in the thematic analysis of text, in Cassell, C. & Symon, G. (Eds.) *Essential Guide to Qualitative Methods in Organizational Research*. London: Sage.
14. Crabtree B.F, Miller W.L (1992). *Doing qualitative research*. London: Sage.

15. Welsh D.H.B, & Carraher S.M. (2009). An examination of the relationship between the mission of the university and the business school. *Journal of Applied Management and Entrepreneurship*
16. Stein S, Lauer Y, & El Kharbili M. (2009). Using Template Analysis as Background Reading Technique for Requirements Elicitation.
17. Spencer D & Warfel T. (2004). Card sorting: a definitive guide. Accessed on 18th May 2010 http://www.bboxesandarrows.com/view/card_sorting_a_definitive_guide
18. Christian, M.S, Edwards, B.D, & Bradley, J.C. (2010). Situational judgement tests: constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, 63, 83-117.
19. Bergman, M. E, Drasgow, F, Donovan, M. A., Henning, J.B & Juraska, S. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, 14, 223–235.
20. Lievens, F, Peeters, H, & Schollaert, E. (2008). Situational judgement tests: a review of recent research. *Personnel Review*, 37, 426-441.
21. Weekley, J.A, & Ployhart, R.E. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance*, 18, 81-104.
22. Sternberg, R. J, Wagner, R. K, Williams, W. M, & Horvath, J. A. (1995). Testing common sense. *American Psychologist*, 50, 912-927
23. Lievens, F., Buyse, T., & Sackett, P. R. (2005a). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, 58, 981-1007.
24. McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, 86, 730-740.
25. McDaniel, M.A., Hartman, N.S., Whetzel, D.L., & Grubb, W. L. (2007). Situational judgement tests, response instructions and validity: A meta analysis. *Personnel Psychology*, 60, 63-91
26. Patterson, F, Baron, H, Carr, V, Lane, P & Plint, S. (2009). Evaluation of three short-listing methodologies for selection into postgraduate training: the case of General Practice in the UK. *Medical Education*, 43, 50-57.
27. Llievens, F & Peeters, H. (2009). Impact of elaboration on responding to situational judgment test items. *International Journal of selection and Assessment*, 16, 345-355.

28. Chan, D & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance*, 15, 233-254.
29. Clevenger, J, Pereira, G.M, Wiechmann, D, Schmitt, N, & Harvey, V.S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology*, 86, 410-417.
30. Lievens, F. & Sackett, P. R. (2006). Video-based vs. written situational judgment tests: a comparison in terms of predictive validity. *Journal of Applied Psychology*, 91, 1181–1188.
31. Polyhart, R.E, & Erhart, M.G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment*, 11, 1-16.
32. Nguyen, N. T, Biderman, M. D, & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment*, 13, 250–260.
33. Cullen, M.J, Sackett, P.R, & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment*, 14, 142-155.
34. Oswald, F.L, Schmitt, N, Kim, B.H, Ramsay, L.J, & Gillespie, M.A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, 89, 187-207.

Appendix 1: Literature Review Specification & Search Strategy

1. Background

- 1.1 The review will form part of the triangulation approach to the job analysis of the F1 role and will be conducted in two parts, with separate research aims. The literature review will be included as part of the job analysis report.

2. Research Aims

- 2.1 The first part of the review will involve a review of official documents and all relevant publications in the literature to assist in identifying the key skills, abilities and professional attributes of a FY1 doctor. This review will allow for the identification of specific contextual issues and elements which may be incorporated into the proposed framework.
- 2.2 The second part of the review will consist of a review of Situational Judgement Tests and their use in high stakes selection in medical education and other professional bodies. This will assist in placing the purpose of the job analysis in context.

3. Sources

- 3.1 A range of sources will be used to assist in the reviews. These include databases, journals, reports, web searches and expert contacts. Further details are specified in the table below.

Databases: These will be searched in accordance with the guidance developed for the Best Evidence Medical Education systematic reviews and are likely to include:	Medline
	BREI
	ERIC
	CINAHL
	Psychinfo
	PUBMED
	Embase
	Web of Science
	BioMed Central
	Clinical Evidence
	BIOME –collection of evidence based gateways
Journals: for identifying the key skills, abilities and professional attributes of a Foundation doctor	Medical Education
	Academic Medicine
	Medical Teacher
	Advances in Health Sciences Education
	British Medical Journal
	Journal of Public Health
	British Journal of Clinical Practice
	Journal of Medical Education
	Clinical Medicine
	Journal of the Royal Society of Medicine
	Teaching and Learning in Medicine

	Medical Journal of Australia
	Occupational Medicine
	BMC Medical Education
Occupational Journals: for the review of Situational Judgement Tests	Journal of Applied Psychology
	International Journal of Selection and Assessment
	International Journal of Human Resource Management
	Personnel Psychology
	Journal of Occupational and Organisational Psychology
Reports: Including clinical guidance and regulatory policy and reports from other bodies. To include; Foundation Programme Curriculum, the New Doctor, Tomorrow's Doctors, Foundation Programme person specification (2010), Bologna Process. Editorials and opinion pieces will also be examined where these add to the analysis/argument	
Web searches: Websites of relevant key organisations will be searched for additional papers/documents of interest/links to additional information or references e.g. UK government departments, medical societies and committees. Filip Lievens' (expert in SJTs) website will also be searched.	
Expert contacts: Key individuals identified through the review and discussions with the research team may be contacted and asking for feedback on the draft report, specifically seeking advice on any missing references and asking for information on any unpublished or in press papers/work in progress.	
Hand searches: Hand searches will be carried out if appropriate	

4. Search Dates

4.1 The search date will be limited to twenty years and will cover the years 1990 to 2010.

5. Search Terms

5.1 Search terms/key words have been identified by their relation to the Research Aims and are outlined in the table overleaf:

Foundation Programme Doctor Review	SJT Review
Foundation doctors	Situational Judgement Test
Professional attributes	Recruitment
Skills	Selection
Competencies	Assessment
Foundation year	High Stakes Settings
Medical student	Response Instructions
Non-cognitive attributes	Response Formats
Personal qualities	Scoring Key
Interpersonal qualities	Domains
Capabilities	Predictive validity
Professional Competence	Personality
Medical Undergraduate	
House Officer	
PRHO	
Post Graduate Year One	
Interns/internship	
Knowledge	
Skills	
Attitudes	

5.2 A range of search terms for each of the identified search concepts will be considered, including:

- o Synonyms, e.g. *doctor, practitioner, consultant*
- o Acronyms, e.g. *F1, Dr*
- o Differences in terminology across national boundaries
- o Differences in spellings
- o Old and new terminology
- o Lay and medical terminology

5.3 Search terms will be used individually initially and then in combination. Appropriate truncation for textword searches will be used where applicable, e.g. *nurs\$* to find *nurse, nurses, nursing* etc.

6. Database Search Strategies

6.1 The search strategies created will be designed for maximum sensitivity (recall) to ensure that effort is taken not to overlook any papers of significance.

6.2 A pilot search in each database will be carried out. This will review search terms as necessary before conducting any final searches as initial searches can highlight other appropriate search terms, words or phrases and database subject headings.

6.3 Appropriate subject headings for each database used will be identified. Correct use of subject headings improves the accuracy of results and is essential to an effective search. The coverage, scope and definition of each subject heading will be checked.

6.4 The search will use a combination of textwords (free text) and subject heading searches. The search will avoid restricting database subject heading searches using the '*major*

descriptors or *subheading* options in the first instance (to avoid missing relevant material). The strategy will use specific functions in the relevant databases, for example the PubMed 'related articles' function and lists of references in CINAHL.

- 6.5 Advanced search options in databases will also be used to increase sensitivity. This includes using search operators to vary the combination of pre-determined search terms e.g. using AND to combine two different concepts; foundation doctor AND professional attributes, using OR to search for similar concepts; personal OR interpersonal qualities, using NOT to restrict search to particular concepts; foundation doctor NOT medical student etc.

7. Search Methodology

- 7.1 The search methodology will involve three main stages; a) search, b) filtering by abstract, c) detailed review.

- a) Databases will be searched using the database search strategies outlined above. The web search will include scanning websites of relevant key organisations for additional papers, articles and links to further sources. Contact will also be made with expert individuals to seek advice on unpublished or in progress work which may be of relevance in order to obtain most up to date and current material.

Citation tracking will be carried out; checking reference lists for other relevant journal articles and following up references and contacts emerging from documents identified through the initial stages of the search.

- b) All documents/articles will be scanned for relevance on the basis of title and abstract; non-relevant papers will be rejected at this stage according to the inclusion criteria:
- Relevant to identified topic
 - Adequate methodology
 - Informed opinion which adds to the picture of the topic
 - Peer Reviewed
 - English Language

If a paper is felt to satisfy the criteria, or in cases where an abstract is not available but the title suggests it might do, the full paper will be obtained from electronic journals, library hard copies, or inter-library loan.

- c) During the detailed review, the obtained papers will be read by a member of the research team. Data extraction sheets will be used to record all relevant information. The data extraction sheets will comprise of a table recording the following: method of identification (e.g. database), author, title, source (e.g. journal, conference), year, volume, pages, institutional affiliation, field, type of paper (e.g. empirical, conceptual or descriptive), research aim, study design, theory, analysis, key findings/points. If a paper is not felt to be relevant following this review, a note will be made to this effect.

Appendix 2: Job Analysis Participant Briefing

Confirmation of Assistance with Foundation Year One Doctor Job Analysis

Thank you for agreeing to take part in a semi-structured interview as part of the F1 doctor role job analysis. I can confirm that the interview will take place at <time> on <date>. The interviewer, <name>, will contact you at the specified time on the telephone number you have provided. If you are unable to make the appointment for whatever reason, please let myself know as soon as possible.

The interview should take no longer than one hour. During the interview a trained interviewer will ask you about the knowledge, skills and abilities an F1 doctor should possess, the tasks that they carry out and scenarios that they commonly face. Your input will be of great value in the design and development of the Situational Judgment Test (SJT) items and it will help ensure that the scenarios are pitched at an appropriate level. We would also like to consult with you, if possible, around the specification of the SJT, for example appropriate length, types of items/responses. All the responses you provide will be anonymised for issues of confidentiality. The outcomes of the job analysis will be made available in a summary report after the job analysis has been completed.

I have provided some background information to the project below for your information.

As you may be aware, *Next Stage Review: A High Quality Workforce*¹² recommended a review of the method by which final year medical students are selected into foundation training. The Option Appraisal of Selection to the Foundation Programme was a collaborative venture between fifteen UK organisations involved in medical education and assessment. The recommendations, informed by national and international evidence, expert advice and stakeholder feedback, are for:

1. Situational Judgment Tests (SJTs) to assess professional attributes, judgement and employability for a Foundation programme post (to replace white space questions)

In combination with

2. A measure of Educational Performance of applicants at their medical school to assess clinical knowledge/ skills. (evolved from the academic quartiles)

As part of a pilot to design and develop a Situational Judgement Test, the Work Psychology Group and the University of Cambridge Psychology Department are conducting a job analysis of the F1 role with those familiar with the role, including Foundation School Directors, Educational Supervisors, Medical School Educational Leads, F1 doctors and other relevant groups. The outcomes of the job analysis will be used to assist in the development of the scenarios which will be the basis of the Situational Judgement Test.

Situational Judgement Tests (SJTs) are designed to assess individuals' judgement regarding situations encountered in the workplace. Candidates are presented with a set of hypothetical work-based scenarios and asked to make judgements about possible responses. Candidates' responses are

¹² DH (2008) *A High Quality Workforce: NHS Next Stage Review*, available from http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_085840

evaluated against a pre-determined scoring key to provide a picture of their situational judgement in that particular context. SJT scenarios are based on extensive analysis of the job role, to ensure that test content reflects the most important situations in which to test candidates' judgement and are concerned with testing attitudes and ethical values rather than knowledge or clinical skills. SJTs have become increasingly popular over the last 20 years and are used mostly in large-scale selection processes, often at the shortlisting stage, but can also form part of workplace assessment to highlight employee development needs. In the UK, SJTs are used nationally to select GP registrars and in other 'high stakes' selection. The research literature indicates that SJTs have significant validity in predicting job performance and can offer incremental validity over methods such as ability tests and personality questionnaires.

Appendix 3: FY1 Doctor Job Analysis Questions

(information to be used by interviewer)

The *Next Stage Review: A High Quality Workforce* recommended a review of the method by which final year medical students are selected into foundation training. The Option Appraisal for Selection to the Foundation Programme was a collaborative venture between fifteen UK organisations involved in medical education and assessment. The recommendations, informed by national and international evidence, expert advice and stakeholder feedback, are to develop a Situational Judgment Test for selection to the FY1 post (to replace white space questions) together with an educational performance measurement.

Together with the University of Cambridge Psychology Department and the Medical Schools Council, the Work Psychology Group is conducting a job analysis of the F1 role with those familiar with the role, including Foundation School Directors, Educational Supervisors and Medical School educational leads. Focus groups will also be conducted with F1/2 doctors and other relevant groups at a later date. The outcomes of the job analysis will be used to assist in the development of the scenarios which will be the basis of the Situational Judgement Test.

NB: we already have an F1 person specification. The domains that appear to arise from the person spec are; Working effectively with others, Communication, Resilience/ability to cope with pressure and Integrity (probity and understanding equality and diversity) and these may form the basis of the SJT.

The interview should last no longer than an hour and will be split into three parts;

- 1 – job analysis of the F1 role
- 2 – common scenarios that F1 doctors may face. These may form the basis of the SJT items
- 3 – views on different formats of situational judgement tests

Intro

Ask if interview can be recorded for practical purposes e.g. so don't have to write everything down.

Job analysis questions

- What are the specific duties/responsibilities of the F1 position?
- What are the critical tasks of the position?
- How much time does it take to do each of the important tasks? How often do they perform each task?
- What methods or processes are used to perform the tasks in the job?
- What are the discrete outcomes of the job for which the individual will be held accountable and evaluated on?
- What behaviours, skills, knowledge, abilities and interpersonal attributes are the most important in achieving the key tasks and outcomes?

Scenarios

Situational Judgement Tests (SJTs) are designed to assess individuals' judgement regarding situations encountered in the workplace. Candidates are presented with a set of hypothetical work-based scenarios and asked to make judgements about possible responses. Candidates' responses are evaluated against a pre-determined scoring key to provide a picture of their situational judgement in that particular context.

- Could you please describe some common scenarios/situations that an F1 doctor may face on a regular basis where the doctor may exhibit the key skills/attributes/knowledge or behaviour that we discussed earlier?
- Could you please describe some specific examples of scenarios/situations that you have observed/been involved with, that involved an F1 doctor making a judgement or decision.

This may be something that an F1 doctor has done particularly effectively or not so effectively i.e. did or did not demonstrate key skills/attributes/behaviour/knowledge that we discussed earlier. Please could you describe the following; pre-cursor, people involved, dialogue, outcome, how could this have been done differently.

SJT item specification

Explain to interview that there are various item types and response instructions that can be used in SJT items and would like to ask their views if possible. (This may depend on how familiar they are with SJTs, may be more suitable for stakeholders than F1 doctors)

- Item question type "what would you do" (behavioural) or "what should you do" (knowledge)
- Item response instructions "pick the best", "pick the worst", "pick the best and worst answer" "rate effectiveness of item" "rank items from most effective to least effective" "choose best three (or other number) responses"
- Number of responses
- Length of question
- Number of items/length of test

Appendix 4: Patient Representative Briefing

Dear

Thank you for agreeing to take part in a semi-structured interview as part of the F1 doctor role job analysis. I can confirm that the interview will take place at <time> on <date>. The interviewer, <name>, will contact you at the specified time on the telephone number you have provided. If you are unable to make the appointment for whatever reason, please let myself know as soon as possible.

Please find below further information relating to the Improving Selection to the Foundation Programme project and the interview.

As you may be aware, *Next Stage Review: A High Quality Workforce*¹³ recommended a review of the method by which final year medical students are selected into foundation training. The Option Appraisal of Selection to the Foundation Programme was a collaborative venture between fifteen UK organisations involved in medical education and assessment. The recommendations, informed by national and international evidence, expert advice and stakeholder feedback, are for:

1. Situational Judgement Tests (SJTs) to assess professional attributes, judgement and employability for a Foundation programme post (to replace white space questions)

In combination with

2. A measure of Educational Performance of applicants at their medical school to assess clinical knowledge/ skills. (evolved from the academic quartiles)

As part of a pilot to design and develop a Situational Judgement Test, the Work Psychology Group and the University of Cambridge Psychology Department are conducting a job analysis of the F1 role with those who may have some experience or knowledge of the role, including Foundation School Directors, Educational Supervisors, F1 doctors and patients. The outcomes of the job analysis will be used to assist in the development of the scenarios which will be the basis of the Situational Judgement Test (SJT).

SJTs are machine-markable tests designed to assess individuals' judgement regarding situations encountered in the workplace. Candidates are presented with a set of hypothetical work-based scenarios and asked to make judgements about possible responses, which are scored according to a pre-determined key. SJTs have become increasingly popular over the last 20 years and are used mostly in large-scale selection processes, often at the shortlisting stage. In the UK, SJTs are used nationally to select GP registrars and in other 'high stakes' selection. The research literature indicates that SJTs have significant validity in predicting job performance and can offer incremental validity over methods such as ability tests and personality questionnaires.

The interview should take no longer than 30 minutes and will relate to the role of a Foundation Doctor. A doctor will enter the Foundation Programme after five years at medical school. The

¹³ DH (2008) *A High Quality Workforce: NHS Next Stage Review*, available from http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_085840

Foundation Programme lasts for two years and during this time a Foundation Doctor will undertake a number of rotations within different specialties. At the end of the two year Foundation Programme, a doctor will usually go on to specialise in a chosen specialty.

During the interview a trained interviewer will ask you about your knowledge of or experience with Foundation Doctors. This may include specific interactions you have had with Foundation Doctors, knowledge you possess with reference to their role within the medical profession or your understanding of the professional attributes that are necessary for a Foundation Doctor to carry out their role in an effective manner.

Your input will be of great value in the design and development of the Situational Judgment Test (SJT) items and it will help ensure that the scenarios are relevant and cover a broad range of perspectives. All the responses you provide will be anonymised for issues of confidentiality. The outcomes of the job analysis will be made available in a report after the job analysis has been completed.

Appendix 5: Observation Observe and Record Sheet

[illegible]

Appendix 6: Foundation Year One Doctor Professional Attributes Questionnaire

Please rate each professional attribute on the scales provided below by indicating **ONE** number from 1 to 6 that best corresponds to your evaluation of the attribute.

PROFESSIONAL ATTRIBUTE	DESCRIPTION	How important is this professional attribute for an FY1 doctor?					
		little importance	-----	extremely important			
Commitment to Professionalism	<i>Displays honesty, integrity and awareness of confidentiality & ethical issues. Is trustworthy and reliable. Demonstrates commitment and enthusiasm for role. Willing to challenge unacceptable or unsafe behaviour when appropriate. Takes responsibility for own actions.</i>	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Coping with Pressure	<i>Capability to work under pressure and remain resilient. Demonstrates ability to adapt to changing circumstances and manage uncertainty. Remains calm when faced with confrontation. Develops and employs appropriate coping strategies and demonstrates judgement under pressure.</i>	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Effective Communication	<i>Actively and clearly engages patients and colleagues in equal/open dialogue. Demonstrates active listening. Communicates verbal and written information concisely and with clarity. Adapts style of communication according to individual needs and context. Able to negotiate with colleagues & patients effectively.</i>	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Learning & Professional Development	<i>Demonstrates desire and enthusiasm for continued learning, takes responsibility for own development. Willing to learn from others and from experience. Is open and accepting of feedback. Demonstrates a desire and willingness to teach others.</i>	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Organisation and Planning	<i>Manages and plans workload effectively, displaying efficient time management and delivering tasks on time. Able to prioritise effectively and re-prioritise where appropriate. Is conscientious and maintains accurate records.</i>	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Patient Focus	<i>Ensures patient is the focus of care. Demonstrates understanding and appreciation of the needs of all patients, showing respect at all times. Takes time to build relationships with patients, demonstrating courtesy, empathy and compassion. Works in partnership with patients about their care.</i>	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐

Problem Solving and Decision Making	<i>Demonstrates an ability to assimilate a range of information and identify key issues. Engages with the wider issues and thinks creatively to solve problems and reach appropriate decisions. Is proactive and demonstrates initiative. Is able to attend to detail.</i>	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Self Awareness and Insight	<i>Demonstrates awareness of the boundaries of their own competence and willing to seek help when required, recognising that this is not a weakness. Exhibits appropriate level of confidence and accepts challenges to own knowledge.</i>	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Working Effectively as Part of a Team	<i>Capability & willingness to work effectively in partnership with others and in multi-disciplinary teams. Demonstrates a facilitative, collaborative approach, respecting others' views. Offers support and advice, sharing tasks appropriately. Demonstrates an understanding of own and others roles within the team and consults with others where appropriate.</i>	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐

Using the definitions given above, please rate each professional attribute on the scales provided below by indicating **ONE** number from 1 to 6 as to how important you feel it is to evaluate the professional attribute at selection.

Professional Attribute	Important to evaluate at selection					
	not at all -----desirable ----- essential					
Commitment to Professionalism	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Coping with Pressure	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Effective Communication	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Learning & Professional Development	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Organisation and Planning	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Patient Focus	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Problem Solving and Decision Making	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Self Awareness and Insight	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
Working Effectively as Part of a Team	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐

<i>Are there any additional professional attributes or comments you would like to add?</i>	

Thank you for your time

Appendix 7: Foundation Programme Commencing August 2010

Person Specification

	ESSENTIAL CRITERIA	DEMONSTRATED BY
Eligibility	Applicants must meet the requirements set out in the Foundation Programme 2010 Eligibility Criteria.	Eligibility checking
Qualifications	The applicant must have achieved, or expect to achieve, a primary medical qualification as recognised by the General Medical Council (GMC) by the start of the Foundation Programme 2010.	Eligibility checking
Clinical Knowledge & Skills	The applicant must be familiar with and be able to demonstrate an understanding of the major principles of the GMC's Good Medical Practice (2006) including: • Good clinical care • Maintaining good medical practice • Teaching and training, appraising and assessing • Relationships with patients, and can apply this understanding • Working with colleagues, and can apply this understanding • Probity • Health The applicant must demonstrate an understanding of the outcomes to be achieved in the Foundation Programme as set out in The New Doctor (2007).	Application/ pre-employment screening Clinical assessment (where appropriate)
Language &	The applicant must have demonstrable skills	Application/ pre-

Communication Skills	in listening, reading, writing and speaking in English that enable effective communication about medical topics with patients and colleagues, as set out in paragraph 22 of the GMC's Good Medical Practice (2006).	employment screening Clinical assessment (where appropriate)
Attributes	The applicant must demonstrate: • an understanding of the importance of the patient as the central focus of care; • the ability to prioritise tasks and information appropriately; • an understanding of the importance of working effectively with others; • the ability to communicate effectively with both colleagues and patients; • initiative and the ability to deal effectively with pressure and/or challenge • an understanding of the principles of equality and diversity	Application/pre-employment screening
Probity	The applicant must demonstrate appropriate professional behaviour, i.e. integrity, honesty, confidentiality as set out in the GMC's Good Medical Practice (2006). The applicant must have criminal records clearance at the appropriate level subject to prevailing UK legislation.	Application/pre-employment screening

Appendix 8: FY1 Literature Review Table

1. Knowledge, Skills and Abilities of the Year 1 Foundation Doctor

1.1 Good Clinical Care	• The Foundation Programme Curriculum (2010) • The New Doctor (2009) • The Foundation Doctor Person Specification (2011) • Tomorrow's Doctors (2003; 2009) • Good Medical Practice (2006)	F1 doctors must: • Be aware of personal and professional limits • Practice with professionalism (i.e. with integrity, compassion, altruism, respect of cultural and ethical diversity, regard to the principles of equity and aspire to excellence and continually improve) • Know and follow guidance set out in Good Medical Practice • Understand the importance of the patient as the central focus of care • Demonstrate that they are taking increasing responsibility for patient care (under supervision and with appropriate discussions with colleagues) • Demonstrate they recognise and manage acutely ill patients (under supervision) • Promote, monitor and maintain health and safety in the clinical setting • Manage their own time under supervision • Demonstrate that they are able to take appropriate action if their own health, performance or conduct or that of a colleague puts patients, colleagues or the public at risk. • Promote health and prevent disease • Know how errors can happen and how to manage risk • Be able to perform clinical and practical skills safely • Recognise personal and professional limitations and be willing to ask for help when necessary • Prioritise tasks and information effectively • Reflect on practice and be self critical • Use research skills to develop greater understanding • Solve problems, analyse and use numerical data	F1 doctors must know about the following subjects: • History taking • Examinations • Diagnosis and clinical decision-making • Safe prescribing • Medical record-keeping (e.g. medical notes, discharge letter/summaries, outpatient letters, prescriptions) • Time management and organisational decision-making • Quality of care and Patient safety • Infection control • Clinical governance • Nutritional care • Health promotion, patient education and public health • Ethical and legal issues • Palliative care • Principles of treatment • Health issues within communities • Patient-centred care • Principle of equality and diversity • Diagnosis and management of clinical presentations • Biomedical scientific principles (e.g. anatomy, biochemistry, cell biology, genetics, immunology, microbiology, molecular biology, nutrition, pathology, pharmacology and physiology)

	• Show initiative and deal effectively with pressure and/or challenge • Develop critical thinking and professional judgement • Offer assistance in an emergency • Prescribe drugs safely, effectively and economically • Apply biomedical scientific principles, psychological principles, social science principles and medical practice principles to practice as well as scientific methods and approaches to medical research.	
1.2 Maintaining Good Medical Practice	F1 doctors must: • Develop a portfolio that includes a variety of evidence relating to work-based assessments, involvement in educational and clinical teaching sessions and reflections on experiences with patients and colleagues. • Be able to explain how to contribute towards audit. • Internalise the importance of CPD and self-directed learning (demonstrated through effective assessments) • Keep up to date with the developments in their field • Be able to adapt to changing circumstances throughout their professional life • Be willing to respond constructively to the outcome of appraisal, performance review and assessment. • Seek out role models and learn from the behaviours of the best clinical practitioners and leaders	F1 doctors must know about the following subjects: • Self-directed life-long learning • Research, evidence and guidelines • Use of audit to improve patient care. • Working, organisational and economic framework in which medicine is practice. • Current developments and guiding principles in the NHS
1.3 Teaching and Training, Appraising and Assessing	F1 doctors must: • Teach their peers and other students (if required to do so) • Contribute to appraisal, assessment or review of students and other colleagues they work with • Develop the skills, attitudes and practices of a competent teacher • Use different techniques to record, organise and present information including computers and IT resources	F1 doctors must know about the following subjects: • Teaching responsibilities to colleagues, patients and relatives.
1.4 Relationships with Patients	F1 doctors must: • Ensure effective relationships with patients are established and maintained (e.g. creating an environment where the doctor can encourage and support the patient to share all information relevant to the consultation). • Introduce themselves to patients and colleagues with appropriate confidence and authority	F1 doctors must know about the following subjects: • How to develop and maintain successful relationships with patients - o Within a consultation - o When breaking bad news - o When dealing with complaints - o Communicating with patients with mental illness

	• Recognise that patients are knowledgeable about themselves and the effect their health has on their daily life (and they should be aware that relatives and carers often have to be included in the information given to patients). • Be able to communicate clearly, sensitively and effectively with patients and their relatives (e.g. listen to patients, ask for and respect their views about their health; keep them informed about the progress of their care; make sure patients are informed about how information is shared within teams) • Be sensitive and respond to needs and expectations of patients • Respect and uphold patients' rights to refuse treatment or take part in research • Accept the moral and ethical responsibilities involved in providing care (e.g. demonstrate sound knowledge of confidentiality) • Respect patients regardless of their lifestyle, culture, beliefs etc • Work in partnership with patients – give patients the information they want in a way they can understand and support patients in caring for themselves • Be non-judgemental	- o Communicating with patients with mental and physical disabilities • The rights of patients
1.5 Working with Colleagues	F1 doctors must: • Work effectively as a member of a multidisciplinary team (supporting others and efficient handovers) • Respect everyone they work with • Communicate in different ways (including spoken, written and electronic methods). • Communicate clearly, sensitively and effectively with colleagues from a ranges of healthcare services • Share information where necessary • Ensure safe continuing of care of patients on handover between shifts, on-call staff or with 'hospital at night' team • Listen and take account of other health professionals and agencies • Be able to demonstrate effective teamworking and leadership skills • Be willing to lead when faced with uncertainty and change • Be aware of work pressures on others and take responsibility	F1 doctors must know about the following subjects: • Communication with colleagues and teamwork • Interfacing with different specialties and with other professionals • Relevance of outside bodies

	to help reorganise workloads	
1.6 Probity	F1 doctors must: • Be honest in their relationships with patients, colleagues and employers • Be able to complete or arrange for the completion of legal documents correctly (e.g. certifying sickness and death) • Demonstrate knowledge of and be able to apply relevant legislation to their day-to-day activities • Act with integrity (e.g. act without delay if they have good reason to believe they or a colleague may be putting a patient at risk; never abuse a patient's trust).	F1 doctors must know about the following subjects: • Doctor-patient relationship • Health and handling stress
1.7 Health	F1 doctors must: • Demonstrate knowledge of their responsibilities to look after their health (e.g. maintain suitable work-life balance and knowing how to deal with personal illness to protect patients) • Be responsible for their own health in the interest of public safety	F1 doctors must know about the following subjects: • Health hazards of medical practice
1.8 Acute Health	F1 doctors must: • Have core skills in relation to acute illness	F1 doctors must know about the following subjects: • Management of acutely ill patients
1.9 Resuscitation	F1 doctors must: • Be able to recognise critically ill patients • Take part in advanced life support • Feel confident to initiate resuscitation • Lead the team where necessary • Use protocol for deciding when not to resuscitate patients	F1 doctors must know about the following subjects: • Resuscitation
1.10 Management of the 'take'	F1 doctors must: • Be able to function safely in an acute 'take' team	F1 doctors must know about the following subjects: • 'Take' management
1.11 Discharge Planning	F1 doctors must: • Be able to plan discharge for patients, starting from the point of admission and taking into account the effects of any chronic disease.	F1 doctors must know about the following subjects: • Discharge planning
1.12 Selection and Interpretation of Investigations	F1 doctors must: • Be able to select appropriate investigations and interpret results	F1 doctors must know about the following subjects: • Investigations

1.13 Practical Procedures		F1 doctors must: Be confident and competent to do common practical procedures needed to diagnose and manage adult patients and some common procedures for child patients.	F1 doctors must know about the following subjects: Practical procedures
2. Professional Competence	Epstein & Hundert (2002)	Highlight how there is no agreed-upon definition of competence that encompasses all the important domains of professional medical practice. Both physicians and trainees should have professional competence, which they define as “the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice for the benefit of the individual and community being served” (pp.226).	
	Stern et al (2005)	Found that conscientiousness and self assessment accuracy predict outcomes of professional competence in medical students as rated by faculty members.	
	Schon (1983)	Argues that professional competence is the ability to manage ambiguous problems, tolerate uncertainty, and make decisions with limited information.	
	Wright & Carrese (2001)	Looked at what values and attitudes physicians tried to pass on to medical trainees to encourage their professional development. The four main categories were Caring (encompassing empathy, compassion, caring, concern, understanding and consideration); Respect (encompassing patient dignity, patient preferences, being non-judgemental, respect, tolerance and confidentiality); Communication (encompassing body language, openness, sitting down when speaking with patients, silent observation, avoidance of confrontation and effective listening); and Integrity (encompassing trust and honesty).	
3. PRHO Job Role and Learning Outcomes	Bogg et al (2001)	90 PRHOs from three teaching hospitals in the Northwest of England completed a diary of work-based daily activities, a measure of perceived control in the work environment and a Burnout inventory. Eighteen were then interviewed as a follow-up to the questionnaires. Results showed the following breakdown of tasks: CME (took up 5% of their time); management aspects of their job such as delegation (57%), breaking bad news (37%); time management (43%); and routine administrative tasks (20%). More than 50% of junior doctors perceived that they required additional training in the technical and management aspects of the job prior to the pre registration year. Findings also indicated that 25% of participants were experiencing burnout – supporting other research that has suggested pre registration year is the most stressful period of medical practice (Kapur et al, 1999).	
	Stewart et al (1999)	Identified the tasks of preregistration doctors using the Delphi technique where a panel of experts take part in a series of rounds to identify, clarify, refine and gain consensus on a particular issue. The panel were clinical tutors from the Northern Deanery who had at least 2 years experience of supervising PRHOs. The final tasks were agreed upon by 50-95% of the experts and included: communication (e.g. establishing and maintaining good working relationships with other staff; giving advice on individual patient care, communicating information between hospital and community, talking to patients and relatives and handling difficult patient interactions); on call patient care (e.g. differentiate between simple and complex causes of symptoms, perform cardiopulmonary resuscitation); routine patient care (e.g. taking history, prescribing drug regimens); lab investigations (e.g. in liaison with more senior doctor decide which lab investigations are required, take samples); clinical investigations (e.g. in liaison with more senior doctor decide which clinical investigations are required, perform tests); practical procedures (e.g. bladder catheterisation, insert chest drains under supervision); self management (e.g. time management, knowing when to contact more senior doctor for help, understand roles of others).	
	Hesketh et al (2003)	Interviewed 57 PRHOs throughout Scotland regarding learning outcomes of the PRHO year. The following were mentioned by over 60% of those interviewed: Clinical skills (history taking, clinical examinations); Communication (with patients, relatives and healthcare professionals); Appropriate attitudes, ethics and legal (adapting to different disciplines); Appropriate decision-making skills, clinical reasoning and judgement (develop competence in decision making); Role of	

4. Deficiencies in Competency Levels of Medical Graduates Going into Foundation Year or Equivalent

the doctor within the health service (working in a team, adapting to different levels of support, accepting principles of collective responsibility, adapting to different working protocols); **Personal development** (be aware of limitations and seek help; manage time well, recognize self education/personal development as a lifelong process, accept/cope with increased responsibility, develop confidence, develop further competence from clinical experience on job, develop appropriate attitudes to personal health and well-being).

Hannon et al (2000)

Found that in a 25% random sample of doctors from five medical schools across Ireland, over 90% of interns considered themselves to be incompetent with respect to the skills and competencies required of an intern.

Burch et al (2005)

Used an OSCE process to assess the procedural skills of 58 South African graduates at the point of entry into the preregistration year (e.g. phlebotomy and bladder catheterisation). None of the students achieved an average score equivalent to reach proficiency in any one skill. Similar findings across other countries (e.g., Kowlowitz et al, 1990; Langdale et al, 2003; Joorabchi & Devries, 1996; Ringsted et al, 2001; Moercke et al, 2002)

Prince et al (2004)

Found that junior doctors felt confident about their skills and diagnostic knowledge but had difficulty coping with some of the other aspects of their work such as the sudden increase in responsibilities, lack of practical procedural knowledge, contact with other healthcare workers, and teamwork.

Goodfellow & Claydon (2001)

Report a lack of core clinical skills training in the PRHO year

Moss (1997)

Suggest that acquiring key skills in communication, team working and learning can be difficult due to PRHOs overwhelming involvement in service delivery.

Williams et al (2001)

Interviewed PRHOs about the doctor-patient relationship and found that most felt communication skills were important but difficult to develop due to both work overload and fears that seniors would perceive them as working too slowly.

Fox et al (2000)

Found that 22 PRHOs at the Whittington and Chelsea & Westminster Hospitals were deficient in basic clinical skills as measured by 2 OSCE stations. Furthermore, neither the PRHOs nor their educational supervisors were aware of these deficits.

5. Evaluating Undergraduate Courses in Relation to Trainees' Experiences as a Foundation Year Doctor/PRHO

Prince et al (2005)

Compared the responses of 1150 junior doctors regarding key competencies required for professional practice. These junior doctors had graduated from either a problem-based learning (PBL) college or a non problem-based learning college in the Netherlands. PBL college graduates tended to use expert knowledge, profession-specific skills and communication skills to a greater degree than non-PBL college graduates. All students rated **expert knowledge, profession-specific skills, communication skills, teamwork skills, independence and accuracy** to be of great importance in their current pre-registration year.

Lempp et al (2003)

Asked 16 PRHO to comment on their specific experiences in the first few months as a PRHO as part of an evaluation of a newly introduced Year 5 curriculum in a London medical school. Both rewards and difficulties emerged across responses. The majority found it stressful with **new found responsibilities** (e.g. stress was related to competence in emergency situations, ignorance of routine hospital procedures and clinical and administrative load; they also recognised their professional responsibility for patients, colleagues and the hospital). Most talked positively about their acquisition of a significant role within a **team**. The PRHOs had doubts they would ever reach the same competency levels of their seniors (i.e. **lacked confidence in their role**). Many commented on the adoption of **coping strategies** to deal with heavy demands and clinical tasks e.g., prioritising work and asking for help when needed. Almost all the PRHOs felt they had **improved relationships with patients** and some highlighted their **reluctance to delegate responsibilities**.

O'Neill et al (2003)

Evaluated the new curricula reform set out by the GMC's Tomorrow's Doctors (1993) by asking PRHOs (across two hospitals) to reflect on their 'problem-based' undergraduate course in relation to challenging incidents in their current role. Key challenges that PRHOs faced surrounded: **clinical situations** (e.g. dealing with complex medical problems); **limitations of competence** (e.g. knowing when to ask for help); **emotional involvement** (e.g. dealing with the death of a patient); and **communication** (e.g. breaking bad news). Fewer PRHOs who had been taught under the new curriculum brought up difficulties relating to managing uncertainty, recognising personal limitations and asking for help.

Brennan et al (2010)

Recognised that previous research evaluating aspects of the Tomorrow's Doctors curriculum largely relied on attitude questionnaires and whilst more recent research has used focus group and interviews, almost all of these studies were conducted in the context of the PRHO training system as opposed to the new Foundation years structure (see above for examples). This study subsequently aimed to ascertain how 31 medical graduates, across five hospitals, found the transition from medical school to foundation year using semi-structured interviews and an audio diary approach. Key experiences included: feelings of **stress and anxiety** concerning their levels of competence particularly regarding prescribing; having to **manage uncertainty** and **knowing when to ask for help**; **learning from prior experiences** helped ease the transition; **working with others**, particularly with senior doctors and the multidisciplinary team; and having to **deal with dying patients**. They conclude that the transition from medical school to foundation year is still considered highly stressful and students require more meaningful time with patients throughout their undergraduate years.

6. Frameworks used in

Simpson et al

Developed a set of 12 learning outcome domains that define the abilities of a competent medical student at the end of their

Other Countries for Equivalent Foundation Year Doctors

(2002) - Scottish
Domains

undergraduate training/competencies required as a PRHO. Carried out in collaboration with the Scottish Dean's Medical Curriculum Group:

Domain 1 - Outcomes for **clinical skills** (e.g. take a history from patients, relatives and others, undertake physical examinations)

Domain 2 - Outcomes for **practical procedures** (e.g. measuring and recording [radial pulse rate, blood pressure]; administering and doing [first aid, basic resuscitation])

Domain 3 - Outcomes for **patient investigation** (e.g. general principles of patient investigation i.e. appropriate choice and use of investigation)

Domain 4 - Outcomes of **patient management** (e.g. general principles of patient management i.e. patient-centred, holistic approach)

Domain 5 - Outcomes for **health promotion and disease prevention** (e.g. ability to implement, where appropriate, risk-reducing strategies for individual patients)

Domain 6 - Outcomes for **communication** (e.g. with patients, colleagues, media)

Domain 7 - Outcomes for **medical informatics** (e.g. IT/computing skills)

Domain 8 - Outcomes for **basic, social and clinical services and underlying principles** (e.g. causes of diseases and the ways in which these diseases affect the body)

Domain 9 - Outcomes for **attitudes, ethical understanding and legal responsibilities**

Domain 10 - Outcomes for **decision-making and clinical reasoning and judgement** (e.g. clinical reasoning; critical thinking)

Domain 11 - Outcomes for the **role of the doctor within the health service** (e.g., code of conduct and required attributes; the doctor as researcher)

Domain 12 - Outcomes for **personal development** (e.g. self awareness; self-learner; self care; career choice; motivation)

Accreditation
Council for
Graduate Medical
Education
(ACGME)

Endorsed six general competencies that postgraduate residents should demonstrate:

1. **Patient care** (e.g. provide patient care that is compassionate, appropriate and effective of health promotion)
2. **Medical Knowledge** (e.g. knowledge of biomedical, clinical, epidemiological and social-behavioural sciences)
3. **Practice-Based Learning and Improvement** (e.g. ability to investigate and evaluate their care of patients and improve patient care based on constant self evaluation and life long learning)
4. **Interpersonal and Communication Skills** (e.g. communicate effectively with patients, families and health professionals)
5. **Professionalism** (e.g. compassion, integrity, respect, accountability)
6. **Systems-based Practice** (e.g. awareness of larger context and systems of health)

CanMEDS Physician competency framework from Canada (2005)	CanMEDS is designed for multiple stakeholders including postgraduate educators and trainees. It describes the 7 key roles of the physician and the relevant competencies required within each role: 1. The Medical Expert (e.g. Function effectively as consultants, integrating all of the CanMEDS roles to provide optimal, ethical and patient-centered medical care) 2. The Communicator (e.g. develop rapport, trust and ethical therapeutics relationships with patients and families) 3. The Collaborator (e.g. participate effectively in an interprofessional healthcare team) 4. The Scholar (e.g. Maintain and enhance professional activities through ongoing learning) 5. The Health Advocate (e.g. Respond to individual patient health needs and issues as part of patient care) 6. The Manager (e.g. Participate in activities that contribute to the effectiveness of their healthcare organizations and systems) 7. The Professional (e.g. Demonstrate a commitment to their patients, profession, and society through ethical practice)

FY1 Literature Review References

Burch, V. C., Nash, R. C., Zabow, T., Gibbs, T., Aubin, L., Jacobs, B., & Hift, R. J. (2005). A structured assessment of newly qualified medical graduates. *Medical Education*, 39, 723-731.

Bogg, J., Gibbs, T., & Bundred, P. (2001). Training, job demands and mental health of pre-registration house officers. *Medical Education*, 35, 590-595.

Brennan, N., Corrigan, O., Allard, J., Archer, J., Barnes, R., Bleakley, A., Collett, T., & Regan de Bere, S. (2010). The transition from medical student to junior doctor: today's experiences of Tomorrow's Doctors. *Medical Education*, 44, 449-458.

Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. *JAMA*, 287, 226-235.

Fox, R.A., Ingham Clark, C.L., Scotland, A.D., & Dacre, J.E. (2000). A study of pre-registration house officers' clinical skills. *Medical Education*, 34, 1007-1012.

General Medical Council (2006). Good medical practice. Retrieved 18 September 2009, from http://www.gmc-uk.org/guidance/good_medical_practice/GMC_GMP.pdf

General Medical Council (2009). The New Doctor. Retrieved 13 May 2010, from http://www.gmc-uk.org/New_Doctor09_FINAL.pdf

General Medical Council (2009). Tomorrow's doctors. Retrieved 13 May 2010, from http://www.gmc-uk.org/education/undergraduate/tomorrows_doctors_2009.asp

Goodfellow, P.B., Claydon, P. (2001). Students sitting finals: ready to be house officers? *Journal of the Royal Society of Medicine*, 94, 516-20

Hannon, F. B. (2000). A national medical education needs' assessment of interns and the development of an intern education and training programme. *Medical Education*, 34, 275-284.

Hesketh, E. A., Allan, M. S., Harden, R. M., & Macpherson, S. G. (2003). New doctors' perceptions of their educational development during their first year of postgraduate training. *Medical Teacher*, 25, 67-76.

Hoffman, K., Hosokawa, M., & Donaldson, J. (2009). What criteria do faculty use when rating students as potential house officers? *Medical Teachers*, 31, e412-e417. Lempp, H., Cochrane, M., Seabrook, M., & Rees, J. (2003). Impact of educational preparation on medical students in transition from final year PRHO year: a qualitative evaluation of first year training following the introduction of a new year 5 curriculum in a London medical school. *Medical Teacher*, 26, 276-278.

Moss, F. (1997). The pre-registration year. *Medical Education*, 31 (S1), 10-11

- O'Neill, P.A., Jones, A., Willis, S.C., & McArdle, P.J. (2003). Does a new undergraduate curriculum based on *Tomorrow's Doctors* prepare house officers better for their first post? A qualitative study of views of pre-registration house officers using critical incidents. *Medical Education*, 37, 1100-1108.
- Prince, K.J.A.H., Van de Wiel, M.W.J., van der Vleuten, C.P.M., Boshuizen, H.P.A., & Scherpbier, A.J.J.A (2004). Junior doctors' opinions about the transition from medical school to clinical practice: a change of environment. *Education for Health*, 17, 323–31.
- Prince, K.J.A.H., van Eijas, P.W.L.J., Boshuizen, H.P.A., van der Vleuten, C.P.M., & Scherpbier, A.J.J.A. General competencies or problem-based learning (PBL) and non-PBL graduates. (2005). *Medical Education*, 39, 394-401.
- Schon, D.A. (1983). *The Reflective Practitioner*. New York, NY: Basic Books.
- Simpson, J.G., Furnace, J., Crosby, J., Cumming, A.D., Evans, P.A., Friedman Ben David, M., Harden, R.M., Lloyd, D., McKenzie, H., McLachlan, J.C., McPhate, G.F., Percy-Robb, I.W., & Macpherson, S.G. (2002). The Scottish doctor – learning outcomes for a medical undergraduate in Scotland: a foundation for competent and reflective practitioners. *Medical Teacher*, 24, 136-143.
- Stern, D.T., Frohna, A.Z., & Gruppen, L.D. (2005). The prediction of professional behaviour. *Medical Education*, 39, 75-82.
- Stewart, J., O'Halloran, C., Harrigan, P., Spencer, J. S., Barton, R., & Singleton, S. J. (1999). Identifying appropriate tasks for the preregistration year: modified Delphi technique. *BMJ*, 319, 224-229.
- The Department of Health (2010). *The Foundation Programme Curriculum*. Retrieved 15 May 2010, from <http://www.foundationprogramme.nhs.uk/pages/home/key-documents#fp2011ps>
- The Department of Health (2011). *The Foundation Programme – Person Specification*. Retrieved 7 May 2010, from <http://www.foundationprogramme.nhs.uk/pages/home/key-documents#fp2011ps>
- The Department of Health (2007). *The Foundation Programme: Rough Guide*. Retrieved 15 May 2010, from <http://www.ehealthnurses.org.uk/pdf/rgfp.pdf>
- The Royal College of Physicians and Surgeons of Canada. (2005). *The CanMEDS 2005 Physician Competency Framework*. Retrieved 5th May 2010, from <http://rcpsc.medical.org/canmeds/CanMEDS2005/index.php>
- Williams, C., Cantillon, P., & Cochrane, M. (2001). The doctor-patient relationship: from undergraduate assumptions to pre-registration reality. *Medical Education*, 35, 743-747.

Wright, S.M., & Carrese, J.A. (2001). Which values do attending physicians try to pass on to house officers? *Medical Education*, 35, 941-945.

Appendix 9: Situational Judgement Test - Literature Review Table

1. Definition

Christian et al (2010)	<i>"An SJT is a measurement method typically composed of job-related situations or scenarios that describe a dilemma or problem requiring the application of relevant knowledge, skills, abilities, or other characteristics (KSAOs) to solve" (pp.84).</i>
Banki & Latham (2010)	SJTs require respondents to mentally simulate that they are faced with the situation and select from a list of actions, the ones they would most likely and least likely perform.
Lievens, Peeters & Schollaert (2008)	SJTs are hypothetical work-based scenarios presented to applicants, who make judgements about possible responses, and these responses are then assessed against a pre-determined scoring key. Response alternatives can be presented in either a written (low fidelity) or a video-based (medium fidelity) format. Applicants are not asked to show any actual behaviour (high fidelity) unlike in assessment centre simulations.
Bergman et al (2006)	Scenarios are created on the basis of job analysis findings and critical incidents assessment. Applicants can be required to make their judgements in the format of choosing the best option, worst option, or asked to rate effectiveness of alternatives, and responses are typically scored by comparing them to the judgements of experts in the respective field.
McDaniels & Nguyen (2001)	To develop SJTs, it is common to collect data from incumbents or other subject matter experts (SMEs) regarding critical incidents (stories about situations encountered at work). Sometimes the SMEs are not offered any guidance on the critical incidents on the topic, in other situations they are directed to write items based on competencies derived from a job analysis.

2. Popularity

Motowidlo, Dunnette, & Carter (1990)	The popularity and use of situational judgement inventories has grown over the past 15 years because of the advantages associated with the use of SJTs in a selection context. SJTs tend to be relatively easy and inexpensive to develop, administer, and score.
Weekley & Ployhart (2005)	<i>"The increased popularity of SJTs is undoubtedly due to research showing these tests can have validity on par with cognitive ability tests, whereas usually producing lower levels of adverse impact and more positive test-taker reactions" (pp.82).</i>
Lievens et al (2008)	There are a number of benefits from using SJTs, in that they are relatively easy and cost-effective to develop, administer and score. SJTs are also versatile in their presentation, and can be written, web or video-based.
Landy (2007)	<i>"They seem to represent psychometric alchemy (adverse impact is down, validity is up), they seem to assess practically important KSAOs [knowledge, skills, abilities, and other characteristics], and assessees like them" (p. 418).</i>
Whetzel, MsDaniel & Nguyen (2008)	Highlight two reasons for the increasing popularity of SJTs: firstly they have useful levels of validity in terms of predicting job performance and secondly they are often viewed as both face and content valid as they describe work-related situations.

3. Relevant Theory

Motowidlo et al (1990)	SJTs predict job performance because of the behavioural consistency principle: past behaviour is the best predictor of future behaviour.
Motowidlo et al (1990)	SJTs measure an applicant's intentions and goals; SJT items will be more predictive when they are written to reflect the criterion in terms of target (what the behaviour is directed towards), action (actual behaviour), context (situational context surrounding the behaviour), and time (when the behaviour will take place) (TACT; Ajzen & Fishbein 1980).
Motowidlo et al (2006)	Introduced implicit trait policy, which posits that individual differences in personality traits affect judgements in situations that may require the expression of those personality traits. For example, if the responses to SJT items that express more agreeableness are truly more effective, those candidates who are high in agreeableness will weigh those responses more heavily than those low in agreeableness.

Whetzel & McDaniel (2009)	Personnel selection research typically emerged from an applied need to screen job applicants, therefore it tends to lack a broad theoretical basis.

4. Reliability

i. Internal Consistency

Whetzel and McDaniel (2009)	Highlight the inherent difficulty of estimating the reliability of SJTs. SJTs typically measure multiple constructs and are constructed to be heterogeneous at the item level (McDaniel & Whetzel, 2005). For example, the same item could be targeting both conscientiousness and/or intelligence. Due to this Cronbach's alpha is suggested to be an inappropriate reliability index (Cronbach, 1949, 1951). Test-retest and parallel form reliabilities are more appropriate but rarely reported in the literature and thus many researchers provide internal consistency estimates with or without acknowledging that they underestimate the reliability (e.g. Pulakos & Schmitt, 1996, Pulakos et al, 1996).

McDaniel, Morgeson, Finnegan, Campion & Breverman (2001)	Conducted a meta-analysis of 39 different SJTs and found the internal consistency coefficients across tests varied between .43 and .94.

Legree (1995)	Highlighted how the reliability estimates for SJTs are surprisingly modest given their length (.50-.80).

ii. Parallel Form

Chan & Schmitt (2002)	Estimated a parallel form reliability of .76 for their SJT.

Lievens, Buyse & Sackett (2005a)	Found a reliability of .66 between two alternate forms of their SJT which were developed for the medical admission exams in Belgium and completed by 250 introductory psychology students with a 1 week interval.

iii. Test Retest

Bledlow & Frese (2009)	Found adequate stability of responses over time for their SJT measuring personal initiative.

Ployhart & Ehrhart (2003)	Found test-retest reliability ranging from .20 - .92 depending on the types of response instructions used for their SJT (.20 = rate the effectiveness of each response and .92 = rate how likely you would perform each option).

5. Criterion related validity

Weekley & Jones (1997)	Found their two video-based SJTs correlated with performance criteria at .22 (using a sample from a retail company) and .24 (using a sample from a nursing home operator) and both of these increased to mid .30s after correcting for estimated criterion unreliability. They also reported incremental validity relative to experience and cognitive ability.

McDaniel et al (2001)	Conducted a meta-analysis and based on 102 validity coefficients, estimated that SJTs have an overall criterion-related validity of .34. This is comparable to ACs (Gaugler, Rosenthal, Thornton, & Benson, 1987), employment interviews (McDaniel et al, 1994) and biodata (Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990). When corrected for restriction of range and criterion unreliability, the validity of SJTs increased to .56, which "rivals the validity of the best predictors available to industrial-organisational psychologists in an employment situation" (Clevenger et al, 2001). McDaniel et al (2001) also found that those SJTs based on a job analysis had higher validities than those not based on a job analysis (.38 vs. .29).

McDaniel et al (2007)	In a more comprehensive meta analysis, the overall validity of SJTs across 118 coefficients was found to be .26 (N=24,756) regardless of instruction type.

O'Connell,	Found that SJTs have slight incremental validity over cognitive and personality measures and a composite of both when

Hartman, McDaniel, Grubb & Lawrence (2007)	predicting task performance (r change: .03, .04 and .01 respectively) but only had incremental validity over cognitive ability when predicting contextual performance (.04).
Hedlund et al (2001)	Revealed that a SJT (tacit knowledge inventory) had incremental validity over Graduate Management Admission Test scores among Masters of Business Administration students.
Lievens, Buyse & Sackett (2005b)	Examined the incremental validity of a SJT over traditional cognitively oriented measures in an operational student admission context. Their aim was to measure skills other than cognitive ability (i.e. interpersonal and communication skills). The SJT exhibited incremental validity over the cognitively oriented predictors for curricula with a substantial interpersonal skill component but not for curricula with a minimal interpersonal skill component. Interestingly, the SJT validity increased through the 4 academic years. This is consistent with research showing that non-cognitive predictors become more important when the criterion data are gathered later on (Goldstein, Zedeck & Goldstein, 2002; Jansen & Stoop, 2001). The authors conclude that their study demonstrates that SJTs can be a useful and valid complement to traditional student admission tests even in an operational high-stakes context.
Oswald et al (2004)	Their SJT had “ <i>potentially useful levels of criterion-related and incremental validity over and above SAT/ACT scores (cognitively oriented measures) and existing measures of the Big Five</i> ”, although it failed to predict college GPA.
Chan & Schmitt (2002)	They found that the SJT measured a stable individual difference attribute that was not associated with job experience and that was distinct from cognitive ability and the Big Five personality traits. The SJT had substantial validity in predicting task performance (core technical proficiency), contextual performance (job dedication, interpersonal facilitation) and overall job performance. Moreover, relative to the validity offered jointly by cognitive ability, all Big Five personality traits and job experience, the SJT provided substantial incremental validity in predicting performance dimensions (.24 for predicting core technical proficiency, .30 for job dedication, .17 for interpersonal facilitation, and .21 for overall job performance).
Lievens & Coetsier (2002)	Found their SJTs used in a Flemish medical admissions context correlated with students’ first year scores (.23). Furthermore, they found that out of the 4 SJTs they developed, those that had situations which the applicants had previously experienced had higher predictive validities than those that were unfamiliar (i.e. a video-based lecture and written text from a core text book over a physician-patient interaction and medical expert discussion).
Weekley & Ployhart (2005)	Their SJT had incremental validity over experience, ability and personality in terms of predicting performance. The authors conclude that the SJT is therefore not merely a proxy for experience, ability or personality but relates to performance for different reasons.
Borman, Hanson, Oppler, Pulakos & White (1993)	Developed a SJT of supervisory judgement using different situations in the armed services and found a correlation of $r=.22$ with supervisory ratings of performance using a sample of 570 supervisors in the military.
Patterson, Baron, Carr, Plint & Lane (2009)	Compared the validity of three shortlisting methodologies for selection into postgraduate GP training. An SJT was piloted alongside a structured application form (AFQ) and a clinical problem solving test (CPS) in the live selection process in 2006. All three tests predicted selection centre performance but the SJT served as the best single predictor, showing incremental validity over the AFQ and CPS. This is the first study to evaluate a machine-marked SJT to assess non-clinical domains for postgraduate selection.
Bledlow & Frese	Validated a SJT for measuring personal initiative (SJT-PI). Scores on the SJT-PI were significantly correlated with supervisor

6. Construct validity

(2009)	ratings of overall performance (.37) and personal initiative specifically (.48).
Mumford, van Iddekinge, Morgeson & Campion (2008)	Validated a SJT (called the Team Role Test; TRT) which looked at individual differences in team role knowledge. Scores on the TRT correlated significantly with peer ratings of team role performance in both academic and work settings.
Schmitt et al (2009)	In a 4 year longitudinal study carried out across 10 US colleges and universities, an SJT (addressing everyday practical problems a student may encounter both in and out of the classroom) was found to have incremental validity in predicting GPA college scores and absenteeism over and above high school GPAs and standardized tests of cognitive ability in the areas of verbal and mathematical skills (SAT/ACT; n=1165).
Patterson, et al (2009)	Carried out two validation studies across two consecutive years looking at the use of a critical problem solving test (CPS) and a SJT for selection into Core Medical Training (CMT) in the UK. The average uncorrected correlation with CMT interview scores and the SJT was .52. Furthermore the SJT (and the CPS) were found to have significant incremental validity over the traditional application form shortlisting process (17% for the SJT and 22% for the SJT and CPS combined).
Christian et al (2010)	Reviewed the literature regarding the construct validity of SJTs and claimed that there is a huge need to disentangle the effects of the measurement method from the constructs trying to be measured. They suggest the need for a construct-oriented paradigm for SJT research and this will depend upon researchers reporting detailed construct information. By doing this, it will be easier to compare different types of stimulus material (e.g. paper and pencil, computerized), different response modalities (e.g. written, oral) or different scoring strategies (subject matter experts, consensus based scoring) whilst holding the construct constant.
Weekley & Jones (1999)	Present two perspectives on SJTs, the first suggests that SJTs predict performance because they reflect the influence of other constructs that are themselves related to job performance e.g. cognitive ability, job experience. The second perspective suggests that SJTs measure a unique construct providing incremental validity over traditional predictors of performance. In their 2 studies they found that cognitive ability and job experience were related to their SJTs. They also found that their SJTs added incremental validity in the prediction of performance over cognitive ability and experience in one of their studies but not the other. The authors subsequently conclude that <i>"SJTs are multidimensional manifestations of several job-relevant skills and abilities...it is probably best at this point to conclude that SJTs represent a method and not a construct"</i> (pp.17).
Weekley & Ployhart (2005)	<i>"Although SJTs are increasingly popular for practical reasons, there remains much uncertainty as to why they actually work and what they actually measure."</i> (pp.82)
Clevenger, Pereira, Wiechmann, Schmitt & Harvey (2001)	Found their SJTs to have incremental validity over and above cognitive ability, personality tests, job knowledge and job experience. They propose that SJT items that require knowledge about problem solving, planning and decision making will be related to cognitive ability measures; Whereas those situations that require knowledge of dealing with difficult interpersonal situations will not be highly related to cognitive ability but more highly related to personality constructs.
Lievens et al (2005b)	Suggest that there is an emerging consensus that SJTs are essentially measurement methods that can be designed to measure a variety of constructs (both cognitive and non cognitive). They argue that researchers are moving away from general discussions of predictors as "valid" to considerations of "valid for what"? They consider it useful to consider the construct domain underlying a given SJT rather than viewing the SJT has a single entity. They also consider that expectations for the validity of a given SJT require attention to the criterion in question, for example, an SJT with an interpersonal skill focus would not be theoretically expected to predict a highly cognitively loaded criterion.
Sternberg,	SJTs form the basis for measuring tacit knowledge, a form of intelligence or 'street smarts'. Tacit knowledge is considered to be a

	Wagner, Williams & Horvath (1995)	construct independent of cognitive ability and is derived from experience.
	Lievens et al (2008)	Highlight the sheer number of different constructs that SJTs have been used to measure including: teamwork knowledge (McClough and Rogelberg, 2003; Morgeson et al., 2005; Stevens and Campion, 1999), aviation pilot judgment (Hunter, 2003), employee integrity (Becker, 2005), call-centre performance (Konradt et al., 2003), or academic performance (Oswald et al., 2004).
	Weekley & Ployhart (2005)	Using path analysis, they found that their SJT correlated with aspects of ability, experience, and personality that most related to managerial experience (e.g. cognitive ability, general work experience, tenure, job-specific training experience, conscientiousness, emotional stability and extroversion).
	Westring et al (2009)	Propose that construct validity can be better understood by isolating out the effects of the situational factors and individual differences directing participant responses. In their study they specifically designed a SJT of achievement goals (SJT-AG) to estimate individual and situational influences on the types of achievement goals students report they would adopt in a range of academic situations. Using structural equation modelling, they found that situational factors accounted for over three times as much variance as individual difference factors.
i. Personality	Oswald et al (2004)	SJTs have much in common with personality constructs because SJTs rely on individuals' subjective judgments of response-option appropriateness.
	McDaniel and Nguyen (2001)	Examined relations between SJT scores and personality characteristics according to a bare-bones meta-analytic strategy that did not make adjustments for unreliability. The aggregated correlations with SJT scores were .25 for agreeableness, .26 for conscientiousness, .31 for emotional stability, .06 for extroversion, and .09 for openness. Excluding results from one particularly large study that found exceptionally strong relations with agreeableness and conscientiousness, correlations aggregated over the remaining studies were .13 for agreeableness and .17 for conscientiousness.
	Mullins and Schmitt (1998)	Reported that an SJT was most strongly correlated with the Conscientiousness ($r = .26$) and Agreeableness ($r = .22$) factors of the NEO-PI-R (Costa & McRae, 1992).
	Smith and McDaniel (1998)	Found that their SJT was correlated with a measures of Conscientiousness ($r = .32$) and Neuroticism ($r = .22$). They did not find a relationship between their SJT and Agreeableness.
	Clevenger et al (2001)	Highlight that although the pattern of results is not completely consistent across previous studies investigating the relationship between personality and SJTs, there does appear to be sufficient evidence to suggest that SJTs are correlated with important aspects of personality. They suggest that <i>"because conscientiousness is also the personality construct most consistently and highly related with job performance (Barrick & Mount, 1991), the empirical validity of SJTs may be partially a function of their relationship with conscientiousness"</i> .
	Motowidlo, Hooper, and Jackson (2006); Motowidlo & Beier (2010)	They suggest that the procedural knowledge typically measured in an SJT is made up of implicit trait policies (ITPs) and specific job knowledge. ITPs refer to implicit beliefs about the costs and benefits of expressing certain personality traits in situations described in SJT items (e.g. knowing that generally being more agreeable is likely to be better across a range of situations). ITPs are more general than specific job knowledge which is more about effective action that may be unique to a given job (e.g., it is only once you have been in the job that you can learn when, for example, it may be more beneficial to be disagreeable in a certain work situation). As such, a response to an SJT question is determined by both of these types of knowledge: ITPs can compensate for when you have not had experience within the job and job knowledge can compensate for inaccurate ITPs. In 2010, Motowidlo and Beier expanded the model to include cognitive ability and its direct effect upon both ITPs and specific job knowledge. They

have found preliminary support for their model. Two key implications from this work are that an SJT can be valid for selecting novice applicants with no relevant job experience so long as the SJT items tap into their ITPs and a scoring key can be developed from effectiveness judgements by judges with no relevant job experience depending on three key conditions: 1. The SJT items must express personality traits that correlate with their effectiveness 2. Judges must have had experiences outside the specific job context that taught them the utility and value of various trait expressions 3. Judges must be able to accurately estimate the extent to which the SJT responses express those traits. However, they conclude that if experts are available, a scoring key based on their judgements is still preferential.

	Lievens & Coetsier (2002)	Found their SJT score significantly correlated with a composite score from the NEO-PI-R (.26) particularly the openness to experience scale.
	McDaniel et al (2007)	Meta analysis revealed that SJTs had incremental validity over personality by 6-7%.
ii Cognitive Ability	McDaniel et al (2001)	Meta-analysis which reported a mean correlation of SJTs and cognitive ability of .46. However, there was a great deal of variability in these correlations (nearly .0 to 1.0) and examination of moderators of the relationship indicated only modest success in accounting for this variability (e.g. SJTs based on a job analysis were more highly correlated with cognitive ability .50 vs. .38 and SJTs with more detailed questions were more highly related to cognitive ability .56 vs. .47)..
	Chan & Schmitt (2002) and Clevenger et al (2001)	Found SJT scores to be uncorrelated with cognitive ability
	Chan & Smith (2002)	Similar to Clevenger et al (2001), they suggest that the variability in estimates of the correlation between cognitive ability and SJT are due to the nature of the situations presented. Those that are more cognitive in nature (e.g. planning, organisational ability, analytic problem solving) will correlate higher with cognitive ability than those which refer to more interpersonal issues (e.g. leadership); in these situations personality will play a greater role.
	Lievens & Coetsier (2002)	Found their SJTs to have incremental validity over cognitive ability tests.
	McDaniels et al (2007)	Meta analysis revealed that SJTs provide incremental validity over cognitive ability varying from 3 to 5% and a small degree of incremental validity over both the Big 5 and cognitive ability (1 to 2%).
	Weekley & Jones (1997)	Found significant incremental validity of SJTs over cognitive ability.
	McDaniel & Whetzel (2005)	The observed validities of SJTs (.21, .25) are approximately the same as the observed validities of cognitive ability tests (about .25).
iii Job Knowledge	Schmidt (1994)	Suggest that SJTs are nothing more than tests of job knowledge.
	Weekley & Jones (1997)	Reported a correlation of .20 between a rationally scored SJT and job experience.
	Weekley & Jones (1999)	Reported a correlation of .23 between SJT scores and a broad measure of general work experience and a correlation of .02 between job tenure and the same SJT measure.
	Quiñones et al (1995) and	Argued that experience is a multifaceted construct and different operationalisations of it may result in different findings regarding SJTs and job experience (e.g. general experience vs. highly job-specific experience).

7. Fairness and adverse impact	Tesluk & Jacobs (1998)	
	Weekley & Jones (1999)	Suggest some SJTs may be more strongly related to general experience measures than to highly job-specific experience measures
	McDaniel & Nguyen (2001)	In their 'bare-bones' meta-analysis, job experience was found to have a small positive correlation with SJTs among all the correlates examined (mean $r=.05$). Despite such a small correlation, McDaniel and Nguyen (2001) claim that this provides evidence that SJTs are measures of job knowledge. However, they also suggest that correlations with job experience would be greater with less experienced applicant pools (as one learns most of the job knowledge needed for job performance in the early years of experience and each additional year contributes less and less job knowledge).
	Clevenger et al (2001)	In their study, correlations between the SJT and job experience were all low and non-significant. The authors argue that previously found correlations between job experience and SJTs may be influenced by the variability in experience as well as the types of previous experiences. Likewise different types of job knowledge may be targeted in SJTs, for example technical job knowledge vs. contextual job knowledge.
	Weekley & Ployhart (2005)	Conclude that their SJT reflects general forms of knowledge as opposed to job specific forms of knowledge (as there SJT completely mediated the effects of general work experience and cognitive ability whilst remaining unrelated to training experience and job tenure).
	Lievens et al (2005b)	Their SJT measuring interpersonal skills in a physician-patient interaction was found to have predictive validity even though it was administered to students who had never conducted an interview with a patient. They therefore highlight how this finding is relevant in the context of the common assumption that SJTs are primarily measuring job knowledge or experience. The authors argue that what is involved here is not job-specific knowledge but, rather, more general knowledge of effective behaviour in interpersonal settings.
	Motowidlo et al (1990)	Found standardized mean differences on an SJT between African Americans and Caucasians of .14 and .29 in two different samples.
	Motowidlo & Tippins (1993)	Reported a difference on a SJT between African Americans and Caucasians approximately equal to one third of a standard deviation. They also found that women outperformed men with effect sizes ranging from .11 to .32.
	Pulakos & Schmitt (1996)	Reported a standardised mean difference on an SJT of .41 for African Americans and Caucasians and .02 for Hispanic Americans and Caucasians.
	Hough et al (2001) and Weekley & Jones (1999)	Both report moderate subgroup differences favouring Whites (d 's ranging from .52 to .85) and smaller subgroup differences favouring women (d 's ranging from .19 to .31).
	Nguyen, McDaniel & Whetzel (2005)	Found a difference in mean scores between White and Black candidates of .38 of a standard deviation in favour of White candidates and a difference in mean scores between males and females of .10 standard deviations in favour of females.
	Lievens & Coetsier (2002)	Found that women scored significantly higher than men in their SJTs. However, these differences were diminished when the SJTs were paired with cognitive ability measures; the gender differences are balanced out as men typically perform better on cognitive ability tests.
	O'Connell,	Found subgroup differences between Black and White participants of .38 in favour of White participants. They also found a

	Hartman, McDaniel, Grubb & Lawrence (2007)	gender difference of .27 in favour of females.
	Banki & Latham (2010)	Found no adverse impact by sex on their SJT.
	Clevenger et al (2001)	Analyses of subgroup differences for the SJT were encouraging, in that, relative to other predictors (cognitive ability, job knowledge) the size of such differences was relatively small. Only the Conscientiousness measure displayed evidence of smaller subgroup differences than the SJT. The authors suggest that SJTs can provide validity comparable with cognitive ability tests, yet produce smaller subgroup differences and more favourable applicant reactions.
	Jensen (1998)	Differences in mean SJT scores across race are typically smaller than cognitive ability tests.
	Oswald et al (2004)	Racial subgroup mean differences were much smaller on the SJT than they were on the standardised tests and GPA
	Weekley & Jones (1999)	Gender differences could be because of the interpersonal nature of SJTs.
	Lievens et al (2005b)	Suggest that SJT research in personnel selection has generally found lower subgroup differences for SJTs than for cognitive ability tests. SJTs with a lower cognitive loading seem to have less adverse impact than SJTs that are more g-loaded.
	Whetzel et al (2008)	Provided a systematic review of mean race and sex differences in SJT performance. Their findings revealed that, on average, White test takers performed better on SJTs than Black ($d = .38$), Hispanic ($d = .24$) and Asian ($d = .29$) candidates. Female candidates performed slightly better than male candidates ($d = -.11$). They also looked at two moderators: cognitive and personality loadings and response instructions. Typically race differences were explained by the cognitive loading of the SJT i.e. those SJTs that had higher correlations with cognitive ability had larger mean race differences. Regarding the effect of personality loadings of SJTs on race differences, Black-White and Asian-White differences were smaller to the extent that SJTs were correlated with emotional stability and Hispanic-White differences were smaller to the extent that SJTs were correlated with conscientiousness and agreeableness. Cognitive loading had minimal impact on gender differences, however the more the SJT correlated with agreeableness and conscientiousness, the greater the sex differences favouring women. Regarding the effect of the response instructions, knowledge instructions produced larger race and gender differences favouring White participants and women (due to higher correlations with cognitive ability).
8. Candidate reactions	Kanning et al (2006)	In two studies, the authors investigated the subjects' reactions to different kinds of situational judgment items. The item formats differed with regard to two variables: interactivity (non-interactive: in each item, the subject is confronted with a new situation vs. interactive: the situation develops according to the answer given and the subject is asked once more about the new situation) as well as modality of presentation (stimulus and response components of the items are given in the form of a video vs. in text form). The authors expected the degree of interaction and the presentation via video to have a positive effect on the subjects' evaluation (in terms of usefulness, emotional reaction, transparency, job-relatedness, acceptance and fairness). In accordance with the researchers' expectations, interactive situational judgment items using videos in the stimulus as well as in the response component received the highest ratings.
	Chan & Schmitt (1997)	SJTs were perceived as favourable.

9. Test format

Olson-Buchanan & Drasgow (2006)	Video-based SJTs help us “better understand how the assessee will interpret verbal and nonverbal behaviours of others in the workplace and choose to respond” (p.253).
Weekley & Jones (1997)	There are lots of advantages to using video-based SJTs – they can depict richer and more detailed behavioural incidents increasing their fidelity; they require no reading so provide no means for assessing candidates who would otherwise perform poorly on written tests; watching a video may be more natural for many people; they may provide a more realistic job preview; they can be administered to a large group; have high face validity. However, there are some downsides, for example they are costly to develop and may insert irrelevant contextual information and unintentionally bring more error into SJTs.
Chan & Schmitt (1997)	Conducted a laboratory experiment, showing that a video-based SJT had significantly less adverse impact than a written SJT. The authors suggest that this can partly be explained by subgroup differences in both reading comprehension ability and face validity perceptions (with Black candidates perceiving the written SJT to have lower face validity than White candidates).
Richman-Hirsch et al (2000)	Found that students reacted more favourably to a multimedia format of a conflict resolution skill SJT as compared with a computerized and written format of the same test (in a laboratory study). The students found the multimedia SJT more face valid, more enjoyable and more modern than the other two versions.
Lievens & Sackett (2006b)	Examined whether video-based SJTs have higher predictive validity than written SJTs (keeping verbal content constant). The samples consisted of 1,159 students who completed a video-based version of an SJT and 1,750 students who completed the same SJT in a written format. The study was conducted in a high stakes testing context in Belgium. The video-based version of an interpersonally oriented SJT had a lower correlation with cognitive ability than did the written version. It also had higher predictive and incremental validity for predicting interpersonally oriented criteria than did the written version. Interestingly, there were no significant differences between the face validity perceptions of the video-based and written SJTs contrary to prior laboratory studies (e.g. Chan & Schmitt, 1997; Richman-Hirsch et al., 2000).
Siu & Reiter (2010)	Argue that since SJTs aim to provide additional measures of non-cognitive qualities, the written SJT has less added value compared to the video-based SJT.
Whetzel & McDaniel (2009)	One must weight up the cost of video-based SJTs (e.g. cost of actors, videographers, studios etc) against the predictive and incremental validity over cognitive ability in predicting performance in high stakes settings.
Funke & Schuler (1998)	Compared various stimulus components (video, orally-presented questions) and response components (multiple choice, written, orally-given replies) of SJTs looking at occupational social competency. They found that the stimulus component alone does not account for additional validity when holding the response format constant. Rather, it is the response format that is most important, with higher fidelity response formats leading to higher validities (e.g. orally-given responses). Written response formats (compared to multiple choice and orally-given responses) and orally-presented questions (compared to video stimulus) were also more difficult for participants (n=75 students).

10. Response instructions

McDaniel & Nguyen (2003)	Highlight the range of response instructions that can be used for SJTs: • What would you most likely do? • What would you least likely do? • What would you most likely do? And then What would you least likely do? • Pick the best answer • Pick the best answer and then pick the worst answer • Rate each response on effectiveness “What would you most/least likely do” and “Pick the best/worst answer” are the most commonly used in SJT literature.

Ployhart & Ehrhart (2003)	Found that response instructions had a large effect on SJT reliability and validity. In general, behavioural tendency instructions (i.e. asking what one 'would do') showed increased criterion-related validities (relating to GPA, self ratings and peer ratings) and higher test-retest reliabilities compared to knowledge based instructions (i.e. asking what one 'should do'). The response instructions did not have a large impact upon internal consistency reliabilities.
McDaniel et al. (2007)	Provide a comprehensive review on the moderating effects of response instructions. They found that SJTs requiring knowledge responses measure maximal performance. Similar to cognitive ability, knowledge questions ask candidates whether they know what the most effective answer is (knowledge response formats are more highly correlated with cognitive ability and less highly with personality measures; McDaniel, Hartman & Grubb, 2003). On the other hand, SJTs requiring behavioural responses measure typical performance as they require candidates to report how they typically behave, similar to non cognitive inventories (behavioural tendency formats have been shown to correlate less highly with cognitive ability and more highly with personality – particularly agreeableness, conscientiousness and emotional stability; McDaniel, Hartman & Grubb, 2003). SJTs with behavioural instructions are saturated with non-cognitive variance and so are susceptible to self deception and impression management. Therefore the construct validity varies as a function of the response instructions. When SJT content is held constant, McDaniel et al (2007) make a tentative conclusion that predictive validities tend to be higher with knowledge instructions (which is in contrast to Ployhart and Ehrhart's (2003) findings). The authors advise that test developers of SJTs should use behavioural tendency instructions if they wish to emphasize the assessment of personality constructs and those wishing to maximise cognitive ability should use knowledge instructions.
Nguyen & McDaniel (2003)	Found that subgroup differences existed in both behavioural tendency and knowledge based instructions with the behavioural tendency instructions producing slightly less subgroup differences. They also found that the knowledge based instructions had a higher cognitive loading explaining why this produced slightly more subgroup differences.
Motowidlo, Crook, Kell & Naemi (2009)	Developed a SJT with a single response option (i.e. rate the effectiveness of each incident on a 7 point scale from 1 = 'very ineffective' to 7 = 'very effective') as opposed to a multiple response option (i.e. select the best-worst option or the option you would most/least likely perform). The SJT was designed with three scales: work effort, professionalism and personal skills. The work effort scale was the only one to relate to supervisory ratings of work effort. The authors conclude that a single response SJT format might be a useful alternative for test developers.
Lievens, Sackett & Buyse (2009)	Highlight the limitations of previous work that has explored the effects of response instructions; namely the use of incumbents who are in a familiar domain and have little motivation to fake. They carried out their study in a high stakes medical admissions process in Belgium wherein applicants were assigned papers with behavioural tendency instructions or knowledge based instructions. Firstly they found no difference in mean scores across both sets of responses sets. Secondly, despite the SJTs with knowledge based instructions correlating more highly with cognitive ability, this was a very small effect and thus any effects upon adverse impact would also be small. They also found no difference in criterion-related validity across instruction sets. The authors conclude that <i>"Given the minimal differences between formats, we suggest that knowledge instructions should be preferred, because they make faking a nonissue, and thus do not create a moral dilemma (should I fake or not?) for applicants"</i> (pp.1100).
11. Susceptibility to faking	Nguyen, Biderman, & McDaniel (2005) They administered a SJT with both knowledge and behavioural tendency instructions to a sample of students who had to complete the measures under 'honest' and 'fake good' conditions. Findings showed that SJTs with behavioural tendency instructions were fakable (average d = .25). However, the fakability of the knowledge based instructions only existed when students completed the 'faking good' condition before the 'honest' condition. The authors conclude that knowledge based SJT instructions <i>'are relatively immune from faking'</i> .

12. Coaching /Practice effects	Peeters & Lievens (2005):	Studied the fakability of a SJT using college students. Their SJT consisted of 23 items related to student issues (e.g. teamwork, studying for exams, organizing, accomplishing assignments, interpersonal skills, social responsibility, perseverance, and integrity) and students were asked how they would respond (a behavioural tendency instruction). Their results showed that students in the fake condition (N=153) had significantly higher SJT scores than students in the honest condition (N=138; and this had a particularly large effect size $d = .89$). They also found that the criterion-related validity in the faking group was significantly lower ($r = .09$) than in the honest group ($r = .33$). Furthermore in the faking group, the SJT did not add any incremental validity over GPA, whereas it added a further 5.1% explained variance in the honest group.
	Lievens & Peeters (2009)	When 249 students were asked to elaborate on their responses to an SJT related to college student success (i.e. provide a short reason why they had chosen their option out of the list of alternatives), those who were asked to fake their responses (by getting the highest score they could) had lower overall SJT scores compared to those who were not asked to elaborate. Furthermore, this effect was only apparent with familiar items (i.e. those items where students had encountered the situation in the past). These findings therefore confirm that prior familiarity with items is an important factor in determining the effects of elaboration on mean scores. Familiar items were also more easily faked than non familiar items. Lievens and Peeters (2009) conclude that elaboration may be a useful strategy to decrease faking when SJTs are primarily made up of familiar items (e.g. for licensing purposes for job incumbents or job seekers with considerable experience). In contrast, elaboration is less useful for SJTs with non-familiar items (e.g. use of SJTs for college admissions or training contexts). Finally, it was found that elaboration received more favourable participant reactions in terms of allowing them to present themselves more accurately and show their KSAs.
	Hooper et al (2006)	Reviewed existing research on faking and SJT scores. They concluded that SJTs might be prone to faking good ($d = .08-.89$). They suggest that this large variability could be explained by moderating variables such as SJT instructions, the constructs measured, the transparency of the items and the study design used. They also concluded that SJTs were less fakable than personality inventories.
	Vasilopoulos, Reilly & Leaman (2000)	SJT test takers who have relevant job knowledge and/or experience have been found to be better able to fake than inexperienced test-takers.
	Kanning & Kuhne (2006)	Studied the effect of social desirability on SJT items. They included an 11-item SJT using behavioural tendency instructions and a measure of social desirability in a test battery designed to select police officers. They used three groups consisting of real applicants, students instructed to present a positive impression, and students instructed to respond honestly (N=55 for each group). They found that the means were nearly equal in the two student samples (honest and faking). Further, the number of students who "passed/did not pass" the battery was not that different between the honest vs. faking condition.
	Whetzel & McDaniel (2009)	In their review of the literature they conclude that typically those instructed to fake can – they subsequently suggest using knowledge based instructions.
	Oswald et al (2004)	In their study, their SJI was correlated with self-deception and impression-management measures of social desirability. Whether this correlation is performance related (i.e., students who are better at impression management also do better in their academic work) or this is a form of bias has been debated in the employment literature; Oswald et al (2004) propose that most likely both are partial explanations. They argue that the relevant question is not whether students can be coached to get better scores on the tests (they believe they can) but whether the higher scores invalidate the utility of these measures and whether researchers and practitioners can take steps to minimize the effects from coaching or faking these measures.
	Cullen, Sackett & Lievens	Examined the coachability of two SJTs (the College Student Questionnaire; CSQ; and the Situational Judgement Inventory; SJI). They developed a 'coaching course' for these two SJTs identifying test taking strategies. They found that scores increased for the

	(2006)	CSQ but decreased for the SJI (compared to controls). In an attempt to understand these findings, they administered surveys asking participants about the level of effort required to understand the test taking strategies, and those for the SJI were considered more difficult. They concluded that SJTs, typically constructed from SME judgements or empirical keying rather than defined rules are less coachable.
	Lievens, Buyse & Sackett (2005b)	Looked at the effects of re-testing in an operational setting, namely a Belgian medical admissions context. They categorized retest effects into four categories: 1) measurement error, 2) true change in construct of interest, 3) criterion-relevant change (e.g. familiarity of the test) and 4) criterion-irrelevant change (e.g. tricks) and explored both within and between subject effects. They found significant retest effects for the SJI. However, using various techniques, they found that the change could be categorised in both 2) and/or 3) because retaking the test did not result in any significant change to validity.
13. Developing alternate forms	Lievens & Sackett (2007)	Compared three different methods from item generation theory to decide how similar an SJI needs to be to provide consistent measurement across alternate forms. They conducted the study in a high stakes setting in Belgium and employed the 'radicals-incidentals approach' to identify three different perspectives about key features of SJTs (radicals are structural item features that determine difficulty and incidentals refer to surface characteristics like wording or grammar which have no effect upon difficulty). The first perspective (random assignment) assigns the targeted domain/construct as a radical and therefore only holds the targeted domain constant across SJTs changing all other features – this produced very low consistencies across the SJTs. The second approach (incident isomorphic) assigns the domain and critical incidents built into the items (CIs) as radicals – this also produced low consistencies across SJTs. The third approach (item isomorphic) assigns the domain, CIs and the context within which the CIs and responses are embedded as radicals (e.g. stem wording and response options). They found empirical support for this perspective, however, they warn that the item isomorphic approach could cause retest effects. Consistent with item generation theory, none of the alternate form strategies produced a change in construct or criterion-related validity.
14. Scoring SJTs	Lievens, Peeters & Schoolaert (2008)	Describe the various scoring methods for multiple choice SJTs. They are typically broken down into 'rational' and 'empirical'. Rationally scored items are based on experts' judgements concerning the effectiveness of responses or best and worst options. When items are empirically scored, they are administered to a large pilot sample. Items are selected and/or weighted according to evidence that they differentiate between individuals who score at different levels on a criterion variable (e.g. job performance). Options that are scored relatively often by high performing individuals are scored as correct etc.
	Bergman, Drasgow, Donovan, Henning, & Juraska (2006)	Looked at 11 different types of scoring using a leadership skills assessment SJI including empirical scoring (items or options are scored according to their relationships with a criterion measure; Hogan, 1994), theoretical scoring (items are constructed to reflect theory, or theory can be used to identify the best and worst options) expert-based scoring (subject matter experts make judgement about the items) and hybridized scoring (a combination of different scoring methods). Results showed that four keys accounted for additional variance in leadership ratings beyond cognitive ability and personality: empirical key (2.3%); hybrid initiating structure (3.0%); hybrid participation, (1.7%); and SME, (4.9%). They also suggest that as the length of a SJI gets longer, the scoring could get more complex. They concluded that the validity of a SJI depends in part on its scoring and that poor choices could lead to the conclusion that SJTs are not valid when it may actually be down to the scoring key that is not valid.
	Legree, Psotka, Tremble & Bourne (2005)	Suggest Consensus Based Measurement (CBM) may be a suitable alternative to expert-based scoring. Their CBM procedure allowed examinee responses to be scored as deviations from the consensus defined by the response distributions of the examinee sample. They reason that "...knowledge domains may exist that are lodged in opinion and have no objective standard for verification other than societal views, opinions, and interpretations." (p. 103). In their earlier work (Legree, 1995; Legree & Grafton, 1995), they compared expert ratings and mean ratings across examinees using an SJI and found a correlation of .72 between the two scoring methods. Similarly high correlations between experts and examinees have been found (e.g. Hedlund et al, 2003).

Whetzel & McDaniel (2009)	In their review of the literature, they conclude that different scoring keys can lead to a SJT assessing different constructs, although more research is needed. They also suggest using a Likert scale to rate each response for effectiveness by recoding responses into dichotomies such that responses are either effective or ineffective.
Phillips (1992, 1993)	Found that due to experts often not agreeing, a number of experts opinion is required to produce a credible scoring key, for example 75% of 20 experts must agree that a specific response is 'most appropriate' for that scenario to be included in a SJT.
Schmitt & Chan (2006)	Suggest using a scoring guide rather than a scoring key because real-world job demands often have multiple solutions that go well beyond technical task knowledge.
Legree (1995)	Found that using a Likert scale as opposed to a forced choice method increased the reliability of their SJT.
Weekley & Jones (1997)	Successfully used customers to develop a scoring key for their video-based SJT implemented in a nursing home organisation.
Schubert et al (2008)	Concluded that the best way to validate 'correct' responses on a professional behaviour SJT was to take "a multidimensional approach including a sound theoretical concept, clearly defined educational objectives, the definition of the level of available evidence for each question, and an expert survey...(which) keeps experts from assigning the same rank to more than one answer...and the possibility for the experts to comment on each question." (pp.533)
Wang et al (2009)	Used latent class analysis (LCA) to ensure the validity of expert scoring for their teamwork SJT. As described by Wang et al (2009), "LCA identifies qualitatively different groups of cases based on consistencies in response patterns. An LCA of SJT items can determine whether there are distinct groups of test takers showing different patterns of response on the SJT. If discrete groups of test takers have higher or lower expert-derived scores, this provides evidence that expert scoring is valid" (pp.111).
Ployhart & Erhart (2003)	Employed three different scoring methods: (1) For the SJTs that asked participants to pick one response (e.g. what option would you choose?) the item was scored a '1' if it matched the best response, a '-1' if it matched the worst response and a '0' if their response was one of the other two options. (2) For those SJTs that asked to give a good and bad response (e.g. what would you most/least likely do?), the good response was scored a '1' if it matched the best response, a '-1' was given if it matched the worst response and a '0' was given to all other options. (3) Finally, for those SJTs asking for ratings of each item on a 1 to 5 scale, the score for the best response was added to the score for the worst response. These different types of scoring methods affected the internal consistency: the last scoring method (ratings) led to the highest internal consistencies (.65 to .73). The second scoring method (most/least likely) had lower internal consistencies (.30-.65), and the first scoring method (choose one response) led to the lowest internal consistencies (.24-.65).

15. General Issues

i. Research vs. operational settings

Lievens, Buyse & Sackett (2005b)	Highlight how the vast majority of studies are concurrent in design and do not involve the use of SJTs in operational settings (i.e. SJTs were not used for making actual selection decisions, and no predictive validation design was used). In the meta-analysis of McDaniel et al (2001), for example, only 6 studies used a predictive validation design, whereas 96 used a concurrent design.
Whetzel & McDaniel (2009)	They also highlight how most SJT validity studies rely on concurrent designs which involve respondents who are incumbents and therefore typically have little motivation to distort their responses compared to job applicants who may be motivated to fake their responses (Lievens and colleagues' research provide a key exception). Therefore, because SJT research primarily use concurrent studies, it is possible that some of the conclusions may not hold for SJTs used to screen job applicants. They suggest that response instructions may become less important because some applicants may ignore the behavioural tendency instructions and report the most desirable behaviour (which is what Lievens et al (2009) found when using a SJT in a high stakes selection context).
MacKenzie et al	Found that cognitive ability was more strongly related to SJT scores in incumbent contexts than applicant contexts.

	(2010)	
	Ployart, Weekley, Holtz & Kemp (2003)	Looked at whether there was an effect of context (applicant vs. incumbent) upon the psychometric properties of a SJT. Found that the applicant context showed less normal distributions, higher means and less variance. They conclude that applicants typically indicate what they think is the best answer whereas incumbents draw upon what they did in the past.
ii. SJTs vs ACs	Lievens et al (2008)	SJTs are known to have lower 'physical fidelity' than Assessment Centres or work sample assessments as they do not allow candidates to behave in a way they would in the actual job. (However, criterion-related validities are often comparable; McDaniels et al, 2001).
iii. Cross cultural applicability	Banki & Latham (2010)	Their SJT was found to have criterion-related validity in an Iranian company. The authors support its global applicability, whilst emphasizing the need to tailor the questions and scoring key to the societal culture.
iv. Use in non-medical high stake settings	Hunter (2003)	Developed a SJT to target pilot judgement. Their SJT had good internal consistency and significantly correlated with the Hazardous Event Scale (HES) which measures the number of times a pilot has experienced an accident or a hazardous in-flight event during a specified period. Their study also found equivocal findings across pencil and paper and internet-based administrations.
v. Proctored paper and pencil vs. internet administration	Ployhart, Weekley, Holtz & Kemp ((2003)	Compared a web-based SJT, personality measure and bio data measure to a paper and pencil administration of the same tests in a selection context. They found that the web-based measure tended to have more normal distributions, lower means, higher internal consistency reliabilities, more variance and higher correlations than the paper and pencil measures.

References for Situational Judgement Literature Review

Ajzen, I., & Fishbein, M. (1980). *Understanding Attitudes and Predicting Social Behavior*. Englewood Cliffs, NJ: Prentice- Hall.

Banki S., & Latham G.P., (2009). The Criterion-Related Validities and Perceived Fairness of the Situational Interview and the Situational Judgment Test in an Iranian Organisation, *Applied Psychology: An International Review*, 59, 124–142.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, 14, 223–235.

Bledow, R., & Frese M., (2009). A Situational Judgment Test of Personal Initiative and its Relationship to Performance. *Personnel Psychology* 62, 2, 229.

Borman, W.C., Hanson, M.A., Oppler, S.H., Pulakos, E.D., & White, L.A. (1993). Role of early supervisory experience in supervisor performance. *Journal of Applied Psychology*, 78, 443-449.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, 82, 143-159.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance*, 15, 233-254.

Christian, M.S., Edwards, B.D., & Bradley, J.C. (2010). Situational judgement tests: constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, 63, 83-117.

Clevenger, J., Pereira, G.M., Wiechmann, D., Schmitt, N., & Harvey, V.S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology*, 86, 410-417.

Cronbach, L. J. (1949). *Essentials of psychological testing*. New York: Harper & Row.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.

Cullen, M.J., Sackett, P.R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment*, 14, 142-155.

Funke, U., & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment*, 6, 115-123.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Benson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, 72, 493-511.

Goldstein, H. W., Zedeck, S., & Goldstein, I. L. (2002). *g*: Is this your final answer. *Human Performance*, 15, 123-142.

Hedlund, J., Plamondon, K., Wilt, J., Nebel, K., Ashford, S., & Sternberg, R. J. (2001, April). *Practical intelligence for business: Going beyond the GMAT*. In J. Cortina (Chair), *Out with the old, in with the new: Looking above and beyond what we know about cognitive predictors*. Symposium presented at the 16th Annual Convention of the Society for Industrial and Organizational Psychology, San Diego, CA.

Hooper, A.C., Cullen, M.J. and Sackett, P.R. (2006). Operational Threats to the Use of SJTs: Faking, coaching, and retesting issues. In: Weekley, J.A. and Ployhart, R.E. (eds), *Situational Judgment Tests: Theory, measurement and application*. Mahwah, NJ: Lawrence Erlbaum, pp. 205-232.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment*, 9, 152-194.

Jansen, P. G. W., & Stoop, B. A. M. (2001). The dynamics of assessment center validity, results of a 7-year study. *Journal of Applied Psychology*, 86, 741-753.

Jensen, A.R. (1998). *The G Factor: The Science of Mental Ability*, Praeger, Westport, CT

Kanning, U. P., & Kuhne, S. (2006). Social desirability in a multimodal personnel selection test battery. *European Journal of Work and Organizational Psychology*, 15, 241-261.

Landy, F. J. (2007). The validation of personnel decisions in the twentyfirst century: Back to the future. In S. M. McPhail (Ed.), *Alternate validation strategies: Developing and leveraging existing validity evidence* (pp. 409-426). San Francisco: Jossey-Bass.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor. *Intelligence*, 21, 247-266.

Legree, P. J., & Grafton, F. C. (1995). *Evidence for an interpersonal knowledge factor: The reliability and factor structure of tests of interpersonal knowledge and general cognitive ability*. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences Technical Report No. 1030).

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. R. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze & R. D. Roberts (Eds.), *Emotional Intelligence: An International Handbook* Cambridge, MA: Hogrefe & Huber.

Lievens, F., Buyse, T., & Sackett, P. R. (2005a). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, 58, 981-1007.

Lievens, F., Buyse, T., & Sackett, P. R. (2005b). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology*, 90, 443-452.

Lievens, F. & Sackett, P. R. (2006). Video-based vs. written situational judgment tests: a comparison in terms of predictive validity. *Journal of Applied Psychology*, 91, 1181–1188.

Lievens, F., Sackett, P.R., Buyse, T., (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context, *Journal of Applied Psychology*, 94, 1095-1101.

Lievens, F. & Coetsier, P. (2002). Situational Tests in Student Selection: An examination of predictive validity, adverse impact and construct validity, *International Journal of Selection and Assessment*, 10, 245-255.

Lievens, F., & Peeters, H. (2009). Impact of elaboration on responding to situational judgment test items. *International Journal of selection and Assessment*, 16, 345-355.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgement tests: a review of recent research. *Personnel Review*, 37, 426-441.

Lievens, F., & Sackett, P. R. (2007). Situational judgement tests in high-stakes settings: issues and strategies with generating alternate forms. *Journal of Applied Psychology*, 92, 1043-1055.

McDaniel, M.A., Hartman, N.S., Whetzel, D.L., & Grubb, W. L. (2007). Situational judgement tests, response instructions and validity: A meta analysis. *Personnel Psychology*, 60, 63-91

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, 86, 730-740.

McDaniel, M.A., & Nguyen, N.T. (2001). Situational judgment tests: A review of practice and constructs assessment. *International Journal of Selection and Assessment*, 9, 103-113.

McDaniel, M. A., & Whetzel, D. L. (2007). Situational judgment tests. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied Measurement: Industrial Psychology in Human Resources Management* (pp. 235–257). Mahwah, NJ: Lawrence Erlbaum and Associates.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 599-616.

Motowidlo, S.J., & Beier, M.E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology, 95*, 321-333.

Motowidlo, S. J., Crook, A. E., Kell, H. J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business and Psychology, 24*, 281-288

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*, 640-647

Motowidlo, S. J., Hooper, A.C., & Jackson, H.L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology, 91*, 749–761.

Motowidlo, S.J. & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology, 66*, 337-344

Mullins, M. E., & Schmitt, N. (April, 1998). *Situational judgment testing: Will the real constructs please present themselves?* Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology.

Mumford, T. V., van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2008). The team role test: development and validation of a team role knowledge situational judgment test. *Journal of Applied Psychology, 93*, 250-267.

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment, 13*, 250–260.

Nguyen, N.T., & McDaniel, M.A. (2003). Response instruction and racial differences in a situational judgment test. *Applied H.R.M. Research, 8*, 33-44.

O'Connell, M.S., Hartman, N.S., McDaniel, M.A., Grubb, W.L., & Lawrence, A. (2007). Incremental validity of situational judgement tests for task and contextual job performance. *International Journal of Selection and Assessment, 15*, 19-29.

Oswald, F.L., Schmitt, N., Kim, B.H., Ramsay, L.J., & Gillespie, M.A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, 89, 187-207.

Patterson, F., Baron, H., Carr, V., Plint, S., & Lane, P. (2009). Evaluation of three short-listing methodologies for selection into postgraduate training in general practice, *Medical Education* 43, 50-57.

Patterson, F., Carr, V., Zibarras, L., Burr, B., Berkin, L., Plint, S., Irish, B., & Gregory, S. (2009). New machine-marked tests for selection into core medical training: evidence from two validation studies.

Peeters, H., & Lievens, F. (2005). Situational judgement tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement*, 65, 70-89

Phillips, J.F. (1992). Predicting sales skills. *Journal of Business and Psychology*, 7, 151 - 160.

Phillips, J.R. (1993). Predicting negotiation skills. *Journal of Business and Psychology*, 7, 403-411.

Polyhart, R.E., & Erhart, M.G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment*, 11, 1-16.

Ployhart, R.E., Weekley, J.A., Holtz, B.C. & Kemp, C. (2003). Web-Based and Paper-and-Pencil Testing of Applicants in a Proctored Setting: Are personality, biodata, and situational judgment comparable? *Personnel Psychology*, 56, 733-752.

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance*, 9, 241-258.

Pulakos, E. D., Schmitt, N., & Chan, D. (1996). Models of job performance ratings: An examination of ratee race, ratee gender, and rater level effects. *Human Performance*, 9, 103-119.

Quinones, M. A., Ford, J. K., & Teachout, M. S. (1995). The relationship between work experience and job performance: A conceptual and meta-analytic review. *Personnel Psychology*, 48, 887-910.

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology*, 85, 880-887.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology*, 75, 175-184.

Schmidt FL, (1994). The future of personnel selection in the U,S, Army, In Rumsey MG, Walker CB, Harris JH (Eds.), *Personnel selection and classification* (pp, 333-350), Hillsdale, NJ: Erlbaum,

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests* (pp. 135–156). Mahwah, NJ: Lawrence Erlbaum.

Schubert, S., Ortwein, H., Dumitsch, A., Schwantes, U., Wilhelm, O., & Kiessling, C. (2008). A situational judgement test of professional behaviour: development and validation. *Medical Teacher*, 30, 528-533.

Smith, K.C. & McDaniel, M.A. (1998, April). *Criterion and construct validity evidence for a situational judgement measure*. Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Siu, E., & Reiter, H.I. (2009). Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. *Advances in Health Sciences Education*, 14, 759-775.

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvath, J. A. (1995). Testing common sense. *American Psychologist*, 50, 912-927

Tesluk, P. E., & Jacobs, R. R. (1998). Toward an integrated model of work experience. *Personnel Psychology*, 51, 321–355.

Vasilopoulos, N.L., Reilly, R.R., & Leaman, J.A. (2000). The influence of job familiarity and impression management on self report measure scale scores and response latencies. *Journal of Applied Psychology*, 85, 50-64.

Wang, L., MacCann, C., Zhuang, X., Liu, O.L., & Roberts R.D., (2009). Assessing Teamwork and Collaboration in High School Students. *Canadian Journal of School Psychology*, 24, 108-124.

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology*, 50, 25-49.

Weekley, J.A. & Jones, C. (1999). Further studies of situational tst. *Personnel Psychology*, 52, 679-700

Westring, A.J.F., Oswald, F.L., Schmitt, N., Drzakowski, S., Imus, A., Kim, B., & Shivpuri, S. (2009). Estimating Trait and Situational Variance in a Situational Judgment Test, *Human Performance*, 22, 44–63.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, 21, 291–309.

Whetzel, D. L., & McDaniel, M., A., (2009). Situational judgment tests: An overview of current research, *Human Resource Management Review*, 19, 188–202.

Weekley, J.A., & Ployhart, R.E. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance*, 18, 81-104.

Appendix 10: Project Team Biographies

Professor Fiona Patterson BSc, MSc, PhD, CPsychol, AcSS, FRSA

Professor Fiona Patterson is a leading expert in the field of assessment, selection and innovation in organisations. She has recently taken the post of Principal Research Associate in the Department of Social and Developmental Psychology, University of Cambridge and she is Professor of Organisational Psychology at City University, London. She is the Executive Director for the Work Psychology Group Ltd, providing research-led consulting on selection and assessment internationally.

Relevant Roles

- Adviser to the *Department of Health and Academy of Medical Royal Colleges* on assessment and selection methodology
- Selection & Psychometric Adviser, *Royal College of Surgeons* (UK)
- Assessment and Psychometric Adviser, *Royal College of General Practice*, UK
- Adviser to Specialty Training Boards, *NHS Education for Scotland & Scottish Government Health Department*

Background

Professor Patterson's research interests are in the field of assessment, selection and innovation within organisations. Professor Patterson's research in medical selection began whilst at the University of Nottingham focusing on GP selection. This became the start of a long-term research programme. In collaboration with the senior figures in the medical community this research has delivered new selection methodologies across all medical specialties in the UK. The impact of the research has been fundamental in changing the way in which doctors are selected and assessed internationally. Her research and publications have had major impact on policy and practice, based on robust research evidence. Professor Patterson is now a leading academic advising on high stakes selection methodology across a range of settings internationally.

Professor Patterson's programme of research uniquely involves the development of theoretical models of important non-cognitive skills required of doctors (such as empathy, communication and integrity) and the related design, implementation and validation of assessment methodologies. Through her research team Professor Patterson's activities have provided evidence for the introduction of innovative selection methodologies in medical selection such as situational judgement tests (SJT) and selection centres. These methods are now being piloted by medical Royal Colleges in the UK (via the Academy of Medical Royal Colleges) and the Medical Schools Council. Her research is now being adopted internationally (e.g. Royal College of Surgeons, Australasia; GP Education and Training in Australia, Netherlands, Denmark etc).

Full publication list on request

Vicki Archer

BA, MSc, CPsychol

Vicki is a Member of the British Psychological Society (BPS) and a chartered member of the Division of Occupational Psychology (DOP). She holds a first class BA (Hons) degree in Psychology and an MSc with Distinction in Occupational Psychology.

Vicki works as a Consultant Psychologist for the Work Psychology Group. Her role involves designing bespoke innovative assessment and development solutions, delivering individual and group based training and conducting validation studies to determine the quality of different assessment and development interventions. Recent projects Vicki has been involved in include the development and psychometric evaluation of SJTs for UK General Practice training, design of selection process for a leading financial institution, the design and implementation of behaviour change interventions to improve performance of doctors in early training and the design, delivery and analysis of employee surveys.

Vicki previously worked for the National Policing Improvement Agency (NPIA). The NPIA is an agency that provides assessment solutions for all ranks within the police service. As a Higher Psychologist, Vicki designing and delivering assessment solutions to identify those individuals who will undertake the most senior roles within the Police Service. Vicki also worked on the Recruit Assessment Team, designing and delivering of the national selection process for Police Constables, Police Community Support Officers (PCSOs) and Special Officers, including the design and development of an SJT.

Dr. Máire (Maura) Kerrin

BSc, MSc, PhD, CPsychol

Dr Kerrin is a founding Director of the Work Psychology Group & Visiting Senior Lecturer at City University, London. Over the last 15 years, she has been involved in the management of large scale research projects based in a wide range of organisations in both the public and private sector. She specialises in selection and assessment, evaluation, and innovation with her most recent work in medical selection and assessment. Dr Kerrin is a registered Chartered Occupational Psychologist and a full member of the British Psychological Society.

Previously Dr Kerrin worked as a Senior Consultant with the Institute for Employment Studies and has held a variety of academic posts. These have included Senior Lecturer in Organisational Psychology at City University, and Lecturer in Organisational Psychology at the University of Nottingham Business School.

Research interests and activities

Dr Kerrin's research and practice is in four key areas: 1) Selection 2) Evaluation & 3) Work Based Assessment 4) Innovation

Selection: Medical selection eg design, implementation & validation of selection methods in to specialty training (specialties include GP, Obstetrics & Gynaecology, Surgery, Public Health) and graduate entry to Medicine; on-line recruitment and selection (e-recruitment); executive assessment; design and implementation of selection processes for 'volume' recruitment

Evaluation: Evaluation of training & development activities (in particular leadership development and on-line applications); Development of evaluation methodologies

Work Based Assessment: Evaluation of use of 360 degree feedback on performance; review of work based assessment methods

Innovation: Recently funded research by the National Endowment for Science Technology and the Arts (NESTA) and produced a report on Everyday Innovation: How to enhance innovative working in organisations and employees.

http://www.nesta.org.uk/publications/reports/assets/features/everyday_innovation_how_to_enhance_innovative_working_in_employees_and_organisations

Full publication list on request

Victoria Carr

BA, MSc

Victoria has been a Consultant Research Psychologist with Work Psychology Group (previously Work Psychology Partnership) since January 2005. Prior to this she was a Research Associate at the University of Nottingham and also worked for several public and private sector organisations, including local government and police recruitment. Victoria is a Graduate Member of the British Psychological Society (BPS) and a member of the Division of Occupational Psychology (DOP), currently submitting for Chartership as an Occupational Psychologist. She holds a first class BA (Hons) degree in Psychology and an MSc with Distinction in Occupational Psychology. Victoria is a qualified test user, holding both the BPS Level A Certificate of Competence in Occupational Testing and the BPS Level B (Intermediate) Use of Personality Questionnaires.

Victoria has particular research expertise in selection, assessment and evaluation, specialising in the development and evaluation of methodologies for high stakes assessment. She has developed specialist knowledge of the design and validation of psychometrics in the medical and other high stakes contexts, with a particular interest in Situational Judgement Tests (SJTs). Since joining Work Psychology Group, she has worked extensively in medical selection, contributing at a strategic level to many high-profile organisational change projects and working with all UK medical specialties. Recent and ongoing relevant research projects include:

- Design, ongoing development and psychometric validation of national shortlisting tests (clinical problem solving and SJT) for entry into UK general practice training (COGPED).
- Development, validation and quality assurance of national assessment centre methodology for entry into UK GP training (COGPED).
- Development of SJT and assessment centre methodologies for entry into UK surgical training (Joint Committee for Surgical Training, Royal College of Surgeons of England & Edinburgh).
- Development and psychometric support for shortlisting tests (clinical problem solving and SJT) for entry into UK anaesthesia/acute specialties training (South West Peninsula Deanery).
- Development & piloting of a SJT for entry into UK public health training (Faculty of Public Health).
- Design & piloting of a SJT for selection into UK Dental Foundation Programme (Oxford Deanery).

Full publication list on request

Louise Faulkes

BSc, MSc

Louise has held the role of Consultant Psychologist for Work Psychology Group since 2008. Louise graduated with BSc (Joint Honours) in Psychology and Philosophy at Warwick University in 2006 and graduated with her MSc in Organisational Psychology from City University in 2008. Prior to joining Work Psychology Group, she gained significant research experience in wellbeing, pharmaceutical and commercial settings. Louise is a Graduate member of the British Psychological Society (BPS) and Division of Occupational Psychology (DOP) and is an active member of the DOP Science and Practice Strategy Group. Louise holds BPS Level A Certificate of Competence in Occupational Testing and the BPS Level B (Intermediate) Use of Personality Questionnaires, and is working towards Chartership as an Occupational Psychologist.

Expertise

- Significant experience in the design, delivery, evaluation and validation of high-stakes selection processes into medical specialty training; particularly working with surgical specialties.
- Significant experience using Qualitative research methods to evaluate key stakeholder experiences of selection processes.
- Development of the Joint Committee on Surgical Training (JCST) 'Good Practice Surgical Selection Toolkit': www.jcst.org/selection_toolkit
- Significant experience in the development of Situational Judgement Tests (SJTs) for selection into medical specialty training for General Practice (COGPED) and Surgery (JCST).
- Development of Training and Development Interventions to enhance job performance.

- Significant experience in personality feedback for development purposes.

Full publication list on request

Phillipa Coan **BSc, MSc**

Having gained extensive experience over the last year at the Work Psychology Group in projects relating to selection, development and innovation, Phillipa has recently been promoted from an intern to an Associate Consultant Psychologist. Over the last year, Phillipa has contributed to work that has won the Division of Occupational Psychology's category prize for Development as well as work that will lead to national policy change for how we recruit and develop our doctors in the UK. Phillipa's exceptional communication skills also led to employment by the University of Birmingham as the youngest Visiting Lecturer amongst industry experts.

Areas of responsibility and key achievements include:

Selection and Assessment

- Job analysis support and development of competency frameworks and behavioural indicators
- Quality assurance and design of bespoke selection exercises for public and private sector clients
- Lead administrator of psychometric and personality assessments
- Lead assessor of assessment centre exercises
- Facilitating the delivery of assessment centres with groups ranging from 5-500

Training and Development

- Design support and comprehensive evaluation of an award winning development programme to support the revalidation of doctors in the UK

Organisational Change

- Lead researcher for a variety of organisational change projects involving state of the art literature reviews
- Supporting the design and implementation of behaviour change interventions to improve performance of doctors in early training

Data Analysis and Evaluation

- Conducting full psychometric reviews of medical selection processes, analysing potential incremental validity of future selection assessments based on quantitative and qualitative analysis
- Structuring and conducting interviews and focus groups and carrying out comprehensive thematic template analysis

- Survey analysis for innovation project for UK government. Included integrating UK-wide survey findings with interview responses and case examples for a final client report. Devised key implications for corporate and government policy

Dr David Good

David Good is a Psychologist at the University of Cambridge where he is in the Department of Social & Developmental Psychology, is a member of the Psychometric Assessment Centre and a Fellow of King's College. He was educated at the Universities of Sussex and Cambridge, and has held full time staff positions at both institutions. He is co-founder of the Crucible network which supports cross-disciplinary collaboration between social scientists and researchers in technology and the sciences. He is a member of the Governing bodies of the University of Cambridge and the Royal College of Art. From 2000-2008 he was Education Director for the Cambridge-MIT Institute where he was responsible for the development and assessment of novel educational programmes designed to enhance the engineering and entrepreneurial skills of students, and through this developed a line of research on techniques for assessing practical and professional skills which are not reliably assessed by standard academic and professional examinations. He is continuing this work on enterprise and engineering with support from the Royal Academy of Engineering and in collaboration with colleagues from King's College, London.

Recent relevant publications

- 2009 *Evaluation of SPEED (Student Placement for Entrepreneurs in Education: Evaluation Report)*. Commissioned by the University of Wolverhampton HEIF3 consortium. (with S.Cooper & I. Daly)
- 2009 *Evaluation of the BEST Programme*. London: Royal Academy of Engineering. (With P. Kutnick & J. Osborne)
- 2008 Developing female entrepreneurial capital for technology entrepreneurship: the case of MEETS. *Proceedings of the Manchester HTSF meeting*, June 2008, Manchester Business School.
- 2008 Design as communication: exploring the validity and utility of relating intention to interpretation. *Design Studies*, **29**(5), 425-457. (with N. Crilly, D. Maltravers & P.J. Clarkson)
- 2007a Success in Distributed Working: Near term effects and how to measure them. *Electronic Journal for Virtual Organisations and Networks*, **9**, 101-118.
- 2007b *Collaboration for Innovation: Lessons from the Cambridge-MIT Institute*. Rotterdam, Sense Publishers. (with R. Cox, S. Greenwald and M. Goldman)

- 2007c The UK Higher Education Landscape and the Cambridge MIT Initiative. In D. Good et al (Eds.), *University Collaboration for Innovation: Lessons from the Cambridge MIT Institute*. Rotterdam: Sense.
- 2007d Languages of Innovation. In Crawford, H. & Fellman, L. (Eds) *Artistic Bedfellows: Collaborative History and Discourse*. Lanham, MD: University Press of America. (with A. Blackwell)
- 2006a Developing Self-efficacy And Entrepreneurial Intent For Technology Entrepreneurship: The Influence Of Role Models. In A. Zacharakis (Ed.), *Frontiers of Entrepreneurship Research: Proceedings of the 26th Annual Entrepreneurship Research Conference*. Babson Park, MA: Babson College. (with F. Cave, S.Cooper, & A.E.Ward)

Improving Selection to the Foundation Programme

Appendix E

Person Specification FP2012

Foundation Programme Commencing August 2012 Person Specification

	ESSENTIAL CRITERIA	DEMONSTRATED BY
Eligibility	Applicants must meet the requirements set out in the Foundation Programme 2012 Eligibility Criteria.	Eligibility checking
Qualifications	The applicant must have achieved, or expect to achieve, a primary medical qualification as recognised by the General Medical Council (GMC) by the start of the Foundation Programme 2012.	Eligibility checking
Clinical Knowledge & Skills	The applicant must be familiar with and be able to demonstrate an understanding of the major principles of the GMC's Good Medical Practice (2006) including: • Good clinical care • Maintaining good medical practice • Teaching and training, appraising and assessing • Relationships with patients, and can apply this understanding • Working with colleagues, and can apply this understanding • Probity • Health The applicant must demonstrate an understanding of the outcomes to be achieved in the Foundation Programme as set out in The New Doctor (2009).	Application/ pre-employment screening Clinical assessment (where appropriate)
Language & Communication Skills	The applicant must have demonstrable skills in listening, reading, writing and speaking in English that enable effective communication about medical topics with patients and colleagues, as set out in the GMC's Good Medical Practice (2006) .	Application/ pre-employment screening Clinical assessment (where appropriate)

Attributes	The applicant must demonstrate: • an understanding of the importance of the patient as the central focus of care; • the ability to prioritise tasks and information and take appropriate decisions; • an understanding of the importance of working effectively with others; • the ability to communicate effectively with both colleagues and patients; • initiative and the ability to deal effectively with pressure and/or challenge • commitment to learning and continued professional development; • self-awareness and insight into the boundaries of their own abilities; • an understanding of the principles of equality and diversity.	Application/pre-employment screening Clinical assessment (where appropriate)
Probity	The applicant must demonstrate appropriate professional behaviour, i.e. integrity, honesty, confidentiality as set out in the GMC's Good Medical Practice (2006). By the start of the programme, the applicant must demonstrate criminal records clearance at the appropriate level and complete all other pre-employment requirements according to current government legislation	Application/pre-employment screening

Improving Selection to the Foundation Programme

Appendix F

Report of the SJT pilots



Work Psychology Group
Thinking differently



**UNIVERSITY OF
CAMBRIDGE**

Design and Evaluation of a Situational Judgment Test for Selection to the Foundation Programme

Final Report

Improving Selection to the Foundation Programme

Work Psychology Group and the University of Cambridge

Professor Fiona Patterson

Vicki Archer

Dr Máire Kerrin

Dr David Good

Victoria Carr

Louise Faulkes

Helen Stoker

July 2011

Acknowledgements

This project was conducted by the Work Psychology Group in collaboration with the University of Cambridge on behalf of the *Improving Selection to the Foundation Programme (ISFP)* Project Group. The ISFP Project Group, Chaired by Professor Paul O'Neill, includes representatives from fifteen organisations: the Academy of Medical Royal Colleges (AoMRC), the British Medical Association (BMA) Medical Students Committee, the Conference of Postgraduate Medical Deans (CoPMED), the four UK Departments of Health, the General Medical Council (GMC), the Medical Schools Council, the Northern Ireland Medical and Dental Training Agency (NIMDTA), NHS Education for Scotland (NES), NHS Employers (NHSE), the Scottish Board for Academic Medicine, the UK Foundation Programme Office (UKFPO) and clinical tutors. The ISFP Project is funded by the Department of Health.

The Co-directors of the project to develop a Situational Judgement Test for ISFP are Professor Fiona Patterson and Dr David Good. The core project team consisted of Professor Fiona Patterson, Vicki Archer, Dr Maire Kerrin, Victoria Carr, Louise Faulkes and Helen Stoker in conjunction with the Medical Schools Council project team including Professor Paul O'Neill, Professor Tony Weetman, Dr Katie Petty-Saphon, Dr Denis Shaughnessy, Siobhan Fitzpatrick, Amy Stringer and Oliver Watson.

Contents

1.	Executive Summary.....	6
2	Introduction	9
2.1	Purpose and Structure of the Report.....	9
2.2	Background	9
2.3	Expectations and Responsibilities.....	9
	Part One: Item Development & Review	11
3	Initial Design of SJT Specification.....	11
3.1	SJT Literature Review.....	11
3.2	Micropilot Reviews	12
3.3	Summary of SJT Design Specification	18
4	Item Development & Review.....	20
4.1	Evaluation Criteria	20
4.2	Process Overview.....	20
4.3	Item Writing Workshops.....	21
4.4	Initial Item Review	29
4.5	Lead Item Review.....	29
4.6	Focus Groups	30
4.7	Concordance Panel	34
4.8	Test Construction.....	35
	Part Two: Pilot Analysis & Evaluation	38
5	Pilots Overview	38
5.1	Purpose and Specifications of the Pilots.....	38
5.2	Development of Pilot Documentation.....	39
5.3	Pilot Administration	39
5.4	Evaluation Overview	40
5.5	Scoring Convention.....	41
6	Initial Pilot	43
6.1	Pilot Test Specification.....	43
6.2	Sample	43
6.3	Overall Summary of Results.....	47
6.4	Test Level Analysis	47

6.5	Item Level Analysis.....	50
6.6	Group Differences.....	52
6.7	Construct Validity.....	54
6.8	Participant Feedback	54
6.9	Conclusions and Recommendations from Initial Pilot.....	56
7	Large scale Pilot	59
7.1	Pilot Test Specification.....	59
7.2	Sample	60
7.3	Overall Summary of Results.....	69
7.4	Test Level Analysis	70
7.5	Item Level Analysis.....	74
7.6	Group Differences.....	75
7.7	Computer Versus Paper Tests.....	78
7.8	Construct Validity.....	79
7.9	Criterion-related Validity	82
7.10	Participant Evaluation.....	85
7.11	Conclusions from the Large Scale Pilot.....	88
	Part Three: Conclusions & Recommendations	91
8	Conclusions	91
9.	Recommendations	91
9.1	Overview	91
9.2	Test Specification	91
9.3	Analysis & Evaluation.....	93
9.4	Principles of Standard Setting.....	93
9.5	Principles of Test Equating.....	94
9.6	Item Development Methodology	94
9.7	Developing the Item Bank	95
9.8	Equality and Diversity	96
10.	References	97
11.	Appendices	99
	Appendix A: Additional Technical Detail.....	99
	Appendix B: Example SJT Items used as Feedback Questions.....	107
	Appendix C: Item Level Results for Initial Pilot.....	108

Appendix D: Item Level Results for Large Scale Pilot.....	110
Appendix E: Peer Review: Professor Chris McManus.....	113
Appendix F: Peer Review: Professor Val Wass	114
Appendix G: Response to Quality Assurance Peer Review Summary	125

1. Executive Summary

1.1 In 2009, the Department of Health in England (DH), on behalf of the four UK health departments, commissioned the Medical Schools Council (MSC) to lead a cross stakeholder steering group to design, develop and pilot new arrangements for the selection of trainee doctors into the Foundation Programme (FP). The FP is a two-year generic training programme which forms the bridge between medical school and specialist/general practice training.

1.2 Following a nine month Options Appraisal, the DH accepted that piloting should take place of:

- a Situational Judgement Test (SJT) to replace white space questions as a measure of aptitude
- an Educational Performance Measure (EPM) to replace academic quartiles as a measure of academic performance

This report is concerned with the design, development and evaluation of the SJT through a series of pilot studies.

1.3 A literature review, together with a series of student micropilot reviews and a job analysis, assisted in the development of the initial design specification of the SJT. The initial design specification was used to inform the item writing process and outlined that the test would contain both ranking and multiple choice items, it would target five professional attribute domains based on the Job Analysis (commitment to professionalism, coping with pressure, effective communication, working effectively as part of a team and patient focus), and it would consist of 60 items to be taken within a two hour period.

1.4 Seven item writing workshops were run, with a total of 89 item writers attending the Continuing Professional Development (CPD) accredited training, outlining item writing principles. A total of 453 items were written by the item writers.

1.5 All items went through a thorough review process that included initial review by a psychometrician, lead item review by a clinician and review by Foundation Year 1/2 doctor focus groups, where each item was examined for fairness and realism.

1.6 Concordance panels were conducted to identify a high level of consensus between experts on the item answer keys. Six concordance panels were run, with a total of 60 subject matter experts attending. Those items that exhibited high levels of consensus (200) were included in the pilots. Those items exhibiting low levels of consensus were stored for further review.

1.7 An initial pilot was run between October and November 2010. This pilot was carried out as a proof of concept to determine whether the SJT had sufficient potential to warrant investment in a larger pilot in spring 2011.

1.8 A total of 455 applicants participated in the initial pilot from four UK medical schools. Demographic data indicated that the sample was representative in terms of sex and ethnicity.

- 1.9 The psychometric analysis of the initial pilot indicated that the SJT is a reliable measurement methodology, demonstrating internal reliability of $\alpha=0.74$. Analysis also showed that the SJT is able to differentiate sufficiently between participants and item-level analysis showed that the majority of the SJT items worked well.
- 1.10 Differences in performance on the SJT between males and females and between ethnic groups were found. On average males scored significantly lower than females ($p<0.01$), and the Black and Minority Ethnic group scored significantly lower than the White ethnic group ($p<0.01$). These findings are similar to those found for other selection methods and examinations.^{12,13,14} Caution should again be employed when interpreting these results due to the small sample size.
- 1.11 Participant feedback was sought following the test. The majority of participants felt that the content of the test seemed relevant to the Foundation Programme, that the scenario content seemed appropriate for their training level and that the content of the test appeared to be fair.
- 1.12 A large scale pilot was carried out in the spring of 2011. The purpose of this pilot was to undertake further psychometric analysis to be able to make recommendations as to the design and operational (live) test specification. The pilot also provided the opportunity to continue to pilot administrative and logistical arrangements and to pilot SJT items for use in live selection.
- 1.13 A total of 639 applicants participated in the initial pilot from 13 UK medical schools. Demographic data indicated that the sample was representative in terms of sex and ethnicity.
- 1.14 The large scale pilot consisted of three pilot papers. This was to ensure that the maximum number of new SJT items could be piloted (135), whilst also ensuring that sufficient sample sizes were possible for each paper.
- 1.15 The psychometric analysis indicated that the SJT is a reliable measurement methodology, demonstrating overall test paper reliability of between $\alpha=0.74$ and $\alpha=0.76$. Item level analysis showed the majority of the SJT items worked well. When the item level analysis was adjusted to take account for items that did not have sufficient psychometric qualities and test length, reliabilities for the three papers were between $\alpha=0.79$ and $\alpha=0.85$, sufficient for that required for a SJT being used in a high stakes assessment ($\alpha=.80$).
- 1.16 There was a wide range of scores (minimum 658 and maximum 922) and results show a close to normal distribution, indicating that the SJT is capable of differentiating sufficiently between applicants.
- 1.17 Differences in performance on the SJT between males and females were found for only one of the papers (Paper One, $p>0.01$). Differences were found in performance between ethnic groups for all three papers. On average Black and Minority Ethnic group scored significantly lower than the White ethnic group ($p<0.05$ - $p<0.01$). Caution should again be employed when interpreting these results due to the small sample size. Findings of this nature and

magnitude are typical with many selection methods and examinations^{12,13,14}. A richer understanding of the implications of the observed groups differences in practice (for sex and ethnicity) is needed as an impetus for future research. Without detailed systematic research in this area, causal factors cannot be reliably identified¹⁷.

- 1.18 In order to examine the construct validityⁱ of the SJT, the participants at one medical school were invited to complete a NEO-PIR personality measure to examine the associations between scores on the SJT and measures of personality. The correlations suggest that those individuals who have higher activity levels (Extraversion domain), higher openness to values (Openness to Experience domain) and higher levels of achievement striving (Conscientiousness domain) perform better on the SJT than those with lower levels of these facetsⁱⁱ.
- 1.19 To assist with establishing the criterion related validityⁱⁱⁱ of the SJT, analysis was carried out using other measures of performance from one of the medical schools. The findings from the predictive validity study indicate that the SJT correlates with all aspects of the examination data. The SJT correlates most highly with the OSCES (a behaviour based exam) and also correlates, to a lesser extent, with the written exam. *In broad terms, this supports the proposition that the SJT is an aptitude test for clinical practice* (i.e. high fidelity assessment that replicates what trainees will face in FY1), rather than a replication of the knowledge testing (i.e. written exams) completed throughout medical school and is good preliminary evidence of criterion-related validity. However further longitudinal validation work will allow us to confirm this finding.
- 1.20 Participant feedback was sought following the test. The majority of participants felt that the content of the test seemed relevant to the Foundation Programme, that the scenario content seemed appropriate for their training level and that the content of the test appeared to be fair.
- 1.21 Although the number of participants in the large scale pilot was lower than planned, the datasets **prove the concept, demonstrate reliability and validity of the new SJT** and support the conclusion that **an SJT is an appropriate method of selection in this context**.
- 1.22 Based on the evaluation of the SJT, recommendations are provided regarding the SJT design specification as well as regarding item writing methodology. Commentary is also provided regarding the principles of standard setting and test equating.

ⁱⁱ This is establishing that the selection test or procedure is measuring the constructs (or attribute/skill/proficiency) that is claims to be measuring. For example, that the SJT is actually measuring the competencies that underlie successful performance in the job.

ⁱⁱ When undertaking large numbers of correlations, it is possible that significant correlations found may be artificially inflated. As such, a conservative approach was undertaken and a Bonferroni correction was carried out. This resulted in there being no significant correlations between overall SJT score and facet levels of personality.

ⁱⁱⁱ This is when the outcomes of, or scores on, a selection test or procedure predict measures of performance that are already accessible.

2 Introduction

2.1 Purpose and Structure of the Report

2.1.1 The purpose of this report is to provide details of the design, development and evaluation of a Situational Judgement Test (SJT) as part of the Improving Selection to the Foundation Programme (ISFP) project.

2.1.2 The report is divided into three main parts:

- Part One describes the design of the initial SJT specification and the item development and review process.
- Part Two describes the results and analysis of the pilots that were carried out.
- Part Three provides a detailed outline of the recommended SJT design and operational specification as well as recommendations going forward.

2.2 Background

2.2.1 In 2009, the Department of Health in England (DH), on behalf of the four UK health departments, commissioned the Medical Schools Council (MSC) to lead a cross stakeholder steering group to design, develop and pilot new arrangements for the selection of trainee doctors into the Foundation Programme (FP). The FP is a two-year generic training programme which forms the bridge between medical school and specialist/general practice training.

2.2.2 This steering group recommended the pilot of a Situational Judgement Test (SJT) and Educational Performance Measure (EPM) and that these two assessments in combination should be used for selecting applicants and allocating them to foundation schools. The SJT must therefore be developed and validated in accordance with accepted best practice, so that it provides an effective, rigorous and legally defensible method of selection.

2.2.3 The authors were commissioned to develop and validate an SJT for use in selection to the FP, informed by a detailed job analysis of the role of the Foundation Programme doctor^{iv}.

2.3 Expectations and Responsibilities

2.3.1 The authors were responsible for aspects of the work required to produce the SJT deliverables as follows:

- defining the specification for the SJT against the criteria determined through the Job analysis (detailed in section 3.3)
- producing training materials and guidance for medical schools and applicants
- training volunteers and facilitating item writing workshops to produce the necessary numbers of items to meet the SJT specification (detailed in section 4.3)

^{iv} The results of the Job Analysis are detailed in a separate report

- reviewing and quality assuring items to ensure they meet the SJT specification (detailed in section 4.4-4.7)
- proposing the specifications for the pilots, including the number of pilots to be run, the number of medical schools and applicants to participate in each pilot, and the evaluation criteria for the pilots (detailed in section 5.1)
- analysing the data to assess the effectiveness of the tests in relation to the agreed evaluation criteria in line with best practice (detailed in sections 6 and 7)
- producing a report conforming to the agreed format and structure and making recommendations for future work, specifically covering any changes to the SJT specification

2.3.2 The Medical Schools Council was responsible for a number of aspects of the work as follows:

- scheduling of item writing workshops pilots, and the booking and provision of venues and facilities
- recruiting participants for item writing and pilots
- managing communications with participants
- managing the contribution of individual medical schools to the process
- arranging independent quality assurance activities
- gaining any necessary consent from those participating in the pilots
- providing access to other agreed data for assessment purposes
- incentives for participants

Part One: Item Development & Review

3 Initial Design of SJT Specification

3.1 SJT Literature Review

- 3.1.1 Situational Judgement Tests (SJTs) are designed to assess individuals' judgement regarding situations encountered in the workplace. Applicants are presented with a set of hypothetical work-based scenarios and asked to make judgements about possible responses. Applicants' responses are evaluated against a pre-determined scoring key to provide a picture of their situational judgement in that particular context. SJT scenarios are based on extensive analysis of the job role, to ensure that test content reflects the most important situations in which to test applicants' judgement and are concerned with testing non-cognitive attributes rather than knowledge or practical skills. SJTs have become increasingly popular over the last 20 years and are used mostly in large-scale selection processes, often at the shortlisting stage, but can also form part of workplace assessment to highlight employee development needs.
- 3.1.2 A literature review was carried out to assist in the development of the initial SJT design specification and to provide background information about the research evidence available on SJT development and evaluation. The literature review can be found in the Job Analysis report. A summary of the relevant evidence is provided below.
- 3.1.3 Item stems (questions) can vary in length and detail. Longer and more detailed questions tend to provide more work-specific information, thus the detail may enhance the importance of work-specific knowledge. Length may also increase the [cognitive loading](#) of a question. McDaniel et al (2001)¹ suggests that tests with detailed questions have higher correlations with work performance; however some tests with less detail also had slightly higher validity.
- 3.1.4 Response instructions can either have a 'knowledge' format (what should you do) or a behavioural based format (what would you do). McDaniel et al (2007)² noted that knowledge response instructions measure maximal performance and behavioural measures typical performance. McDaniel et al's (2007)² meta-analysis showed that knowledge based response instructions tend to be correlated more positively with cognitive ability compared to behavioural instructions, and behavioural response instructions tend to be more positively correlated with personality traits compared to knowledge instructions. However, it should be noted that a correlation with either of these constructs, does not mean that the SJT *measures* either cognitive ability or personality, rather that there is common variance in the relationship.
- 3.1.5 One issue to consider with response instructions is faking. McDaniel et al (2007)² indicate that SJTs with behavioural instructions are saturated with non-cognitive variance and so are susceptible to self deception and impression management. Knowledge based response instructions have been found to be less prone to faking than behavioural based response instructions^{3,4} and therefore it is recommended that knowledge based response instructions be used where possible so faking becomes a non issue and does not create a

moral dilemma⁵. On a similar vein, behavioural based instructions are thought to favour 'test wise' applicants. 'Low fidelity' assessments, such as SJTs and interviews, as much as they try, can never truly tap into candidate actual behaviour, rather as to what an applicant decides to represent about themselves. Therefore 'test wise' applicants would still answer what they 'should' do, where as less 'test wise' applicants would answer what they 'would' do. As the correct or keyed answer must reflect what an applicant 'should' do, this would favour these 'test wise' applicants. In high stakes settings such as these, it would seem sensible to minimise faking issues as far as possible.

- 3.1.6 With regards to criterion-related validity, knowledge based response instructions have been found to have higher criterion related validity than SJTs with behavioural instructions² i.e. an applicant demonstrating knowledge of the correct answer has been shown to predict performance in the role. In contrast Ployhart & Erhart (2003)³ found that behavioural tendency showed increased criterion-related validities and higher test-retest reliabilities compared to knowledge based instructions.
- 3.1.7 There are a variety of different response formats that can be used in SJTs; for example, pick best, pick best/worst, pick best three, rate effectiveness and rank options. 'Effectiveness' response format algorithms have been found to yield higher reliability and validity than pick best/worst and rank, with ranking yielding the second highest³. However, it has been noted that pick best/worst will be at a disadvantage where reliability is concerned as only two responses are rated.
- 3.1.8 Weekley (2004)⁶ reported the findings of their study examining response types. Across four SJTs, rank-ordered scoring showed a significant improvement in performance prediction in two of the four SJTs. The rank-order scoring method for the leadership SJT showed validity approximately 1.5 times greater than the most commonly used method (pick the best/worst). Ranking items are also less likely to be prone to coaching, compared to pick best/worst due to the complexity of ranking responses. The choice of response format is likely to reflect the scenario content and the appropriate format to both provide and elicit the information needed.
- 3.1.9 There are various scoring methods for multiple choice SJTs. They are typically broken down into '*rational*' and '*empirical*'. Rational scoring of items is based on experts' judgements concerning the effectiveness of responses or best/worst options. When items are empirically scored, they are administered to a large pilot sample. Items are selected and/or weighted according to evidence that they differentiate between individuals who score at different levels on a criterion variable (e.g. job performance). Options that are scored relatively often by high performing individuals are scored as correct etc⁷. One study carried out demonstrated a relatively high correlation between the two scoring methods ($r = .72$)⁸.

3.2 Micropilot Reviews

- 3.2.1 To assist in the development of the initial SJT design specification, a series of 'micropilots' (i.e. medical school student reviews) were carried out in several universities across the UK. The aim of the micropilot reviews was to assist in establishing the appropriateness of the

SJT design specification including test length, appropriateness of content, response format and response instructions.

- 3.2.2 The micropilots were also used as an opportunity to seek feedback on the use of SJTs for selection to the FP. For example, do participants feel that an SJT is a fair, relevant and appropriate selection method in this context? The micropilots also gave the opportunity for the researcher to provide information to students on the ISFP project.
- 3.2.3 A total of ten micropilot review sessions were conducted. (University of Cambridge, two sessions; University of East Anglia, two sessions; The University of Sheffield, three sessions, and The University of Manchester, three sessions). A total of 32 participants attended the micropilots. Participants' demographic details can be found in Table 1 below.

Sex	Male	12
	Female	20
	Not Stated	0
Age group	25 and under	25
	26-35	7
	36-45	0
	46-55	0
	56-65	0
	66 and over	0
	Not stated	0
Ethnicity	White	15
	Black	0
	Asian	8
	Mixed	2
	Chinese	3
	Other	3
	Not Stated	1

Table 1: Demographic details of micropilot participants

- 3.2.4 Prior to attending the micropilot review, participants received a briefing document that outlined the background to the project and the aim of the micropilot. Participants were also asked to sign a confidentiality declaration to ensure security of the items reviewed.
- 3.2.5 The micropilots lasted for between 1.5 and 2.5 hours and each micropilot review session consisted of between three and six participants. During the micropilot reviews, participants were presented with a sample SJT test paper consisting of 16 example items. The items had been developed by an experienced item writer, using scenarios from interviews with Foundation doctors. It should be noted that these items were for illustrative purposes only. The SJT items covered a range of the possible item types and response formats that could be used in SJT design:

- Item Type: Eight items were outlined as '*what would you do*' and eight items were '*what should you do*'
- Response Format: four response formats were used:
 - rating of effectiveness (rate effectiveness of response on scale of 1 to 5. There are no constraints as to rating each option e.g. each option could be rated 1)
 - ranking of five options
 - multiple choice (pick three answers out of eight)
 - choose best/worst options

3.2.6 Participants were asked to complete the items as if under examination conditions. They were provided with 30 minutes to complete the 16 example items. All participants completed the test within this time. Their responses to these items were not marked as there was no formal scoring key agreed in advance as the items had not gone through the full development and review process.

3.2.7 After completing the sample items, the researcher facilitated a short focus group discussion with the participants where feedback was requested on the SJT items that they had completed. Questions in the focus group covered the test instructions, length of time allowed for the test (e.g. is two hours and 60 questions sufficient or too long?), item format, as well as the relevance of the scenarios used in the item content. In particular, participants were asked if they felt that the content was appropriate for them at the stage in the education pathway (e.g. do you feel able to answer items that used scenarios staged in Foundation Training?). All responses were audio recorded and transcribed.

3.2.8 Table 2 presents the key questions from the micropilot reviews with example quotations.

Response format Which of the four response formats did you like/feel was most appropriate?	"I like Multiple Choice the best, because in real life you end up doing more than one thing at the same time" "I thought choose best/worst was the best, because there is always one you would do and one you would not, but then perhaps it is too easy..." "I found ranking confusing, because you could potentially write down the wrong thing when reading across, although I appreciate that it provides you with more information about an individual" "I think ranking is better for discriminating between people" "I found the ranking harder. Easy to identify the best and worst but in the middle is much more difficult to decide" "If you have a ranking then you know exactly what you're doing. With the rating you are tempted to put the middle ground if you don't really know, where as with the ranking you don't have that option which makes it better. It forces you to make a decision rather than sitting on the fence". "Ranking is most challenging, but also think it would be the most frustrating to do." "Multiple Choice was probably the best one" "I thought rating was the best in terms of differentiation and making you think" "Rating effectiveness – if I wasn't comfortable answering it, I would give it a 3. It's a bit of a cop out." "Choosing the best/worst is so clear cut – not sure how to differentiate candidates" "I thought the ranking was most appropriate – good way of assessing" "Ranking and multiple choice are the best in terms of getting us to think about the questions" "The reason why I liked multiple choice and ranking made sense was because you are seeing a thinking process; you are getting a more global understanding of why that person has thought like that."
Response instructions Do you prefer 'should' or 'would'?	"I feel more comfortable answering questions when it asks what you should do, rather than what you would do. With 'Would' are they trying to get at what you would do or what the right thing to do is. Was confused when answering the question at what they were trying to get at, and felt was trying to catch me out and didn't really like that. It is a bit sneaky. Would punishes honesty"

	"I would feel more comfortable with should. It's an exam so I am going automatically circle the one I think I should do but not what I might do in the real situation" "I like the 'would' – it puts personality into it"
Appropriateness of context Do you feel setting the context in FY1 is appropriate?	"I think when you get to 5th year you begin thinking like an FY1 regardless. I didn't think I couldn't answer the questions" "In your final year you are treated like a FY1, quite a lot of your experiences are in your final year which would help you with this" "Generally yes, I think the level is fine" "I found myself going back to the experience I have had in the hospital. Having clinical experience helps and is quite important"
Length of test Is 60 questions in two hours fair/reasonable?	"Yes. I think so" "Anything between 90 minutes to 2 hours will be fine" "Yes, could drag it out to 4 hours, you just have to go with your gut feelings"
Feedback on SJTs in general	"I think this does seem fairer, because you can get a standardised set of questions and everyone is marked the same. White space is subjective, so in this way I think it is fairer" "Struggle with these as can't discuss them...I can talk to you about why I chose one or why I didn't choose one, but don't get that option" "If you are looking at clinical judgement then this is great. A great way to assess by putting you in lots of scenarios – if you don't have the judgement, then you will fall short so it's good" "A lot of people will feel that they demonstrate abilities quite well, but I'm not comfortable demonstrating professionalism, team work etc, on paper, but given the limitations I think it is the best alternative to an interview etc"

Table 2: Summary of micropilot feedback

3.2.9 All participants were asked to complete an evaluation form. The participants were asked to rate six questions about the example SJT items on a five point Likert-type scale (where 1 = Strongly Disagree and 5 = Strongly Agree). A free text box was also provided at the end of the form for them to add any further comments not covered in the focus group.

Respondents (N=32)				
Strongly Disagree %	Disagree %	Neither %	Agree %	Strongly Agree %
The level of difficulty of the SJT items was appropriate for those applying to the Foundation Programme				
The content of the SJT items is relevant to those applying to the Foundation Programme				
I enjoyed taking the SJT today				
The SJT is fair method of selection for those applying to the Foundation Programme				
The SJT is an appropriate method of selection for those applying to the Foundation Programme				
The SJT is able to measure the attributes necessary for the Foundation Programme				

Table 3: Micropilot evaluation form responses

3.2.10 The responses from the evaluation show an overall positive response pattern. In relation to the fourth question regarding fairness, the majority of participants agreed that the SJT is a fair method of selection for those applying to the Foundation Programme. However, 37.5% were unsure. This is not a surprising result as although information was provided about the evidence regarding the use of SJTs in selection and how this would apply to the Foundation Programme, some participants may want more information to allow them to form their opinion. Similarly, this same reasoning can be applied to the last two questions. Qualitative feedback from the workshops indicate that these views reflect apprehension that the SJT does not allow an opportunity to justify their answer, and that an interview would be a preferred way of assessing professional attributes.

3.3 Summary of SJT Design Specification

- 3.3.1 Based on information gathered from the literature review and the micropilot reviews, an initial SJT design specification was developed and is outlined below. In particular, evidence from the SJT used for selection to GP training is drawn upon as this is the only SJT that has been used successfully for selection into postgraduate training in medicine. One of the purposes of the pilots was to evaluate the data in relation to these areas to identify if the specification is appropriate for this context and purpose.
- 3.3.2 The SJT is designed to be staged in the context of the FP, with the applicants assuming the role of a Foundation Year 1 (FY1) doctor. Feedback from the micropilot reviews indicated that participants involved in the micropilots felt this was appropriate.
- 3.3.3 Based on the evidence from the literature reviews and the micropilot reviews, a knowledge based response instruction (*'what should you do'*) was used for the SJT rather than a behavioural based response instruction (*'what would you do'*). As discussed in 3.1.4, knowledge based instructions are more appropriate for high stakes selection contexts where faking and coaching are potential threats. Micropilot evaluation indicated that a knowledge based instruction *'what should you do'* was viewed favourably. Additionally, in the context of professional behaviour and with the GMC putting a high premium on probity, it seems more appropriate to frame the response instruction as 'should'.
- 3.3.4 Two types of response format were used within the SJT; ranking and multiple choice. The rationale for using two response formats is based on scenario content and subject matter, rather than to enable analysis of separate traits or skills within the SJT, and is informed by experience in designing successful high stakes SJTs in other contexts. For example, the nature of some scenarios and the possible responses to them lend themselves to ranking items (requiring ability to differentiate between singular actions in response to a scenario that vary in appropriateness) whereas some scenarios lend themselves to multiple choice items (where it is necessary to do more than one thing/tackle more than one aspect in response to a scenario). Using both response formats enables a fuller range of item scenarios to be used, rather than forcing scenarios into a less appropriate item type and potentially reducing item effectiveness (e.g. asking applicants to rank three equally correct options). Evaluation of the information from the micropilot reviews indicated that the ranking and the multiple choice response instructions were viewed favourably by most participants and evidence from the literature review also supports their inclusion.
- 3.3.5 Evidence and previous experience in other contexts⁹ suggests that approximately two thirds of the items used in the test should have the ranking answer format, and one third should have the multiple choice format. Using a majority of ranking items allows for greater granularity in the data due to the more complex response/scoring format, however ranking items are typically more time-consuming to complete and are not appropriate for all scenarios, therefore a balance with multiple response items is needed. This structure also follows the validated methodology from the SJT for use in shortlisting to General Practice training posts⁹ and tends to reflect the proportion of items produced of each type by item writers (see 4.3.22).

- 3.3.6 The SJT items were mapped against five professional attribute domains identified from the FY1 Job Analysis as follows; Commitment to Professionalism, Coping with Pressure, Effective Communication, Patient Focus and Working Effectively as Part of a Team. Although the items are mapped against these domains for item development and test construction purposes, due to the nature of SJTs as encompassing complex, realistic work scenarios, each item does not exclusively measure any single domain and the SJT should not be viewed as assessing five individual constructs or as a 'blueprint'. The benefits of this are that not only does the complexity of the item increase, but the possibility of a 'template answer' being provided by applicants is avoided. The target domains can be seen more as facets relating to the professional attributes required for a Foundation doctor, with the test as a whole providing an overall measure of this group of relevant professional attributes.
- 3.3.7 Within and across target domains, the item content should be designed to cover a range of relevant topics e.g. dealing with a difficult colleague; being asked to undertake an unfamiliar procedure.
- 3.3.8 The test is designed to last for two hours and contain 60 items. 60 items was anticipated to be sufficient to cover the five target attribute domains in a sufficiently reliable and broad way without overloading applicants. 60 items also allows sufficient distribution of scores considering the large number (7000+) of applicants who will be taking the test. 60 items is a slightly greater number than similar tests used in other areas of medical selection, but was planned to ensure a sufficient level of test reliability. In this context, the SJT, in combination with the EPM, is the only method of selection into the FP. This is different from where SJTs are used in specialty selection (e.g. UK GP training), where the SJT is used as one of two tests for short listing and successful applicants then go on to further assessments at a Selection Centre. The score on the FP SJT is therefore proportionally more of an applicant's overall assessment of suitability for the role compared to the GP SJT.
- 3.3.9 Allowing two minutes per item is consistent with previous evidence and experience which has demonstrated good reliability and validity⁹. Analysis of the pilot data identifies whether the number of items is sufficient to achieve reliability and whether two hours is an appropriate time, based on analysis of participant reactions.

4 Item Development & Review

4.1 Evaluation Criteria

4.1.1 The technical evaluation criteria defined in the project brief for the development of SJT items are outlined below:

- Items must be content valid (i.e. items assess the attributes/skills required for successful performance within the Foundation Year One (FY1) role, and seem realistic in relation to what is expected of a FY1). This will be established through using the target attribute domains to assist with item writing, using Subject Matter Experts to develop the items, and through item review and FY1/2 focus groups.
- Items must be face valid ('looks right' to the applicant i.e. that it would be clear that what is being assessed is relevant to the role of a FY1). This will be established through the use of Subject Matter Experts developing the items, item review and FY1/2 focus groups.
- Items must not unfairly discriminate against any one particular group. This will be achieved through careful consideration of equality and diversity issues throughout test design, including analysis of group differences at test and item level.
- Items must be acceptable to stakeholders and applicants and perceived as fair. This will be achieved through consultation with stakeholders and applicants throughout the design and development process, and obtaining applicants' reactions, and dispelling any misconceptions.

4.2 Process Overview

4.2.1 The SJT item development and review process was held over two phases. Phase One ran between April and October 2010, and Phase Two ran between November and March 2011. The purpose of the second phase was to develop more items for the large scale pilot and undertake further psychometric analysis to be able to make recommendations as to the design and operational (live) test specification. The pilot also provided the opportunity to continue to pilot administrative and logistical arrangements, and to pilot SJT items for use in live selection.

4.2.2 The SJT item development and review process consists of a series of stages. Figure 1 summarises the stages of the SJT development and review process.



Figure 1: Item development and review process

4.3 Item Writing Workshops

- 4.3.1 A total of seven item writing workshops were held between April and December 2010. Five of the workshops were held in Phase One, between April and May and two further workshops were held in Phase Two, between November and December.
- 4.3.2 Invitations for the item writing workshops were sent via the UKFPO foundation school directors email list, and they were asked to cascade it to relevant colleagues. Invitations were followed up at later date with contacts in the Society for Academic Primary Care (SAPC) and a contact from Psychiatry to cascade the invitation to community clinicians. Invitations were also sent to medical school tutors via the MSC Assessment Alliance.
- 4.3.3 The person specification for volunteer item writers was as follows:
- familiar with the FY1 role, including responsibilities and tasks undertaken as well as the necessary skills and abilities required for the role
 - to have worked, within the last two years, with FY1 doctors
 - may have been involved in the process of marking the ‘white space’ questions
 - motivated to produce high quality and effective SJT items
 - available to commit one day to attend an item writing workshop

- willing to commit time to develop 6-8 SJT items and their scoring keys within a specified time (3-6 weeks) after the item writing workshop, working closely with another item writer
- willing to respond to feedback during the period of Item review

4.3.4 A total of 89 item writers attended the workshops. The number of item writers at each of the workshops is detailed in the table below.

London (22 nd April)	22
Manchester(4 th May)	9
London (13 th May)	15
Birmingham (17 th May)	12
Glasgow (26 th May)	12
London (11 th Nov)	11
London (6 th Dec)	8

Table 4: Item writing workshops

4.3.5 Previously trained item writers were also invited to attend the second phase of workshops. A total of nine previously trained item writers attended the workshops.

4.3.6 The breakdown of the demographics of the item writers, their job role and primary speciality is provided in the tables below.

Sex	Male	53
	Female	34
	Not Stated	2
Age group	25 and under	0
	26-35	7
	36-45	38
	46-55	26
	56-65	15
	66 and over	0
	Not stated	3
Ethnicity	White	57
	Black	0
	Asian	20
	Mixed	0
	Chinese	3
	Other	5
	Not Stated	4

Table 5: Demographics of item writers

Clinical Tutor	6
Clinical/Educational Supervisor	44
Foundation School Director	15
Lay Representative	2
Medical School Director	4
Other	16
Not stated	2

Table 6: Item writers job role

Specialty	No.	Specialty	No.
Anaesthetics	7	Genito-Urinary Medicine	1
Breast surgery	1	Haematology	1
Diabetes	1	Intensive Care	1
Diabetics	1	Neurology	2
Education	2	Neurosurgery	1
Emergency Medicine	2	Obstetrics & Gynaecology	3
Endocrinology	2	Orthopaedic Surgery	2
Gastroenterology	1	Paediatrics	7
Gastrointestinal Medicine	2	Psychiatry	2
General Medicine	8	Respiratory	1
General Surgery	4	Rheumatology	3
Geriatric Medicine	5	Urology	2
General Practice	10	Not stated	17

Table 7: Item writers' primary speciality

- 4.3.7 Information was sent to item writers prior to their attendance of item writing workshops outlining the background to the project and the use of SJTs for selection, the purpose of the workshops, their role on the day and some example SJT items. Item writers were also provided with some initial information about item writing principles and were asked to complete some pre-work prior to attending the workshops. They were also requested to complete follow on work. On the day, item writers were required to sign a confidentiality and code of conduct form.
- 4.3.8 Prior to attending the workshop, each item writer was asked to think of at least three possible scenarios or critical incidents that could form the basis of SJT items. They were advised that these should be incidents that involved some dilemma and need for judgement, therefore avoiding straightforward incidents that would elicit text book responses. Item writers were also asked to think about possible responses to the scenarios that should be actions that could plausibly be taken by the applicant in response to the scenario.

- 4.3.9 Item writers were given guidance on the SJT target domains and were advised that the scenarios should be relevant to one of these domains. This ensures that the scenario is relevant to the professional attributes that are expected of a FY1 doctor and that the content of the scenarios maps on to the person specification. However, the nature of SJT items (consisting of a scenario and a number of response options) means that a single SJT item will not exclusively measure one target domain. The content of the scenario and the correct response keys are likely to relate to the targeted domain, but a number of the other response options are likely to tap into one or more of the other SJT target attribute domains (see issues regarding construct validity above). The benefits of this are that not only does the complexity of the item increase, but the possibility of a ‘template answer’ being provided by applicants is avoided.
- 4.3.10 The SJT target domains which were presented in the pre-information sent to item writers and at the item writing workshops are outlined in the table below.

Commitment to Professionalism	<i>Displays honesty, integrity and awareness of confidentiality & ethical issues. Is trustworthy and reliable. Demonstrates commitment and enthusiasm for role. Willing to challenge unacceptable behaviour or behaviour that threatens patient safety, when appropriate. Takes responsibility for own actions.</i> <i>Demonstrates desire and enthusiasm for continued learning, takes responsibility for own development. Willing to learn from others and from experience. Is open and accepting of feedback. Demonstrates a desire and willingness to teach others.</i>
Coping with Pressure	<i>Capability to work under pressure and remain resilient. Demonstrates ability to adapt to changing circumstances and manage uncertainty. Remains calm when faced with confrontation. Develops and employs appropriate coping strategies and demonstrates judgement under pressure.</i> <i>Demonstrates awareness of the boundaries of their own competence and willing to seek help when required, recognising that this is not a weakness. Exhibits appropriate level of confidence and accepts challenges to own knowledge.</i>
Effective Communication	<i>Actively and clearly engages patients and colleagues in equal/open dialogue. Demonstrates active listening. Communicates verbal and written information concisely and with clarity. Adapts style of communication according to individual needs and context. Able to negotiate with colleagues & patients effectively. Demonstrates courtesy, empathy and compassion.</i>
Working Effectively as Part of a Team	<i>Capability & willingness to work effectively in partnership with others and in multi-disciplinary teams. Demonstrates a facilitative, collaborative approach, respecting others’ views. Offers support and advice, sharing tasks appropriately. Demonstrates an understanding of own and others’</i>

	<i>roles within the team and consults with others where appropriate.</i>
Patient Focus	<i>Ensures patient is the focus of care. Demonstrates understanding and appreciation of the needs of all patients, showing respect at all times. Takes time to build relationships with patients, demonstrating courtesy, empathy and compassion. Works in partnership with patients about their care.</i>

Table 8: Professional Attribute SJT target domains

4.3.11 The introduction to the one day workshops covered the background to the project (presented by an ISFP representative), an overview of the test specification (including SJT target domains) and an introduction to item writing principles. Each workshop was facilitated by a minimum of two facilitators.

4.3.12 The key principles for writing the FY1 SJT items are summarised below.

- Items should not require in-depth clinical knowledge. Measuring clinical knowledge is outside the scope of the SJT. The purpose of the SJT is to measure professional attributes of a Foundation doctor, as identified by the job analysis of the Foundation role. The SJT will be used in combination with the EPM which will cover any elements of clinical knowledge required for a Foundation doctor.
- Items should not require specific knowledge of the practices/procedures/protocols of a particular Trust/geographical area beyond that expected of all applicants.
- Items should not assume that applicants are knowledgeable about the roles and responsibilities of a FY1 doctor as applicants may come from different backgrounds (e.g. overseas training).
- The language used should not unfairly discriminate against non-UK trained applicants.
- Scenarios should be set in the context of the Foundation Programme and the scenario content should be relevant to the work of a FY1 doctor, i.e. relate to situations they might be expected to deal with. These should be clinically based scenarios (but not require clinical knowledge). However, scenarios should be familiar to a final year medical student.
- Items should relate to a target domain as identified by the job analysis.
- Scenarios should ideally be fairly short (typically 30-60 words and no more than 80 words) but should provide enough detail to provide dilemma and complexity.
- Scenarios and responses should be described clearly and unambiguously.

4.3.13 Item writers were requested to write two formats of items; ranking items and multiple choice items. Ranking items ask applicants to “rank in order the following actions in

response to this situation (1= most appropriate; 5= least appropriate)” and multiple choice items ask applicants to “choose the three most appropriate actions to take in this situation.”

- 4.3.14 Item writers were then asked to work in pairs creating initial scenarios. Items were then group reviewed and more time was spent developing new items and reviewing items.
- 4.3.15 In the second phase of workshops, small alterations were made to the content of the workshop, based on feedback and learning from the proceeding item review process.
- Item writers were specifically advised of topic areas that from previous reviews had been considered as not appropriate to cover within the scenarios due to issues with procedural knowledge or legal knowledge.
 - Item writers were also advised of areas to avoid due to a large number of items already existing within this topic area.
 - Greater emphasis was also placed on ensuring that the scenarios did not require clinical knowledge or knowledge of particular procedures that may vary in different hospitals. As requested from past item writers, examples of ‘poor’ SJT items were also provided as part of the training.
- 4.3.16 In the second phase of workshops, previously trained item writers were also invited to attend. Previously trained item writers attended a separate session to the newly trained item writers. Prior to attending the workshop, item writers were asked to re-familiarise themselves with the SJT item writing principles. The training consisted of a brief re-cap of SJT item writing principles and a summary of the issues outlined in 4.3.12 above. Item writers were also asked to spend some time within the workshop reviewing previously written items as a group. During this phase, item writers were asked to write between 3-5 items. This is fewer than the number of items requested to be written in the previous five workshops. This was to encourage item writers to produce better quality items and spend more time trying to refine and develop each item. This was particularly emphasised during the workshops. Reviews of some items from previous workshops indicated that although some item writers submitted a large number of items, many were not able to be taken forward.
- 4.3.17 Seven of the 89 item writers did not submit any items either at or following the workshops (all from the first phase of workshops). This is a return rate of 92%. It is not clear why these seven item writers attended the workshops, but did not submit any items. For the second phase of item writing workshops it was ensured that the item writers were clear about the purpose of the workshops and the commitment that was required, as outlined in the item writer person specification.
- 4.3.18 In total, over the seven item writing workshops, 453 items were written, by 82 item writers. This equals an average of 5.5 items per item writer. However, there was great variation in the number of items written by each item writer which ranged from one to 19.

4.3.19 A greater number of items were submitted in the first phase of workshops (376 items, an average of six items per item writer) than in the second phase of workshops (77 items, an average of four items per item writer).

4.3.20 The breakdown of items relevant to each of the target domains written at or following the workshops was as follows:

- Commitment to Professionalism - 180
- Coping with Pressure - 91
- Effective Communication - 68
- Patient Focus - 46
- Working Effectively as Part of a Team - 68

4.3.21 The breakdown of items regarding item format was as follows:

- Ranking - 264
- Multiple Choice - 189

4.3.22 This split is in line with the initial design specification of the SJT, which required that two thirds of the items used in a test should have the ranking answer format, and one third should have the multiple choice format (although item writers were not asked specifically to write more ranking items than multiple choice).

4.3.23 Item writers were asked to complete an evaluation form following the workshops. A total of 72 item writers returned the questionnaire (a response rate of 81%). Item writers were asked to respond to four questions and were also invited to provide some free text responses.

4.3.24 In total, 97.2% of item writers felt that the background information was useful reading material.

	No. of Delegates	Percentage of Delegates
Very Useful	44	61.1%
Useful	26	36.1%
A Little Useful	1	1.4%
Not Useful	0	-
Not declared	1	1.4%

Q1. The Introduction and pre-work document was designed to give you some background to Situational Judgement Tests and item writing principles. How useful was this reading material?

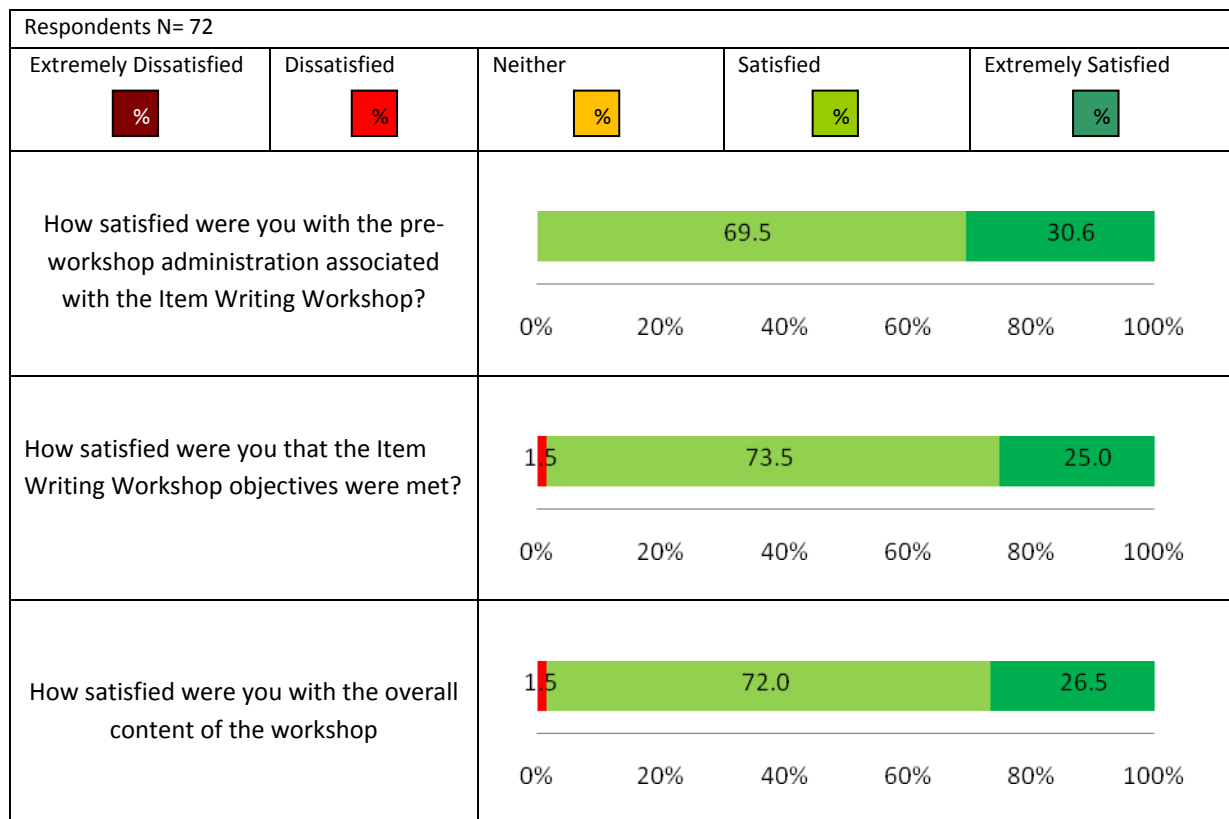


Table 9: Item writing workshop evaluation form responses

4.4 Initial Item Review

- 4.4.1 All submitted items were logged on a spreadsheet which indicated item writer, the date the item was submitted, type of item, target domain, answer key and a short one line summary of the item. The logging spreadsheet also recorded dates the item was reviewed, when it was returned to the item writers and any comments about the item.
- 4.4.2 All items from the item writing workshops were reviewed by the core team of item reviewers from Work Psychology Group. The majority of items were returned within 10 working days to the item writer with comments and suggestions about the item. If necessary, item writers were asked to review their items and resubmit them in order to retain ownership of the process. A number of iterations were necessary in some cases to reach a version of an item that was ready for the next stage of review. All amendments resulting from review were highlighted clearly in the text, comments were added as annotations, and each version of the item was saved.
- 4.4.3 During the first phase of workshops, 42 item writers returned their items after the initial review; a total of 262 items. Twenty one item writers did not return their items after the initial review; a total of 114 items. The items that were not reviewed by item writers were finalised by the core team of item writers. Again it is not clear why these item writers did not return their items.
- 4.4.4 During the second phase of workshops, 17 of 19 item writers returned their items after initial review and worked in conjunction with the item of item writers to review their items; a total of 73 items. Two item writers did not return their items; a total of four items. It is encouraging that a larger proportion of item writers were willing to complete follow on work and review their items. This is likely to be the result of clearer expectations being communicated to the item writers.
- 4.4.5 During the first phase of workshops, following the initial review, 290 items were ready for the next stage of review and 115 items were deemed not suitable to be taken any further; an attrition rate of 31%. An item was deemed not suitable if it did not align with the item writing principles outlined in 4.3.12. If the item was deemed not suitable and not recoverable, then the item writer was informed that their item was not being taken forward to the next stage of review.
- 4.4.6 During the second phase of workshops, 70 of the 77 items were ready for the next stage of review. Seven items were deemed not suitable and were not taken any further in the process; an attrition rate of 9%. This is a great improvement from the first set of workshops and may be due to the improvements made to the workshop and the motivation on the part of the item writers.

4.5 Lead Item Review

- 4.5.1 Following the initial review, all items were reviewed by a lead item reviewer. Twelve lead item reviewers were identified during and following the workshops as those individuals

who appeared to have fully understood the item writing principles and were willing to be involved further in the process.

4.5.2 Lead item reviewers were asked to consider the following points when lead item reviewing:

- Target domain: do the scenario and response options reflect the target domain?
- Scenario: is it clear and unambiguous. Are you clear about what is being asked and what is happening in the scenario?
- Scenario: is it realistic and relevant. Is the scenario something that an FY1 is likely to encounter? Would you expect them to deal with the situation?
- Options: are all options plausible (although not necessarily appropriate)? Are all options realistic options?
- Answer key: please review the key and suggest an alternative if you think it is not correct, giving your reasons for any corrections.
- Language: is the language and terminology clear? Are the terms commonly used across UK hospitals?

4.5.3 For all items written in phases one and two lead item reviewers were sent no more than 25 items each to ensure that no individual had access to too many items, although each had signed a confidentiality form. Lead item reviewers reviewed the items and then returned them by email. They were asked to use track changes to suggest any alterations to the items and make any comments in the comments box. These items were then reviewed by the core team of item reviewers and finalised ready for the next stage of review.

4.5.4 A number of items were also 'lead item reviewed' by previously trained item writers as part of the second phase of item writing workshops.

4.5.5 Lead item writers reviewed 360 items. At this stage, 54 items were not deemed suitable to go any further, leaving 306 items to go forward to the next stage.

4.5.6 At this stage, items were only reviewed in light of the item writing principles. Other issues such as the number of items within each format and coverage of target domains were considered as part of the test construction.

4.6 Focus Groups

4.6.1 The purpose of the FY1/FY2 focus groups was to gain feedback from job incumbents about the relevance and fairness of the items that were due to go forward to the concordance panel. This was to help ensure that the content of the items was realistic and assisted in establishing content validity^v and face validity^{vi}.

^v This is when a selection test or procedure appears to cover a representative sample of the domain being assessed e.g. the test appears to assess the competencies/skills required for successful performance within a role.

^{vi} This is when a selection test or procedure 'looks right' to the applicant or others i.e. the link between the test and performance on the job is clearly evident.

- 4.6.2 Focus groups took place in six locations over the two phases of development; Cambridge, Nottingham, London, Leeds, Keele and Belfast. Other foundation schools were approached for assistance in running the workshops but were unable due to time constraints. Twenty focus groups were held with between two and five individuals in each focus group, with a total of 63 individuals taking part. A breakdown of FY1s/FY2s' demographic details can be found in the table below.

Sex	Male	16
	Female	46
	Not Stated	1
Age	Mean age	26.3
	Range	23-39
Ethnicity	White	32
	Black	2
	Asian	19
	Mixed	3
	Chinese	3
	Other	2
	Not Stated	2

Table 10: Focus Group participants' demographic details

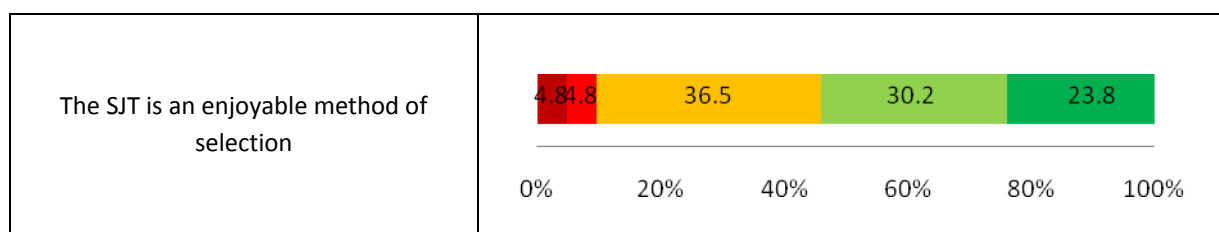
- 4.6.3 All participants who volunteered to take part were sent briefing material outlining the purpose of the focus groups and their role on the day. All participants also completed a confidentiality and code of conduct form.
- 4.6.4 Participants were asked to review the test instructions and discuss whether they were clear and easy to understand. All trainees found the instructions clear and easy to understand. Trainees were then handed the SJT items and asked to consider the scenario content and the response options in their review. They were also asked to provide a possible answer key. For security reasons no individual had access to more than 20 items. Their comments and suggestions were recorded by the facilitator. Trainees were also asked about the timing of the test (60 questions in 120 minutes) and whether they felt this was appropriate. All trainees felt this would be an appropriate amount of time to complete 60 questions.
- 4.6.5 A total of 306 items were reviewed during the focus groups; comprising 201 ranking items and 105 multiple choice items. Below is the breakdown of the items by target domain:
- Commitment to Professionalism - 117
 - Coping with Pressure - 56
 - Effective Communication - 45
 - Patient Focus - 41
 - Working Effectively as Part of a Team - 47
- 4.6.6 Following the focus groups, 20 of the 306 items were deemed not suitable to go forward to the next stage of review. Concerns were raised about a further 11 of the items. Their

comments were considered alongside the concordance comments and results (all 11 items did not go forward to the pilot).

- 4.6.7 All participants were asked to complete an evaluation form comprising eight questions. Participants were asked to rate each statement on a five point Likert-type scale (where 1 = strongly disagree and 5 = strongly agree). A free text box was also provided at the end of the form for them to add any further comments not covered in the focus group.

- 4.6.8 Table 11 below outlines the responses to the evaluation form.

Respondents N=63				
Strongly Disagree %	Disagree %	Neither %	Agree %	Strongly Agree %
The level of difficulty of the SJT items reviewed today was appropriate for those applying to the Foundation Programme				
The content of the SJT items reviewed today was relevant to those applying to the Foundation Programme				
Overall, the content of the SJT items reviewed today is fair to those applying to the Foundation Programme				
The SJT is a fair method of selection for those applying to the Foundation Programme				
The SJT is an appropriate method of selection for those applying to the Foundation Programme				
The SJT is able to measure the attributes that are necessary for the Foundation Programme				
The SJT is able to discriminate between different candidates				

**Table 11: Focus group evaluation form responses**

4.6.9 The comments from the free text box on the evaluation form were themed and the results are outlined below.

Theme	No. of comments	Example comment
Ability to differentiate themselves	7	“There definitely needs to be some scope for some kind of personal statement to take into account personality and extracurricular activity” “Slightly concerned about lack of opportunity within the form to show individual character”
Ambiguity	7	“A few questions seen today remain ambiguous in the stem”
Use of SJT in combination with other methods	6	“SJT’s would be appropriate and fair in selection to Foundation Programme but as part of a wider selection process, e.g. interviews/white space questions”
Comparison to White Space Questions	5	“Much better than open questions used in the past”
Response instruction	5	“Candidates will answer according to what they should do rather than what they would do”
Relevant & fair	4	“I think that the situations are relevant and the questions appropriate”
Familiarity of scenario content/reliance on experience	4	“Questions are realistic to FY1 but may be hard given the limited clinical exposure that medical students have, especially as this varies across medical schools”
Traits assessed	2	“The SJT only tests certain traits and not others, i.e. leadership. Does not consider moral/legal aspects of medicine”
Timing/dates of test	2	“Ensure dates for this are fairly allocated”
Response options	1	“There are often a lot of appropriate answers and most can be justified”
Discriminatory	1	“I think SJTs are a great way of assigning jobs. If you score well it shows that you have had adequate hospital experience as a student and will be a

		competent FY1"

Table 12: Themes of evaluation form free text box

- 4.6.10 Two of the main concerns appear to be around whether the SJT will allow them to express "their character" or "to get credit for extra-curricular activities", as in the current system and as participants feel they may be able to do in interviews.
- 4.6.11 Two participants were concerned that the SJT only tests certain characteristics and not others e.g. leadership. The extensive job analysis study did not reveal leadership to be a professional attribute to target at this level in the training pathway and therefore is not a target domain for the SJT. One comment also refers to the fact that the SJT does not measure the "moral aspects" of medicine. However, it is proposed that this aspect is already covered within the Commitment to Professionalism domain.
- 4.6.12 Some trainees were concerned about ambiguity, although it is important to note that the items viewed were not finalised test items and were still in the process of development. However, ambiguity within stems was addressed as a result of the feedback provided.

4.7 Concordance Panel

- 4.7.1 In order to further validate the SJT items, concordance panels were conducted. Concordance panels involve Subject Matter Experts, in this case clinicians, completing an SJT consisting of trial items. Following best practice in SJT design, the aim of a concordance stage is to identify a high level of consensus between experts on the item keys. Those items that exhibit high levels of consensus go forward to the pilot. Those items exhibiting low levels of consensus are put to one side for further review with changes made if necessary.
- 4.7.2 The answer key provided by the concordance panel was used in combination with information from item writers and focus groups to determine the rational scoring key for the pilot data, however this may not reflect the final key as information will be used from the pilot to further develop the items and their keys. For example, if good performing candidates consistently provide a different key to the established key, then the key will be reviewed with the assistance of subject matter experts.
- 4.7.3 The criteria for Subject Matter Expert involvement in the concordance panel was that the individuals work closely with FY1 doctors and are very familiar with the responsibilities and tasks, as well as the necessary skills and abilities required for the role. The concordance panel was recruited via the UKFPO Rules Group^{vii}.

^{vii} The Recruitment Rules Group exists to develop the national selection process for the Foundation Programme. The Group exists to ensure that the national Foundation Programme recruitment and selection processes are open, transparent, legally robust and in line with good HR practice; to develop new, and/or refine existing, recruitment processes; and to determine the components of ranking applicants, giving due consideration to feedback. Membership and Terms of Reference can be found on http://www.google.co.uk/url?sa=t&source=web&cd=1&ved=0CBgQFjAA&url=http%3A%2F%2Fwww.foundationprogramme.nhs.uk%2Fdownload.asp%3Ffile%3Dtor_rulesgroup.pdf&rct=j&q=UKFPO%20Rules%20Group&ei=PivyTd29FYKKhQfy8YA&usq=AfQjCNHjDptRLIVJYG6ONeyfioJBUUIEYg

- 4.7.4 Four concordance panels were held over the two phases of development. The panels consisted of between eight and 19 individuals (total n=60). Ideally, for a concordance panel, a minimum of 10 individuals should be involved to ensure robust results. Four papers were reviewed; one of the papers consisted of 80 items, two papers consisted of 70 items and one paper consisted of 66 items (a total of 286 items). At this stage, the tests were not constructed as final tests i.e. no consideration was given as to spread of item topics or domains as the aim of the concordance panels was to analyse individual items.
- 4.7.5 After completing a confidentiality and code of conduct form, the panel was asked to complete the SJT items under test conditions. There was no time limit, although the panels were told the test would take no more than two hours 30 minutes to complete.
- 4.7.6 Feedback on the item content was provided by the panel, and this resulted in some minor alterations to a small number of items to provide clarification. No item was altered sufficiently to affect the interpretation of the question or the answer key.
- 4.7.7 Following the concordance panel meeting, a concordance analysis was undertaken to analyse the experts' level of agreement over the keyed response for each trial item. Using established criteria of acceptance levels, items were deemed either to have acceptable levels of concordance (259) or unacceptable levels of concordance (27)^{viii}.
- 4.7.8 Items were selected to go forward to the pilot based on a number of considerations. Those items with above satisfactory levels of concordance were deemed suitable for the two phases of piloting. However, consideration was also given to feedback from the concordance panel when selecting items. Twenty one items (all that demonstrated satisfactory levels of concordance) were removed from the test paper due to feedback from the concordance panel. Those that had lower, but still satisfactory, levels of concordance were also considered in light of the evidence from the item writer and focus group key. Broadly speaking, when the concordance key differed with the item writer and focus group key, then it was not seen as appropriate for the item to be used in the pilot.
- 4.7.9 Following the concordance panel stage, 200 items were selected for the pilots over two phases^{ix}.^x

4.8 Test Construction

- 4.8.1 When constructing test papers either for piloting purposes, or for operational tests, there are a number of considerations:
- There should be roughly two thirds ranking and one third multiple choice items within a test paper.

^{viii} Please see appendix (A.1) for established criteria of acceptance levels and further statistical information

^{ix} The numbers of items that went to concordance but not to pilot does not equal 86; this is because all design and review stages were carried out over two phases and many items from phase one were reviewed and entered the review process of phase two at various points

^x Please see appendix B for an example of a ranking SJT item and a multiple choice SJT item used as feedback questions within both pilots

- There should be, as far as possible, an equal split between the domains being targeted as identified in the job analysis. As far as possible, there should also be an equal split of between the domains within each of the item types. As outlined in 3.3.6, items do not measure a domain exclusively. Target domains are used as a way to ensure item writers cover the professional attributes outlined in the person specification, therefore allowing items to be mapped against the five target domains. As such although it is important to get a roughly equal spread of target domains, it is not necessary for this to be an exact equal split.
- Each item, as well as being mapped against a domain, is also assigned a 'topic' number. Scenarios were also assigned to three broad topics of a 'colleague' topic, a 'patient' topic or a 'personal' topic. Within these three broad topic areas, scenarios were further assigned to different topic areas. Examples of topics may be; a colleague making a mistake, an angry or violent patient, or there being difficulty attending teaching. A good range of topic areas should be used within each paper.
- Items should be spread throughout the paper to ensure they were not grouped together by 'topic' or 'domain'.
- Items that include gravely serious scenarios (e.g. those involving deaths or serious mistakes by the FY1) should not be used at the beginning of the paper, as applicants may find these scenarios distressing initially; placing these items later on in the test allows the applicant time to familiarise themselves with the context of the test and what is expected of them. However, there are no fixed rules or procedures determining this aspect.

4.8.2 When creating multiple papers for operational tests there are a number of considerations to take into account:

- There should be the same number of ranking items and the same number of multiple choice items on each of the papers, to ensure the same range of scores and to use a single answer sheet.
- The spread of domains/topic areas should be as equal as possible between the papers.
- The difficulty of all items (mean score) in the item bank will be known following piloting. The ranges of item difficulty should be as equal as possible between the papers.
- The quality (discrimination) of all items in the item bank will be known. The range of quality should be as equal as possible between the papers and should reflect the range of item quality in the item bank as a whole.
- Multiple papers will need to be equated to ensure that results from different papers are comparable. In order to equate papers, there needs to be a proportion of anchor items within each paper; these can either be common to all papers or common to particular pairs of papers taken within a single application period. Depending on the

equating strategy used, between 25-40% anchor items are recommended for test equating purposes, independent of the number of operational papers required. The proportion would be the same for all papers. Note that the number of anchor items may also depend on number of items in the item bank.

Part Two: Pilot Analysis & Evaluation

5 Pilots Overview

5.1 Purpose and Specifications of the Pilots

- 5.1.1 A small scale initial pilot and a large scale pilot were carried out. The purpose of the initial pilot held at four medical schools in autumn 2010 was to test the approach on a relatively small scale before making an investment in the main pilot involving 13 UK medical schools in spring 2011. Specifically, the initial pilot was intended to:
- check that there could be sufficient confidence in the psychometric characteristics of the SJT to justify a larger pilot
 - inform the design and development of an SJT for selection to the Foundation Programme
 - expose any unforeseen issues with the administrative and logistical arrangements before the larger pilot
- 5.1.2 Only the first two of these three objectives are considered in detail in this report. The third has already been covered in a separate report to the ISFP Project Group.
- 5.1.3 The purpose of the large scale pilot was to:
- undertake further psychometric analysis to be able to make recommendations as to the design and operational (live) test specification
 - continue to pilot administrative and logistical arrangements
 - pilot a larger number of SJT items
- 5.1.4 The original project specifications outlined that the SJT would either be a paper-based machine markable test, or a computer delivered test, depending upon the outcome of a separate investigation by MSC into the relative advantages of the two options. A third option for test delivery was also piloted with one school, whereby the answers were entered onto electronic handsets. After the initial review, examination of this issue was incorporated as a tertiary purpose of the large scale pilot. While acknowledging that using two media may impact upon applicant reactions to the test and thus the interpretation of the data, the importance of exploring the logistics of test delivery was recognised. It was recommended that only one medium was used per paper; using two media for a single paper can create unsystematic error that is difficult to explain.
- 5.1.5 The proposed specification for the small scale pilot, as defined by the authors, was a minimum of 400 participants over a small number of medical schools. The proposed specification for the large scale pilot was to try to obtain 400 participants per paper, with a targeted sample size of 2000+, sampling a wide range of medical schools. This would allow the psychometric properties of the SJT to be fully analysed and evaluated and would mean that greater confidence could be placed in the results.

- 5.1.6 Another requirement of both pilots was that the pool of participants was representative of the wider pool of applicants. For any voluntary pilot, this is likely to be a challenge as the individuals will be self selecting and may represent only a sample of the population. An associated issue relevant to the pilot population is that of test taking motivation. In any piloting of a selection test, the motivation of those taking the test may differ to that of applicants for a live selection test. Motivational issues may impact upon the time taken to complete the pilot test (with individuals completing the test in a quicker time than may be anticipated on a live test), the number of items completed (with fewer items being completed than on a live test) and the quality of the answers provided.
- 5.1.8 As the pilots were carried out in a number of schools, the possible influence of other factors is required to be controlled for as much as possible to ensure that any differences found are due to the test content/applicant ability rather than other factors. This includes ensuring that the tests are invigilated in a standardised way (using instructions provided) and are conducted in similar environments (i.e. test halls).

5.2 Development of Pilot Documentation

- 5.2.1 An information flyer was sent out to all medical schools taking part in the pilot, to be cascaded to students. This flyer outlined what was expected of participants, the role of the pilot in developing the SJT and the ISFP project. Workshops were held at participating schools to outline the purpose of the feedback and provide information on the project which proved to be useful. In the initial pilot, due to timing issues, it was only possible to conduct a workshop at one school. However, in the large scale pilot eleven workshops were held in ten medical schools. Those schools unable to host a workshop were sent relevant briefing material to cascade to their students.
- 5.2.2 A participant evaluation questionnaire was developed in order to capture feedback. Whilst feedback was anonymous, participants were asked to declare their RA number so that feedback could be linked to answers where relevant to the analysis.
- 5.2.3 Guidance for invigilators was developed to support the standardised administration of the pilots at the pilot sites. Prior to the pilot, an administration document which included a background to the project, instructions for administration on the day, invigilator checklists, a briefing for participants and a feedback sheet was sent to all pilot leads. More information regarding feedback from the pilot and the administration of the events is provided in the End of Stage Report.

5.3 Pilot Administration

- 5.3.1 All pilots were taken under invigilated conditions. Participation in the pilots was voluntary, although for participants from School C in the initial pilot, this was during a compulsory teaching session.
- 5.3.2 As part of the initial pilot, all participants were also asked to complete a work style measures questionnaire in order to help establish the construct validity of the SJT. Analysis of these data can be found in Section 6.8.

- 5.3.3 One medical school piloted the use of handheld devices. Participants completed the pilot test on paper and then subsequently entered the data into the handheld devices following completion of the test so as not to interfere with the pilot. The evaluation of the use of handheld devices will be covered in the Final Report.
- 5.3.4 Incentives to participate in the pilots were provided; including feedback on the SJT and answers to a subset of items, the experience of taking part in an SJT as preparation for future selection processes, and a cash prize draw (one per school) with names drawn at random.
- 5.3.5 A feedback session was held following each of the pilots with a member of the ISFP Project Team or a member of medical school staff delivering a standardised slide set. The purpose of this feedback session was to provide participants with the correct responses to five of the questions in the test and to provide a forum in which participants could ask questions about the SJT and the pilot process.

5.4 Evaluation Overview

- 5.4.1 This section outlines the psychometric analysis for the pilot. Any high stakes, high profile test needs to meet exacting psychometric standards in terms of the quality of individual items and of the test as a whole, including reliability, validity and fairness. The main analysis and evaluation activities reported here include:
- test level statistics, including reliability^{xi} and scoring distributions
 - item level statistics, including item facility^{xii} and effectiveness; only those items with sufficient psychometric properties will be used in an operational test
 - analysis of group differences at a test level to explore fairness
 - evaluation of participant reactions
- 5.4.2. Internal reliability (Cronbach's alpha) was calculated based on the total score for each item, therefore meeting the assumption of independence of items when examining reliability.
- 5.4.3 Internal reliability refers to the extent of consistency of results across items within a test. SJTs can be viewed as 'broad band' uni-dimensional tests, in that the test overall aims to measure a single, broad construct as determined by the design specification. In this context, the SJT aims to provide an overall measure of the professional attributes required of an F1 doctor; within this are domains, or facets, relating to this overall construct. Factor analysis carried out on an operational test of similar specification supports this showing one overall single factor¹¹. As such, one statistic for internal reliability is presented, rather than for individual target domains, as it is not the intention to provide a separate assessment of each individual domain. Similarly, reliability is not reported for the two different response formats. Two separate response formats are used within the test not to tap into different skills or abilities, rather to allow a greater variety of scenarios to be sampled.

^{xi} See Appendix A.4 for details of reliability estimates recommended

^{xii} See Appendix A.4 for details of item partial criteria

- 5.4.4 Within this report a psychometric item level analysis is reported to provide a broad indication of item quality. However, as the initial pilot is a primary stage in the development of the items, items, including the scoring keys, are not finalised and are still subject to review following the pilot.
- 5.4.5 Additional analysis carried out for the initial small scale pilot included analysis of the relationships between Social Desirability and SJT scores.
- 5.4.6 Additional analysis was also carried out for the large scale pilot and includes:
- analysis of group differences at an item level to explore fairness
 - relationships between application form scores, quartiles and the SJT
 - criterion related validity analysis between exam scores and the SJT for a small sample
 - relationships between the SJT and scores on the NEO-PIR personality measure
- 5.4.7 All additional analysis carried out following both pilots used the entire test i.e. poorly performing items were not removed when analysing group differences, relationships with other assessments or criterion related validity.

5.5 Scoring Convention

- 5.5.1 Following best practice, the SJT scoring key is determined through:

- consensus at the item review stage of item writers and initial subject matter experts
- expert judgement in a concordance panel review
- review and analysis of the pilot data. Consideration is given to how the participants keyed the items and the consensus between this and the key derived from the first two stages

Keyed Rank	Candidate Rank				
	1	2	3	4	5
1	4	3	2	1	0
2	3	4	3	2	1
3	2	3	4	3	2
4	1	2	3	4	3
5	0	1	2	3	4

Figure 2: Scoring for ranking items

- 5.5.2 For ranking items, a total of 20 marks are available for each item. For each of the five response options up to four marks are available. Marks are awarded for near misses and the framework for scoring can be seen in Figure 2. If a participant ties two options, then

they receive 0 marks for each of the options they tied, as they have not answered the question. There is no negative marking employed.

- 5.5.3 Although 20 points are available for ranking items, if a participant offers a complete answer to the item (i.e. does not tie ranks or have missing data), the lowest possible available score is 8 and the highest possible score is 20 due to having a scoring system that allows for 'near misses'. Participants who struggle to complete the question (due to time or lack of ability) may only provide, for example, a best and/or worst response and will legitimately receive fewer marks for this.
- 5.5.4 For multiple choice items, four points are available for each keyed option correctly identified, making a total of 12 points available for each item. If a participant chooses more than three options, then they score 0 for that item. There is no negative marking employed. More points are available for ranking items compared to multiple response items due to the nature of the response formats. Ranking is a more complex format where the individual has to analyse and make a forced choice ranking of five pieces of information, therefore ranking items are likely to demand more cognitive effort than multiple response items.
- 5.5.5 Participants who do not complete items have these items scored as 'missing' and receive 0 for that item. Participants are not credited with random guesses on the items that they **did not** complete, nor are they encouraged to guess at any of the items as this is not part of the scoring algorithm.
- 5.5.6 Scoring SJTs is a complex process and the choice of methodology for how to score the items is not straightforward. There are inevitable consequences of using two response formats and this will result in difficulties in trying to ensure that the scoring convention is a reflection of the input from the applicant. Alternatives for this have been considered i.e. not providing points for near misses, or systematically subtracting eight from each ranking item. This latter option has been trialled in the SJT used for Acute specialties and demonstrated reliability and has been able to differentiate between applicants¹⁰, however any changes to the scoring convention need to be carefully considered.

6 Initial Pilot

6.1 Pilot Test Specification

- 6.1.1 The pilot paper consisted of 65 items; 42 ranking and 23 multiple choice. The design specification recommended that the test consisted of 60 items. Five extra questions were added to the test to allow feedback to be provided to participants, including providing the scoring key.
- 6.1.2 As far as possible, an equal number of items from each of the five target domains were selected for the paper, in order to ensure sufficient and broad coverage of the relevant professional attributes. The breakdown of items by domain is outlined below:
- Commitment to Professionalism - 14
 - Coping with Pressure - 15
 - Effective Communication - 12
 - Patient Focus - 10
 - Working Effectively as Part of a Team - 14
- 6.1.3 A spread of item topics was also aimed for. Ideally a greater range of topic areas would have been in the pilot test; however this was constrained by the number of available items. For an operational (live) test, a greater number of items in a broader range of topic areas will be available.
- 6.1.4 The total score available for the paper was 1116.
- 6.1.5 As outlined in the literature review, there are various scoring methods that can be employed when analysing SJT results; *rational keying* and *empirical keying*. In this report, the results are provided using both types of scoring in order to obtain a full picture of the results, to inform the development of the SJT scoring methodology, and for future piloting.

6.2 Sample

- 6.2.1 A total of 455 participants completed the pilot SJT paper in autumn 2010. The table below outlines the number of participants involved from each medical school.

	No. of participants	Percentage of Overall Sample
School A	72	15.8%
School B	53	11.6%
School C	224	49.2%
School D	106	23.3%
Total	455	100%

Table 13: Number of participants at pilot sites

6.2.2 Demographic data about participants was collected by cross referencing to their application to the Foundation Programme (except for 21 participants (4.6%) for whom the FPAS score was unknown, likely to be because they had applied to the academic programme which is via a separate system).

6.2.3 Table 14 outlines the breakdown of participants by sex. Overall, more females participated in the pilot (275, 60.4%) than males (159, 34.9%).

	Male	Female	Not declared
No. of participants	159	275	21
% of participants	34.9%	60.4%	4.6%

Table 14: Number of participants at pilot sites

6.2.4 Table 15 outlines the breakdown of participants by ethnicity. The majority of participants reported their ethnicity as white (319, 70.1%), the smallest proportion of participants (10, 2.2%) reported their ethnicity as Black and Chinese.

	White	Asian	Black	Chinese	Mixed	Other	Not declared
No. of participants	319	67	10	10	12	7	30
% of participants	70.1%	14.8%	2.2%	2.2%	2.6%	1.5%	6.6%

Table 15: Ethnicity (6 groups) of participants

6.2.5 Table 16 outlines the breakdown of participants' ethnicity (white and Black and Minority Ethnic (BME) groups). Three hundred and nineteen (70.1%) participants reported themselves as white and 106 (23.3%) participants reported themselves as being from Black and Minority Ethnic groups.

	White	BME	Not declared
No. of participants	319	106	30
% of participants	70.1%	23.3%	6.6%

Table 16: Ethnicity (two groups) of participants

6.2.6 Table 17 outlines the breakdown of participants' nationality. Overall, the majority of participants (378, 83.1%) reported themselves as British. Thirty (6.6%) participants reported their nationality as being something other than British.

	British	Non-British	Not declared
No. of participants	378	30	47
% of participants	83.1%	6.6%	10.3%

Table 17: Nationality of participants

- 6.2.7 The mean age of the sample was 24.7 years, with a range of 22 – 43 years.
- 6.2.8 Data were taken from participants' applications to the Foundation Programme. The quartile scores and competency based application form (white space question) scores for the participants were examined. Table 18 below outlines the descriptive statistics for these scores; mean, range and standard deviation^{xiii}. Application form data are provided for both questions 1-6 (career history and competency questions) and questions 2-6 (competency questions only). Ideally, the application scores should be normally distributed. The quartile score should show a good spread of scores with roughly equal numbers in each quartile, as would be expected in a full operational sample. It should be noted that this information was only available for 408 participants due to incorrectly completed identification numbers.

	Mean	Range	Standard Deviation (SD)
Application form Qs 1-6	40.0	16-55	6.2
Application form Qs 2-6	38.2	16-50	5.7
Quartile score	37.1	34-40	2.2

Table 18: Descriptive statistics for application form scores and quartiles

- 6.2.9 The descriptive statistics, and examination of the histograms in Figures 3 and 4 below, indicate that the data are normally distributed and shows a good spread of scores.

^{xiii}Standard deviation indicates how much variation there is from the mean. A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data is spread out over a large range of values.

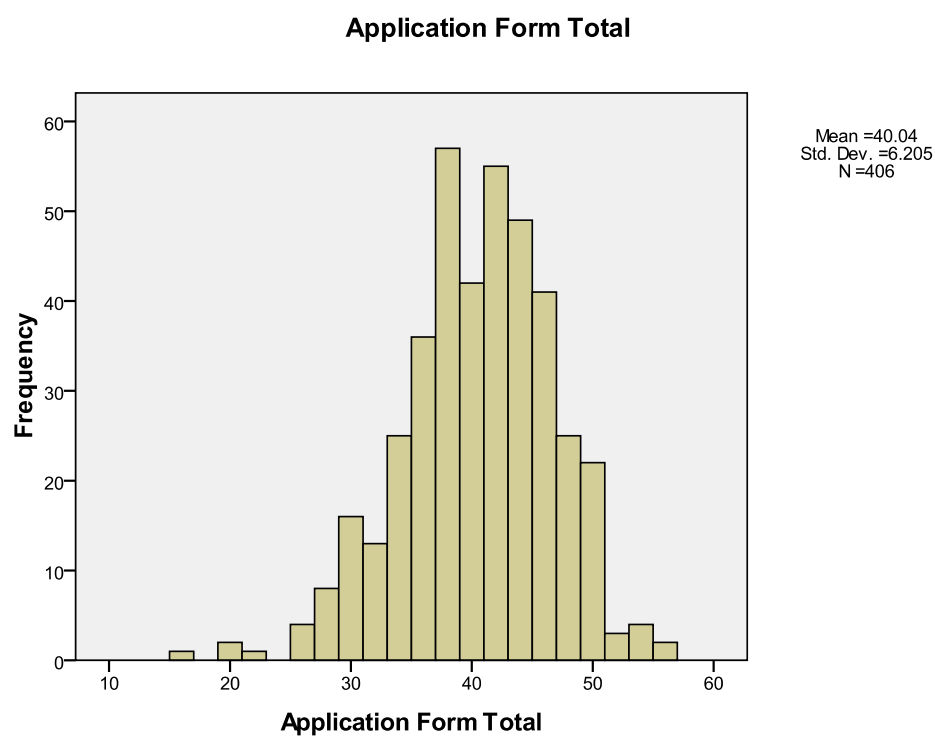


Figure 3: Histogram for Application Form Total

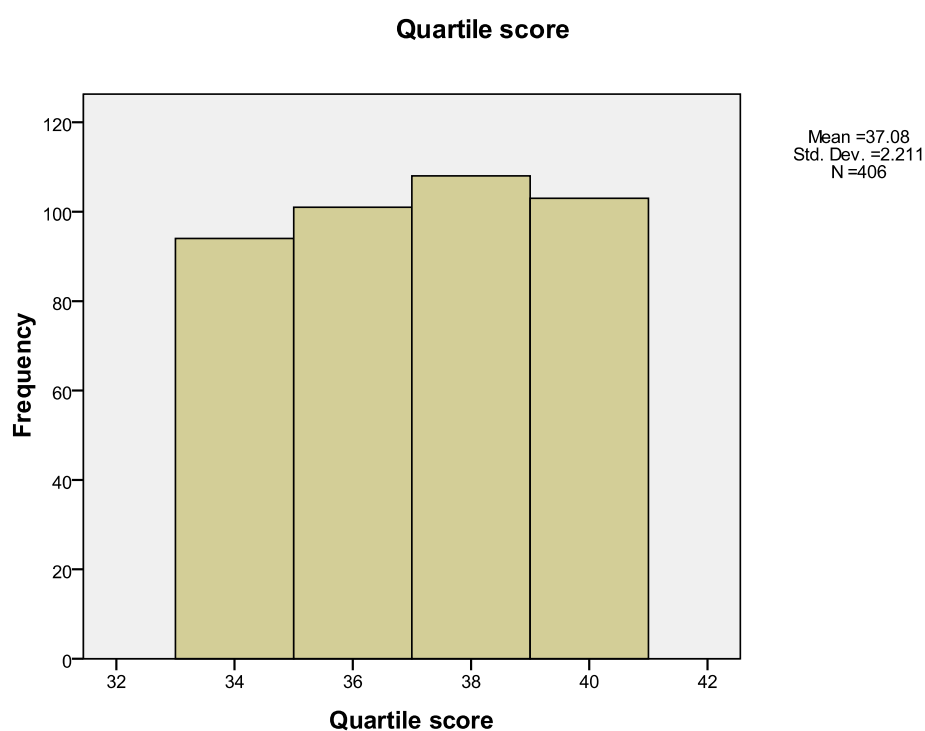


Figure 4: Histogram for Quartile scores

6.3 Overall Summary of Results

- 6.3.1 Overall, 5.1% of participants (23) did not finish the test, categorised by not completing item 65. 4.4% of participants (20) did not complete the last five questions and 1.9% of participants (8) did not complete the last 10 questions. This indicates that the test can be considered to be a 'power test' rather than a speed test and the majority of the population (94.9%) completed the test (characterised by completing the last question) within the two hours.
- 6.3.2 All items were completed by 89% of participants (405), with no clear pattern of missing data (with the exception of the missing items outlined above). 4.4% of participants (20) missed more than five questions.
- 6.3.3 On the participant evaluation form, participants were asked to record the time taken to complete the test if they completed it in less than two hours. The evaluation form was completed by 428 of the 455 participants. 3.1% (14) of participants reported completing the test in 60 minutes or less. The shortest time reported to complete the test was 45 minutes (last item attempted). This is an extremely short time to complete the pilot test and it is likely to represent a lack of test taking motivation. 28.4% (129) of participants reported completing the test between 60 and 90 minutes, and 36.7% (167) of participants reported completing the test between 90 minutes and 120 minutes. Therefore, 68.1% (310) of participants indicated that they completed the test in *less* than the two hours available. 0.9% (4) of participants reported that they completed the test in 120 minutes and 1.3% (6) individuals reported that they completed the test in more than 120 minutes with the longest time reported of 150 minutes (due to extra time allowed or reasonable adjustment). 135 participants did not provide information regarding how long it took them to complete the test.
- 6.3.4 After an initial review of the results, two participants were removed from the analysis. One individual's response was removed as only the first part of the test was completed. Another individual's response was removed due to an erratic answer pattern (only choosing one answer in each of the multiple choice items, instead of three). The sample figures we have therefore presented aim to give the most useful view of the pilot results by excluding only those participants with unusually high levels of missing data and/or erratic response patterns. In an operational test, analysis would be completed on the entire sample, and figures reported as such.

6.4 Test Level Analysis

- 6.4.1 As outlined above, results are reported for two scoring methods; rational scoring (using concordance data) and empirical scoring (using live data). Table 19 below shows a summary of the test level results and Figures 5 and 6 provide an illustration of the distribution of scores.

Scoring	N	Reliability (α)	Mean	Mean%	Skew	SD	Min	Max
Rational	453	0.74	870.1	78.0%	-0.94	37.3	682	952
Empirical	453	0.77	899.6	80.6%	-0.85	39.7	710	992

Table 19: Test level analysis results

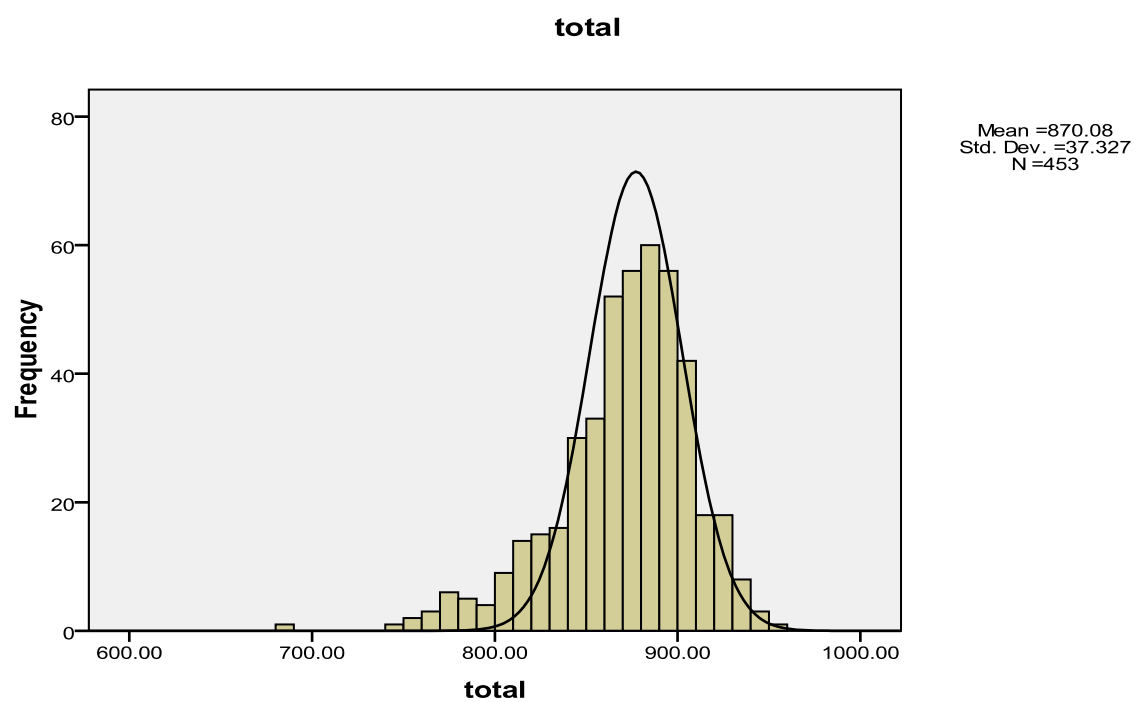


Figure 5: Histogram of rational scoring

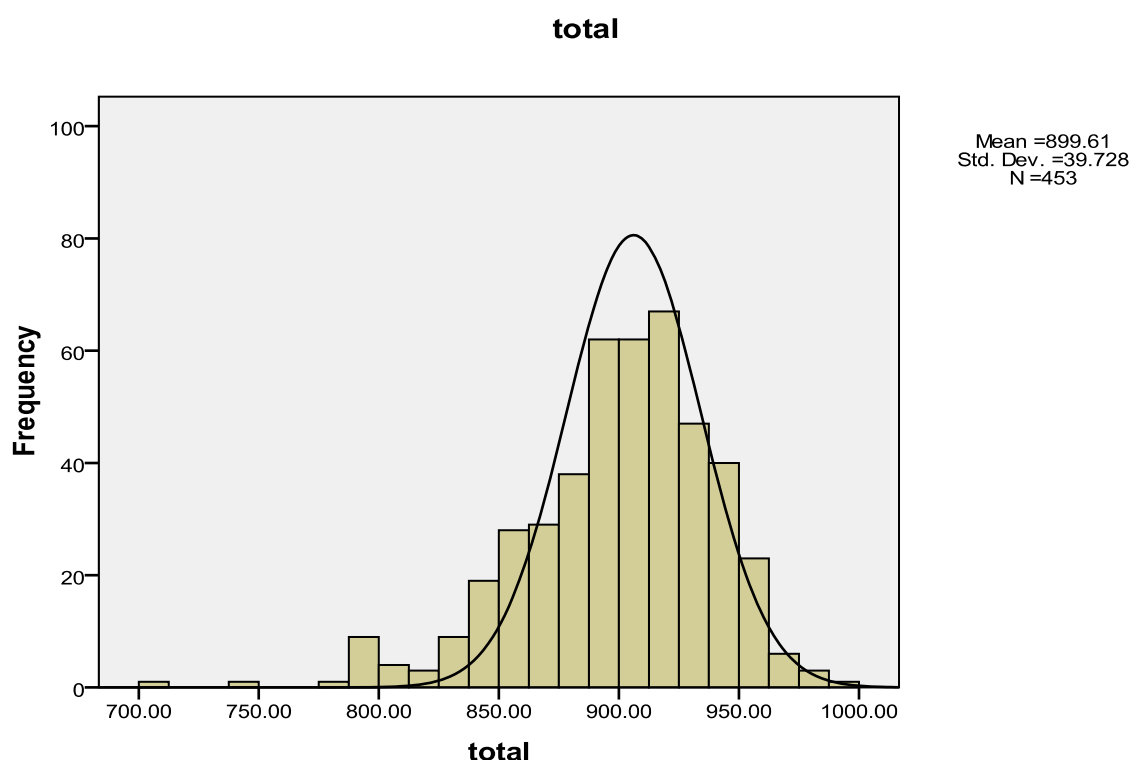


Figure 6: Histogram of empirical scoring

- 6.4.2 For both scoring methods, test statistics showed a high level of internal reliability^{xiv} ($\alpha=0.74$ - 77), only slightly under the level of internal reliability required for a high stakes assessment ($\alpha=0.80$). This is very encouraging considering it is the first time that an SJT has been used in this context.
- 6.4.3 The mean score represents between 78.0% and 80.6% of the total marks available (maximum possible score was 1116).
- 6.4.4 In terms of scoring distribution, there was a wide range of scores (minimum 682 and maximum 952 for rational scoring and minimum 710 and maximum of 992 for empirical scoring). This range of scores provides a good spread and shows that overall the test is able to differentiate between applicants.
- 6.4.5 Overall the distribution of scores is very slightly negatively skewed (see Figures 5 & 6). However, results show a close to normal distribution and indicate that the SJT is capable of differentiating between applicants, although inevitably there are a number of tied scores.
- 6.4.6 Analysis of response patterns suggests that the several low scores observed are not due to participant response format errors such as tying ranks (for ranking items) or providing more than three responses (for multiple response items) but are due to participants not finishing the test (see 6.3.2).

^{xiv} Refers to the extent of consistency of results across items within a test

- 6.4.7 Correlation analysis showed a high level of agreement between the rational keying and empirical keying ($r = .96$, $p < .001$). This suggests that rational keying closely represents empirical keying, and supports the intended use of rational keying for the operational test. If the correlation between the two were low, (i.e. conflict between the two responses) it may indicate that that rational approach would need to be reviewed.

6.5 Item Level Analysis

- 6.5.1 Item analysis was used to look at the difficulty and quality of individual SJT items. Together these can help identify how well the items differentiate between applicants and the results are used to further refine the items and scoring keys.
- 6.5.2 Item facility (difficulty) is shown by the mean score for each item (out of a maximum of 20 for ranking items and 12 for multiple response items). If the facility value is very low, then the item may be too difficult and may not yield useful information. This may also indicate that the scoring key for these particular items may need to be examined again, as there may be little consensus between participants and the concordance panel. If the facility value is very high, then the item may be too easy and may not provide useful information or differentiate between applicants. Broadly speaking, a range of item facilities is needed for an operational test, with few very easy (characterised by a mean score of greater than 90% of the total available score) or very difficult (characterised by a mean score of less than 30% of the total available score) items. However, this specification depends on the purpose of the test; if the aim was only to select the very best performers, then more difficult items would be needed. However, in this context where applicants are being ranked and few if any are to be deselected, it is not necessary to have many very difficult items. Within this test, 'very easy' equates to a score of 18 for ranking and 10.8 for multiple choice, and 'very difficult' equates to 6 for a ranking and 3.6 for a multiple choice.
- 6.5.3 The standard deviation of scores for an item should also be considered. If an item's standard deviation is very small (i.e. most candidates are scoring close to the mean for that item), it is likely to not be differentiating between applicants. The standard deviation for an item should be at least 1.0.
- 6.5.4 When using rational scoring, item facility for ranking items ranged from 14.3 to 18.8 with a mean of 16.2. Four items had a mean score of 18 or above. This suggests that these items may be too easy to provide any useful information about applicants. Item facility for multiple response items ranged from 6.1 to 11.5 with a mean of 8.2. Two items had a mean score of 10.5 or above and consideration should be paid by a psychometrician to removing these items as they may be too easy to provide any useful information about applicants. Each item produced a standard deviation of between 1.5 and 3.3. For all ranking and multiple choice items, a proportion of candidates achieved the maximum possible score for the item, as would be anticipated.
- 6.5.5 When using empirical scoring, item facility for ranking items ranged from 14.2 to 18.8 with a mean of 16.7. Six items' mean scores were 18 or above and consideration should be paid to removing these items as they may be too easy to provide any useful information about applicants. Item facility for multiple response items ranged from 6.1 to 11.5 with a mean of

8.6. Two items had a mean score of 10.5 or above and consideration should also be given to removing these items as they may be too easy to provide any useful information about applicants. Each item produced a standard deviation of between 1.5 and 3.3. For all ranking and multiple choice items, some participants received the maximum score available for that item. These results only differ slightly from the rational scoring, with slightly higher means for both ranking and multiple choice items, as would be anticipated with empirical keying. This supports the intended use of rational scoring as the scoring method.

- 6.5.6 Item quality was determined by the correlation of the item with overall SJT score, not including the item itself (item partial). This analysis compares how individuals perform on a given item with how they perform on the test overall. It would be expected that high scoring participants overall would select the correct answer for each item more often than low scoring participants, i.e. the item discriminates between good and poor applicants. This would show a good to moderate correlation/partial. A poor correlation would indicate that performance on the individual item does not reflect performance on the test as a whole and the item requires further review/revision before being used again^{xv}.
- 6.5.7 For rational scoring, item partials ranged from -0.03 to 0.43, with a mean of 0.17. The item analysis showed that 25% (16) of items were classified as good (item partial ≥ 0.25), 23% (15) as satisfactory (item partial 0.17-0.24), 14% (9) were classified as moderate (item partial 0.17-0.13) and 38% (25) were classified for further review in future (item partial < 0.13). This level of item success (62% acceptable) is typical for SJT development and further review will help to identify the reasons for poor performance of particular items. Five items showed negative item partials, suggesting possible key errors. In this case, removal of any of the less well performing items did not increase the reliability of the test and all items were retained for the remainder of the analysis.
- 6.5.8 For empirical scoring, item partials ranged from 0.04 to 0.42, with a mean of 0.19. The item analysis showed that 22% (14) of items were classified as good (item partial ≥ 0.25), 32% (21) as satisfactory (item partial 0.17-0.24), 19% (12) were classified as moderate (item partial 0.17-0.13) and 28% (18) were classified for further review (item partial < 0.13). No items were negatively correlated with the test overall (negative item partial), which is typical of empirical scoring. Again, here the removal of any of the less well performing items did not significantly increase reliability. Further breakdown of the item level results can be found in Appendix C.
- 6.5.9 The item level results indicate that well over half of the items work well and differentiate sufficiently between applicants. This is an encouraging result for the first use of a test. Where the empirical key differs from the rational key, agreement is being sought with Subject Matter Experts as to whether the empirical key is appropriate. If so, these items can then be placed in the item bank and used in the live test without the need for further piloting.

^{xv} See Appendix A.5 for further details of item partial acceptability criteria

6.5.10 Analysis at an item level indicates that some items do not effectively differentiate between applicants in their present form. This could be for a number of reasons, e.g. weaker participants are guessing the correct response, an item does not assess what is being assessed in the rest of the test, an item is too easy or too difficult to differentiate between applicants. Negative partial correlations indicate that the current item key may be incorrect. When judging the overall quality of the test it should be noted that it is not necessary for all test items to differentiate greatly between applicants in order to create a reliable and effective test, although this would be ideal, so long as there is sufficient differentiation in the test overall. There are other important considerations in test design such as item content and ensuring a sufficient coverage of the target professional attribute domains in the test specification. For those items identified as having low capacity to differentiate between applicants in this pilot, further reviews will be conducted to explore the reasons for the results and updates will be made where appropriate.

6.6 Group Differences

6.6.1 In order to examine fairness issues regarding the use of a SJT for selection in to the FP, group differences in performance within the participant sample were analysed on the basis of sex, age and ethnic group using the rational scoring method.

6.6.2 Table 20 shows group differences in performance on the pilot SJT based on sex. Female participants scored slightly higher than male participants (by approximately 0.3 SD) and a t-test showed that this difference was statistically significant ($p < 0.01$). This is a typical finding with many assessment methods across many occupational groups and comparable with other selection methods.^{12,13}

Sex	N	Mean	SD	Sig Difference
Male	158	862.3	38.5	p<0.01
Female	274	874.0	35.3	

Table 20: Group differences by sex

6.6.3 Table 21 shows group differences in performance on the pilot SJT based on ethnicity. Caution should be employed when interpreting these results due to the very small sample sizes of the majority of the groups. The highest performing group was 'Chinese' and the lowest performing group was 'other ethnic group'. Further analysis of ethnic group differences should be examined in future.

Ethnicity	N	Mean	SD
White	318	876.5	31.8
Asian	66	845.7	36.4
Black	10	864.5	30.9
Mixed	12	882.2	40.2
Chinese	10	884.6	30.1
Other Ethnic Group	7	799.4	70.3
Not Stated	12	838.5	34.5

Table 21: Mean scores by ethnicity

6.6.4 Table 22 shows group differences in performance on the pilot SJT based on ethnicity by white and Black and Minority Ethnic (BME) groups. White participants scored higher than BME participants (by approximately 0.6 SD) and a t-test showed that this difference was statistically significant ($p < 0.01$). This is a typical and comparable finding with many selection methods and examinations¹⁴. Further research here is essential for more fully understanding the issues for the individuals involved and, most importantly, for the contexts within which these observed findings occur. Without detailed systematic research in this area, causal factors cannot be reliably identified¹⁵.

Ethnicity	N	Mean	SD	Sig Difference
White	318	876.5	31.8	$p < 0.01$
BME	105	852.3	43.3	

Table 22: Group differences by ethnicity

6.6.5 Table 23 shows group differences in performance on the pilot SJT based on ethnicity (white and BME groups) and sex. The descriptive statistics suggest that both sex and ethnicity are contributing to the observed score differences, with white females the highest performing group and BME males the lowest performing group.

	Male			Female		
Ethnicity	N	Mean	SD	N	Mean	SD
White	107	869.2	35.4	211	880.2	29.2
BME	44	848.7	42.2	61	854.9	44.3

Table 23: Group differences by ethnicity and sex

6.6.6 In terms of age, there was no significant correlation (Spearman's rho) between age and scores on the SJT.

6.7 Construct Validity

- 6.7.1 All participants in the initial pilot were asked to complete a measure of Social Desirability (Marlowe Crowne Social Desirability scale, Crowne & Marlowe, 1960). Social Desirability bias is a term used in scientific research to describe the tendency of respondents to reply in a manner that will be viewed favorably by others. This will generally take the form of over reporting good behavior or underreporting bad behavior. If a participant answers in a socially desirable manner on the Social Desirability measure, then they may be answering similarly on the SJT and ‘faking good’^{xvi}.
- 6.7.2 The Social Desirability measure consists of 10 items on a 5 point Likert scale. Six of the items are reverse coded. The total maximum score is 50. A high score indicates high levels of social desirability; that is participants have a higher need for social approval.
- 6.7.3 57 participants had to be removed from the analysis due to the lack of identification numbers and/or lack of participation in the measure. Analysis of the measure was carried out and the descriptive statistics can be seen in Table 25 below. A histogram shows the measure to be normally distributed.

Measure	N	Mean Score	Min	Max	SD	Skew	Reliability
Social Desirability	396	34.4	19	47	5.0	-0.30	0.62

Table 25: Descriptive statistics for Social Desirability

- 6.7.4 The mean score (34.4) indicates that participants were, on average, not responding to the Social Desirability scale in a socially desirable way. From this, it can be suggested in principle that participants were not responding to the SJT in a socially desirable way or ‘faking good’. A correlational analysis was also carried out between the social desirability scale and the SJT score; this was not significant ($r=-.04$, $p=0.43$), indicating that those that did demonstrate higher levels of social desirability, were not performing better on the SJT. From this exploratory research, it can be concluded that the SJT is less susceptible to faking than other selection methods and is also evidence that the SJT should be resilient to coaching.

6.8 Participant Feedback

- 6.8.1 All participants were asked to complete an evaluation questionnaire regarding their experience and perception of the SJT. A total of 428 participants completed the questionnaire (94.1% response rate).
- 6.8.2 Participants were asked to indicate their level of agreement with several statements regarding the content of the SJT paper, the results of which are shown below.

^{xvi} This is a term used by researchers for false answers given by interviewees and respondents to questionnaires in an attempt to present themselves in a favourable light.

Participants N=428				
Strongly Disagree %	Disagree %	Neither %	Agree %	Strongly Agree %
The information I was given about the pilot was clear and helpful		19.1 13.8 44.6 34.6 0% 20% 40% 60% 80% 100%		
The instructions for the test were clear and easy to understand		22.2 15.8 44.1 45.5 0% 20% 40% 60% 80% 100%		
The test seemed well-run and well-invigilated		22.2 15.8 44.1 45.0 0% 20% 40% 60% 80% 100%		
The content of the test seemed relevant to the Foundation Programme		28.6 11.7 45.7 33.0 0% 20% 40% 60% 80% 100%		
The scenario content seemed appropriate for my training level		23.1 15.4 50.2 26.6 0% 20% 40% 60% 80% 100%		
The level of difficulty of the test was appropriate		26.2 16.8 51.3 21.2 0% 20% 40% 60% 80% 100%		
The content of the test appeared to be fair		16.9 21.5 49.4 18.3 0% 20% 40% 60% 80% 100%		
The results of the test should help selectors to differentiate between weaker and stronger candidates		11.2 24.9 31.5 24.9 7.2 0% 20% 40% 60% 80% 100%		

Table 26: Pilot participant evaluation form responses

6.8.3 78.7% of participants agreed or strongly agreed that the content of the test seemed relevant to the Foundation Programme, 76.8% of candidates agreed or strongly agreed that

the scenario content seemed appropriate for my training level and 72.5% of participants agreed or strongly agreed that the level of difficulty of the test was appropriate. Participants were asked if they disagreed with the previous statement to indicate if they felt the test was too hard or too easy. Fifty two participants responded; 32.7% reported that they felt the test was too easy and 67.3% responded that the test was too hard. 67.7% of participants agreed or strongly agreed that the content of the test appeared to be fair. Only 32.1% of participants agreed or strongly agreed that the results of the test should help selectors to differentiate between weaker and stronger participants, however 31.5% neither agreed nor disagreed with this statement.

6.8.4 In the free text box, 311 people provided a comment (73% of those who completed the questionnaire). A total of 416 separate issues were captured. Common themes found in responses can be summarised as follows and are also outlined in Appendix A.3. Many of the comments refer to queries that were answered as part of the feedback session and the responses to these queries as discussed in the feedback sessions are also outlined in the table in Appendix A.2.

- Many participants felt that the SJT was a **good idea, relevant, useful and fair and better than the current process**. It was commented that the **scenarios were realistic** and relevant to foundation training.
- Many participants felt that the SJT **could be used in the selection process** alongside other measures, for example the current application process, medical exams or **interviews**. A small number of participants felt the test **should not be used at all**.
- Several participants commented that the SJT did not allow participants to **demonstrate the reasoning/thought process** behind their responses, and others felt the test was **impersonal** and did not allow them to show their personality.
- Some participants felt that the scenarios reflected **experience they would not gain until FY1**. Some felt that some scenarios were **not realistic, ambiguous or subjective**.
- A number of participants felt that the test consisted of **repetitive items** and that the test should be **shortened** as it resulted in fatigue.
- Some participants felt that they **needed more time**, although some reported that the **two hours was sufficient** to complete the questions.

6.9 Conclusions and Recommendations from Initial Pilot

6.9.1 The psychometric analysis of the initial pilot presented is evidence that the SJT is a reliable measurement methodology in this context and is able to differentiate sufficiently between applicants. Item-level analysis showed the majority of the SJT items worked well.

6.9.2 Analysis of the results from the initial pilot indicates that 60 items is a sufficient number to ensure a reliable test. Feedback indicated that some participants felt that the piloted test was too long, produced fatigue, and contained repetitively similar questions. It is therefore recommended that the tests in the large-scale pilot should each contain 60 items, rather than 65, and further analysis of the test length carried out.

- 6.9.3 The time allowed for the test should remain at two hours to ensure it is a 'power' test rather than a 'speed' test. Given the number of professional attribute domains to be tested a two hour test is optimum to ensure broad coverage of domains. In the initial pilot the majority of participants completed the test (94.9%); however participant feedback was mixed as to whether there was too much or too little time allowed.
- 6.9.4 The test should continue to consist of roughly two thirds ranking and one third multiple-choice, as outlined in the design specification. The response instruction 'what should you do', rather than 'what would you do', should also be retained.
- 6.9.5 It is recommended that rational keying, rather than empirical keying is employed in the large scale pilot. The high correlation ($r = .96$, $p < .001$) between the two methods supports this view and evidence suggests that rational keying is a valid approach to scoring SJTs¹⁵.
- 6.9.6 The concordance panel held in September 2010 produced only satisfactory levels of concordance on some items, possibly because some of the participants did not have sufficient recent practical experience of working closely with FY1 trainees in a work setting. The selection and recruitment of participants for the forthcoming concordance panel meetings should ensure that all participants meet the target profile.
- 6.9.7 The FY2 focus group reviews were productive in ensuring item content was appropriate and relevant. However, organising and facilitating separate review meetings was labour intensive, so the option of combining them in some way with the concordance panel events to improve resource efficiency should be explored.
- 6.9.8 The demographic data about participants, which was collected during the pilot from the Foundation Programme Application System (FPAS) system, was sufficient to allow analysis of group differences. However, the FPAS data was not available for a small number of the participants, as some had not made FPAS applications, and others had provided incorrect FPAS identification numbers and could not be matched with their FPAS data. For the main pilot, the option should be explored of collecting demographic data directly from the participants during the SJT (via additional boxes on the answer sheet).
- 6.9.9 In any piloting of a selection test, the motivation of those taking the test may differ to that of applicants for a live selection test. Motivational issues may have impacted upon the time taken to complete the pilot test (with individuals completing the test in a quicker time than may be anticipated on a live test), the number of items completed (with fewer items being completed than on a live test) and the quality of the answers provided. To increase motivation in future pilots it is essential that participants are fully informed as to the purpose and the background of the SJT, including some of the key issues that were raised during the feedback sessions in the initial pilots.
- 6.9.10 The diversity of individuals involved in the development of the SJT was monitored throughout the process. Results indicate that a diverse range of individuals, with regards to sex, age and ethnicity, were involved in the process. Efforts should continue to be made to

ensure that there is a diverse range of individuals involved in the development of the SJT in the future.

- 6.9.11 Differences in performance between males and females and between ethnic groups were found in the SJT pilot; on average males scored slightly lower than females, and the Black and Minority Ethnic (BME) group scored significantly lower than the white ethnic group. Findings of this nature and magnitude are typical with many selection methods across many occupational groups.^{12, 13, 14} Although during test development item writers were asked to ensure that the language used did not unfairly discriminate against participants who were not trained in the UK, an additional language and diversity check is recommended to be carried out as part of the development of the current design methodology to help ensure that the content of the items does not unfairly discriminate against any particular group due to a factor unrelated to the job specification. This is in line with UK employment law. Analysis of group differences in test scores and selection process outcomes should be continued to be monitored for future tests.

7 Large scale Pilot

7.1 Pilot Test Specification

- 7.1.1 The large scale pilot consisted of three pilot papers. This was to ensure that the maximum number of items could be piloted, whilst also ensuring that a sufficient sample size in terms of number of participants where possible for each paper.
- 7.1.2 Each paper consisted of 60 items; 40 ranking and 20 multiple choice. Within each of the papers, 15 of the items were anchor items common to each paper. This was to assist with analysis of the three papers; i.e. whether the anchor items performed the same or differently in the three tests. At this stage, it was not possible to use these anchor items to equate the tests as this was the first time the items had been piloted. In addition, five items from the initial pilot were also included as part of the 60 items. These were the items that were used for feedback to participants with the aim of using these for feedback to participants again. Therefore, overall there were 20 anchor items and these were all included in all further analysis. An outline of the spread of the domains for the 20 anchor items is outlined below.

Anchor items

- Commitment to Professionalism - 7
 - Coping with Pressure - 3
 - Effective Communication - 5
 - Patient Focus - 2
 - Working Effectively as Part of a Team - 3
- 7.1.3 As far as possible, an equal spread of target domains were selected for each paper, however the proportion of items from each target domain is also a reflection of the number of items written within each domain. However, this also had to be balanced with item response format and topic when constructing the papers. An outline of the spread of domains for each of the papers is outlined below:
- a) Paper One
- Commitment to Professionalism - 21
 - Coping with Pressure - 12
 - Effective Communication - 10
 - Patient Focus - 8
 - Working Effectively as Part of a Team - 9
- b) Paper Two
- Commitment to Professionalism - 22
 - Coping with Pressure - 11
 - Effective Communication - 11
 - Patient Focus - 7
 - Working Effectively as Part of a Team – 9

c) Paper Three

- Commitment to Professionalism - 24
- Coping with Pressure - 10
- Effective Communication - 11
- Patient Focus - 6
- Working Effectively as Part of a Team - 9

7.1.4 There were a total of 1040 marks available for each version of the SJT paper.

7.2 Sample

7.2.1 There were a total of 639 participants in the large scale pilot. As outlined in 5.1.5, it was recommended that 400 participants per paper participated in the large scale pilot. This would provide sufficient numbers to allow the psychometric properties of the SJT to be fully analysed and evaluated and would mean that greater confidence could be placed in the results. However, a minimum of 200 participants taking each test would be sufficient to obtain meaningful results and to have confidence in the results.

7.2.2 A breakdown of the number of participants who sat each of the three papers can be seen in Table 27 below. Only one version of a paper was undertaken at each school for logistical reasons, and to minimise security risk with the items. As two papers had less than 200 participants, caution should be used when interpreting the results, particularly for Paper One and Paper Three.

7.2.3 There was an unequal split of participants between the three papers. Although attempts were made to ensure an equal as possible split between the three papers, the number of participants within each of the schools meant that this was not possible.

	No. of participants	Percentage of Overall Sample
Paper One	138	21.6%
Paper Two	344	53.8%
Paper Three	157	24.6%

Table 27: Number of participants taking each paper

7.2.4 Thirteen medical schools participated in the pilots and the breakdown of the number of participants from each medical school is outlined in Table 28.

	No. of participants	Percentage uptake from school	Percentage of Overall Sample	Paper Taken
School E	100	29.4%	15.6%	2
School F	11	7.3%	1.7%	3
School G	57	43.8%	8.9%	2
School H	116	43.0%	18.2%	2
School I	53	16.1%	8.3%	1
School J	9	15.0%	1.4%	1
School K	15	5.2%	2.3%	1
School L	36	14.4%	5.65	3
School M	64	25.6%	10.05	3
School N	71	47.3%	11.1%	2
School O	45	14.1%	7.0%	3
School P	41	24.8%	6.4%	1
School Q	20	8.3%	3.1%	1

Table 28: Number of participants at pilot sites

7.2.5 Participant demographic data were collected directly. Table 29 outlines the breakdown of participants by sex. Overall, more females participated in the pilot (401, 62.8%) than males (228, 35.7%) (reflecting the male/female split of medical students) and the proportion of males and females is roughly equal across all three papers.

		Male	Female	Not declared
Overall	No. of participants	228	401	10
	% of participants	35.7%	62.8%	1.6%
Paper One	No. of participants	47	89	2
	% of participants	34.1%	64.5%	1.4%
Paper Two	No. of participants	122	217	5
	% of participants	35.5%	63.1%	1.5%
Paper Three	No. of participants	59	95	2
	% of participants	37.8%	60.9%	1.3%

Table 29: Participant sex by paper

7.2.6 Table 30 outlines the breakdown of participants by ethnicity. Overall, the majority of candidates reported their ethnicity as white (436, 68.2%) with the smallest proportion of candidates (14, 2.2%) reporting themselves as being black. This reflects the medical student profile in the UK. The proportion of ethnic breakdown is roughly the same across

all three papers, with Paper Two having a slightly higher proportion of white participants than Paper One and Paper Three.

		White	Asian	Black	Chinese	Mixed	Other	Not declared
Overall	No. of participants	436	89	14	31	23	24	22
	% of participants	68.2%	13.9%	2.2%	4.9%	3.6%	3.8%	3.5%
Paper One	No. of participants	92	22	1	6	7	5	5
	% of participants	66.7%	15.9%	0.7%	4.3%	5.1%	3.6%	3.6%
Paper Two	No. of participants	241	45	6	12	13	16	11
	% of participants	70.1%	13.1%	1.7%	3.5%	3.8%	4.7%	3.2%
Paper Three	No. of participants	103	22	7	13	3	3	5
	% of participants	66.1%	14.1%	4.5%	8.3%	1.9%	1.9%	3.2%

Table 30: Participant ethnicity (5+1 groups) by paper

7.2.7 Table 31 outlines the breakdown of participants' ethnicity (White and Black and Minority Ethnic (BME) group). 436 (68.2%) participants reported themselves as white and 181 (28.3%) participants reported themselves as being from Black and Minority Ethnic groups. Paper Two had a slightly higher proportion of white participants (70.1%) than Paper One (66.7%) and Paper Three (66.0%).

		White	BME	Not declared
Overall	No. of participants	436	181	8
	% of participants	68.2%	28.3%	1.3%
Paper One	No. of participants	92	41	5
	% of participants	66.7%	29.7%	3.6%
Paper Two	No. of participants	241	92	11
	% of participants	70.1%	26.7%	3.2%
Paper Three	No. of participants	103	48	5
	% of participants	66.0%	30.8%	3.2%

Table 31: Participants ethnicity (2 groups) by paper

7.2.8 Table 32 outlines the breakdown of participants' nationality. Overall, the majority of participants (553, 86.55%) reported themselves as British. 69 (10.8%) participants reported

their nationality as being other than British. The breakdown across the three papers is roughly the same, with slightly more participants reporting themselves to be British in Paper One compared to Paper Two (2.5%) and Paper Three (4.5%).

		British	Non-British	Not declared
Overall	No. of participants	553	69	6
	% of participants	86.5%	10.8%	0.9%
Paper One	No. of participants	123	12	3
	% of participants	89.1%	8.6%	2.2%
Paper Two	No. of participants	298	37	9
	% of participants	86.6%	10.7%	2.6%
Paper Three	No. of participants	132	20	4
	% of participants	84.6%	12.8%	2.6%

Table 32: Participant nationality by paper

7.2.9 The mean age of the sample was 24.8 years, with a range of 21 – 48 years.

7.2.10 Participants were asked if they had a disability, as outlined by the Disability and Equality Act (2010). 14 (2.2%) participants stated that they had a disability.

	Yes: Disability	No: Disability	Prefer not to disclose	Not declared
No. of participants	14	588	12	25
% of participants	2.2%	92.0%	1.9%	3.9%

Table 33: Participants by disability

7.2.11 In addition to reporting whether they had a disability, on the evaluation form provided to all participants on completion of the test, participants were asked that if they had received extra time, having being classified as having a disability according to the Disability and Equality Act (2010), how much extra time they had been provided with whether they had been provided with. Fifty eight participants said they had extra time, ranging from five minutes to 90 minutes, however it is likely that participants did not complete this form accurately as 30 minutes (n=24) is the recommended extra time for participants with a disability (25% of two hours) and in addition, only 14 participants overall were reported as having a disability. This data therefore has limited use.

7.2.12 To assist with establishing the equivalence of those taking the three papers, the quartile scores and competency based application form (white space question) scores for the participants taking the three tests were examined. Ideally, each population's scores should be normally distributed and should show a good and comparable spread of scores for both the application form and the quartile score as would be expected in a full operational sample. It should be noted that this information was not available for all participants due to incorrectly completed identification numbers.

7.2.13 The results show that participants who took Paper One had slightly higher overall application form scores (Q1-6 and Q2-6) than those participants who took Paper Two and Paper Three and t-test analyses shows that these differences are significant. There was no difference in quartile scores. It may therefore be expected that that the overall score for Paper One would be higher than for Paper Two or Three, as these participants have scored significantly higher on the application form. This is based on the assumption that the application form is a reliable and valid assessment method which correlates positively with the SJT.

	Paper One (n=128)			Paper Two (n=310)			Paper Three (n=133)		
	Mean	Range	SD	Mean	Range	SD	Mean	Range	SD
Application Form Qs 1-6	42.2	29-56	6.2	40.1	21-57	7.0	39.8	14-52	6.9
Application Form Qs 2-6	39.2	27-48	5.4	37.7	21-50	6.1	37.4	13-48	6.4
Quartile score	37.7	34-40	2.1	37.4	34-40	2.2	37.1	34-40	2.4

Table 34: Application form and quartile scores by paper

7.2.14 The range of scores on the application form is less for Paper One, than either Paper Two or Paper Three. The standard deviation is also slightly lower for Paper One compared to the other two papers. The standard deviation indicates how much variation there is from the mean. A low standard deviation indicates that the data points tend to be very close to the mean, whereas a higher standard deviation indicates that the data is spread out over a large range of values. Further examination of the spread of scores (illustrated in Figures 9,10 & 11 below) shows that Paper One is less normally distributed than Paper Two, and Paper Three is slightly skewed.

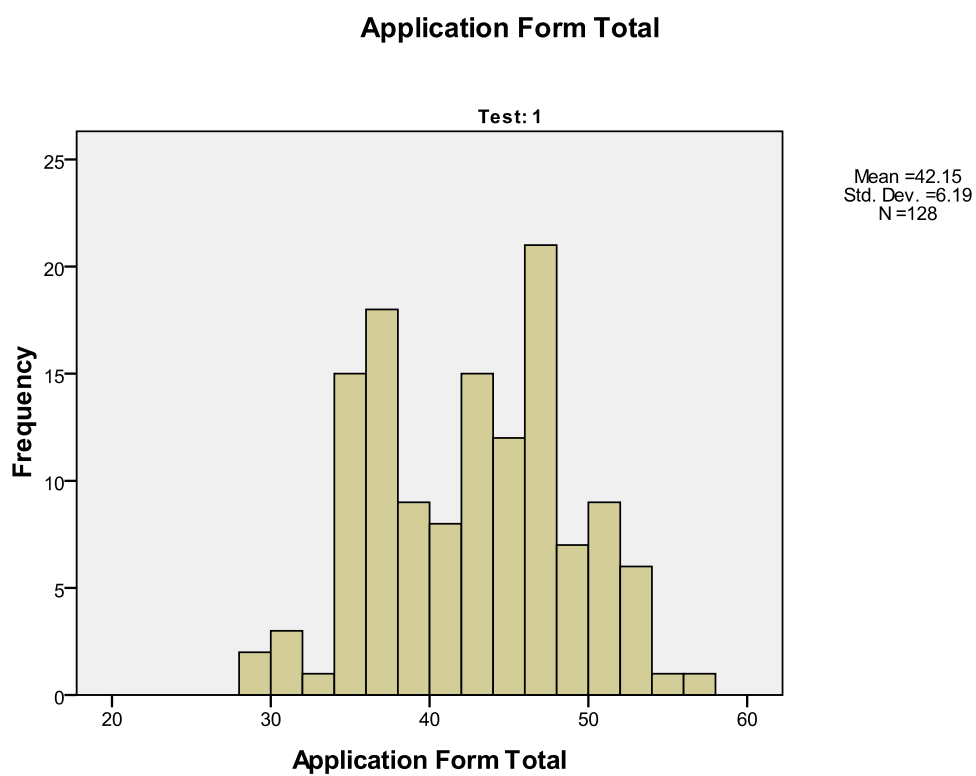


Figure 9: Histogram for Application Form Total for Paper One

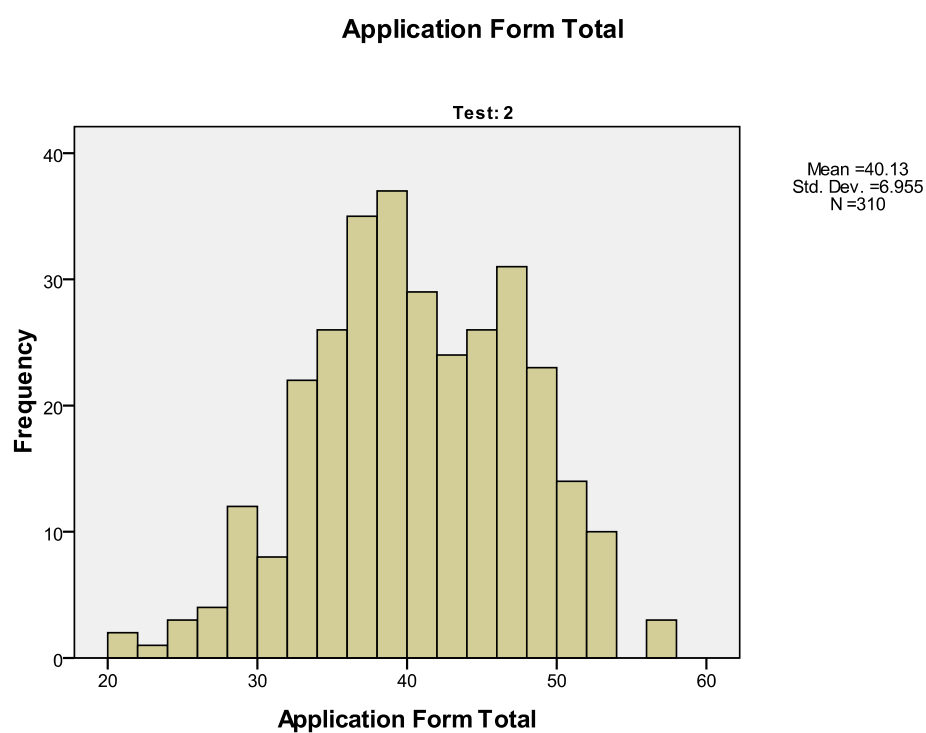


Figure 10: Histogram for Application Form Total for Paper Two

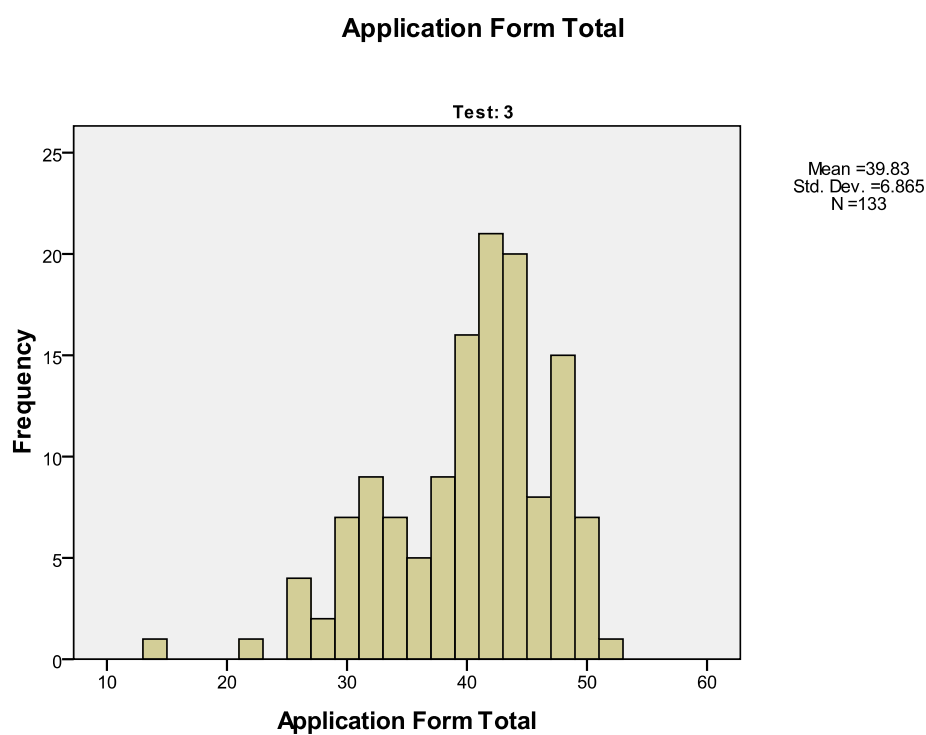


Figure 11: Histogram for Application Form Total for Paper Three

7.2.15 The range for the quartiles is the same for all three papers with similar standard deviations. However, further examination of the data shows that the distribution of scores for the three papers is not fully comparable with there being a higher proportion of participants in Paper One scoring at the higher end of the quartiles with Papers Two and Three having more of an even spread.

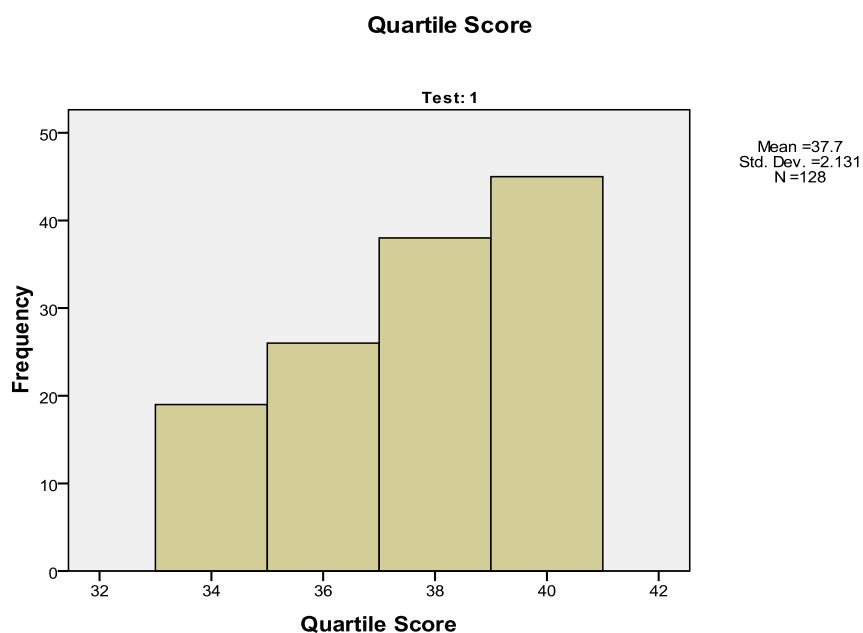


Figure 12: Histogram for Quartiles for Paper One

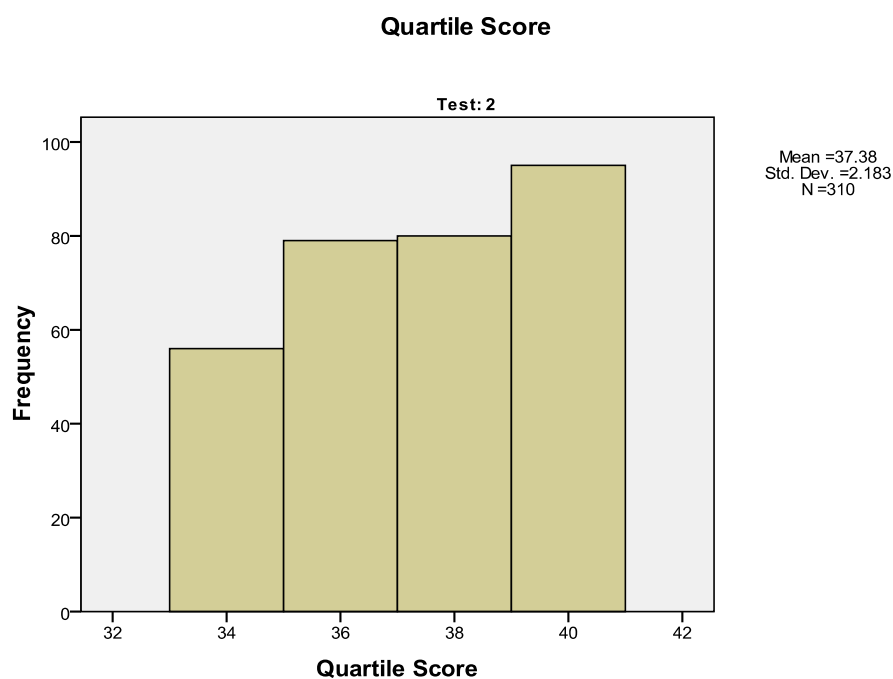


Figure 13: Histogram for Quartiles for Paper Two

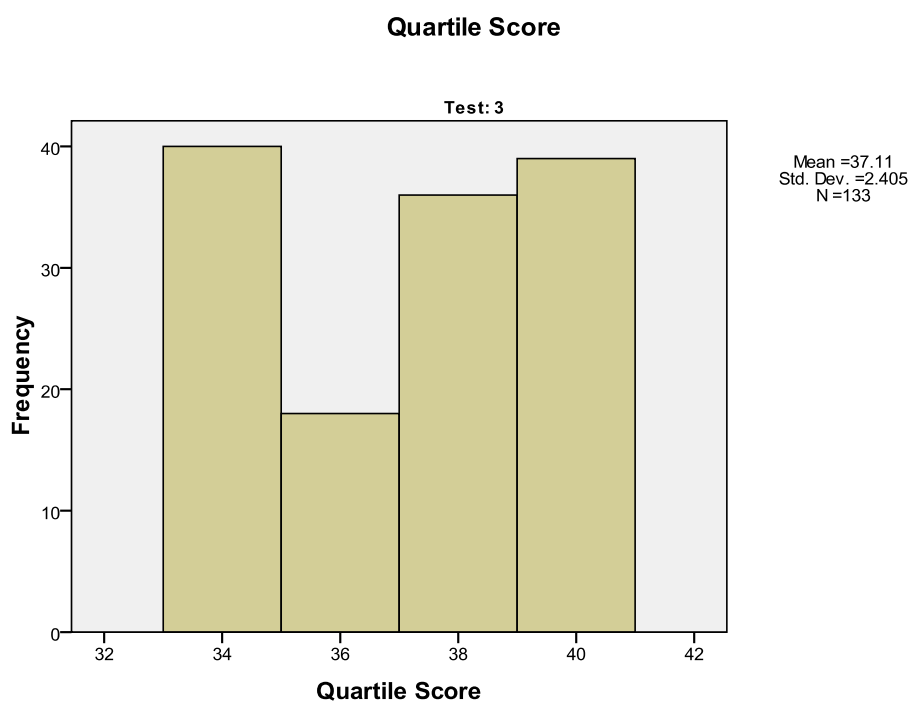


Figure 14: Histogram for Quartiles for Paper Three

7.3 Overall Summary of Results

- 7.3.1 Overall, 18 participants did not finish the test, categorised by not completing item 60. This represents only 2.8% of the entire population. In Paper One, only one candidate (0.7%) did not complete the final question, in Paper Two, 12 people (3.5%) did not complete the final question and in Paper Three, five participants (3.2%) did not complete the final question. This is very encouraging and following the initial pilot, reconfirms that the SJT is a power test, rather than a speeded test with 97.2% of participants completing the test. This indicates that 120 minutes is an appropriate length of time to complete 60 questions. Analysis of candidate's opinions with regards to the length of the test will be used in conjunction with these results to make a recommendation as to the appropriate test length.
- 7.3.2 Ten participants (1.6%) did not complete the last five questions and six participants (0.9%) did not complete the last 10 questions. This was roughly equal over the three tests (Last 5 questions = Paper One = 1, 0.7%, Paper Two = 6, 1.7%, Paper Three = 3, 1.9%) (Last 10 questions = Paper One = 1, 0.7%, Paper Two = 3, 0.9%, Paper Three = 2, 1.3%). This indicates that the test is not a speeded test as the majority of the population (97.8%) completed the test (characterised by completing the last question) within the two hours.
- 7.3.3 Overall, 614 participants (96.0%) completed all items. Twenty five participants (Paper One = 4, 2.8%, Paper Two = 12, 3.5%, Paper Three = 9, 5.8%) had some missing data, ranging from missing 1-28 items. Eleven participants (1.7%) missed more than five questions.
- 7.3.4 On the participant evaluation form, candidates^{xvii} were asked to indicate how long it took them to complete the test if they completed the test in less than two hours. This section was not completed by 129 participants (20.2%). For the remaining participants there was a range of values (5-200 minutes). Some of these values indicate that the form was not completed accurately. Seven participants reported times above 120 minutes, but only three of these indicated that they had been awarded extra time so this data is either inaccurate, or there were issues with administration. Issues with administration are unlikely as an ISFP representative was present at most pilots.
- 7.3.5 Removing outliers (any data below 30 minutes and data above 120 minutes that is not accompanied by extra time), the Table 35 below outlines how long candidates reported it took them to complete the test.

^{xvii} The number of candidates reported here is 652; this is how many candidates completed the candidate evaluation form in the large scale pilot, plus in Malta and Dublin

Time	Number	Percentage of candidates
30-59	16	2.4%
60-89	183	28.1%
90-104	219	33.6%
105-119	93	14.3%
121-135	3	0.5%
Missing	129	19.8%
Removed	9	1.4%

Table 35: Reported time taken to complete test

- 7.3.6 The reported time taken to complete the test indicates that 120 minutes is sufficient time to complete the test. In total 78.4% of participants reported that it took them less than 120 minutes to complete the test. However, this must be analysed in the context of this being a pilot test. Test taker motivation is likely to be much lower than on a real test and participants' would not necessarily take as much time as they would for an operational test.
- 7.3.7 After initial review of the results, some participants were removed from the analysis. For Paper One, 1 participant was removed due to a large number of missing data, for Paper Two, 3 candidates were removed from the analysis due to a large amount of missing data and from Paper Three, 4 participants were removed from the analysis. Three due to large numbers of missing data, and 1 participant was removed due to an erratic answer pattern (tying ranks for all 40 ranking items). The sample figures we have therefore presented aim to give the most useful view of the pilot results by excluding only those participants with unusually high levels of missing data and/or erratic response patterns. In an operational test, analysis would be completed on the entire sample, and figures reported as such.

7.4 Test Level Analysis

- 7.4.1 Test level analysis was carried out for all three papers separately. The figures below illustrate the test level descriptives along with further detail in Table 36. Data are not provided for all three tests combined as the tests have not been equated and the data would therefore be meaningless. It is not possible to equate the tests until following piloting of all items.

	Paper	N	Reliability (α)	Mean	Mean %	Skew	Kurtosis	SD	Min	Max	SEM
Paper One	60 items	137	0.71	842.5	81.0%	-2.3	10.8	32.9	658	913	17.7
	33 items	137	0.75	-	-	-	-	-	-	-	-
	33 items corrected*	137	0.85	-	-	-	-	-	-	-	12.7
Paper Two	60 items	341	0.74	851.8	81.9%	-1.2	3.9	33.7	659	916	17.2
	48 items	341	0.75	-	-	-	-	-	-	-	-
	48 items corrected*	341	0.79	-	-	-	-	-	-	-	15.4
Paper Three	60 items	153	0.76	849.1	81.6%	-.85	1.5	36.3	700	922	17.8
	44 items	153	0.79	-	-	-	-	-	-	-	-
	44 items corrected*	153	0.84	-	-	-	-	-	-	-	14.5

Table 36: Test level descriptives by paper

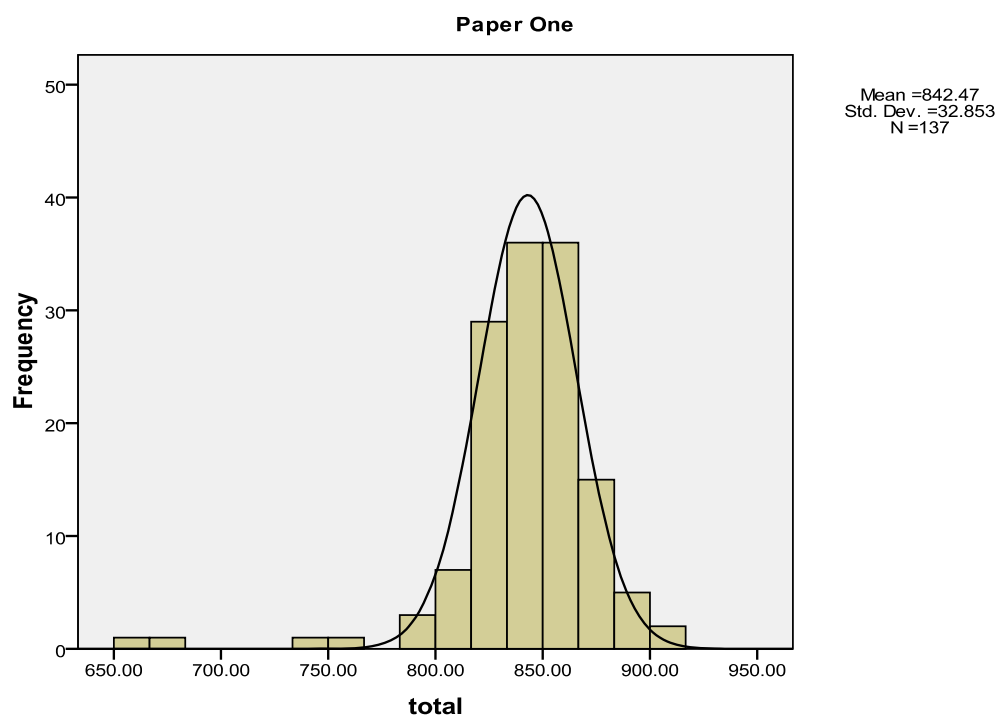


Figure 15: Distribution statistics for Paper One

* Corrected using Spearman Brown's formula; a formula relating psychometric reliability to test length and used by psychometricians to predict the reliability of a test after changing the test length

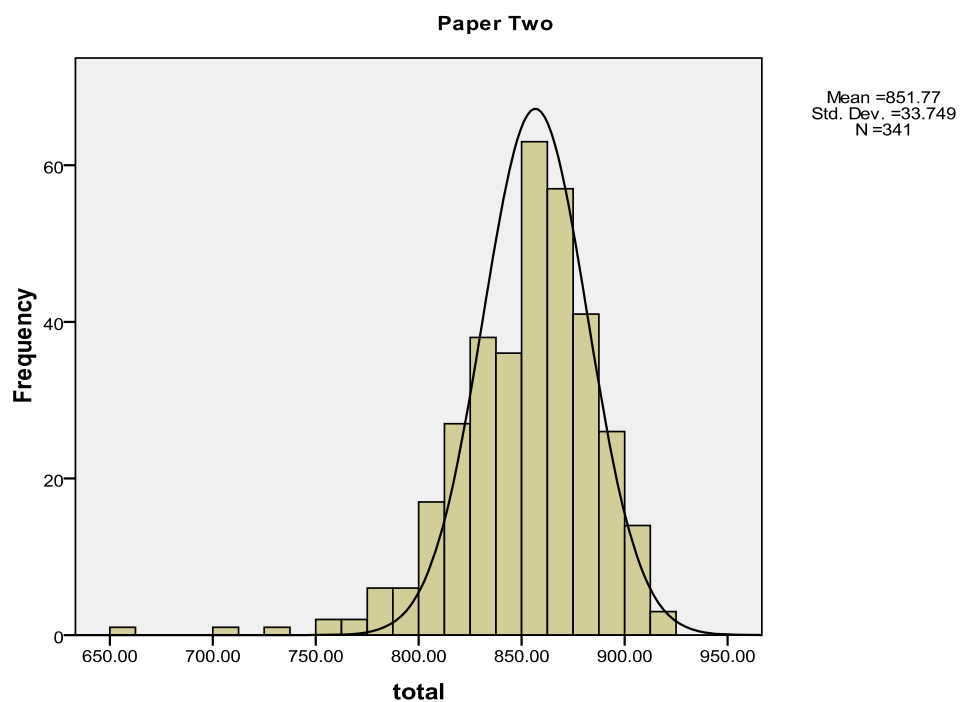


Figure 16: Distribution statistics for Paper Two

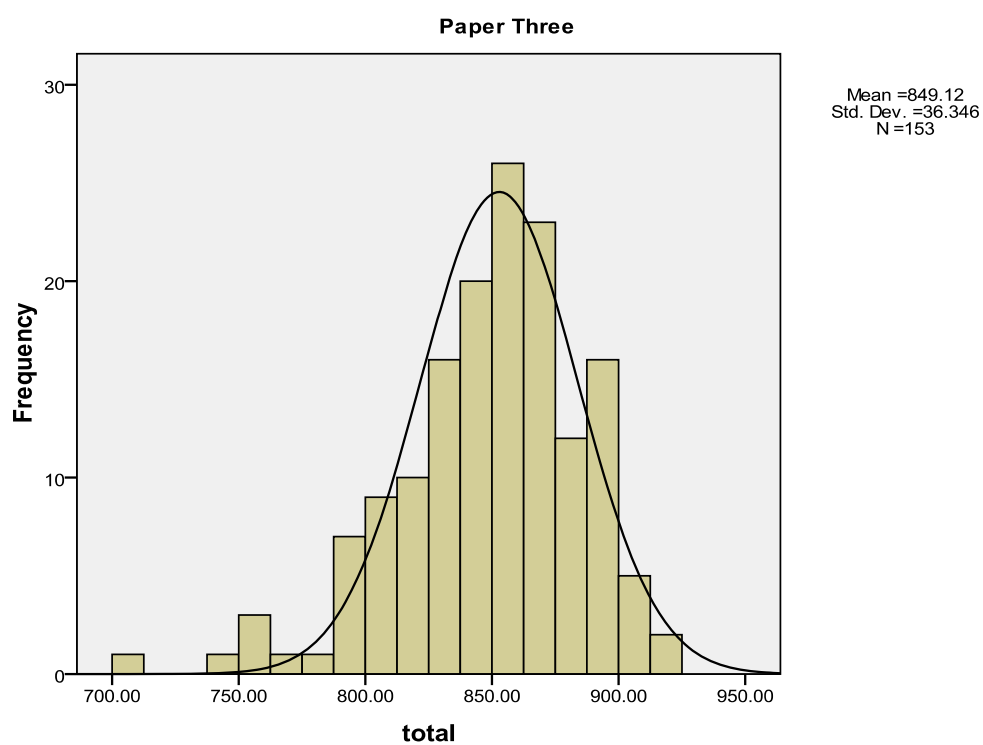


Figure 17: Distribution statistics for Paper Three

- 7.4.2 The mean score for all three papers is similar. Paper One shows the lowest mean score (842.5) and Paper Two shows the highest mean score (851.8). The standard deviation for all three papers is broadly comparable, with Paper Three having a slightly higher standard deviation than Paper One or Paper Two. The kurtosis figure for Paper One indicates that more of the variance is the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations.
- 7.4.3 The mean scores represent between 81.0% and 81.9% of the total marks available (maximum possible score of 1040). This is comparable with the mean percentage score for the initial pilot (78.0%). It is possible that the five feedback questions were known to participants. However, analysis of the mean scores of these items from the two pilots would indicate that this is not the case (Appendix D).
- 7.4.4 In terms of scoring distribution, there was a wide range of scores (minimum 658 and maximum 922). This range is similar to that achieved in the initial pilot (minimum 682 and maximum 952). This range of scores provides a good level of spread and shows that overall the test is able to differentiate between participants.
- 7.4.5 Overall the distribution of scores is slightly negatively skewed. However, results show a close to normal distribution and indicate that the SJT is capable of differentiating between participants.
- 7.4.6 Analysis of response patterns suggests that the several low scores observed are not due to participant response format errors such as tying ranks (for ranking items) or providing more than three responses (for multiple response items) but are due to participants not completing all items (see 5.3.3)^{xviii}.
- 7.4.7 One of the aims of the pilot was to pilot new items to establish whether they have sufficient psychometric properties to enter the item bank. Only those items with sufficient psychometric properties will be used in a live test. Therefore, as well as presenting test reliability for all items within the test (including those with poor psychometric properties), we also provide test reliability for only those items which have sufficient psychometric properties (0.13 and above) and would therefore be used in a live operational test. The Spearman Brown formula is then applied to provide an estimation of the reliability of a 60 item test with similar quality of items.
- 7.4.8 For all three papers, test statistics showed a relatively high level of internal reliability ($\alpha=0.71 - .76$), before being corrected by excluding poorly performing items. When only those items that have good psychometric properties are analysed and corrected for test length using the Spearman Brown formula, all three papers show good reliability, sufficient for that required for a SJT being used in a high stakes assessment ($\alpha=.80$)^{xix}.

^{xviii} See Appendix A.3 for further technical information

^{xix} See Appendix A.4 for further commentary on reliability statistics

7.5 Item Level Analysis

- 7.5.1 Item analysis was used to look at the difficulty and quality of individual SJT items. Together these can help identify how well the items differentiate between participants and the results are used to further refine the items and scoring keys^{xx}.
- 7.5.2 Item facility (difficulty) is shown by the mean score for each item (out of a maximum of 20 for ranking items and 12 for multiple response items). If the facility value is very low, then the item may be too difficult and may not yield useful information. This may also indicate that the scoring key for these particular items may need to be examined again, as there may be little consensus between participants and the concordance panel. If the facility value is very high, then the item may be too easy and may not provide useful information or differentiate between participants. A range of item facilities is needed for an operational test, with few very easy (characterised by a mean score of greater than 90% of the total available score) or very difficult (characterised by a mean score of less than 30% of the total available score) items. However, this specification depends on the purpose of the test; if one was only selecting the very best performers, then more difficult items would be needed. However, in this context where participants are being ranked, it is not necessary to have many very difficult items. Within this test, 'very easy' equates to a score of 18 for ranking and 10.8 for multiple choice, and 'very difficult' equates to 6.0 for a ranking and 3.6 for a multiple choice.
- 7.5.3 The standard deviation of an item should also be considered. If an item's standard deviation is very small, it is likely to not be differentiating between participants. The standard deviation for an item should be at least 1.0.
- 7.5.4 For Paper One, item facility for ranking items ranged from 13.9 to 18.6 with a mean of 16.5. Three items' mean scores were 18 or above and consideration should be paid to these items as they may be too easy to provide any useful information about participants. Item facility for multiple response items ranged from 7.4 to 11.8 with a mean of 9.0. Two items scored above 10.5 or above and consideration should be paid to these items as they may be too easy to provide any useful information about participants. Each item produced a standard deviation of between 0.8 and 3.1. Those items that had a standard deviation of below 1 are unlikely to be differentiating, so should be reviewed in light of this. Four items received the maximum score available.
- 7.5.5 For Paper Two, item facility for ranking items ranged from 13.5 to 18.3 with a mean of 16.7. Only one item's mean score was 18 or above and consideration should be paid to this item as it may be too easy to provide any useful information about participants. Item facility for multiple response items ranged from 7.9 to 10.7 with a mean of 9.2. Three items scored above 10.5 or above and consideration should be paid to these items as they may be too easy to provide any useful information about participants. Each item produced a standard deviation of between 1.5 and 3.0. Four items received the maximum score available.

^{xx} See Appendix D for item level results

- 7.5.6 For Paper Three, item facility for ranking items ranged from 14.2 to 18.4 with a mean of 16.6. Two items' mean scores were 18 or above and consideration should be paid to these items as they may be too easy to provide any useful information about participants. Item facility for multiple response items ranged from 7.6 to 11.7 with a mean of 9.2. Three items scored above 10.5 or above and consideration should be paid to these items as they may be too easy to provide any useful information about participants. Each item produced a standard deviation of between 1.2 and 3.1. Four items received the maximum score available.
- 7.5.7 Item quality was determined by the correlation of the item with overall SJT score, not including the item itself (item partial). This analysis compares how individuals perform on a given item with how they perform on the test overall. You would expect that high scoring participants overall would select the correct answer for each item more often than low scoring participants, i.e. the item discriminates between good and poor participants. This would show a good to moderate correlation/partial. A poor correlation would indicate that performance on the individual item does not reflect performance on the test as a whole^{xxi}. Table 36 below outlines how items performed for each of the three papers and overall^{xxii}. Paper One has a higher number of items that require further review compared to Paper Two and Paper Three.

	Overall	Paper One	Paper Two	Paper Three
Range of Item Partial	-0.10 – 0.46	-0.10 - 0.46	-0.09 – 0.31	-0.05 – 0.43
Mean Item Partial	0.19	0.16	0.20	0.20
Good (≥ 0.25)	41	13 (22%)	9 (15%)	19 (32%)
Satisfactory (0.17-0.25)	63	15 (25%)	29 (48%)	19 (32%)
Moderate (0.13-0.17)	21	5 (8%)	10 (17%)	6 (10%)
Item requires further review (< 0.13)	55	27 (45%)	12 (20%)	16 (26%)

Table 37: Item level statistics

7.6 Group Differences

- 7.6.1 In order to examine fairness issues regarding the use of a SJT for selection into the FP, group differences in performance within the participant sample were analysed on the basis of sex, ethnicity, nationality and age for each of the three papers. Group differences for all three papers overall is not provided as the three tests had not been equated and therefore the data would not be meaningful.

^{xxi} See Appendix A.4 for further details of item partials acceptability criteria

^{xxii} See appendix A.5 for item level analysis of anchor items

7.6.2 Table 38 shows group differences in performance on the SJT based on sex. There were no significant differences in the mean SJT scores for males and females for Paper Two and Three. Female participants scored slightly higher than male participants (by approximately 0.5 SD) on Paper One and a t-test showed that this difference was statistically significant ($p < 0.01$). Findings of this nature and magnitude are typical with many selection methods across many occupational groups.^{12,13}

	Sex	N	Mean	SD	Sig Difference
Paper One	Male	46	831.4	43.7	$p < 0.01$
	Female	89	848.1	24.4	
Paper Two	Male	120	851.7	32.7	ns
	Female	216	853.3	31.3	
Paper Three	Male	57	843.6	34.8	ns
	Female	94	852.3	37.0	

Table 38: Mean scores by sex

7.6.4 Table 39 shows group differences in performance on the SJT based on ethnicity by White and Black and Minority Ethnic (BME) groups. White participants scored higher than BME participants on all three tests and t-tests show that this difference was statistically significant ($p < 0.05$ - $p < 0.01$). Findings of this nature and magnitude are typical with many selection methods and examinations¹⁴. A richer understanding of the implications of the observed groups differences in practice (for sex and ethnicity) is needed as an impetus for future research. Without detailed systematic research in this area, causal factors cannot be reliably identified¹⁷.

	Ethnicity	N	Mean	SD	Sig Difference
Paper One	White	92	845.9	29.3	$p < 0.05$
	BME	40	832.9	39.9	
Paper Two	White	239	859.9	24.0	$p < 0.01$
	BME	91	835.5	36.7	
Paper Three	White	102	860.3	29.0	$p < 0.01$
	BME	47	824.1	39.1	

Table 39: Mean scores by ethnicity

7.6.5 Table 40 shows group differences in performance on the SJT based on ethnicity (white and BME groups) and sex. The descriptive statistics suggest that the relationships between ethnicity, sex and SJT performance varied somewhat across the three papers. For Papers One and Three, the highest performing group was white females, whereas for Paper Two, white males and females performed very similarly. For Papers One and Two, the lowest performing group was BME males, whereas for Paper Three, it was BME females.

		Male			Female		
	Ethnicity	N	Mean	SD	N	Mean	SD
Paper One	White	27	832.5	40.3	65	851.4	21.4
	BME	17	827.5	51.5	22	836.4	29.7
Paper Two	White	82	861.2	23.6	157	859.2	27.3
	BME	36	829.5	40.0	55	839.5	34.1
Paper Three	White	33	850.2	31.7	68	865.9	26.0
	BME	22	832.5	38.3	25	816.7	39.1

Table 40: Mean scores by ethnicity and sex

7.6.6 Table 41 shows group differences in performance on the SJT based on nationality (British and non-British). British participants scored higher than non-British participants on all three papers and t-tests show that these differences were statistically significant ($p < 0.01$) for Paper Two and Three. However, any interpretation of these results should be treated with caution as numbers in the non-British group are small.

	Nationality	N	Mean	SD	Sig Difference
Paper One	British	122	843.4	31.1	ns
	Non-British	12	834.2	50.1	
Paper Two	British	295	855.4	30.9	$p < 0.01$
	Non-British	37	830.8	31.5	
Paper Three	British	130	854.7	32.3	$p < 0.01$
	Non-British	19	813.7	41.4	

Table 41: Mean scores by nationality

7.6.7 In terms of age, there was a negative correlation ($r = -.13$ - Spearman's rho, $p < 0.01$) between age and scores on the SJT; that is younger participants scored significantly better on the SJT than older participants. Analysis by paper shows that this is non-significant for Paper One, is significant for Paper Two ($r = -.12$, $p < 0.05$) and significant for Paper Three ($r = -.21$, $p < 0.01$).

7.6.8 Differential Item Functioning (DIF) was examined at an item level. The DIF analysis is a procedure used to determine if test items are fair and appropriate for assessing the ability of various demographic groups. It is based on the assumption that test takers who have similar ability (based on total test scores) should perform in similar ways on individual test items regardless of their sex or ethnicity. DIF is a necessary but not sufficient condition for bias: bias only exists if the difference is illegitimate, i.e., if both groups should be performing equally well on the item. An item may show DIF but not be biased if the difference is due to actual differences in the groups' ability to answer the item, e.g. if one

group is high proficiency and the other low proficiency, the low proficiency group would necessarily score much lower.

- 7.6.9 DIF, undertaken using a multiple regression analysis, was used to examine whether the demographic variable (e.g. sex) significantly predicts performance on each item once overall test performance has been controlled for (i.e. is there a difference in item performance beyond that which expected due to differences between groups on the test overall?).
- 7.6.10 28 (15.6%) items were flagged for sex differences (14 favoured males and 14 favoured females). 6 of these items were anchor items (appeared in all papers), 7 of the items were in Paper One, 7 of the items were in Paper Two and 8 of the items were in Paper Three. 15 (8.3%) items were flagged for ethnicity differences (8 favoured white candidates and 7 favoured BME candidates). One of these items was an anchor item, 2 of the items were in Paper One, 7 items were in Paper Two and 5 items were in Paper Three. As items favouring males and items favouring females, and items favouring white and BME candidates are present in equal proportions, strongly suggests that the test is not biased.
- 7.6.11 These items will be reviewed in light of these results following the pilot to identify whether there appears to be any bias in the item content. Once reviewed, if the items do appear to demonstrate bias (as outlined above, DIF is a necessary but not sufficient condition for bias), items will either be adjusted and re-piloted or will be removed from the item bank.

7.7 Computer Versus Paper Tests

- 7.7.1 The SJT was undertaken both in paper and pen format (n=413), and using a secure server on computer (n=218). Primarily, the two media were piloted to evaluate the logistical possibilities; however it is also interesting to compare how participants performed on the SJT using either medium. Both Paper One and Paper Two were completed using both media, whereas Paper Three was completed solely in paper format.
- 7.7.2 Analysis of the mean scores of computer versus paper and pen (irrelevant of test version), show that the medium in which the SJT was completed did not appear to influence the results (Paper = 848.8, SD = 33.8, Computer = 849.7, SD = 35.6), however this result should be treated with caution as this sample comprises two different test versions. Descriptive statistics for the two methods for each paper are outlined in Table 42 below. As outlined, there is no significant difference in mean scores between the two media for Paper Two. For Paper One there is a significant difference between mean scores, with those undertaking the paper on a computer scoring significantly higher than those taking the paper using paper and pen ($p < 0.05$). This is likely to reflect, at least in part, differences between the candidate groups who used each medium, although it could reflect aspects of the media, e.g. some individuals using paper and pen not completing some items within the test.

Paper	N	Reliability (α)	Mean	Mean %	Skew	SD	Min	Max	Sig Diff
Paper One (Paper and Pen)	74	.76	836.9	80.5%	-2.65	37.8	658	891	p<0.05
Paper One (Computer)	63	.54	849.0	81.6%	0.21	24.6	797	913	
Paper Two (Paper and Pen)	186	.64	853.3	82.0%	-0.52	28.7	761	916	ns
Paper Two (Computer)	155	.81	850.0	81.7%	-1.43	39.0	659	916	

Table 42: Mean scores by medium

7.7.3 A number of practical issues were reported using the computer delivered format which may have caused variability, making it difficult to infer conclusions about the results. Some of these are summarised below:

- At one school screen resolution meant about two thirds of participants had difficulty looking at the questions, making scrolling necessary. There was some frustration about this which may be reflected in performance in the test.
- At the same school, participants had to use an old version of the software (the newest wasn't installed), so the test items were all the same but a few of the issues like font size were not addressed and some of the administrative control features didn't work fully .
- At another school, there were some software problems at the beginning of the test and the computers crashed for five minutes. Participants talked during this time and led to the test starting informally. In addition, some participants filled in details and were then logged out and had to re-fill everything in. This caused some problems as then some computers were logged in twice, although at no point was the test security compromised.
- In another school there were problems with setting up the computers so there was no briefing and led to the test starting informally and people starting at different times.

7.8 Construct Validity

7.8.1 Previous research shows SJTs to be correlated with personality¹⁷, cognitive ability¹⁸, and job knowledge¹⁹. To attempt to examine the construct validity of the SJT, participants at one school, were invited to complete the NEO-PIR personality measure after completing the SJT to examine the associations between scores on the SJT and measures of personality.

7.8.2 The NEO-PIR is a well-established measure of five major dimensions of personality (or domains). Each of the five domains consists of six facet scales. The five domains and their facets are outlined below:

- **Emotional stability;** the general tendency to experience negative affect such as fear, sadness or embarrassment.
 - Anxiety
 - Frustration
 - Despondency
 - Self Consciousness
 - Impulsiveness
 - Stress Proneness
- **Extraversion;** denotes where individuals tend to obtain their energy from. Extraverts like excitement and stimulation and tend to be cheerful in disposition. Introversion may be seen as the absence of extraversion, rather than its opposite and introverts tend to be more reserved and prefer to spend time alone.
 - Friendliness
 - Sociability
 - Assertiveness
 - Activity
 - Excitement Seeking
 - Positive Emotions
- **Openness to experience;** open individuals are often intellectually curious and are willing to entertain novel ideas and unconventional values. Those who score lower tend to be more conventional in behaviour and conservative in outlook.
 - Fantasy
 - Aesthetics
 - Feelings
 - Actions
 - Ideas
 - Values
- **Agreeableness;** those who are high on agreeableness tend to be fundamentally altruistic and sympathetic to others. Those lower on the scale tend to be more disagreeable, sceptical and competitive rather than cooperative.
 - Trust
 - Straightforwardness
 - Altruism
 - Compliance
 - Modesty
 - Compassion

- **Conscientiousness**; those who score high on this domain are purposeful, strong-willed and determined, those who score lower tend to be less exacting and less rigid in working towards their goals.

- o Competence
- o Order
- o Dutifulness
- o Achievement Striving
- o Self Discipline
- o Deliberation

7.8.3 A total of 108 of the 118 participants at School H completed the NEO-PIR measure. 37.1% (43) of these participants were male and 56.0% (65) were female. 6.9% (8) did not provide their sex. The mean age of the participants was 23.6 years and 77.6% (90) of the sample was white and 14.6% (17) reported themselves as from Black and Minority Ethnic (BME) groups. 7.8% (9) did not provide their ethnicity. The mean overall SJT score for the sample was 861.1^{xxiii}.

7.8.4 Correlation analyses were carried out between overall SJT scores and the five major dimensions of personality. All correlations were non-significant (Emotional Stability - $r=-.03$, $p=.78$, Extraversion - $r=.13$, $p=.18$, Openness to Experience $r=.13$, $p=.17$, Agreeableness - $r=-.04$, $p=.65$, Conscientiousness - $r=.16$, $p=.11$) indicating that at a dimension level, there is no relationship between the SJT and personality.

7.8.5 Significant correlations were found between overall SJT scores and three of the individual facets; Extraversion 4 - Activity ($r=.21$, $p<0.05$), Openness to Experience 6 - Values ($r=.23$, $p<0.05$) and Conscientiousness 4 - Achievement Striving ($r=.26$, $p<0.01$). The definitions of each of these facets are outlined below, together with the finding from the analysis:

- **E4 - Activity**: A high activity score is seen in rapid tempo and vigorous movement, in a sense of energy and in a need to keep busy. Active people lead fast paced lives. Low scorers are more leisurely and relaxed in tempo, although they are not necessarily sluggish or lazy. High scores on the SJT were associated with high scores on Activity.
- **OE 6 - Values**: Openness to values means the readiness to re-examine social, political and religious values. Closed individuals tend to be less open to new ways of working. High scores on the SJT were associated with high scores on Values.
- **C4 - Achievement Striving**: Individuals who score high on this facet have high aspiration levels and work hard to achieve their goals. They are diligent and purposeful and have a sense of direction in life. Low scorers are lackadaisical and

^{xxiii} See Appendix A.6 for descriptives and group differences

perhaps even lazy. They are not driven to succeed. High scores on the SJT were associated with high scores on Achievement Striving.

7.8.6 The correlations found with these data suggest that those individuals who have higher activity levels, higher openness to values and higher levels of achievement striving perform better on the SJT than those with lower levels of these facets. These findings are consistent with expectations and the job analysis results. For example, the role of a Foundation doctor (and beyond) requires an individual to deal with a fast paced work environment which is constantly changing. An ability to thrive in this context is important. Similarly, training to be a doctor is a long and hard process; individuals who excel at this must have high aspirations and work hard to achieve their goals. Those who score highly on Values are less likely to judge people and are accepting of people despite the way society may judge them. This attribute may be seen as essential for doctors, who must treat and respect people whatever their political or social values or situations. These results provide preliminary evidence of construct validity of the SJT.

7.8.7 When undertaking large numbers of correlations, it is possible that significant correlations found may be artificially inflated. As such, a conservative approach was undertaken and a Bonferroni correction^{xxiv} was carried out. This resulted in there being no significant correlations between overall SJT score and facet levels of personality.

7.9 Criterion-related Validity^{xxv}

7.9.1 To assist with establishing the criterion related validity of the SJT, analysis was carried out using the data from one of the participating medical schools, which shall remain anonymous for the purpose of confidentiality. The participants at this school had taken SJT Paper Two. Recent examination data (written and OSCE data) is used as the criterion (outcome) variable and SJT, application form and quartile scores as the predictor variables. The quartile score is calculated using exam data up to year 4 and using both the written and OSCE data; the examination data analysed below (Year 5) is not used to calculate the quartile score.

7.9.2 Criterion related data were available for 96 candidates; two candidates were removed due to high numbers of missing data in the SJT and following peer review, a further outlier was also removed from the analysis. 93 candidates were therefore included in the analysis.

7.9.3 Examination data were available for both the written exam and the OSCE. Details of the variables analysed are outlined below.

- % raw score for written exam (out of 150)
- % raw score for Single Best Answer component of written exam (out of 30)

^{xxiv} The Bonferroni correction is a multiple-comparison correction used when several dependent or independent statistical tests are being performed simultaneously. In order to avoid a lot of spurious positives (as high as 1 in 20), the significance level is lowered to account for the number of comparisons being performed. This conservative analysis strategy was adopted so as not to over-interpret the results.

^{xxv} See Appendix A.7 for further details of criterion-related validity

- % raw score for Extended Matching Questions component of written exam (out of 120)
- % raw score for OSCE exam (17 stations)
- % raw score for Part One OSCE exam (35% of total, **clinical examination** stations)
- % raw score for Part Two OSCE exam (45% of total, **communication skills**)
- % raw score for Part Three OSCE exam (20% of total, **practical skills**)

7.9.4 The descriptive exam data for the SJT sample can be seen in Table 43 below. Analysis of the data for those that undertook the SJT and those that did not, indicate that the samples performed similarly on the examinations.

	Mean	SD	Min	Max	Skew
Written Overall	77.6	7.40	54.7	92.0	-0.86
Written SBA	64.2	10.30	30.0	83.3	-.62
Written EQM	80.9	7.49	58.3	94.2	-.97
OSCE Overall	79.9	6.30	62.0	90.4	-.54
OSCE Part 1	84.1	8.90	50.5	98.1	-1.05
OSCE Part 2	75.7	8.43	53.1	91.8	-.32
OSCE Part 3	81.9	6.09	66.0	92.8	-.50

Table 43: Descriptive statistics for criterion data

7.9.5 All SJT participants involved in this analysis had taken Paper Two. The mean score for this sample was 845.5, which is slightly lower than the mean score for Paper Two (851.8) with an internal reliability of $\alpha=.74$. The data are slightly negatively skewed but is otherwise normally distributed (see Figure 18). The scores on quartile and application forms were also analysed for the sample population whose scores are further reviewed as part of criterion related validity testing. The results show that the scores of this sample are comparable to those of the whole SJT sample that sat Paper Two, although this sample had slightly higher application form scores (43.7) than the rest of the sample (40.1).

	N	Mean	SD	Min	Max	Skew
SJT	93	845.5	32.86	754	906	-.44
Quartile	86	37.5	2.10	34	40	-.23
Application form	86	43.7	6.65	25	57	-.52

Table 44: Descriptive statistics for predictor variables

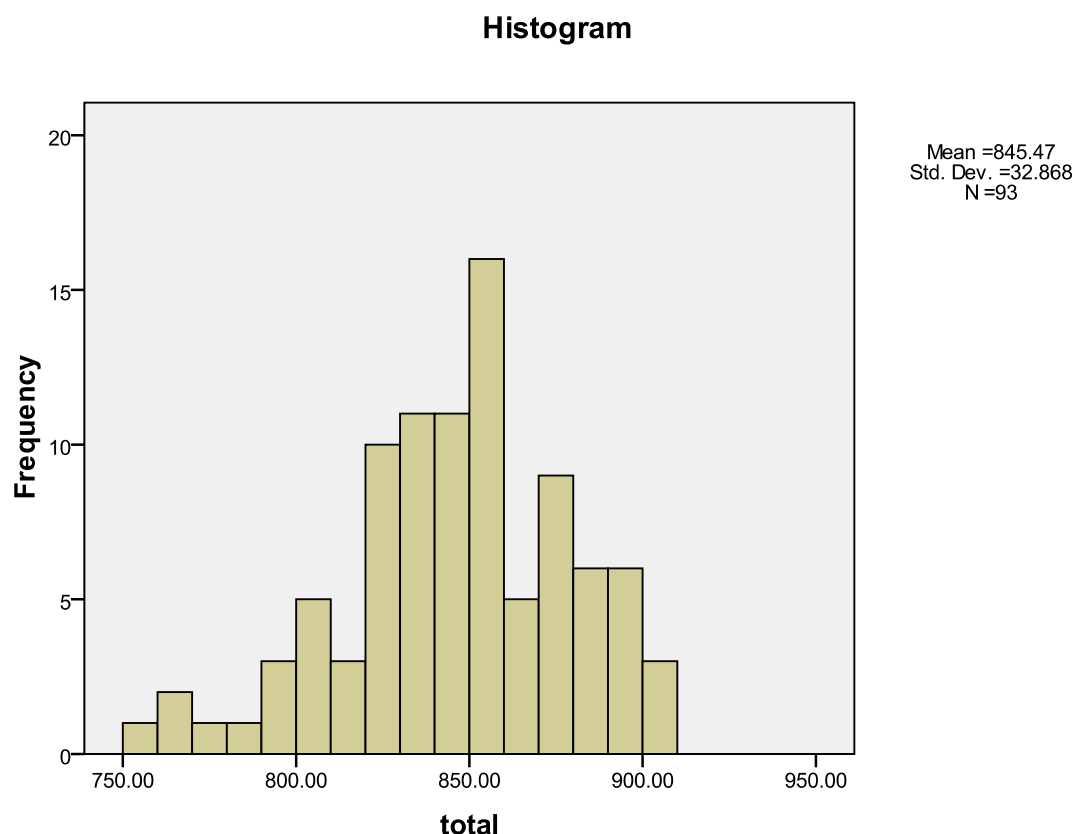


Figure 18: Distribution statistics for SJT total scores

7.9.6 Analysis of the associations between examination scores and the SJT score was conducted. The SJT total score was significantly and positively related to all seven criterion related variables, as well as the quartile score and application form scores. Analysis of the application form scores was also conducted. For the total application score (q1-6) significant correlations were found for all but two aspects of the OSCE (Part 1: clinical examination and Part 3: practical skills). The table below shows the correlation results and the highest correlation for each measure (applications form, quartiles and SJT) is highlighted in bold. The SJT correlates most highly with the OSCE overall and OSCE Part 2 - Communication skills. This is in line with expectations as the SJT is designed to target a range of professional (non-cognitive) attributes identified in the job analysis. These results provide preliminary evidence of the predictive validity of the SJT.

	Application Form q1-6	Quartile	Written Overall	Written SBA	Written EQM	OSCE overall	OSCE Part 1 Clinical Exams	OSCE Part 2 Comm Skills	OSCE Part 3 Practical Skills
Application Form q1-6	-	.29**	.33**	.29*	.31**	.33**	ns	.32**	ns
Quartile total	.36**	-	.69**	.54**	.66**	.59**	.32**	.52**	.49**
SJT total	.40**	.36**	.30**	.38**	.25*	.49**	.36**	.42**	.32**

Table 45: Correlations between predictor and criterion variables^{xxvi} * = p<.05; ** = p<.01

7.9.7 To further explore predictive validity, a series of multiple regression analyses was conducted for both criterion measures (written and OSCE). First, a combination of quartile scores together with application form scores explained 33% of the variance in overall OSCE performance. However, a combination of quartile scores together with SJT scores explained 38% of the variance in OSCE performance, indicating that the combination of the SJT and quartile scores is an improved predictor of OSCE exam performance than the combination of application form and quartile scores. For the criterion measure of written overall score, quartile scores together with application form scores explained 48% of the variance in written exam performance. A combination of quartile scores together with SJT scores also explained 48% of the variance in written exam performance indicating there is no difference in the predictive validity of the two combinations of methods for written examination performance.

7.9.8 The findings from the predictive validity study indicate that the SJT correlates with all aspects of the examination data. The SJT correlates most highly with the OSCES (a behaviour based exam) and also correlates, to a lesser extent, with the written exam. *In broad terms, this supports the proposition that the SJT is an aptitude test for clinical practice* (i.e. high fidelity assessment that replicates what trainees will face in FY1), rather than a replication of the knowledge testing (i.e. written exams) completed throughout medical school and is good preliminary evidence of criterion-related validity. However further longitudinal validation work will allow us to confirm this finding.

7.10 Participant Evaluation

7.10.1 All participants who participated in the main pilot – as with the initial pilot - were asked to complete an evaluation questionnaire regarding their experience and perceptions of the SJT. A total of 652 participants completed the questionnaire^{xxvii}.

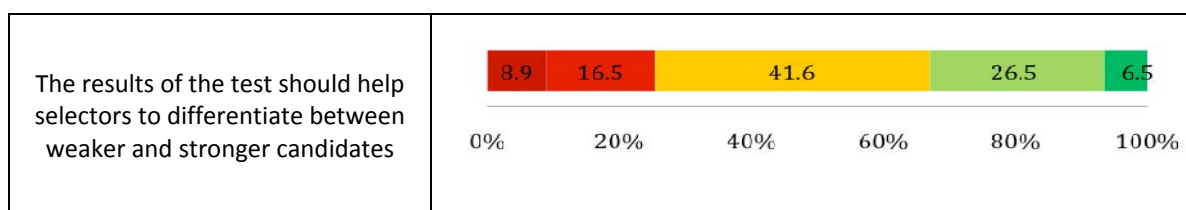
7.10.2 Participants were asked to indicate their level of agreement with several statements regarding the content of the SJT paper, the results of which are shown in the Table 46.

^{xxvi} *Correlation is significant at the 0.05 level (2-tailed)

**Correlation is significant at the 0.01 level (2-tailed)

^{xxvii} This includes some participants from a non UK sample who undertook the paper from the initial pilot but were not included in analysis due to limited sample size



**Table 46: Participant evaluations**

7.10.3 89% of the participants felt that the test instructions were clear and easy to understand, while 85% thought that the information given about the pilot was clear and helpful. 85% of participants also agreed or strongly agreed that the content of the SJT seemed relevant to the FP. 80% felt that the scenario content was appropriate for their level of training and 75% considered that the difficulty level was appropriate. If they felt that the level of difficulty was not appropriate, participants were asked to indicate whether they felt that the test was too hard or too easy. Forty-three (8.5%) participants responded; four participants felt that the test was too easy and 39 felt that it was too difficult. Overall, 73.3% of participants agreed or strongly agreed that the content of the test was fair, but when considering whether the results of the test would help differentiate between the strong and weak participants, 42% neither agreed nor disagreed.

7.10.4 Participants were invited to provide qualitative comments on the SJT pilot prior to the feedback session that provided further information about the purpose of the ST and relevant details regarding its use and validity. Some of the comments refer to queries that were answered as part of the feedback session. Additional comments were provided by 441 (67.6%) participants. A total of 770 separate issues were captured. The comments are detailed below around common themes. A table of all themes and number of comments can be found in Appendix A.8.

7.10.5 SJT Overall

- A large number of participants commented that they found the SJT overall to be a **good, fair, and reasonable** addition the application process and a **better method compared to the current system**.
- A number of participants commented that the **SJT should be combined with other selection methods** rather than used on its own.
- There was disagreement as to whether the length of the test was appropriate. Some participants (20) stated that the **test time was too long**. However, some participants (15) commented that there was **not enough time to complete the test**.
- Some participants felt **unsure whether the SJT would differentiate between the participants** however, an almost equal number of participants felt that the SJT was **better than white space questions at differentiating between 'stronger and weaker' applicants**.

7.10.6 SJT Item Content

- It was commented by some participants that the scenarios **were relevant to FY1 doctors**, that the **scenarios were realistic** and that the SJT assessed the **professional attributes** required for a Foundation doctor.
- Some participants felt that there was a hint of **ambiguity** within some item scenarios and that the scenarios and responses were **subjective**. Some participants felt that **more information** was **required** in order to make an informed judgement, however several participants felt that there was **too much content** in each SJT item, and that it took a long time to read and digest the question.
- A small number of participants commented that some of the **items and responses** were **similar to each other**. It was also commented that participants saw some **responses as equally important**, which made it difficult to answer items.
- Some participants commented that they **might respond differently** in a clinical setting compared to the responses they provided **to the SJT items**.
- A number of participants felt that it was difficult to answer the SJT items due to **lack of experience** with the scenarios.
- Some participants were concerned about a risk of future applicants **being coached how to answer the SJT**, as they felt that it would be easy to be taught how to answer the SJT items from unofficial, external sources.

7.10.7 Logistics of running an SJT

- **Computer related issues** were commented upon. Some participants felt that the **user interface was difficult** to use as participants thought that the **drag and drop system was not user friendly**; however others found the computer system easy to use and preferred this to paper-based test delivery.
- The participants who completed the SJT using pen and paper said that the test **might work better on a computer**.
- Some participants commented that the SJT was run in a **satisfactory way**, however a minority of participants felt that the **examination conditions were poor**, such as the cramped space and the number of invigilators in the room.

7.11 Conclusions from the Large Scale Pilot

7.11.1 The psychometric analysis presented in this report is evidence that the SJT is a reliable measurement methodology in this context and is able to differentiate sufficiently between applicants. Item-level analysis showed the majority of the SJT items worked well, and identified the areas where other items need further review.

7.11.2 Reliability estimates were provided for each test, both including all items and only for those items that had sufficient psychometric properties. Reliability estimates were reasonable when all item were included in the analysis, however when only items that had sufficient psychometric properties were included, all three papers showed good reliability, sufficient for a high stakes selection setting ($\alpha=.80$). As only items that have sufficient psychometric

properties will go into the item bank and be used in operational tests it is concluded that the SJT demonstrates good reliability in this context.

- 7.11.2 Three papers were undertaken in this pilot to maximise the number of items that could be piloted (135 new items). However, the use of three papers was dependent on having sufficient numbers of participants taking each test to be able to have confidence in the results of the psychometric analysis. A total of 639 participants took part in the main pilot. Paper Two had sufficient numbers to be able to obtain meaningful results (n=344), although ideally numbers would have been above 400 to ensure robust results. Paper One and Paper Three each had less than 200 participants taking the test (n=138 and n=157 respectively). This means that any conclusions and interpretations made based on this data should be treated with caution. However, a minimum sample size for the psychometric analyses for each paper is N=100 and so given the sample size for each paper there can be a reasonable degree of confidence in the results.
- 7.11.3 The psychometric analysis indicates that three pilot SJT papers did not perform the same; in particular Paper One appears to have performed quite differently from Paper Two and Paper Three. This might be expected as it was the first time many of these items had been piloted with candidates. In particular, the psychometric analyses showed that Paper One had lower reliability and the two response formats in this paper did not correlate as highly as the other two papers. This could indicate that there are differences in how the anchor items in Paper One relate to the rest of the items in the paper (as also shown by the item partials). This could be due to differences in the content of the items or issues with administration. For a live operational test, the psychometric properties of the item would be known in advance.
- 7.11.4 To examine the construct validity of the SJT, participants at one medical school were invited to complete the NEO-PIR personality measure after completing the SJT to examine the associations between scores on the SJT and measures of personality. The correlations found with these data suggest that those individuals who have higher activity levels, higher openness to values and higher levels of achievement striving perform better on the SJT than those with lower levels of these facets. However, when a conservative approach was taken, due to a large number of correlations, no significant correlations were found between overall SJT score and facet levels of personality.
- 7.11.5 To assist with establishing the criterion related validity of the SJT, analysis was carried out using the data from one of the participating medical schools. The findings from the analysis indicate that the SJT correlates with all aspects of the examination data. The SJT correlates most highly with the OSCES (a behaviour based exam) and also correlates, to a lesser extent, with the written exam. *In broad terms, this supports the proposition that the SJT is an aptitude test for clinical practice* (i.e. high fidelity assessment that replicates what trainees will face in FY1), rather than a replication of the knowledge testing (i.e. written exams) completed throughout medical school and is good preliminary evidence of criterion-related validity. However further longitudinal validation work will allow us to confirm this finding.

- 7.11.6 Two test formats were used in the large scale pilot; paper and pen and computer. In the future it is possible that the SJT may be undertaken on computer, and as such, the large scale pilot was used as an opportunity to pilot both formats. Logistically, the pilot gave opportunity to understand the process of taking the test on computer further and allowed any teething problems or issues to be understood and addressed. Ideally, any single paper would only be undertaken using one format. The reason for this is that using two formats creates unsystematic error or 'noise' that makes psychometric evaluation challenging. As such, some of the variability in the psychometric properties of Paper One and Paper Two may be due to the two different response formats being used, rather than due to the SJT items and/or the individuals taking the test.
- 7.11.7 The pilots were undertaken at 13 UK medical schools. Administration and invigilator guidelines were provided to all schools and an ISFP representative was present at each pilot, to answer questions from participants and ensure the smooth running of the test. Overall, the administration appeared to go well, however there were a small number of administrative issues which may have affected performance in the test, and must be addressed in delivering the SJT for live selection.
- 7.11.8 There were also a number of issues with administration of the computer delivered sessions and these are summarised in 7.7.3 above.

Part Three: Conclusions & Recommendations

8 Conclusions

- 8.1 The authors were commissioned to support the development and validation of the SJT. Specifically the requirements were to design a SJT according to the specification and analyse the effectiveness of the test according to specified evaluation criteria.
- 8.2 A thorough and detailed design process was undertaken that produced sufficient items to be piloted and analysed.
- 8.3 The psychometric analysis presented in this report is evidence that **the SJT is a reliable measurement methodology in this context and is able to differentiate sufficiently between applicants**. Item-level analysis showed the majority of the SJT items worked well. **Analysis of examination data (both written test formats and OSCEs) shows early evidence of criterion related validity.**
- 8.4 Although the number of participants was smaller than proposed at the outset for the large scale pilot (639), we are confident that the datasets were sufficient to prove the concept, demonstrate reliability and preliminary criterion-related validity of the SJT. The results show that **the SJT is an appropriate method of selection in this context.**
- 8.5 Considerations around test taking motivation should also be taken into account. In any piloting of a selection test, the motivation of those taking the test may differ from that of participants for a live selection test. Motivational issues may have impacted upon the time taken to complete the pilot test (with individuals completing the test in a quicker time than may be anticipated on a live test), the number of items completed (with fewer items being completed than on a live test) and the quality of the answers provided.

9. Recommendations

9.1 Overview

- 9.1 Based on the evidence provided in the report and the evaluation of the SJT, the authors are providing a number of recommendations for the stakeholders led by the Medical Schools Council and the Department of Health to consider, given our understanding of the specification and requirements of the project.

9.2 Test Specification

- 9.2.1 Following on from the rationale outlined in the initial design specification, it is recommended that the test should consist of 60 items and last for two hours. There should also be allowances for applicants entitled for additional support including extra time. This is supported by the results of the initial and large scale pilot. In both the initial pilot and the large scale pilot, the majority of participants completed the test (as defined by completing the last question) within the two hours (94.9% and 97.2% respectively) indicating that this is not a speeded test but a power test. Analysis of the data indicates that a test length of 60 items is required to achieve the desired level of reliability.

- 9.2.2 Feedback from participants from the large scale pilot indicated that only a minority of participants (15, 2.3%) commented that the test was too long. Only 13 (1.9%) individuals also commented that the test contained too many items. In the large scale pilot 78.4% of participants reported that it took them less 120 minutes to complete the test indicating that the test is not a speeded test. However, these results should be considered in light of test taking motivation.
- 9.2.3 Two response formats were piloted; ranking and multiple choice. Evidence from the micropilot reviews indicated that these two response formats were acceptable to participants, and tested the appropriate skills. As outlined in the initial design specification, the use of two response formats allows for a greater variety of scenarios (that are mapped to the professional attribute target domains) to be sampled as each of the response formats lends itself to different subject matter content.
- 9.2.4 In both pilots, the division of response formats has been roughly two thirds ranking and one third multiple choice. This enables a balance between the greater granularity of data offered by the more complex ranking format and the quicker completion time of the multiple choice format, whilst ensuring a full range of scenarios suited to each item type can be incorporated into the test, therefore maximising use of the item bank. This balance of item types has been shown to work effectively in other validated medical SJTs (e.g. UK GP).
- 9.2.5 The response instruction 'what should you do', rather than 'what would you do', has been employed throughout the pilot papers, based on the rationale outlined in the initial design specification. Although some comments have been raised from participants about the validity of using this type of response instruction, it is recommended that this continues to be used. This is supported by research evidence suggesting that this type of response instruction is less susceptible to coaching in high stakes assessments¹⁹.
- 9.2.6 It is recommended that items continue to target the five target attributes outlined in the design specification, which reflect the UKFPO person specification. As outlined previously, these domains should be seen more as facets relating to the overall construct (professional attributes of FY1 doctors), rather than independent sub-tests. As such the test should always be treated as a uni-dimensional test, which aims to provide an overall measure of a single broad construct, with the target domains used predominantly as a guide for item writers to ensure a good spread of facets within the person specification is sampled.
- 9.2.7 The scoring convention employed in the SJT provided results that demonstrated good reliability and differentiated between participants. It is recommended that analysis is carried out in the future, modelling different scoring conventions. However, it is advised that robust evidence and a clear rationale is established before changes are made to the current scoring convention.
- 9.2.8 It is recommended that rational keying, rather than empirical keying is employed. The high correlation ($r = .96$, $p < .001$) between the two methods supports from the initial pilot supports this view and evidence suggests that rational keying is a valid approach to scoring

SJTs¹⁵. In addition to the research evidence supporting this approach, rational keying is more defensible. For example, just because better scoring candidates thought that an answer was correct, does not make this so and indeed an answer deemed as right by better performing candidates, may in fact have serious consequences. In high stakes settings such as this, rational scoring is recommended.

- 9.2.9 The large scale pilot was delivered in both paper and pen format and computer format. Unfortunately, the small sample sizes for this, and other unpredicted factors administration and useability of the computer software, mean that clear recommendations are unable to be made as to future administration of the SJT. However, there was no obvious effect on participants' answers to the SJT according to the mode of delivery.

9.3 Analysis & Evaluation

- 9.3.1 It is recommended that further predictive validity studies are conducted in future using live operational data and following trainees in practice. The criterion-related validation conducted on the pilot data reported here indicates that the SJT predicts performance and further predictive validation studies would continue to help establish the long-term validity of the SJT.

9.4 Principles of Standard Setting

- 9.4.1 Standard setting is the process used to define required levels of performance in an assessment and identify cut scores corresponding to those levels. In contexts such as licensing examinations, it is clearly important to identify a cut score which reflects the minimum required standard for registration so that those below this standard do not proceed. However, in selection settings there may be no clear need for a cut score or minimum standard, e.g. if all participants are potentially appointable/trainable and are to be ranked from the top down until all available places are filled. A judgement by the project group needs to be made regarding whether a minimum acceptable standard/cut score is needed/meaningful for the SJT and the potential implications of this (e.g. unfilled FY1 places).
- 9.4.2 If it is determined that a cut score is needed for the SJT, an appropriate standard setting methodology needs to be chosen to identify this cut score in an effective and defensible manner. It is likely to be more difficult to agree on an appropriate minimum standard for a broad measure of professional attributes such as the SJT as opposed to e.g. a clinical knowledge test, where the expectations at a particular level may be much clearer. Standard setting methods may involve detailed expert content review of the SJT (e.g. an Anghoff type procedure) and/or utilise data gained through a predictive validation study.
- 9.4.3 An alternative to establishing a cut score would be to 'flag' any participant who scores more than 2 standard deviations below the mean and their performance further reviewed. However, it would not be recommended to identify individuals who perform less well in any single domain. As outlined in section 5.4; SJTs can be viewed as 'broad band' uni-dimensional tests, in that the test overall aims to measure a single, broad construct as determined by the design specification. In this context, the SJT aims to provide an overall

measure of the professional attributes required of an F1 doctor; within this are domains, or facets, relating to this overall construct. Factor analysis carried out on an operational test of similar specification supports this showing one overall single factor¹¹. As such, an overall should be used, rather than scores for individual target domains, as it is not the intention to provide a separate assessment of each individual domain.

9.5 Principles of Test Equating

- 9.5.1 If the intention is to administer the SJT across multiple dates within one year/cohort (either as multiple single-day administrations throughout the year or as a single multiple-day testing window), multiple versions of the SJT will be required and these will need to be equated so that results from different versions are comparable. The number of versions required and how these are best allocated across the testing dates will depend on the number and schedule of testing dates employed.
- 9.5.2 The construction of the multiple test versions will need to adhere to a number of rules to ensure that the versions are as equivalent as possible before equating, e.g. same number of ranking/multiple choice items in each version, similar spread of domains/topic areas, item difficulties and item qualities across versions.
- 9.5.3 In order to equate the papers statistically, there needs to be a proportion of anchor/shared items within each paper; these can either be common to all papers or common to particular pairs of papers and this may depend upon the test administration schedule. We recommend somewhere between 25-40% anchor items for test equating purposes, independent of the number of operational papers required and depending on the equating design employed. The proportion of anchor items would be the same for all papers. Different test versions would need to be statistically equated post-administration using an appropriate technique, e.g. linear equating.

9.6 Item Development Methodology

- 9.6.1 It is recommended in the longer term that an alternative approach to the production and review of SJT items is trialled. The current approach has depended upon contributions from a large number of voluntary subject matter experts. Results suggest that the commitment from some volunteers has been lacking; for example, a small number of participants did not submit any draft items after their training workshop, and others did not respond to review comments about their SJT items. This is in contrast with other SJT development projects for specialty and GP selection, where all the participants have shown commitment to drafting and reviewing items.
- 9.6.2 There could be several reasons for the differing levels of commitment. Possibly, for example, those who write SJT items for entrance into speciality training may feel that the test they are developing will directly influence which trainees will be working in their speciality, whereas for entry to Foundation training, the SJT results will mainly be used for allocation (rather than highly competitive selection), so there is less at stake for the developers of the test. However, we suspect that the most likely reason is that the specialty and GP projects have involved smaller numbers of item-writers, who spend a

proportionally greater time on the work (and are paid for their time in the case of GPs). Given this, the project should evaluate the options for changing the dependency on large numbers of volunteer item-writers.

- 9.6.3 It is recommended focus groups and concordance panels continue to be used during item development for quality assurance purposes and to ensure that the question stem is not ambiguous, and that the answer key is fair and realistic.

9.7 Developing the Item Bank

- 9.7.1 Following both the initial and large scale pilots, items with insufficient psychometric properties should be reviewed by an expert in SJT development and clinicians, also experienced in SJT development. Based on the information for the psychometric analysis, the content of items may be altered and/or the scoring key may be reviewed. Review may result in some of the items needing to be re-piloted before they are ready for the item bank. This process is currently being carried out. Findings from this should be able to be fed back in to improve the item writing process, i.e. are there any patterns with regards to poorly performing items?
- 9.7.2 It is recommended that further items are written to ensure that there are sufficient items for operational tests in the future. The number of items that should be written each year/in the bank will depend on a number of factors including how many papers are required each year and the percentage of items that are re-used each year.
- 9.7.3 All further items written must be piloted to ensure they have sufficient psychometric properties. Pilot items can be piloted alongside an operational test, but are not used as part of the overall mark. It is recommended that around five additional items are piloted per paper, meaning that several versions of papers would need to be created. Additional time would be provided to all candidates, relative to the number of items piloted. The candidates would not be informed of which of the items were pilot items.
- 9.7.4 There are a number of lessons learned from the phases of item writing, for example topic areas that do not work as SJT items, and these should be incorporated in any future item writing processes. Details of these lessons are not provided within this report for security reasons.
- 9.7.5 Items will need to be re-used. Reliable SJT items are expensive to develop, so optimising their value is an important consideration. Previously used items also have the advantage of having psychometric properties based on a large sample, however there are risks surrounding exposure. One may wish to use 're-used' items that have been seen to have very good psychometric qualities in an operational test as anchor items.
- 9.7.6 For any test there are no established rules as to how many times an item can be re-used; this needs to be determined by experts and stakeholders. Issues to consider are:
- content validity and relevance of items
 - psychometric quality following use in an operational test

- the number of items needed each year e.g. how many versions, & how many items can be developed each year
- the length of time that the test has been running i.e. In the third year, a greater proportion of items could be re-used as items would be available from two previous years. A large proportion from same testing year should not be included within the same paper in future
- spread of domains and response format i.e. If there are fewer items in the item bank categorised as Patient Focus, these may need to be re-used more frequently or be a priority for future item writing

9.8 Equality and Diversity

- 9.8.1 A diverse range of individuals, with regards to sex, age and ethnicity, were involved in the process. Efforts should continue to be made to ensure that there is a diverse range of individuals involved in the development of the SJT in the future. This eliminates any unintended bias as a factor of those individuals involved in item-writing and quality assurance.
- 9.8.2 Differences in performance between males and females and between ethnic groups were found in the SJT pilot; on average males scored significantly lower than females, and participants reporting themselves as from Black and Minority Ethnic (BME) groups scored significantly lower than the white ethnic group. This is a typical and comparable finding with many selection methods across many occupational groups^{12,13,14}. Where group differences have been found, analysis should be continued to be monitored for future tests.
- 9.8.3 Although during test development item writers were asked to ensure that the language used did not unfairly discriminate against participants who were not trained in the UK, an additional language and diversity check is recommended to be carried out as part of the design methodology to help ensure that the content of the items does not unfairly discriminate against any particular group due to a factor unrelated to the job specification. This is in line with UK employment law.

10. References

1. McDaniel, M. A., Morgeson, F.P., Finnegan, E.B., Campion, M.A., & Braverman, E.P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, 86, 730-740
2. McDaniel, M.A., Hartman, N.S., Whetzel, D.L., & Grubb, W.L. (2007). Situational judgement tests, response instructions and validity: A meta analysis. *Personnel Psychology*, 60, 63-91
3. Ployhart, R.E., & Erhart, M.G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment*, 11, 1-16
4. Nguyen, N.T., Biderman, M.D., & McDaniel, M.A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment*, 13, 250-260
5. Lievens, F., Sackett, P.R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context, *Journal of Applied Psychology*, 94, 1095-1101
6. Weekley, J.A. (2004) Scoring situational judgement tests: does the middle matter? *Kenexa HR newsletter*
7. Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgement tests: a review of recent research. *Personnel Review*, 37, 426-441
8. Legree, P. J., & Grafton, F. C. (1995). *Evidence for an interpersonal knowledge factor: The reliability and factor structure of tests of interpersonal knowledge and general cognitive ability*. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences Technical Report No. 1030
9. Patterson, F., Baron, H., Carr, V., Lane, P., & Plint, S. (2009). Evaluation of three short-listing methodologies for selection into postgraduate training: the case of General Practice in the UK. *Medical Education*, 43, 50-57
10. Anderson, I., Carr, V., Davies, P., Roberts, M., Patterson, F., Carr, A., Crossingham, G., Sice, P., Lam, H., Langton, J., & Gale, T. (2010) Is a Situational Judgment Test Valid and Reliable in Selection to Anaesthesia and Acute Care Common Stem Programmes in England? *Proceedings of AMEE Annual Conference; 2010 Sep; Glasgow, UK (Poster)*
11. Patterson, F., Baron, H., Carr, V. (2007) Best Practice Use of SJTs in Selection: The business case, design issues & applications. *Proceedings of BPS Division of Occupational Psychology Annual Conference; 2007 Jan; Bristol, UK*
12. Nguyen NT, Biderman MD, McDaniel MA (2005). Effects of Response Instructions on Faking a Situational Judgment Test. *International Journal of Selection and Assessment* 12;13(4):250-260.
13. Dean, M.A., Roth, P.L. & Bobko, P. (2008) Ethnic and sex subgroup differences in assessment center ratings: a meta-analysis. *Journal of Applied Psychology*, 93, 685-691.

14. Woolf, K., Potts, H., & McManus, I.C. (YEAR) Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis
15. Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, 14, 223–235.
16. Patterson, F., Denney, M.L., Wakeford, R., & Good, D. (submitted). Making assessment fair and equal? An interdisciplinary perspective of a future research agenda. *British Journal of General Practice*
17. Mullins, M.E., Schmitt, N. (1998) Situational judgment testing: Will the real constructs please present themselves? Paper presented at the *13th Annual Conference of the Society for Industrial and Organizational Psychology*; 1998
18. Rust, J., & Golombok, S. (1999) *Modern Psychometrics: the science of psychological assessment*. London: Routledge; 1999
19. Weekley, J.A., & Jones, C. (1997) Video-based situational testing. *Personnel Psychology*. 1997; 50 (1):25-49
20. Patterson, F., Archer, V., Zibarras, L., Coan, P., Kerrin, M., & Paul, N. (submitted) Using situational judgement tests for selection into medical education and training: A review of the research evidence. *Medical Education*
21. Everitt, B.S. (2002) *The Cambridge Dictionary of Statistics*, 2nd Edition, CUP

11. Appendices

Appendix A: Additional Technical Detail

1 Concordance Analysis

- 1.1 The criteria for selecting an item for use in the pilot was a significant Kendall's W ^{xxviii}. Following best practice, any item that produces a low and non-significant Kendall's W is removed from the test for further review. An inclusion criterion of approx 0.60+ is also used to assist in selecting items. However, there is a 'tolerance' around this figure and will depend on a number of factors including how many people have taken the concordance. Consideration of the raw statistics must be combined with consideration of the concordance keys versus item writer and focus group keys as well as further feedback gained from the concordance panel. In this context, a Kendall's W of 0.60 or above indicates good levels of concordance, although anything above 0.50 can be described as having satisfactory levels of concordance.
- 1.2 38 items produced high levels of concordance (0.80 and above) and 96 items produced good levels of concordance (0.60 and above). 44 items had satisfactory levels of concordance (0.50 and above) and 77 items showed moderate, but still significant levels of concordance (below $p < .050$). 27 items produced a non-significant level of concordance, and these were not considered for inclusion in the pilot papers. The remaining items produced low, but still significant levels of concordance.
- 1.3 43 items of the items that produced Kendall's W of between 0.3 - 0.50 (moderate concordance) were also deemed suitable to go forward to the test paper. When selecting these items, consideration of the raw statistics was combined with the item writer and focus group key. Broadly speaking, when the concordance key agreed with the item writer and focus group key, then it was seen as appropriate for the item to be used in the pilot.

2. Participant Evaluation of Initial Pilot

- 2.1 The table below outlines the themed issues, number of comments and responses from the initial pilot participant evaluation.

Issue	No. of comments	Response
The test is fair, is reasonable and the scenarios are realistic	44	
Too much repetition of content of items	31	A wider item bank will provide a wider range of questions. Consideration for item review.

^{xxviii} **Kendall's W** (also known as **Kendall's coefficient of concordance**) is a non-parametric statistic. If the test statistic W is 1, then all the survey respondents have been unanimous, and each respondent has assigned the same order to the list of concerns. If W is 0, then there is no overall trend of agreement among the respondents, and their responses may be regarded as essentially random. Intermediate values of W indicate a greater or lesser degree of unanimity among the various responses

Difficult to answer questions with no experience, requires knowledge of protocols and role of supervisors etc	31	One of the item writing principles is that it does not require knowledge of protocols and procedures; review items. Ensure glossary is clear
Better than white space questions	28	
Should be used in combination with other methods such as measure of medical knowledge, or white space questions	26	Misunderstanding about sole use
Test instructions, answer sheet & evaluation form	26	Review of documentation
The test was too long/there were too many questions resulting in fatigue and inability to concentrate towards the end	26	Consideration of reliability balanced with fatigue
Concern over the realism of the test, including the test itself, scenarios & responses	25	Review of items, ensure review by job incumbents
Would prefer another method of assessment, such as interviews or exam	24	Reiterate options appraisal
Felt that some of the items were ambiguous or need more information, felt that test was subjective	23	Review of items to try and reduce ambiguity. That there is an element of subjectivity is part of a SJT
Need more time	19	Consideration of timing
Query over scoring key i.e. concern that will score no points if get one response in wrong order	18	Scoring method clarified
Belief that it will be difficult to differentiate between candidates	14	Psychometric item analysis will explore this issue directly
Want to be able to clarify/justify their answers with a white space box	14	This is possible in theory but responses will be difficult to score reliably and marking will be relatively time consuming and expensive. This would also alter the selection method from an SJT to a different type of selection method.
Difficult to distinguish/rank responses	13	Item writers are asked to address this issue
Concern that the SJT doesn't assess medical knowledge or abilities	11	Clarify what this test is aiming to measure and will be used in combination
Concern that candidates would answer what they should do rather than what they would	10	This is what we ask them and the reasoning outlined in research evidence
Administration; both positive and negative, including Sheffield not being voluntary	8	Review of administration process
Handheld devices	8	Review of use of handheld devices
Is able to assess the abilities and attributes required for a FY1	6	

Concern that the SJT would become subject to coaching	5	Complexity built into item design to try and reduce this
Wanted individual feedback	2	
Can't prepare for the test	2	
Lack of test taking motivation	1	
May favour those whose first language is not English	1	

Table 47: Summary of participant evaluation qualitative comments**3 Further test level analysis**

- 3.1 The correlation between the two response formats for the three test (n= 613) was $r=0.39^{**}$. (Paper 1 – $r=.25^{**}$, $p<0.01$), Paper 2 – $r=.42^{**}$, $p<0.01$, Paper 3 – $r=.46^{**}$, $p<0.01$). The GP SJT follows the same structure (roughly two thirds ranking and one third multiple choice). The GP live test data consistently has correlations between the two response formats of $r=0.60^{**}$ ($p<0.01$) and thus supports the understanding that the FY1 SJT is a uni-dimensional test.
- 3.2 The standard error of measurement^{xxix} (SEM) was calculated for each of the three papers and is shown in table 36 above. Using the SEM, it can be stated that there is a 68% chance that a participant's 'true' score will lie between +/- the SEM score.

4 Reliability & Item Partial

- 4.1 The level of reliability achieved or desirable for SJTs has been much debated. PMETB report that the level of internal reliability required for a test is $\alpha=0.90$. However this is in the context of a licensing exam (summative assessment) and an MCQ; these types of assessments are very different to that of SJTs both in purpose and content and we would advise against treating the two types of assessments as if they were the same. For example, whereas MCQs may be fairly 'narrow-band' assessments of job specific knowledge or cognitive ability, SJTs are typically 'broad-band' assessments sampling a wider construct space (e.g. professional attributes of doctors) and thus expectations around the level of internal consistency achievable/desirable would be lower. Based on previous research^{9,11,19} we would recommend that an internal reliability of a live test of $\alpha=.80$ would be required.
- 4.2 With regards to acceptable levels of correlations for item partials, guidelines suggest in general 0.2 or 0.3 as identifying a good item²⁰. In this process we have used heuristics based on these guidelines and based on identifying items with sufficient level of correlation to be contributing to the reliability of the test. It is likely that the level of correlation needed to identify a successful item will be influenced by sample size and test length (e.g. with longer tests item-total correlations are likely to be lower). It is also recommended

^{xxix} The standard error of measurement (SEm) estimates how repeated measures of a person on the same instrument tend to be distributed around his or her "true" score. The true score is always an unknown because no measure can be constructed that provides a perfect reflection of the true score. Since all measurement contains some error, it is highly unlikely that any test will yield the same scores for a given person each time they are retested.

that item partials are balanced with other considerations, e.g. need to provide coverage of all target domains.¹⁸

5 Analysis of anchor items

- 5.1 15 items were used as anchor items and appeared in all 3 papers. In addition, there were five feedback questions that were common between the three papers, resulting in 20 anchor items in total.
- 5.2 Exploratory analysis of the anchor items was carried out to assist with understanding the thinking around test equating. Comparison of means of these items between the three papers indicates that generally the items performed similarly across the papers.
- 5.3 Analysis of the item partials for the anchor items shows that in Papers 2 and 3, the item partials for the anchor items tend to be similar. However, the performance of the anchor items appears somewhat different in Paper 1, suggesting that there may be a difference in how the anchor items related to the other items in the test in Paper 1 (see appendix C).
- 5.4 Correlation analysis between anchor items (n=20) and other items within the test (n=40) show varied results for the three tests, in particular for Paper One. Paper One shows lower correlations between anchor items and other items within the test, compared to Paper Two and Paper Three. As the anchor items seem to be performing the same in the three papers with regards to the mean and SD, this suggests there are differences in how the anchor items in Paper 1 relate to the rest of the items in the paper (as also shown by the item partials). This could be due to differences in the content of the items (or underlying constructs being assessed), issues with administration at particular schools, or perhaps differences in the nature of the candidate sample with regards to quartile and application form scores.

		Anchor and feedback (20 items)	Anchor only (15 items)
Paper One	Mean	291.37	217.9
	SD	13.7	10.6
	Correlation between anchor & other items (rho)	.283**	.266**
Paper Two	Mean	289.4	218.0
	SD	14.3	11.6
	Correlation between anchor & other items (rho)	.418**	.396**
Paper Three	Mean	293.2	220.2
	SD	14.5	12.0
	Correlation between anchor & other items (Spearman)	.598**	.407**

Table 48: Anchor item statistics

6. Construct Analysis

- 6.1 Table 43 below shows the descriptive statistics for the NEO-PIR domain scores. The maximum and minimum possible score available differs for the five factors but ranges from 120 to 480. A positive score on all scales indicates the presence of the described trait, with the exception of emotional stability, where a high score indicates the absence of emotional stability.
- 6.2 All five factors are normally distributed, although Emotional Stability and Agreeableness are slightly negatively skewed compared to the other three domains; that is responses are skewed slightly towards the sample having relatively low levels of emotional stability and slightly higher levels of agreeableness. The standard deviations for both these domains are also larger, indicating that there is a greater spread of scores around the mean. Conscientiousness is slightly positively skewed. All five domains show reasonable levels of reliability.

Factor	Mean Score	Max	Min	SD	Skewness	Reliability
Emotional Stability	352.2	441	162	50.3	-.92	.83
Extraversion	285.1	336	201	37.3	-.01	.78
Openness to Experience	294.0	421	171	36.0	-.11	.70
Agreeableness	311.8	425	153	49.2	-.62	.80
Conscientiousness	285.0	455	171	44.2	.38	.81

Table 49: NEO domain mean scores

- 6.3 The table below provides the results by sex. Females had a higher mean score on Emotional Stability (363.1) (indicating lower levels of emotional stability), Openness to Experience (296.9) and Agreeableness (327.8). Males had a higher mean score on Extraversion (287.0) and Conscientiousness (286.8) but these differences were not great.

Factor	Sex	Mean score	SD
Emotional Stability	Male	335.6	54.6
	Female	363.1	44.3
Extraversion	Male	287.0	35.3
	Female	283.9	38.8
Openness to Experience	Male	289.7	41.7
	Female	296.9	31.8
Agreeableness	Male	287.6	53.2
	Female	327.8	39.2
Conscientiousness	Male	286.8	42.5

	Female	283.8	45.6

Table 50: NEO domain mean scores by sex

- 6.4 The table below provides the result by ethnicity. Black and Minority Ethnic (BME) candidates scored higher on Emotional Stability, indicating lower levels of Emotional Stability. BME candidates scored lower in the four other domains.

Factor	Sex	Mean score	SD
Emotional Stability	White	351.2	47.5
	BME	357.1	65.8
Extraversion	White	287.7	35.0
	BME	275.0	46.4
Openness to Experience	White	295.7	36.9
	BME	285.5	32.0
Agreeableness	White	311.8	51.5
	BME	309.4	37.0
Conscientiousness	White	287.8	43.3
	BME	274.3	46.4

Table 51: NEO domain mean scores by ethnicity

7. Correlations with Quartiles and Application Form

- 7.1 Correlations were run between SJT total scores (n=631) and current FP selection methods. Current selection methods include quartiles and a competency based application form. Quartiles are calculated using examination scores from preceding years at medical school. Students are awarded one of four quartile scores (34, 36, 38, 40). The application form consists of 6 questions. Question 1 covers career history and qualifications. Questions 2 to 6 are white space competency based questions, based on the Foundation doctor person specification.
- 7.2 Significant correlations were found between SJT scores and quartile scores (Spearman rho: $r=.20^{**}$, $p<0.01$) and SJT scores and competency scores on the application form (Pearson: $r=.12^{**}$, $p<0.01$). Although these correlations are significant, indicating some shared variance/commonality between the assessment methods, there is also a large amount of variance not explained, therefore the SJT appears to be assessing somewhat different constructs to the other methods.
- 7.3 For Paper One, there were no significant correlations found between SJT, quartile scores and application form. For Paper Two, significant correlations were found between SJT scores and quartile scores (Spearman rho: $r=.23^{**}$, $p<0.01$) and SJT scores and competency scores on the application form (Pearson: $r=.22^{**}$, $p<0.01$). For Paper Three, a significant correlation was found between SJT scores and quartile scores (rho: $r=.34^{**}$, $p<0.01$) but

not for application forms. This is likely to reflect the small sample sizes for Paper One and Paper Three.

8. Participant evaluation of large scale pilot

8.1 The table below outlines the themed issues, number of comments and example comments from the large scale pilot candidate evaluation.

Theme	No. of Comments	Example Comments
The SJT is good/fair/reasonable	121	"Test is excellent" "This is a really good way to examine people" "Yes, this is a reasonable part of a selection process" "This is a fair way of assessing adequacy of applicants"
Computer and interface related issues	50	"END button produces an error 'you have not completed all questions'" "The screen layout made the test very difficult to use" "The drag and drop screens were not user friendly"
Use of SJT combined with other methods	47	"Might be ok as part of a more comprehensive assessment but not in isolation"
Better than current system	43	"This is a reasonable part of a selection process and I feel superior to the previous white space questions"
Ambiguity/subjectivity	39	"This is a fair test, although some of the figures are ambiguous"
Grammatical/spelling issues	34	"... some of the questions had spelling mistakes, didn't quite make sense"
Relevance and realism of SJT items	33	"I feel that some of the questions were relevant to the foundation programme" "Realistic scenarios throughout paper"
Suggestions to enhance SJT	33	"Possibly [include] some more emergency/management questions to differentiate between candidates"
Respond differently in clinical setting compared to SJT	32	"Answers given in theory might not actually reflect actual practice"
Responses equally important	29	"Difficult, as some of the questions, I would do all, or a lot of the available options"
Similar scenarios/options	27	"... there seemed to be [many] questions with some similar content and difficult situations"
Lack of experience with scenario content	24	"I was not really able to answer some of them properly as I have not yet encountered any even vaguely similar situations as a medical student"
More information required	23	"I found that the questions were difficult to answer as I felt I needed more information to help me make a decision"
Preparing for the SJT	23	"It would be hard to prepare for a test like this"
Being coached for the SJT	20	"I can see solicit and commercial [organisations training] students to 'take the test' other than to understand underlying principles of appropriate potential behaviour"
Test time too long	20	"Don't need two hours"
SJT did not assess	17	"The test does not indicate clinical knowledge,

medical knowledge or ability		extracurricular activities, or other personal qualities"
Satisfactory administration	17	"The test was run in a satisfactory and appropriate manner"
Blank space to justify choices	15	"I felt I constantly want to explain my choices but have no opportunity"
Unsure of whether the SJT would differentiate between candidates	15	"I am not sure how that would separate the good doctors from the bad doctors"
Test time too short	15	"I felt pushed for time to answer all 60 questions perhaps the test should be longer in time length"
Too many SJT items	13	"There were quite a lot of questions and some of the scenarios were long"
Instructions unclear	13	"I was so confused as to whether we should rank them in accordance to the order we should carry out the task or whether we should at all carry them out"
Too much content	12	"The test was far TOO LONG to read and as a result, I lost time"
Better at differentiating between candidates	12	"This seems to be a better discriminator of good candidates compared to the white space questions because it assess the candidates skills in judging real life situations"
Assessed professional attributes	10	"Very good way to assess candidates ability to deal with practical solutions"
Should not have equal weighting with other methods	10	"I don't think this should have as much weighting as the white space questions"
Did not understand how the SJT was marked	9	"Slightly unclear as to how the test is marked - score clarification would be useful"
Response format	6	"Important to make clear answer as 'should' NOT 'would'"
Poor examination conditions	5	"There were more invigilators than any other exam I've ever sat, they were constantly walking around"
SJT may work better on computer	3	"I think the test would be improved by having a multiple choice computer exam as it is easy to make mistakes manually"

Table 52: Qualitative comments from candidate evaluation

Appendix B: Example SJT Items used as Feedback Questions

- 1.** You are looking after Mr Kucera who has previously been treated for prostate carcinoma. Preliminary investigations are strongly suggestive of a recurrence. The ward is busy. As you finish taking blood from a neighbouring patient, Mr Kucera leans across and says “tell me honestly, is my cancer back?”

Rank in order the following actions in response to this situation (1= Most appropriate; 5= Least appropriate).

- A Explain to Mr Kucera that it is likely that his cancer has come back
- B Reassure Mr Kucera that he will be fine
- C Explain to Mr Kucera that you do not have all the test results yet.
- D Inform Mr Kucera that you will chase up the results of his tests and ask one of your senior colleagues to discuss them with him
- E Invite Mr Kucera to join you and a senior nurse in a quiet room, get a colleague to hold your ‘bleep’ then explore his fears

- 2.** You review a patient on the surgical ward who has had an appendicectomy done earlier on the day. You write a prescription for strong painkillers. The staff nurse challenges your decision and refuses to give the medication to the patient.

Choose the **THREE most appropriate** actions to take in this situation

- A Instruct the nurse to give the medication to the patient
- B Discuss with the nurse why she disagrees with the prescription
- C Ask a senior colleague for advice
- D Complete a clinical incident form
- E Cancel the prescription on the nurse’s advice
- F Arrange to speak to the nurse later to discuss your working relationship
- G Write in the medical notes that the nurse has declined to give the medication
- H Review the case again

Appendix C: Item Level Results for Initial Pilot

Item Level Results for Rational Scoring

Item	Type	Mean	SD	Partial	Quality
1	Ranking	16.9	1.8	-.004	Further review
2	Ranking	17.1	1.5	.174	Satisfactory
3	Ranking	14.1	2.8	.102	Further review
4	Ranking	18.3	1.7	.204	Satisfactory
5	Ranking	15.5	2.3	.079	Further review
6	Ranking	15.1	2.3	.177	Satisfactory
7	Ranking	16.9	2.9	.218	Satisfactory
8	Ranking	18.1	1.9	.157	Moderate
9	Ranking	15.7	2.4	.077	Further review
10	Ranking	16.6	2.4	.086	Further review
11	Ranking	15.9	1.8	.024	Further review
12	Ranking	15.5	2.2	.157	Moderate
13	Ranking	16.5	2.2	.031	Further review
14	Ranking	15.2	2.4	.096	Further review
15	Ranking	15.6	2.5	.017	Further review
16	Ranking	15.2	1.6	.169	Satisfactory
17	Ranking	17.3	1.9	-.009	Further review
18	Ranking	16.2	2.3	.192	Satisfactory
19	Ranking	17.5	1.8	.141	Moderate
20	Ranking	17.4	2.2	.089	Further review
21	Ranking	15.3	2.1	.136	Moderate
22	Ranking	17.4	1.7	.120	Further review
23	Ranking	15.7	2.1	.228	Satisfactory
24	Ranking	15.7	3.3	.154	Moderate
25	Ranking	17.0	2.2	.038	Further review
26	Ranking	16.2	2.6	.196	Satisfactory
27	Ranking	15.0	2.4	.060	Further review
28	Ranking	15.4	2.2	.146	Moderate
29	Ranking	15.9	1.8	-.021	Further review
30	Ranking	16.6	2.5	.186	Satisfactory
31	Ranking	15.4	2.4	.199	Satisfactory
32	Ranking	18.8	1.5	-.023	Further review
33	Ranking	15.0	2.5	.124	Further review
34	Ranking	16.3	1.9	.100	Further review
35	Ranking	16.8	2.4	.095	Further review
36	Ranking	15.0	2.5	.018	Further review
37	Ranking	15.1	2.8	-.035	Further review
38	Ranking	14.1	2.9	.222	Satisfactory
39	Ranking	18.1	1.7	.113	Further review
40	Ranking	15.7	2.8	.281	Good
41	Ranking	16.3	2.4	.286	Good
42	Ranking	16.8	2.7	.248	Good
43	Multiple Choice	11.5	1.5	.142	Moderate
44	Multiple Choice	9.1	2.3	.123	Further review
45	Multiple Choice	7.0	2.3	.131	Moderate
46	Multiple Choice	9.8	2.5	.125	Moderate
47	Multiple Choice	10.8	2.1	.120	Further review
48	Multiple Choice	10.2	2.3	.177	Satisfactory

49	Multiple Choice	7.2	1.9	.124	Further review
50	Multiple Choice	8.3	2.5	.218	Satisfactory
51	Multiple Choice	8.5	2.7	.236	Satisfactory
52	Multiple Choice	8.4	2.5	.231	Satisfactory
53	Multiple Choice	7.7	2.9	.247	Good
54	Multiple Choice	7.8	2.5	.312	Good
55	Multiple Choice	6.1	2.9	.248	Good
56	Multiple Choice	7.5	2.1	.339	Good
57	Multiple Choice	8.4	2.7	.429	Good
58	Multiple Choice	6.6	2.8	.276	Good
59	Multiple Choice	7.7	3.1	.307	Good
60	Multiple Choice	8.2	3.1	.396	Good
61	Multiple Choice	7.9	2.9	.386	Good
62	Multiple Choice	7.4	2.6	.396	Good
63	Multiple Choice	6.3	2.8	.284	Good
64	Multiple Choice	9.1	3.3	.351	Good
65	Multiple Choice	8.0	3.2	.321	Good

Item Level Results for Empirical Scoring

Item	Type	Mean	SD	Partial	Quality
1	Ranking	16.9	1.8	.038	Further review
2	Ranking	18.7	1.5	.215	Satisfactory
3	Ranking	15.4	2.6	.105	Further review
4	Ranking	18.3	1.7	.232	Satisfactory
5	Ranking	17.2	2.3	.218	Satisfactory
6	Ranking	16.0	2.3	.160	Moderate
7	Ranking	16.9	2.9	.203	Satisfactory
8	Ranking	18.1	1.9	.161	Moderate
9	Ranking	16.5	2.2	.135	Moderate
10	Ranking	17.1	2.2	.085	Further review
11	Ranking	17.7	1.9	.065	Further review
12	Ranking	15.5	2.2	.134	Moderate
13	Ranking	17.2	2.1	.064	Further review
14	Ranking	16.3	2.4	.135	Moderate
15	Ranking	14.2	2.8	.070	Further review
16	Ranking	17.5	2.0	.186	Satisfactory
17	Ranking	17.8	2.2	.095	Further review
18	Ranking	16.2	2.3	.185	Satisfactory
19	Ranking	17.5	1.7	.150	Moderate
20	Ranking	17.4	2.2	.119	Further review
21	Ranking	15.9	2.7	.203	Satisfactory
22	Ranking	18.2	1.9	.176	Satisfactory
23	Ranking	16.5	2.1	.175	Satisfactory
24	Ranking	15.7	3.3	.149	Moderate
25	Ranking	16.8	1.9	.191	Satisfactory
26	Ranking	16.2	2.6	.205	Satisfactory
27	Ranking	15.8	2.3	.120	Further review
28	Ranking	16.9	1.9	.107	Further review
29	Ranking	17.4	1.9	.040	Further review
30	Ranking	16.6	2.5	.172	Satisfactory
31	Ranking	15.4	2.4	.207	Satisfactory

32	Ranking	18.8	1.4	.010	Further review
33	Ranking	15.0	2.5	.152	Moderate
34	Ranking	17.5	1.9	.143	Moderate
35	Ranking	16.8	2.4	.087	Further review
36	Ranking	15.3	2.6	.114	Further review
37	Ranking	15.4	2.2	.085	Further review
38	Ranking	15.0	2.9	.125	Further review
39	Ranking	18.7	1.9	.142	Moderate
40	Ranking	15.7	2.8	.231	Satisfactory
41	Ranking	16.3	2.4	.270	Good
42	Ranking	17.8	2.6	.214	Satisfactory
43	Multiple Choice	11.5	1.5	.176	Satisfactory
44	Multiple Choice	9.0	2.3	.115	Further review
45	Multiple Choice	9.4	2.5	.183	Satisfactory
46	Multiple Choice	9.8	2.5	.144	Moderate
47	Multiple Choice	10.8	2.1	.136	Moderate
48	Multiple Choice	7.7	1.2	.083	Further review
49	Multiple Choice	9.7	2.5	.203	Satisfactory
50	Multiple Choice	8.3	2.6	.254	Good
51	Multiple Choice	8.5	2.7	.236	Satisfactory
52	Multiple Choice	8.4	2.5	.229	Satisfactory
53	Multiple Choice	7.7	2.9	.263	Good
54	Multiple Choice	7.8	2.5	.336	Good
55	Multiple Choice	6.1	2.9	.233	Satisfactory
56	Multiple Choice	9.1	2.6	.330	Good
57	Multiple Choice	8.4	2.7	.417	Good
58	Multiple Choice	7.6	2.8	.245	Good
59	Multiple Choice	7.7	3.1	.275	Good
60	Multiple Choice	8.2	3.1	.367	Good
61	Multiple Choice	7.9	2.9	.381	Good
62	Multiple Choice	8.9	3.2	.341	Good
63	Multiple Choice	7.2	3.3	.326	Good
64	Multiple Choice	9.1	3.3	.371	Good
65	Multiple Choice	8.0	3.2	.333	Good

Appendix D: Item Level Results for Large Scale Pilot

	Paper One		Paper Two		Paper Three	
	Mean	SD	Mean	SD	Mean	SD
item1	13.9	3.0	13.5	3.0	14.2	3.1
item2	15.8	2.1	16.2	1.9	16.0	2.1
item3	17.2	2.3	17.5	1.9	17.5	1.9
item4	16.2	2.4	16.6	2.4	17.0	2.1
item5	15.2	2.6	15.8	2.9	15.7	2.8
item6	17.4	1.9	17.3	2.2	17.4	1.9
item7	15.7	2.4	15.4	2.6	15.3	2.7
item8	18.2	2.0	17.8	2.1	18.1	2.5
item9	16.1	2.3	16.4	2.5	16.4	2.6
item10	17.1	1.8	16.4	1.9	16.7	2.1

item11	17.4	1.7	17.1	1.7	17.1	1.7
item12	18.1	1.7	17.7	1.9	17.5	2.7
item13	18.6	1.6	17.9	1.8	16.0	2.8
item14	16.4	2.8	17.3	1.8	16.8	2.7
item15	15.4	2.0	16.6	2.2	17.6	2.5
item16	17.2	2.0	15.6	2.3	15.8	2.2
item17	17.9	1.5	16.8	2.3	17.2	2.3
item18	16.0	2.2	16.5	2.0	17.6	2.0
item19	15.9	2.2	18.3	2.0	17.2	1.8
Item20	15.3	2.6	16.6	2.0	16.8	2.0
Item21	16.6	2.3	17.5	2.2	18.4	2.2
Item22	16.4	2.2	17.7	2.1	15.4	2.3
Item23	15.5	2.4	17.2	1.8	14.7	3.1
Item24	15.8	1.9	15.8	2.2	16.2	2.3
Item25	16.7	1.7	17.1	2.0	16.2	2.9
Item26	17.2	1.9	15.2	2.6	15.4	2.1
Item27	16.5	2.4	16.7	1.7	18.3	2.3
Item28	15.3	2.5	16.3	1.9	16.3	2.5
Item29	17.5	1.8	15.7	2.2	16.6	2.0
Item30	15.5	2.8	15.5	2.8	16.6	1.9
Item31	17.1	2.4	17.7	2.0	16.8	2.7
Item32	17.5	1.8	16.6	2.3	17.3	2.0
Item33	15.8	2.5	15.7	2.5	16.0	1.9
Item34	17.7	2.3	17.5	1.8	17.4	2.5
Item35	16.5	2.3	17.0	2.1	15.8	2.8
Item36	16.5	2.4	16.4	2.1	15.7	2.7
Item37	16.0	3.2	17.5	2.3	17.2	2.8
item38	16.9	3.2	16.8	2.6	16.9	2.9
item39	17.3	2.4	17.8	1.5	17.6	1.9
item40	16.5	2.7	16.6	1.9	16.3	2.2
item41	11.8	0.8	10.7	2.1	11.7	1.2
item42	10.9	1.8	9.6	2.8	10.6	2.4
item43	10.2	2.2	10.5	2.1	10.3	2.1
item44	9.9	2.4	9.7	2.3	10.3	2.2
item45	7.4	2.8	8.3	2.5	8.5	2.7
item46	10.2	2.2	9.5	2.3	9.8	2.2
item47	8.2	2.3	7.9	2.4	7.9	2.4
item48	8.9	2.2	10.2	2.3	10.2	2.3
item49	8.5	2.8	8.7	2.5	8.7	2.5
Item50	8.8	2.4	9.3	2.4	9.3	2.4
Item51	8.3	2.6	9.8	2.5	9.8	2.5
Item52	9.8	2.5	9.2	2.3	9.2	2.3
Item53	9.0	2.6	8.4	2.3	8.4	2.3
Item54	8.4	2.7	9.0	2.4	9.0	2.4

Item55	8.4	2.5	10.6	2.1	10.6	2.1
Item56	9.7	2.5	9.0	2.4	9.0	2.4
Item57	7.5	2.3	8.2	2.8	8.2	2.8
Item58	7.4	3.1	8.3	2.7	8.3	2.7
Item59	7.8	2.4	9.3	2.6	9.3	2.6
Item60	9.8	2.4	8.1	2.7	8.1	2.7

Comparison of performance on five items used for participant feedback

	Initial pilot Mean score	Main pilot Mean score
Item 38	17.0	16.8
Item 39	17.4	17.5
Item 40	17.0	16.4
Item 41	11.5	11.1
Item 42	10.2	10.1

Appendix E: Peer Review: Professor Chris McManus

Final Report on Situational Judgement Test Patterson et al, July 2011

*Chris McManus
University College London
July 2011*

Summary

The SJT has been carefully piloted and well documented. There seems little doubt about the overall acceptability of the test as produced, and it seems likely that it is of reasonable reliability, and probable validity. Both of these can of course be properly evaluated only after implementation.

I have been fortunate to have seen an earlier version of the report, dealing with the initial pilot and those analyses provided an opportunity to recommend additional analyses, etc., which have been carried out. In particular it was important to see the DIF (differential item functioning) analyses of sex and ethnicity differences, and these suggest that whatever the reason for the differences in performance, they are probably not due to bias in the test itself.

Taken overall there would seem to my mind to be little doubt that piloting has probably gone as far as it can reasonably go, and that the next stage has to be a live implementation, in order to see how the SJT works in practice. There may be some risks in that, but they do not seem to me to be unacceptably large, as long as safeguards are in place, with careful monitoring of the test results.

My detailed report raises some technical questions and some minor issues, which will probably be of interest to the authors and to psychometricians more generally. If there is the opportunity to revise the Final Report then comment on these matters may be helpful to readers. Overall though I do not think that any are sufficiently serious to prevent implementation, and most will inevitably be clarified and will resolve themselves over a couple of years of live usage.

Appendix F: Peer Review: Professor Val Wass

Quality Assurance Report:

Design and Evaluation of a Situational Judgement test for selection to the Foundation Programme:

The Work Psychology Group University of Cambridge

Val Wass

BSc (Hons), FRCGP, FRCP, MHPE, FHEA, PhD

Professor of Medical Education

Head of Keele School of Medicine

Executive summary

The Medical School Council was commissioned by the Department of Health to pilot a new model for selection to Foundation Programme (FP) for the four UK Health Departments. Following its success in postgraduate selection, a Situation Judgement Test (SJT) was chosen. The Work Psychology Group and the University of Cambridge, who have prior experience of SJTs, have carried out a well designed and implemented project involving an initial (IP) and large (LP) pilot in UK medical schools. Their very comprehensive high standard report presents full evaluation of the pilots.

Seventeen medical schools participated in the two pilots with a total of 1094 candidates (455 IP, 639 LP). Recruitment to the large pilot was disappointing but sufficient for evaluation purposes.

Quality assurance has been carried out based on the Utility Equation developed for assessment tools. The (i) Validity (ii) Reliability (iii) Educational impact (iv) Acceptability

(v) Feasibility of the SJT are evaluated from the data provided.

Validity: SJTs were designed to measure five domains based on job analysis: commitment to professionalism, coping with pressure, effective communication, team working and patient focus. Candidate performance (tested on two small separate cohorts) correlated positively with high levels of extraversion, openness to values and achievement using a NEO-PIR personality measure and with OSCE performance compared to written tests. It is safe to conclude that the SJT broadly measures attributes relevant to the skills needed for work in the Foundation programme. Full piloting across an entire cohort is needed to evaluate how candidate performance on the SJT ranks against the Educational Performance Measure. Personality measures of high achievers may well reflect high achievement generally.

Reliability: The reliability of a two hour test using 60 items is just about acceptable provided all individual items perform well psychometrically. There is a wide range of scores on both scoring methods which the authors claim is sufficient for discrimination. The mean scores tended to be high around 80%. There are a number of tied scores which will need addressing. It remains difficult to make a full judgement until data is available across a complete school intake and range of ability. Low uptake of pilots may have biased towards the more motivated. There is no scope for reducing the number of items. Extension of the test to pilot items and to allow for poorly performing ones might prove necessary.

Acceptability: Over 95% of participants evaluated the SJT. There is good consensus that it is relevant, fair and better than the current white paper questions. Differences in performance relating to sex and ethnicity will need further attention. Scenarios must be short and written in clear English to limit potential disadvantage.

Feasibility: Item writing and weighting the responses was time consuming and resource intensive. Questions required extensive review and modification and scoring is complicated.

Different levels of commitment from the 89 item writers suggest that alternative ways to resource efficient question writing methods need to be explored. Different performance across the three papers and a heavier weighting toward one domain in the LP suggest careful blueprinting is essential. Exploration of computer versus pen delivery was inconclusive although the former appears feasible if well planned.

Educational impact: Work using the Social Desirability scale suggests that candidates “faking” their responses to reflect their perception of desirable rather than actual response was not a serious issue. In theory candidates cannot prepare for the test. In actuality there is wide spread concern they can be tutored for the test which would be educationally disadvantageous.

Conclusion. This project, which is well designed and thoroughly implemented, supports the proposed replacement of the current white-paper UKFPO selection with a SJT across the UK. Situational judgement testing is certainly more reliable and arguably more valid. Further piloting of a full year cohort is desirable but not essential. More work on item development is needed. It is very resource intensive. This will need to be reviewed and addressed. Methods to pilot new questions need identifying. The proposed multiple delivery of the paper needs review given the restrictions of item development and need for careful blueprinting of papers to ensure consistency of test performance. Delivery by computer seems feasible. A further pilot evaluating whether students can be trained to answer the test might be considered. On balance this is a very positive evaluation.

1: Background

1:1 The Medical School Council (MSC) has commissioned a cross stakeholder Steering Group to develop and pilot new arrangements for selection into the Foundation programme (FP). Two methods have been chosen: (i) a Situational Judgement Test (SJT) and (ii) an Educational Performance Measure (EBM) based on academic performance rankings across the medical course at the individual schools.

1:2 Situational judgement tests have been in use for 20 years and are gaining increasing acceptability for large scale selection processes. They are based on extensive analysis of the job role from which the domains required to test the candidate's cognitive attributes and ability to make judgements in the workplace are determined. They do not assess knowledge and practical skills.

1:3 In more recent years SJTs have been developed for selection into postgraduate medical training programmes e.g. General Practice where evidence is emerging of its efficacy as a selection method. However its main use to date has been for short listing followed by further assessment. This contrasts with the FP which is the main route into the allocation of jobs.

1:4 The report includes a literature review providing evidence to support the development of, and rationale behind, the test methodology. In addition micro pilots were carried out across four UK Universities to explore the most appropriate test specification format for undergraduates. Based on evidence thus accrued a SJT design specification was developed.

1:5 A two hour test of 60 items over two hours covering five domains of job analysis: (i) commitment to professionalism, (ii) coping with pressure, (iii) effective communication, (iv) working effectively as part of a team and (v) patient focus was proposed supported by evidence well laid out in the report. Decisions to use a knowledge based response "*what should you do?*" rather than behavioural "*what would you do?*" and to use two test formats of ranking (66%) and multiple choice (33%) were also robust in light of the literature and micro pilot findings.

1:6 Evaluation of the SJT is based on an initial pilot (IP) and a second larger pilot which included 13 medical schools and 1094 participants. The Utility Equation was used as a framework for quality assuring the assessment methodology and the psychometric data published in the report.

Validity x reliability x acceptability x feasibility and cost x educational impact

2: Development of SJT items.

2:1 Seven item writing workshops were held with 89 writers over a two year period. These were volunteer clinicians drawn from more than 23 primary specialities and prepared to commit a day to the workshop. Experience of working with Foundation year 1 (FY1) doctors was essential. The broad selection of writers seemed to work well.

2:2 Four hundred and fifty three items were written by 82 item writers averaging 5.5 items per writer; 7 failed to write any items. The authors highlight that the level of commitment and output was low relative to their experience with other writers. GP selection, for example, uses a smaller group of seemingly more committed and more productive writers who are reimbursed for their time out of practice. The mode of item writing may need review (see 6:1).

2:3 Review of items by the authors themselves was associated initially with a low return. 50% of authors did not revise their questions and 31% of the initial items proved unsuitable. Later in the project both author reviewing and standard of questions improved. The improvement in commitment and attrition over time argues for a small expert writing group.

2:4 All items were then reviewed by a lead team reviewer. At this stage 54 of 360 (15%) questions were rejected. The attrition rate was high at 32.5%; 306 of 453 (67.5%) of questions were retained. A further 20 items were rejected at a later stage as the result of focus groups with Foundation doctors.

2:5 Four concordance panels were then held using subject matter experts familiar with the work of Foundation doctors. Sixty individuals participated (range of specialities not given). They completed one of 66 – 80 item papers under test conditions and gave feedback on the questions. A concordance analysis of responses identified questions where concordance between experts was satisfactory. Twenty-seven further items were found to have unacceptable concordance. A further 21 questions were removed. Two hundred items were selected for the pilots.

2:6 Overall 40% of initial items were discarded. There was in addition divergence from the original blueprint of equal distribution across the domains. One concludes these are difficult questions to write and very resource intensive (see 8:1).

3 Blueprinting the paper:

3:1 Blueprinting is essential to any test construction to ensure content is relevant to the domains being assessed and standardised. It is of particular importance if multiple papers are envisaged for one year of Foundation entry. Pilots were planned against a format of a fairly equal (but not exact) split across domains. Scenarios were assigned to three broad topics “colleague”, “patient” and “personal”. Items were randomly distributed across the paper and not grouped by domain or topic. Sensitive potentially emotional scenarios at the start of the test were avoided. No data is given on the generalisability of SJTs. One assumes that content specificity is essential and a wide range of contexts is important. The resultant statistics on performance suggest this is the case.

3:2 Each paper had 40 ranking and 20 multiple choice (MCQ) questions. Anchor items i.e. questions used to standardise and compare cohort performance between papers, were

used; 15 per paper. Item difficulty and discrimination could only be controlled once questions had been piloted.

3:2 It is notable that at the workshops questions were not written uniformly across the domains. Over a third (38%) were written in “commitment to professionalism”. The IP papers achieved an equal split. In contrast a third (instead of a fifth) of questions in the LP papers were in the domain preferred by the item writers i.e. “commitment to professionalism”. This skewed the balance of papers used in the LP where the intended equal weighting was lost. This reflected question availability rather than poor blueprinting.

3:3 As in the LP there was a significant difference in the performance of the paper 1 compared to 2 and 3 (see 6:4), careful blueprinting both to inform item writers and to construct the test is essential.

3:4 Two forms of scoring were compared; “rational” based on expert judgement and “empirical” based on item performance on a large pilot sample. The IP compared both types of scoring. Correlation statistics show high levels of agreement between the scores and supported the decision to work with rational keying of scores in the LP.

3:5 The logic for using a 2:1 weighting for ranking and MCQ items is not immediately apparent from the report. The scoring convention for both these formats remains complex. It may not be ideally suited to the proposed use of SJTs in the FP. Other specialties have trialled different scoring conventions. Scoring of necessity needs to remain in the hands of the experts and be open to continuing review as more data on a wider candidate performance becomes available.

4: The Pilots:

4:1 Four schools contributed to IP and 13 to LP. Piloting throughout was confounded by poor uptake in some volunteer schools compared to others. The report does not address the different size of student cohorts across schools where uptake varied considerably. A percentage uptake of the total school year would be helpful e.g. 106 at a small school such as Keele approaches 95%. It is difficult to assess uptake at the other schools in the absence of cohort size. Assessment of response rate and the degree of potential bias is difficult. Can we ask if this data is available?

4:2 Uptake for the LP was disappointing; 639 participants (58% of the total candidates). Almost certainly the pilots will have been biased towards motivated keen students. This will undoubtedly have impacted on test psychometrics e.g. for item difficulty. The high mean scores seen in item performance may reflect this.

4:3 Overall in both pilots the spread of sex and ethnicity allowed comparisons to be made with the exception of non-UK participants of whom 41 sat the IP. This was insufficient to make firm conclusions. In both pilots differences in performance were seen across sex (females performing better than males) and ethnicity where poorer performance on average was seen from ethnic minority students. These findings are consistent with statistics across UK undergraduate and postgraduate medical examinations. We know little

of the impact of cultural diversity on scenarios. It will be crucial to monitor this. Ensuring a good ethnic mix of item writers and short clear scenarios (in case of English as a second language) is essential.

4:4 Significant differences in both uptake of the three papers in LP and the psychometrics were seen. Paper 1 performed less well with a significantly lower standard deviation across performance. This group of candidates had significantly higher application scores which may have been relevant. The mean score and candidate distribution was the same as the other two papers. Blueprinting may also have been relevant (see 6:4).

5: Test validity

5:1 Validity is a conceptual term which, unlike reliability, cannot be expressed as a simple coefficient. It can only be assessed after the test has been delivered. Validity is a measure of whether the test has actually measured what it set out to measure. This is of particular importance for this relatively new test format. The authors have built in a series of methods to do so and are to be congratulated on this. Performance on the test has been compared using different samples across the pilots with (i) the NEO-PIR personality measure (ii) OSCE and written tests (iii) application scores.

5:2 The NEO-PIR personality measure assesses five major dimensions of personality: emotional stability, extraversion, openness to experience, agreeableness and conscientiousness. One hundred and eight participants at one medical school completed the NEO-PIR. SJT performance correlated positively with high levels of extraversion, openness to values and achievement. This suggests that high achieving individuals do better. The findings are consistent with expectations and Foundation job analysis requirements. This supports the validity of the test.

5:3 Data on 96 candidates at one medical school were correlated with their year 4 OSCE and written test results. SJT scores correlated with performance in the clinical skills Objective Structured Clinical Examination (OSCE) but showed limited correlation with the written examinations. Further evidence that the attributes of a good practising doctor are being measured.

5:4 For this cohort, application scores and the academic performance quartiles of students were combined with the examination results for multi-regression analysis. The OSCE correlated most highly with communication skills. This is further evidence that the SJT targets the non cognitive attributes necessary for Foundation placements.

5:5 Further statistical analysis showed that a combination of SJT and the quartile scores (the two scores it is intended to use in Foundation allocation ranking) predicted OSCE performance more reliably than the student application and quartile scores. This is very encouraging evidence for the validity of the SJT and confirmation that it performs better than the current system.

5:6 The various correlations continue to raise questions as to what the SJT as a uni-dimensional test actually measures. Further analysis along these lines, particularly as the

high achievers tend to do best on the SJTs, will be important. It will be interesting to see how SJTs score correlate with the EPM in full student cohorts.

5:7 Twenty focus groups with 63 Foundation doctors were held in six locations. The participants reviewed test instructions and each was shown 20 items only. Their comments confirmed the SJT had face validity. It was felt appropriate to their work and fair. There were some reservations that it only tested certain characteristics omitting areas such as leadership and extracurricular activities.

5:8 Behavioural based tests such as SJTs compared to knowledge based tests are more prone to the “faking” of answers. They cannot tap into candidates’ actual behaviour and assess how they decide to present themselves. A knowledge based response instruction (*‘what should you do’*) was used for the SJT, rather than a behavioural based response instruction (*‘what would you do’*), to reduce this concern. Never the less there is general concern that candidates will train for the test and learn the desired responses. All IP candidates completed a measure of Social Desirability which measures “the tendency of respondents to reply in a manner that will be reviewed favourably by others”. Participant scores suggest they were not “faking good.” There was no correlation between SJT scores and Social desirability scores. Although the authors suggest the SJT is less susceptible to faking and should be resilient to coaching, this remains an area of concern and worthy of further piloting.

6: Item analysis: Reliability

6:1 Reliability is a measure of the consistency and reproducibility of the test result i.e. if the students took the same test on another day would the result be the same. It is expressed as a coefficient (Cronbach alpha) where a coefficient of 1.0 equates with 100% reliability. This test is relatively high stakes in allocating students to posts but does not (and as it stands could not) define minimum standards of performance and hence a pass/fail cut off. The authors’ aim for a reliability coefficient of 0.8 is reasonable and represents the minimum acceptable standard. Lower reliabilities would have unacceptable measurement error and a potential significant impact on ranking. The test must remain legally defensible.

6:2 The conclusion that a test length of 60 items is adequate is acceptable but only “just.” Achieving this was dependent on ensuring all test items performed well. To ensure this thorough piloting of items is essential. Thought needs to be given as to how this would be achieved when the test goes “live”. There is little margin for error in the current test length of two hours. Many high stakes assessments test across three hours to ensure reliability. Careful item preparation and piloting is essential.

6:3 The rationale for apportioning the 2:1 random to MCQ item split for the paper is not clear. Item analysis (Appendix B) shows a strikingly better performance on the “partial item” discrimination measure for the MCQ questions. Reconsideration of the current split and movement to a 50:50 split might significantly help by ensuring good item psychometrics and improved reliability. The current mean discrimination index of 0.19 for items is not bad but could be improved.

6:4 Standardisation across the tests is essential and the results of this pilot cast some doubt on the reliability of offering multiple papers across a year. As outlined in 3:3 the LP demonstrated significant differences across the three papers with low reliability on one paper. This almost certainly relates to blueprinting and item performance variance rather than the cohort ability differences proposed by the authors.

6:5 Anchor items (3:2) were used across the papers to allow comparison of overall cohort performance. The psychometrics of the anchor items used were not known prior to inclusion in the LP as these items had not been piloted. Anchor items are generally selected on prior performance and need to have a good discrimination index. The report's recommendation for "25-40% anchor items per paper for test equating purposes is robust. The use of pre-tested items is in fact essential to ensure good psychometrics and compare cohorts across as well as within years.

6:6 The pilot results suggest there is acceptable discrimination which will permit ranking. There is a wide range of scores on both scoring methods which the authors claim is sufficient. The mean item scores tended to be high around 80% across both pilots. All cohorts demonstrated a reasonable normal distribution of performance. It remains difficult to make a full judgement until data is available across a complete school intake and range of ability. There are a number of tied scores which will need addressing as these will impact on the ease of ranking.

6:7 The authors speculate on the need to define a cut off score or minimum standard of performance of the SJT for registration purposes. The test is designed for selection purposes and not licensing. It would be unwise to redefine the test as a licensing test with a cut off standard in its current proposed format. I would strongly advise against this. Not the least because, as outlined above, we remain uncertain as to what is actually measures. Defining a minimum competence standard would be difficult. Secondly testing would almost certainly need to be extended to ensure a robust standard error of measurement around the pass mark.

6:8 As discussed in 4:3, differences in performance were found across sex and ethnicity not unsurprisingly as this is a common finding. Differential Item Functioning was used to determine whether the test items were appropriate and fair for assessing the ability of different demographic groups. The measurement flags possible bias but is not a stand alone measurement. Twenty-eight items were flagged as possibly discriminating and for further review with adjustment or repiloting as necessary. Overall experience would suggest the differences may be more candidate than item specific.

7: Acceptability

7:1 Over 95% of participants evaluated the SJT. The general consensus was it was relevant, fair and significantly better than the current white paper questions. There were some concerns it would discriminate sufficiently to reliably inform ranking and it measured a rather narrow range of ability excluding parameters such as leadership.

7:2 On the whole the length of the test was acceptable.

7:3 Differences in performance relating to sex and ethnicity will need further attention. There were some comments on the length and complexity of the scenarios. To reduce disadvantage for candidates scenarios must be short, concise, avoid unnecessary distracting information and written in clear English

7:4 Computer versus pen delivery was piloted in the LP using several approaches. Several delivery logistics and unforeseen circumstances impacted on delivery and make firm conclusions difficult. Computer delivery on balance seems acceptable and practical. Further piloting is necessary.

8: Feasibility

8:1 Item writing proved very resource intensive with a relatively low output of 5.5 items per writer. Questions then required review and revision. If the intention is to administer the test over multiple dates each year a large bank of questions will be needed. The resources needed to both generate and pilot a bank of questions sufficiently large for multiple delivery and to ensure the necessary high standard of psychometrics to meet the required level of reliability are significant. Review of the item writing process using a small number of committed writers reimbursed for their time and keeping testing dates to absolute minimum will need to be considered.

8:2 The report raised doubts as to whether multiple opportunities to take the test across a year are feasible. A large bank of items with confirmed good performance statistics and careful blueprinting is essential. The lack of standardisation across the three LP papers would suggest delivery of two papers a year would be more practical. It is not pass/fail and resits do not have to be offered.

8:3 Whether one can coach to the test (5:8) and learn the desired responses remains to be seen. Candidates seem to think coaching will occur.

8:4 The feasibility of computer versus pen needs further piloting (7:4).

9: Educational impact: White paper questions encourage reflection and keeping a portfolio on which to base answers. The SJT is not designed for educational impact. In theory candidates cannot prepare for the test. In actuality there is wide spread concern they can tutored for the test (as for UKCAT) which would be educationally disadvantageous. This could be addressed by a study investigating whether training improves performance.

10 Minor issues to raise with authors:

10:1 Can the authors provide the percentage of participants compared to total numbers in the medical school year?

10:2 p30 4.6.2 States five locations but gives six?

10:3: p33 would be good to have range of specialities of experts used to concordance panel

10:4: p100 3:1 Is the correlation given for paper 1 $r = .04$ an error??

Conclusions

1: This is an outstandingly good evaluation pilot of an assessment tool covering all the essential parameters. It is well designed and implemented and very clearly written. Many sincere congratulations to the working group.

2: Item design and piloting was very resource intensive with 40% attrition of items across a multistaged process and with some poor commitment from writers and a low output per writer. For the test to succeed item writing and piloting needs both revision to produce a more efficient system and significant resourcing. Employing a smaller group of committed, experienced (and possibly reimbursed) writers and ensuring expert input to and review of rational scoring methods seems an attractive alternative.

3: The test blueprint needs careful planning and perhaps revision. A tendency for writers to overproduce questions in one domain and skew the blueprint needs addressing. Is the balance of ranking and MCQ items the correct one? How was it reached and, as MCQ items appear to perform better, is it open to review?

4: Scoring is complex and needs expert input as the methodology is not fully transparent. It should be kept under review and may need future modification as has been done by other specialties. Currently the recommendation is not to change.

5: Validity has been well evaluated using several approaches. The superiority of SJT testing to the current white paper is robustly established. Further work on validity and defining what the test actually measures will be of interest.

6: Sixty items and two hour testing time is "just about" acceptable for the level of reliability needed to ensure the test is legally defensible. It is dependent on robust item psychometrics and not open to including unpiloted questions.

7: The SJT is acceptable to the pilot participants and deemed fairer than the current system.

8: The feasibility of standardising multiple delivery of papers across a year seems under threat given the difficulties experienced across the three LP papers. Establishing a sufficiently large item bank in the time frame is challenging but essential.

Appendix G: Response to Quality Assurance Peer Review Summary

Number	Issue	Summary of changes/comments
CMm1	Provide example SJT items	Examples of each item type are now provided in the appendix (those used for feedback).
CMm2	Scoring key requires clarity	The scoring convention for marking the SJTs requires clear communication to students as it is complex. We suggest that the communications strategy addresses this issue directly with schools and students.
CMm3	Rational vs empirical scoring key	In addition to paragraph 6.10.5 recommending the use of rational keying due, a final recommendation (9.2.8) provides a clear justification for the use of rational keying for scoring.
CMm4	Reliability estimates for candidates who answered all items	Some candidates are removed from the analysis due to substantial missing data and/or erratic scoring patterns. As stated in 7.3.7 (and now 6.3.4) <i>The sample figures we have presented aim to give a robust view of the pilot results by excluding those participants with unusually high levels of missing data and/or erratic response patterns. In an operational test, analysis would be completed on the entire sample, and figures reported as such.</i> Rather than removing all individuals who have missed any items, we recommend the approach of only removing those with high levels of missing data/erratic scoring when evaluating pilot data. However, we have provided some figures below for further information, although these are not in the main report. For the initial pilot, the reliability is $\alpha=.68$ (n=405), compared to $\alpha=.74$ (n=453). For example from the large scale pilot, Paper 2 is $\alpha=.74$ (n=332) which is the same as for n=341. (p.74)
CMm5	Sampling size of UK vs Non-UK	The sample size for these groups is small, so to avoid any potential misinterpretation, this section has now been removed.
CMm6	Correlation results, N and p value for Social Desirability	This data has now been added to the main report (N=396, $r=-.04$, $p=0.43$), p.53.
CMm7	Naming of large scale pilot and combination of the three papers	It is acknowledged that the term 'large scale pilot' may be a little misleading due to the small numbers that actually attended, however we recommend we keep this terminology to avoid confusion, as this term was used in the initial project plan. There was no attempt to combine the papers as this would be problematic in psychometric terms, as the papers contained different items which have not been equated.
CMm8	Consistency of naming of Medical Schools	Anonymity of medical schools is maintained throughout the report.
CMm9	Reporting of mean scores for those with extra time/disability	Only 14 candidates in total reported a disability and therefore the sample size is too restricted to provide any meaningful analysis. The sample size per paper of those with a reported disability was as follows: Paper 1 n= 2; Paper 2 n= 7; Paper 3 n= 5. A total of 58 candidates reported on the evaluation form that they had extra time, but as outlined in 7.2.11, this is likely to be inaccurate as some individuals reported they had up to 90 minutes extra time

		where as the recommended extra time is 25% (30 minutes). We have now clarified this issue in the main report (p.62).
CMm10	Description of DIF requested. N's as a percentage of items requested	A description of our approach to Differential Item Functioning analysis and N as percentage of items is now included in the main report. (p.77)
CMm11	Ensure wording reflects references	The report has now been amended accordingly so interpretation of group differences relates to both selection and exam data (p.75).
CMm12	Personality correlations & Bonferroni corrections applied	Correlations with five domains of personality are now included and the Bonferroni correction was also conducted. Results have now been updated accordingly. (p.78)
CMm13	Predictive validity. Interpretation and removal of outliers	The wording has now been amended to enhance clarity and the outliers removed and discussed in the main report. (p. 81-84)
CMm14	'Assurance that the language used did not discriminate against participants not trained in the UK'	The relevant paragraph reads: <i>Although during test development item writers were asked to ensure that the language used did not unfairly discriminate against participants who were not trained in the UK, an additional language and diversity check was carried out as part of the design methodology to help ensure that the content of the items did not unfairly discriminate against any particular group due to a factor unrelated to the job specification.</i>(p.95) Steps have been taken to ensure the language used does not unfairly discriminate against candidates not trained in the UK. The Project Group recommends this issue continues to be addressed in the next stages of development.
VW1	Reliability of test	It is acknowledged that the reliability for the tests is ' <i>just about acceptable</i> ' ranging from 0.71-0.76 and 0.79-0.85, once corrected using Spearman Brown estimate (p.47 & 70). Challenges including difficulties with the item writing methodology (p.93) and limited sample sizes should be considered.
VW2	Rationale for split of item types	A rationale in the main report is provided as follows; "<i>Evidence and previous experience in other contexts⁹ suggests that approximately two thirds of the items used in the test should have the ranking answer format, and one third should have the multiple choice format. Using a majority of ranking items allows for greater granularity in the data due to the more complex response/scoring format, however ranking items are typically more time-consuming to complete and are not appropriate for all scenarios, therefore a balance with multiple response items is needed. This structure also follows the validated methodology from the SJT for use in shortlisting to General Practice training posts⁹ and tends to reflect the proportion of items produced of each type by item writers (see 4.3.22)</i>". (paragraph 3.3.2, p.17) Ranking is a more complex format where the individual has to analyse and make a forced choice ranking of five pieces of information, therefore ranking items are likely to demand more cognitive effort than multiple response items. Therefore ranking

		items have slightly higher weight because they take a little longer and involve more decisions, however, the weighting of items is actually by their variance and not by numbers of points available.
VW3	Standardisation and use of multiple papers	As this was a pilot, the papers were not equated and therefore we would not plan to standardise until full psychometric properties were known (as outlined in 6.5 of VW report).
VW4	Resourcing of item writing methodology	It is acknowledged that the item writing process described in the report can be resource intensive. A new methodology that makes use of SJT item writing expertise is currently being piloted. It is expected that this will improve the item writing efficiency.
VW5	Cut score/minimum standard of performance	The peer reviewer states <i>'The test is designed for selection purposes and not licensing. It would be unwise to redefine the test as a licensing test with a cut off standard in its current proposed format. I would strongly advise against this. Not the least because, as outlined above, we remain uncertain as to what is actually measures. Defining a minimum competence standard would be difficult. Secondly testing would almost certainly need to be extended to ensure a robust standard error of measurement around the pass mark.'</i> The authors have been asked to present an options paper for standard setting methodology. This will be given to the Project Group to be used as part of their consultation on how to take the issue forward of those who do not meet the person specification.
VW6	Provide percentage of participants compared to total number in medical school	Now included in the main report (p.60).
VW7	Focus group locations	Now included in the main report (p.30).
VW8	Range of specialities for concordance	This data was not collected as this was part of the piloting of the methodology. However, this data will be collected in future pilots alongside other demographic data.
VW9	Correlation for Paper One between item types	Now included in the main report (p.100).

Improving Selection to the Foundation Programme

Appendix G

SJT Pilot Administrators Guidance

SJT Pilots, Spring 2011

Guidance for Administrators

Contents

Contents	378
1. Introduction	379
2. How the SJTs will be run.....	379
3. Planning Ahead.....	380
Responsibilities	380
Venue and Materials	380
Invigilators and administrators	381
Briefing students	381
Security.....	382
Enrolling students	382
4. Final Preparations.....	382
Getting to know the paperwork	382
Receiving the papers	382
Checking the venue	382
5. Running the Test.....	383
Applicant registration and briefing.....	383
During the Test	383
After the Test	384
Appendix B: EXAMPLE Applicant Assessment Rules.....	388

1. Introduction

- 1.1 Thank you for agreeing to take part in the 2011 pilot of the Situational Judgement Test (SJT) for selection into the Foundation Programme.
- 1.2 This guidance is meant to provide the information that medical schools will need from us in order to run a successful SJT. Please read it carefully. Although this version of the guidance is intended to be used in the pilot, most of it is written as if it were to be used for a live recruitment round. One aim of the pilot is to learn more about the practicalities of delivering the test, so that we can iron-out problems and refine the process for live recruitment rounds. This means that your feedback is important, so please let us know if there is anything about the guidance that is unclear or might be improved.
- 1.3 The next section of the guidance presents some general information about the way in which the SJTs are intended to be run. After that, the remaining sections reflect the sequence in which medical schools will need to get things done, namely:
 - Planning ahead- what needs to be done in advance, e.g. to book rooms and invigilators
 - Final preparations- things to be done just before the test
 - Running the test- including invigilation, and instructions for applicants
 - Giving us feedback about your experience of the pilot
- 1.4 There are some appendices at the end of the guidance that contain detailed reference information
- 1.5 If you have any questions about the guidance then please contact the project team via admin@isfp.org.uk or 020 7419 5494

2. How the SJTs will be run

- 2.1 The SJTs will be part of the selection process for recruitment to the Foundation Programme. All applicants will be required to sit an SJT under invigilated conditions.
- 2.2 We expect that UK medical schools will be responsible for running the SJTs in accordance with national guidelines. UK students will take the test at their medical school; there will be special arrangements for overseas applicants to take the test in the UK.
- 2.3 An important priority is that the SJTs must be held in a way that prevents applicants from passing on information which might help other applicants to achieve an improved score. Given this, SJTs will be held at a set time on a small number of dates (expected to be 2-3) according to the national timetable for the recruitment round. The questions used in the SJTs held on one date will differ from those used on another date, so applicants taking the tests on a later date cannot benefit from 'leaked' questions from an earlier date. All applicants taking the test on a given day must start it at the same time, or be subject to suitable 'quarantine' arrangements (discussed further below).

- 2.4 We currently expect that the tests will be delivered on paper, and machine-marked. The papers will be centrally printed and distributed to schools shortly before the test is to be run. The answer sheets will be collected from schools and processed centrally.
- 2.5 The scores from the test will be centrally loaded onto the computer system that is used to administer FP selection. In order to match the results of the test to the right student records on the system, each student's answer sheet will need a unique ID. For the time being we are assuming this will be the applicant's 'RA number' which is assigned to them when they apply on the FPAS system.

3. Planning Ahead

- 3.1 This section covers the activities that need to be done or planned well-ahead of the test.

Responsibilities

- 3.2 Each school should nominate a suitable member of staff to take overall responsibility for the proper running of the test, and to act as a point of contact for liaison with the MSC central team who will be coordinating the tests nationally. This person should also have a suitable stand-in who can take-over the management of the arrangements if required.

Venue and Materials

- 3.3 You will need to book a suitable venue for the tests. The venues you use for other formal assessments, such as finals exams, might be suitable, but please consider the following:
- 3.4 The venue needs to be large enough to hold all of the applicants at a single sitting, or it needs to allow applicants to be quarantined so that those taking an early sitting cannot pass on information to the others.
- 3.5 The applicants will need to sit at individual desks- or at most two to a table- with ample space between them to prevent copying and to allow invigilators to walk between desks.
- 3.6 There will need to be a clock that is visible to all applicants.
- 3.7 PowerPoint and a projector that is visible to all applicants should be available for the feedback session.
- 3.8 Applicants cannot be allowed to keep books, computers, etc with them during the tests, so you will need to find a place to store such items during the test.
- 3.9 The venue should be a suitable one for a high-stakes assessment, with good lighting, low background noise, etc.
- 3.10 Applicants will be asked to bring their own pencils, erasers and sharpeners, but you should have a stock of spares to use as a contingency measure.
- 3.11 You will need to register the applicants before allowing them into the test room, so you will need space for them to congregate, and facilities to check them in.

- 3.12 We will provide you with paper copies of the SJT question papers, OMR forms and evaluation forms. The papers will be delivered by UPS two working days (working days being Monday to Friday) in advance of the pilot, to be signed for by the named Responsible Officer. There will be instructions with the delivery regarding the secure return of the papers (including blanks) and collection by UPS the next working day after the pilot.
- 3.13 We will provide you with a list of applicants and RA numbers for your school, for reference on the day for those students who may have forgotten to bring with them their RA number. If a student does not have an RA number, and nor can their name be located on the list of RA numbers, that student should enter their RA number as 111 followed by their date of birth, e.g. 111 211 080 (for 21st October 1980).

Invigilators and administrators

- 3.14 You will need to make arrangements for a sufficient number of suitable invigilators and administrators to be available for the test, allowing for the possibility that extra hands may be needed in case people drop-out on the day. There should be at minimum one invigilator for every 60 applicants. The invigilators should all be briefed in advance on the standards to be enforced during the test (see below).

Briefing students

- 3.15 Briefing materials for students will be centrally produced, and provided to you to cascade through your local communications networks. The materials will also be published on a central website. In live rounds the national timetable will include the dates on which information for applicants will be published.
- 3.16 For the pilots, the key messages we want to give to students are as follows:
- Taking part in the pilot will be good practice for the future, as students are likely to encounter SJTs when they come to apply for specialty training posts
 - It is important that we get a good turn-out of students in order to have a statistically significant sample size
 - The pilot SJT results of individual students won't be used for anything other than evaluating the effectiveness of the SJT- they won't have any affect at all on a student's application for FP training, nor will they ever be made public. We won't retain any personally-identifiable information
 - There will be a prize draw for students, with a total prize of £100 and £1, 000 will be given to the MedSoc or Grad Ball (by prior agreement) if there is a turnout of 50%+ on the day
 - Each student will need to bring their RA number to the test
- 3.17 In keeping with the British Psychological Society (BPS) ethical guidelines, you must make it clear to students that their participation is entirely voluntary- they don't have to take part if they don't want to
- 3.18 You will need to provide your students with information about the date, time and venue for the test. Students should be asked to arrive at the venue in good time for registration and briefing.

Security

- 3.19 There are two main reasons why the SJT test questions need to be kept secure. Firstly, if students were to access the questions before the test then the results of the pilot might be invalidated. Secondly, some of the questions being piloted are likely to appear in live SJTs in future, so we don't want them to be leaked. Given this we will be taking special steps to keep the tests secure. The papers will be delivered to your school in special security wallets. You will need to make arrangements to keep the papers secure until you are ready to distribute them on the day of the test. You will also need to make arrangements to store the papers, and the completed answer sheets, after the test until we collect them from you.

Enrolling students

- 3.20 We need to be sure that sufficient numbers of students turn up to take the pilot test. Given this, it is important that you should ask your students to confirm in advance whether they are willing to participate, and record their responses, so that you can monitor the rate of take-up to check that it is on-target. We will contact you periodically to check take-up, but if you have any reason to believe that turn-out might be low, please let us know directly.

4. Final Preparations

- 4.1 This section covers the activities that need to be completed shortly before the test.

Getting to know the paperwork

- 4.2 In advance of the test we will let you have samples of the documentation that you will be giving to applicants, such as answer sheets, feedback forms, etc. Please make sure that all those who will be involved in running the test are familiar with the paperwork that they will be expected to deal with.

Receiving the papers

- 4.3 On an agreed date before the test we will deliver your test documentation in a secure pack. Please store the pack, unopened, in a secure place until the you are ready to distribute the papers immediately before the test.

Checking the venue

- 4.4 Make sure that the room is properly laid out.
- 4.5 No rough paper is to be provided to applicants. Applicants may write on the question booklet, but only the answer sheet will be marked.

5. Running the Test

Applicant registration and briefing

- 5.1 You should have a written list of the applicants you are expecting to sit the test on the day. Applicants should not be allowed into the examination room until they have signed-in and provided evidence of their identity. Applicants will be asked to bring their 'RA' numbers with them on the day, as they will need to put their RA number on their answer sheet. We will provide you with a list of the RA numbers for your applicants, in the event that an applicant has forgotten to bring with them their RA number.
- 5.2 Question papers, glossaries and OMR answer forms may be placed on each exam desk before the applicants enter the room. Invigilators should watch to ensure that applicants do not open their papers.
- 5.3 Applicants should not be allowed to take personal belongings to their desks in the examination room. Mobile phones must be switched off in the examination room.
- 5.4 Applicants who arrive after the briefing has started must wait outside until the test is underway. They will then receive the briefing outside the examination room, before being allowed to enter and start the test.
- 5.5 Once all applicants are seated, announce that applicants must not open question papers until they are told to do so.
- 5.6 All applicants must receive a standard briefing before the assessment commences. The applicant briefing and administration procedure is available in Appendix A.
- 5.7 Once the briefing is completed the test may begin.

During the Test

- 5.8 Applicants are not allowed to leave and re-enter the assessment room during the test, except for a toilet break. Applicants wishing to leave and re-enter the assessment room during the test must be escorted.
- 5.9 Applicants should not be allowed to enter the examination room after the starting time, except in exceptional circumstances.
- 5.10 Agreed local procedures should be followed for giving extra time to applicants with known dyslexia or other medical conditions.
- 5.11 If an applicant fails to follow any of the agreed local rules regarding conduct during the assessment (see Appendix B for example rules), they should be informed immediately without disturbing other applicants.
- 5.12 Applicants should be informed 15 minutes before the end of the test and when the full time allowed for the test has been reached. Applicants who started the test late must finish the test at the same time as all of the other applicants, and a note made of this candidate's RA number to inform the MSC team.
- 5.13 It is essential that all question papers and answer sheets are counted and collected by invigilators at the end of the assessment. Once the assessment has finished, ALL paperwork must be collected by administrators before any Applicants are to leave their desks. As administrators collect the answer sheets, they should check that applicants have provided their name and RA number on their OMR answer sheet.

- 5.14 Applicants should be given an applicant evaluation form to complete and return when they have finished the test.. Applicants will notify invigilators of this by raising their hand.

After the Test

- 5.15 All question papers must be collected and returned via courier (UPS), as agreed with the Medical Schools Council.
- 5.16 The original answer sheets should be counted to cross-check against the list of registered applicants to ensure that all papers are present.
- 5.17 The original answer sheets should be placed in the envelope provided and held securely, together with the question papers and applicant evaluation forms, until they are collected by UPS.
- 5.18 For audit purposes you should:
- Retain the list of registered applicants.
 - Record details of any applicants who arrived late.
 - Record details of any actual or suspected breaches of security or procedures, and notify these to the Medical Schools Council.
 - Record the names of the administrators and invigilators who participated in the test, separately identifying those who were responsible for checking that all papers were present after the test.
 - Record details of any applicants who were allowed extra time, and the reasons for the allowance.
- 5.19 Since the purpose of the pilot is to gain insights into the effectiveness of the SJT arrangements, we need your feedback. We will provide a questionnaire for you to complete and return to us. Please try to complete the questionnaire as soon as you can, while the experiences of those concerned are still fresh in their minds.

Appendix A: Applicant Briefing

Please read the following aloud to the applicants once they are all seated:

- Thank you for volunteering to take part in the pilot Situational Judgement Test for Improving Selection to the Foundation Programme. Please do not open the question booklet until I tell you to do so.
- Before we begin, I want to reassure you that your participation in today's pilot will have no bearing whatsoever on your real application to the Foundation Programme. Please note that all of the answers and information you share with us today will remain completely anonymous, and will not be shared outside of the research team who are piloting and evaluating the SJT.
- Please be aware that the content of the SJT question paper is **STRICTLY CONFIDENTIAL**. By taking part in this pilot, you have access to the questions that may be used in future years by applicants to the Foundation Programme. It is essential that you do not reproduce or share any information relating to the contents of the pilot.
- Although today's test is a trial, we want to make it as realistic as possible, so we will be following a formal process for the conduct of the test.
- Please ensure that your mobile phone has been switched off and placed securely with your belongings. If you have any electrical devices or any written material that has not already been declared to an administrator, please raise your hand to declare this now (*allow a few short minutes*).
- If you wish to speak to an administrator during the assessment, please raise your hand and we will come to you.
- If you leave the room once the assessment has started, you will be unable to return. You will not be able to leave the room for the first hour of the test. The only exception is if you wish to go to the toilet, in which case please raise your hand and an administrator will escort you. You will not be able to go to the toilet during the last 15 minutes of the test.
- The test will last two hours. You will be told when there are 15 minutes to go before the end of the test.
- Once the SJT pilot is complete, we will collect the question papers and answer sheets. You will then be asked to complete a short evaluation form. Please raise your hand when you have finished the test and an invigilator will bring an evaluation form to you. Your feedback is important to help validate the pilot paper.
- We will then run a feedback session on today's pilot paper, which you will find useful to attend.
- You should have a question paper, answer sheet and separate glossary on your desk, and you should have brought your own pencil, sharpener and eraser. If you do not have any of these items, please tell me now (*Pause*).
- Please ensure that you complete your RA number and paper number on the answer sheet – you need to mark the paper number on both sides of the sheet. You may make rough notes on the question paper should you wish, but only your answers on the answer sheet will be scored.
- Instructions on how to complete the Situational Judgement Test are on the front page of the question paper. Please follow them now as I read these aloud but do not turn the page until instructed to do so.

Please make sure that you:

- Fill in your paper number on both sides of your answer sheet, this number can be found on the front of your question paper. Please also fill in your nine digit RA number, name and medical school on the answer sheet. Please give your age in years, if you do not wish to give this then please leave it blank. When answering the disability question- if you consider yourself to be disabled, as defined by the Equality Act (2010), please select yes, if you do not, please select no.
- You have 120 minutes to complete this paper.
- There are two parts in this assessment, each with a different type of response format. Please read the instructions for each part carefully before attempting the questions; all questions require you to mark more than one option.
- There are 60 questions in this paper; 40 questions in Part One and 20 questions in Part Two. Please try and assign your time accordingly.
- In this assessment you will be presented with scenarios typical of those that Foundation Year One (FY1) doctors encounter.
- For each scenario, consider yourself to be a Foundation Year One (FY1) doctor.
- Please answer what you should do when responding to the questions.
- You may sometimes feel that you would like more information before answering, however please give your best answer based on the information provided in the question.

Please note:

- Mark your responses on the answer sheet by filling in the lozenges, using pencil only. Please fill in the lozenge as neatly as possible- any stray marks may mean that your answer cannot be read.
- If you need to change a response, rub it out with your eraser and mark your final response. Do not cross out any answers on your answer sheet – this will make it difficult for the machines to read your final answer.
- There is no negative marking; you should therefore attempt all the questions.
- You may write on the question booklet.
- There is a glossary on a separate sheet of paper with definitions of some of the terms that are used within the question paper. The glossary terms are marked with an asterisk (*). You may find it useful to consult this if you do not understand a term.
- The question paper & answer sheet must not be removed from the room.
- When you have finished the test, your question paper and answer sheet will be collected by the invigilator.
- Please turn over the page now.

- On pages two and three you will see two example questions. I will give you a few minutes to read these. Please do not turn on to page four.
- Does anyone have any questions? *(Leave a few minutes for applicants to ask any questions – only answer general questions about the process)*
- You have 120 minutes to complete the Situational Judgement Test, starting now. *(Begin timing 120 minutes).*
- *(After several minutes, administrators should walk around the room, ensuring that all applicants are completing the answer sheets appropriately. Inform applicants if they are not completing the answer sheet appropriately. Administrators should continue to walk around the room periodically throughout the assessment.)*

After 105 minutes, say:

- There are 15 minutes remaining for this test.

After 120 minutes, say:

- Stop working now. Please put your pencils down. That is the end of the test. Please remain in your seats whilst we collect the question paper, answer sheet and glossary.
- We will now give you a short paper to evaluate the SJT pilot. This will take no longer than five minutes to complete – once you have finished the papers, you may leave the room. Please put your hand up and we will collect the papers from you before you leave.
- Please ensure that you complete your RA number and the name of your medical school
- The first paper is labelled 'Evaluation of the pilot'. You should complete this paper within five minutes.
- Your feedback is very important to us and will help us improve this process going forward. Please note that your answers to this paper – as with the pilot – will remain anonymous.

After 5 minutes, say:

- Please remain in your seats whilst we collect the papers.
- Thank you very much for taking part in today's pilot Situational Judgement Test. We really do appreciate your contribution.

Appendix B: EXAMPLE Applicant Assessment Rules

The following are not permitted in the assessment room:

- Electrical devices of any kind (exceptions are made in the case of medical need e.g. hearing aids).

This rule covers:

- Mobile phones
- Personal organisers
- Personal scanners
- Laptops
- Calculators etc

Any devices brought to the assessment must be placed in the box provided by the invigilator (invigilator(s) do not take responsibility for any devices brought in to the assessment room)

- Any written material, including medical dictionaries (Foreign language dictionaries may be used but permission must be sought from the invigilator before the start of the assessment and the book will be checked to confirm that it is a dictionary and not notes/textbooks)
- Food - (drinks only are allowed in a sealed container)

Applicants may be dismissed from the assessment or have other action taken, including possible referral to the regulatory body, for any of the following reasons:

- giving or receiving help from another applicant during an assessment
- using notes, books, any unauthorised notations or other aids
- possession or use of photographic, recording or transmission devices in an assessment
- writing on any material other than that provided by the administration team for the purpose of note taking during the assessment
- removal of assessment materials or notations of any kind from the assessment room or making and removing copies of any part of such papers, answer sheets or assessment materials
- refusal to comply with time allotments or assessment administration procedures
- disruption of the assessment for other applicants
- reproduction or disclosure of assessment content in any manner (including unauthorised notations, engaging in discussion of assessment content with anyone other than assessment personnel during or after an assessment)
- providing and/or disseminating information about the assessment content with a view to assisting current or prospective Applicants whether before or after the assessment
- failure to follow the lead administrator's instruction, after a warning

In any such case, an incident report will be filed by the Responsible Officer and the applicant will be told of this action.

Improving Selection to the Foundation Programme

Appendix H

Report of the EPM pilot

Educational Performance Measurement (EPM)

Report of 2010 Pilot of the EPM draft framework

This document summarises what has been done to pilot the production of EPM scores by UK medical schools, the findings from the pilot, and suggested next steps. The current version is for consideration by the Improving Selection to the Foundation (ISFP) Project Group.

Table of Contents

1. Background	392
2. Draft quality criteria	392
3. Methodology	393
Developing the pilot EPM framework	393
Pilot EPM data collection	394
4. Findings of the EPM pilot	395
Analysis of pilot EPM scores and application data for FP2010	396
Analysis of medical school quartiles and EPM scores	397
Comparison of medical school quartiles and written EPM scores (22 schools)	397
Comparison of medical school quartiles and practical EPM scores (20 schools)	397
Analysis of Question 1A and Question 1B	399
5. Combining scores for two quartiles	399
Achieving 50:50 weighting between written EPM scores and practical EPM scores	400
6. Changing the scoring system	402
Distance from the mean vs distance from the median	402
Increasing granularity (deciles, centiles)	403
Increasing granularity (changing the difference in points awarded to quartiles)	404
7. Other findings	405
8. The way forward	406
Appendix A: Draft Framework for Educational Performance Measurement (EPM)	408
Appendix B: Issues raised by medical schools in adhering to the EPM framework	409
Appendix C: Caveats for interpretation of EPM scores	414

1. Background

Applicants to the Foundation Programme receive a score based on performance at medical school in relation to their cohort, as ranked into four quartiles worth 40, 38, 36 and 34 points. This score is then combined with their score from an online application form, white space questions (total of 60 points), including 10 points for degrees, prizes, publications and presentations.

Concerns about the use of academic quartiles, raised before and during an extensive and detailed Option Appraisal¹, are based around the comparability between applicants from different medical schools and discrimination of applicants at the margins between quartiles. Given issues around the lack of transparency and lack of consistency across medical schools, one of the main drivers to review the quartile system is to ensure defensibility in the event of legal challenge from an applicant. Stakeholder feedback showed strong support for the use of some measure of academic performance as well as non-academic and possibly extra-curricular activities. The advisory international Expert Panel supported the principle of making greater use of information accumulated during medical school, and the development of a standardised measure of educational performance.

The recommendations of the Improving Selection to the Foundation Programme Steering Group were to pilot and evaluate:

1. A Situational Judgement Test (SJT) to assess professional attributes, judgement and employability
In combination with
2. An Educational Performance Measurement (EPM) of applicants at their medical school to assess clinical knowledge/ skills as well as wider personal achievement

The EPM as a selection tool refers to a differential ranking score produced by the applicant's medical school to reflect the applicant's achievements or performance on a range of assessments in relation to their cohort up to the point of application to the Foundation Programme. It was envisaged that the EPM would be derived using a specified and standardised transparent framework of existing performance measures, with a weighting agreed with medical schools in consultation with students and other stakeholders. All UK and non-UK medical schools would be required to provide a local educational performance ranking to the UKFPO derived using the standardised framework.

The ISFP Steering Group envisaged that a uniform transparent framework for an EPM would address some of the current concerns about comparability between applicants of the same quartile from different schools, and that it would enable greater granularity². Depending on the results of the 2010-11 pilots, it is anticipated that the EPM would be used in combination with a second selection tool, resulting in a combined higher level of granularity than would need to be achieved by the EPM alone.

2. Draft quality criteria

The following quality criteria are measures, proposed by the ISFP project team for consideration by the Project Group, against which to evaluate the Educational Performance Measurement (EPM):

- The EPM is a reliable and representative measure of the applicant's performance in their educational progression at medical school up to the point of application
- The EPM is a valid measure in relation to selection to the Foundation Programme (i.e. it is measuring factors that have a bearing on suitability for the job)
- The EPM is a sufficiently granular measure
- The EPM is a fair measure

¹ Medical Schools Council (2009) *Selection to Foundation: An Option Appraisal*

² Granularity is taken to mean the differentiation between applicants (i.e. the number of different scores achieved)

- The EPM is not overly expensive to administer and quality assure
- Medical schools are able to adhere to the framework, and introduce it within suitable timescales
- Where medical schools are not able to produce a reliable EPM, there is a legal and justifiable fall-back
- There is no requirement on any medical school (UK or non-UK) to modify its curriculum in order to comply with the framework
- The framework complies with all relevant legislation

3. Methodology

Developing the pilot EPM framework

Two in-depth consultations in autumn and winter 2009 involving all 30 UK medical schools whose students apply to the Foundation Programme³ informed the development of the draft EPM framework for piloting.

The first consultation, in September and October 2009, gathered evidence around the information currently used to inform quartile rankings, and the assessment information on student performance currently collected and utilised by UK medical schools. The range of information – and the formats – varied widely between schools, as follows;

- The survey revealed a range of assessment types, including clinical skills, curriculum knowledge, Student Selected Components and measures of professionalism, as well as summative, formative and progress-testing. The timing of some assessments is variable between schools.
- The two areas where all schools collect information on student performance are clinical skills and curriculum knowledge, although the number of assessments of each – and the relative weighting (e.g. within course compared with end of stage) – is variable, both within and between schools.
- Between 7% (Manchester) and 100% (Cambridge, Imperial, Oxford) of students at a given school intercalate (n.b. graduate entry students already hold a degree), with a median of 25%. Intercalation is usually competitive entry, based on performance at medical school, and not available to all students (for example graduate entrants).
- The granularity of information available is variable between schools and between assessments within schools. A full ranking within a cohort is sometimes neither feasible nor desirable, as assessments are designed to assess competence, not a spread of marks. Furthermore, as a matter of university policy many schools retain only the grades and bandings awarded for performance on assessments, rather than retaining the more-detailed raw scores. It is difficult to quantify scores for Student Selected Components (usually pass/fail), measures of professionalism (can be pass/fail) and extra-curricular activities (rarely collected systematically by the school, impossible to quantify).

The second consultation, in November and December 2009, consulted around the principles for an EPM framework, the weighting that might be awarded to curriculum knowledge and clinical skills, the proposal for two sets of quartile scores (one for written assessments, and the other for practical assessments), and how additional points for prizes, publications, presentations and previous degrees might be awarded. The survey responses are summarised below:

- All responses recommended that schools should determine how many assessments to include, and the weighting of these, with a general consensus that the weighting between clinical skills and curriculum knowledge should be 50:50.
- Most responses agreed that there should be a spread of scores across all years, with a higher weighting in the later/clinical years. Differences between Graduate Entry (GE) (4 years) and standard entry courses (5/6 years) should be taken into account.
- Responses indicated the need to take into account performance in both written tests of knowledge and those of clinical skills exams. However, some concerns were raised about separating scores for these two areas.

³ Excluding the University of St Andrews. Students transfer to the University of Manchester or to other Scottish medical schools at the end of Year 3, ahead of their application to the Foundation Programme.

- More than four fifths of schools believed that the EPM should award points for presentations, prizes and publications, in addition to points for performance in relation to the cohort. School responses also suggested that additional points could be awarded to recognise the top performing students within the cohort. Responses emphasised that guidance regarding the award of additional points should be standardised and transparent.
- 13 schools (7 GE) thought that there should be additional points for previous degrees as well as those gained during medical school; 7 (4 GE) thought that points should only be awarded for degrees gained during the time at medical school; 6 (3 GE) thought that no points should be awarded for degrees.

A working draft EPM framework for piloting, informed by the feedback from the two consultations, was agreed by the ISFP Project Group in May 2010 and by the Medical Schools Council in June 2010. The draft EPM framework called for students to be given two scores - one relating to written assessments (as an approximation for curriculum knowledge), and the other relating to practical assessments (as an approximation for clinical skills) – according to a specific prescription of weightings between earlier and later years of the course. The two scores would then be combined to provide an overall score.

The pilot EPM framework is attached as Appendix A.

Pilot EPM data collection

UK medical schools were asked to produce a pilot EPM, based on clinical skills and curriculum knowledge up to the point of application to the Foundation Programme 2010 (FP2010), for each of their students who had applied to the FP2010. As the pilot used retrospective data, the EPM scores could then be analysed with the original quartile scores supplied by the medical school, and with the application form scores achieved by applicants during the last recruitment round. At this stage, the framework did not include other aspects of performance such as Student Selected Components (SSCs), academic excellence, previous degrees and/or prizes, publications and presentations, as there was either no consensus around scoring (SSCs, academic excellence) or the information already existed for the cohort (previous degrees, prizes, publications and presentations).

Named contacts at each medical school were sent an individualised Excel template for completion, with the names and email addresses (unique identifier) pre-entered. Schools were tasked to complete the data manipulation, according to the prescription, within one calendar month and to return the data to Jobsite⁴. A two week extension was granted to some schools, with the final anonymised dataset provided to the Medical Schools Council in mid-August.

EPM data were provided as normalised scores, such that comparisons could be made within a single school cohort. The following analysis is based on the EPM scores provided for 5373 applicants to the FP2010, from 22 UK medical schools. Three schools provided EPM scores for a further 442 students in different cohorts, although the scores could not be correlated with other information. Analysis of the EPM data includes the cohort which graduated in 2010, and a small number of applicants who had deferred their application for 1 or 2 years. This analysis excludes applicants to the academic Foundation Programme who did not complete the white space question application form, as there is a separate application process.

The scores were then converted into two sets of quartile ranks reflecting performance in i) written assessments of curriculum knowledge (written EPM scores) and ii) practical assessments of clinical knowledge (practical EPM scores).

Schools were also invited to report back on the issues they encountered in following the pilot EPM framework. A summary of this feedback is attached as Appendix B.

⁴ Jobsite holds application data for the Foundation Programme, by contract with the UKFPO. Anonymised data were correlated and provided to the Medical Schools Council free of charge for the purpose of this pilot.

4. Findings of the EPM pilot

EPM scores were produced by 25 of the 30 UK medical schools with graduating final year students within the timeframe of the pilot. 22 medical schools provided data in a format that enabled analysis of EPM scores with existing data on quartiles and scores on the white space questions. Divergences from the prescribed EPM framework are detailed in Appendix C, and should be referred to when interpreting the EPM data.

Schools reported that the framework required the use of additional assessment data (for example early years assessments), whereas other schools reported that the framework limited the number of assessments that could be used (for example by requiring a split between 'written' and 'practical' EPM scores). EPM scores were provided in raw score format, and Figures 1 and 2 illustrate the range of marks and formats of scores. Whereas in Southampton, written EPM scores consisted of 13 possible scores, for 228 applicants, between 0.5 and 2 (2 being the lowest performing applicant and 0.5 the best performing applicant), Edinburgh provided written EPM scores enabling a complete ranking of all 208 applicants between 65.8 and 88.8 (to 6 decimal places), and Sheffield written EPM scores were provided as 210 possible scores (223 applicants) within a range of 139.6 and 397.6.

8 of 25 schools provided data on written EPM scores which enabled a full rank (no tied scores). In other schools as many as 100 of 228 students had a tied score. For 12 of 25 medical schools, applicants could be ranked into equal quartiles using written EPM scores. Owing to tied scores, for the remaining quartile groups – with the exception of Southampton – there was a variation of between 4 and 10 individuals in the group size.

Figure 1: Mean, median and mode of written EPM scores (25 schools)

School	No. of students	Lower range	Upper range	Mean (2.d.p)	Median	Mode	Mode (no of ties)
Aberdeen	173	9.78	18.33	13.32	13.87	11.67	3
Barts and The London	322	50.28	86.33	66.11	66.71	No ties	No ties
Birmingham	382	51.57	80.00	64.95	64.40	No ties	No ties
Brighton and Sussex	125	57.90	87.92	71.11	71.53	No ties	No ties
Bristol	243	53.02	76.67	63.58	63.31	No ties	No ties
Cambridge	135	35	60	48.90	49	51; 52	12
Cardiff	314	37	86	55.95	77	71.35	9
Dundee	150	31.8	43.9	36.55	36.35	33.60; 34.90; 39.70	3
Edinburgh	208	65.82	88.84	77.22	77.09	No ties	No ties
Glasgow	220	17	42	13.88	28	29	27
HYMS	116	195	279	236.39	239	242; 247	6
Keele	105	602	1024	690.62	907	842; 850	4
King's College London	374	52.95	81.16	68.00	67.71	No ties	No ties
Leeds	261	50.57	94.50	71.77	71.50	66	9
Leicester	203	44.6	86.9	63.33	62.6	60	4
Liverpool	312	66.00	83.03	73.36	74.82	76.1	5
Newcastle	350	12.5	28	22.73	23	21	74
Nottingham	313	51.19	82.17	66.03	65.89	No ties	No ties
Oxford	158	180.83	217.16	199.47	199.70	No ties	No ties
Sheffield	223	139.60	397.60	303.64	302.60	273.5	3
Southampton	228	0.50	2.00	1.26	1.30	1.3	100
St Georges, London	294	37.33	83.82	67.10	66.80	60	4
UCL	329	42.74	89.06	68.21	69.07	69.08	2
Uni. of East Anglia	118	148.60	255.40	210.00	213.70	224.4	3
Warwick	159	1.67	43.75	27.54	29.17	30	15

Three schools provided data which enabled a full rank (no tied scores) for practical EPM scores. For 16 of 25 schools, applicants could be ranked into four equal quartiles using practical EPM scores. Owing to the number of tied marks for the remaining quartile groups there was a significant variation, for example in Newcastle, four times as many applicants were placed in the third quartile (126) as in the fourth (32). For two schools (Leeds and Southampton), there was insufficient granularity based on the practical EPM scores to create quartile bandings.

Figure 2: Mean, median and mode of practical EPM scores (25 schools)

School	No. of students	Lower range	Upper range	Mean (2.d.p)	Median	Mode	Mode (no of ties)
Aberdeen	173	9	20	13.68	13.86	14	8
Barts and The London	322	50.00	75.25	62.00	62.49	No ties	No ties
Birmingham	382	45.53	78.41	63.39	63.52	67.54; 68.41	2
Brighton and Sussex	125	54.00	83.40	70.57	71.50	62	14
Bristol	243	53.63	74.21	63.67	63.67	59.56; 61.61; 62.19; 62.70; 66.16	2
Cambridge	135	108	161	138.20	137	140	10
Cardiff	314	35	82	67.60	68	70	22
Dundee	150	65.30	85.40	73.12	73.10	66.3	3
Edinburgh	208	63.54	90.90	79.53	79.89	No ties	No ties
Glasgow	221	10	28	22.25	22	24	46
HYMS	116	46	75	58.94	58	56	12
Keele	105	539	796	658.70	660	626; 660; 663; 666; 727	4
King's College London	374	50.39	84.83	66.67	67.31	68	3
Leeds	261	0	100	64.52	60	60	131
Leicester	203	31.20	85.40	69.42	69.80	69	6
Liverpool	312	66.84	82.55	72.94	72.80	67	5
Newcastle	350	19	28	23.67	24	24	81
Nottingham	313	48.28	82.87	65.61	65.25	67.52; 62.32	3
Oxford	158	71.28	89.67	80.06	79.97	No ties	No ties
Sheffield	223	350	572	483.40	488	483	6
Southampton	228	0.5	2	1.30	1.5	1.5	115
St Georges, London	294	56.66	88.48	71.04	71.36	60	6
UCLMS	329	53.73	89.51	74.21	76.56	79.37	2
Uni. of East Anglia	118	712	914	821.40	825.25	817.5	3
Warwick	159	20.83	50.00	43.82	45.83	50	15

The number of assessments used to derive a measure of educational performance, the range of possible scores for each assessment, and the weightings applied to different assessments, inevitably affects the range of total scores achieved by individual applicants. For example, the University of Oxford included 14 assessments (8 written, 6 practical), whereas the University of Southampton included 3 assessments (2 written, 1 practical). At the same time, the potential scores available for each assessment can affect the range of total scores achieved. The University of Manchester awards scores based on five grades (Unsatisfactory – Distinction), the University of Aberdeen awards scores on a scale of 9-20, whereas other universities award percentages. Retaining only graded data is a university-wide policy in at least five institutions.

Analysis of pilot EPM scores and application data for FP2010

Figure 3 illustrates the Pearson's Product-Moment Correlation Coefficient⁵ between various pairs of measures relating to the performance of applicants to the Foundation Programme 2010. The first six rows compare measures that were collected during the live recruitment round, namely the original medical school quartile, the application form score

⁵ Pearson's Product-Moment Correlation Coefficient is a measure of dependence between two variables. A correlation of +/-1 is a perfect (linear) correlation, and a correlation of 0 shows no correlation.

(excluding Question 1A and Question 1B), and the scores for Question 1A and Question 1B. The remaining rows include comparisons with and between the pilot EPM scores.

- n.b. Application form score (60 points) – Q1A, Q1B and five ‘white space’ questions, mapped against the Foundation Programme person specification, with a total of ten points each
 Question 1A (total 6 points) – previous degrees and intercalated degrees (national scoring criteria)
 Question 1B (total 4 points) – prizes, presentations and publications (national scoring criteria)

Figure 3: Correlation for pairs of measures for applicants to FP2010 (22 schools)

Measures	Pearson's Correlation
Medical school quartile v application form score	0.18
Medical school quartile v Question 1A score	0.18
Medical school quartile v Question 1B score	0.12
Application form score v Question 1A score	0.08
Application form score v Question 1B score	0.04
Question 1A score v Question 1B score	0.24
Medical school quartile v written EPM score	0.83
Medical school quartile v practical EPM score	0.65
Application form score v written EPM score	0.15
Application form score v practical EPM score	0.16
Written EPM score v practical EPM score	0.51
Written EPM score v Question 1A score	0.18
Written EPM score v Question 1B score	0.11
Practical EPM score v Question 1A score	0.10
Practical EPM score v Question 1B score	0.08

Three pairs of measures indicate significant concurrent validity, namely the medical school quartile and written EPM score, medical school quartile and practical EPM score, and written EPM score and practical EPM score. The correlation with the practical EPM scores is slightly less than with written EPM scores, suggesting that schools may not have previously placed as much emphasis on practical assessments when devising quartiles for recruitment to the Foundation Programme.

Analysis of medical school quartiles and EPM scores

Comparison of medical school quartiles and written EPM scores (22 schools)

- There is a correlation of 0.83, indicating a strong correlation between the assessments – and weightings – used by schools to inform quartiles as to inform written EPM scores
- 67% of students were in the same quartile for written EPM scores as with medical school quartiles. 30% of students moved to one quartile either side of their original rank, and 3% of students moved by more than 2 quartile ranks⁶

Comparison of medical school quartiles and practical EPM scores (20 schools)

- There is a correlation of 0.65, indicating a moderately strong correlation between the assessments – and weightings – used by schools to inform quartiles as to inform written EPM scores.

⁶ i.e. applicants were moved from 1st to 3rd or 1st to 4th; 2nd to 4th; 3rd to 1st; 4th to 2nd or 4th to 1st

- 46% of students were in the same quartile. 44% of students moved to one quartile either side of their original rank, and 10% of students moved by more than 2 quartile ranks

An analysis of some of the correlations by medical school is presented in Figure 4 (the cells are blank where data was not available).

Figure 4: Correlation between medical school quartiles and EPM scores, by school

Medical school	Application form v original quartile	Medical school quartile v written EPM	Medical school quartile v practical EPM
Aberdeen	0.32	0.92	0.76
Barts and The London	0.26	0.93	0.48
Belfast	0.28		
Birmingham	0.37	0.95	0.49
Brighton and Sussex	0.17	0.82	0.53
Bristol	0.33	0.87	0.66
Cambridge	0.28		
Cardiff	0.36	0.83	0.67
Dundee	0.24		
Edinburgh	0.22	0.93	0.84
Glasgow	0.09	0.79	0.69
HYMS	-0.05	0.79	0.59
Imperial	0.21		
Keele	0.06	0.89	0.78
King's College London	0.34	0.80	0.68
Leeds	0.19	0.77	
Leicester	0.19	0.79	0.63
Liverpool	0.21	0.83	0.49
Manchester	0.20		
Newcastle	0.24	0.81	0.57
Nottingham	0.20	0.62	0.76
Oxford	0.12		
Peninsula	0.28		
Sheffield	0.35	0.90	0.75
Southampton	0.34	0.75	
St Georges, London	0.33	0.85	0.65
UCL	0.23	0.80	0.85
Uni. of East Anglia	0.15	0.76	0.78
Warwick	0.01	0.75	0.50

It is interesting to note the different levels of correlation between the two EPM scores and medical school quartiles. Only in one school – Sheffield – is the correlation stronger between medical school quartiles and the practical EPM scores, than with written EPM scores.

Feedback from medical schools in collating quartiles suggested that most of the difference in rank place relates to students at the margin. The EPM framework required schools to use the passmark for any resit assessments; however, subsequent feedback highlighted that many schools use the first-attempt mark. This created clustering at the

margins, as more applicants who had previously received their first-attempt mark, received the pass-mark. The piloted EPM framework also stipulated the weighting of assessments in the early and later years of the course. Where the weightings differed from those used by schools to inform quartiles, this also created movement in the ranked places to reflect student performance.

This analysis shows that the ranking a student achieves varies according to the ranking formula used. A common approach to producing EPM scores would reduce variability between schools whilst taking into account the approach to assessment and progression within schools.

Analysis of Question 1A and Question 1B

Data relating to additional points for previous degrees, and for prizes, publications and presentations, can be correlated with the pilot EPM scores using points awarded to applicants for Question 1A (previous degrees) and Question 1B (prizes, publications and presentations) on the white space application form.

Figure 4 illustrates a wide variation in the degree of correlation between medical school quartiles and application form scores, which have a weak positive correlation (0.22) overall. This suggests that the white space questions may be measuring different qualities and competencies of applicants, particularly for Question 1A and Question 1B. Depending on the results of pilots in 2010-11, the white space questions may be replaced with a Situational Judgement Test (SJT). Thus points awarded to recognise previous degrees, prizes, presentations and publications may only be incorporated within the EPM. The following analysis explores how these components are currently combined.

Question 1A (previous degrees – maximum 6 points)

- 57% of students received points for Q1A, of whom 34% were also in the 1st quartile
- 67% of students in the 1st quartile received points for Question 1A, compared with 48% in the 4th quartile
- Overall 7% of students scored between 1 and 4 points; 35% of applicants scored 5 points (55% of whom were in the 2nd and 3rd quartiles); 15% of students scored 6 points (76% of whom were in the 1st and 2nd quartiles)

Question 1B (prizes, publications, presentations – maximum 4 points)

- 25% of students received additional points for Q1B, of whom 34% were also in the 1st quartile
- 31% of students in the 1st quartile received additional points for Question 1B, compared with 18% in the 4th quartile
- Overall 17% of students received 1 point for Question 1B; 6% 2 points; 2% 3 points; and 0.4% 4 points

This analysis suggests that whilst the minority of applicants gain points for prizes, publications and presentations – and just over half gain points for previous degrees - the award of additional points does provide some additional discrimination at the upper end of the spectrum.

5. Combining scores for two quartiles

One of the aims of the EPM is to improve granularity from the current four quartile ranks⁷. By collecting two sets of quartiles of performance at medical school – written EPM scores and practical EPM scores – the two quartile scores could be combined in some way to achieve additional granularity. In principle, other bandings for example septiles or deciles, would achieve a higher level of granularity, although the feasibility and desirability of bandings at this level would need to be considered.

⁷ Where more than one student achieves the same mark, or there is an odd numbered cohort size, it may not always be possible to divide a cohort into four equal groups.

Figure 5 illustrates the moderately positive correlation (0.51) between the written and practical EPM scores, which shows concurrent validity for the two measures. The cells in the table show the number of applicants with a given combination of practical and written EPM scores. 43% of applicants were placed in the same quartile for performance on written assessments (curriculum knowledge) and practical assessments (clinical skills).

Figure 5: Comparison of written EPM quartile and practical EPM quartile (23 schools)

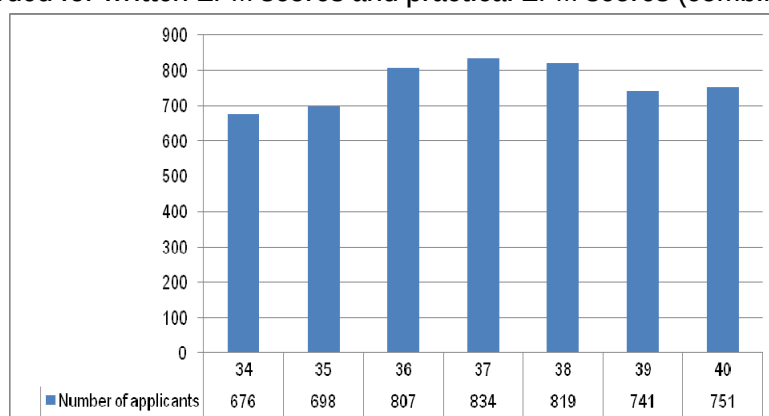
Number of applicants		Practical EPM quartile			
		1 st	2 nd	3 rd	4 th
Written EPM quartile	1 st	751	407	202	89
	2 nd	334	428	344	183
	3 rd	189	338	451	371
	4 th	63	173	327	676

One way of combining two quartiles (assuming a weighting of 50:50, as informed by the two consultations with medical schools) is to award points based on their quartile achieved, as illustrated in Figure 6. The points currently allocated for performance by quartiles is used for illustration. The distribution of students achieving each of the points (34-40) is illustrated in Figure 7, which illustrates additional discrimination at the upper and lower ends of the spectrum.

Figure 6: Points awarded for written EPM scores and practical EPM quartiles (combined by addition)

Points awarded		Practical EPM quartile			
		1 st	2 nd	3 rd	4 th
Written EPM quartile	1 st	40	39	38	37
	2 nd	39	38	37	36
	3 rd	38	37	36	35
	4 th	37	36	35	34

Figure 7: Points awarded for written EPM scores and practical EPM scores (combined)



Achieving 50:50 weighting between written EPM scores and practical EPM scores

Statistical advice regarding the method of combining two scores, and achieving a 50:50 weighting of the two elements, was sought. A greater degree of granularity can be achieved by combining two sets of quartiles, as illustrated in Figures 5, 6 and 7.

In order to achieve a true 50:50 weighting between two measures, dividing applicants into quartiles can only take place after the combination has been achieved. By dividing applicants into two (or more) quartiles before combining

the scores, the actual score (normalised) and position within the banding, is no longer taken into account. In Figure 8, applicant A is ranked 49th /100 (2nd quartile) and applicant B is ranked of 26th /100 (2nd quartile). However, when the quartiles are allocated before combining, both applicants achieve the same score.

Figure 8: Example of two applicants – scores and quartiles

	Written EPM rank/100	Written EPM score	Practical EPM rank/100	Practical EPM quartile	Written EPM score + Practical EPM quartile
Applicant A	49	2 nd	49	2 nd	4 (2+2)
Applicant B	1	1 st	51 st	3 rd	4 (3+1)

Figures 9 and 10 illustrate alternative ways of combining scores based on two quartiles. Neither method takes into account the position within the band, or the underlying score achieved by the applicant, and can thus be rejected on grounds of unfairness. It is interesting to note that multiplying scores for quartile ranks provides greater differentiation between applicants in the 3rd and 4th quartiles. Division of quartiles distorts scores, depending on which EPM score is the denominator, but awards the same number of points for an applicant in the top quartile for both EPM scores as for an applicant in the bottom quartile for both EPM scores.

Figure 9: Points achieved by multiplying the points for quartile rank

Quartile		Practical EPM quartile			
		1 st	2 nd	3 rd	4 th
Written EPM quartile	1 st	1	2	3	4
	2 nd	2	4	6	8
	3 rd	3	6	9	12
	4 th	4	8	12	16

Figure 10: Points achieved by dividing the points for quartile rank

Quartile		Practical EPM quartile			
		1 st	2 nd	3 rd	4 th
Written EPM quartile	1 st	1	0.5	0.33	0.25
	2 nd	2	1	0.67	0.5
	3 rd	3	1.5	1	0.75
	4 th	4	2	1.33	1

Calculations to combine banded scores such as quartiles are less fair as they do not take into account the position within the band. It is not possible to achieve a true 50:50 weighting without taking into account the mark achieved in relation to the pass-mark and/or the cohort, and the difficulty of the exam. Any calculations based on raw scores would need to be normalised to take local factors into account.

Some schools adhere to a university-wide policy that only bandings or gradings are retained for student assessment performance. In this sense, the composite scores for the EPM already diverge from the specified module or assessment weightings. One way to remedy the issue of already graded data, or data with a small range between

scores, is to use a 'basket' of assessments. Whilst different assessments will be measuring different sets of skills, it can be assumed that with more assessments in the 'basket', the more closely converged the average score will be.

6. Changing the scoring system

Quartiles are based on the assumption that graduates from all medical schools are of equal calibre, such that a student in the 1st quartile from one school is of the equivalent standard of a student in the 1st quartile at a different school. Medical school assessments are quality assured by the General Medical Council, and the minimum standard of graduates from any school is assured to be of the same minimum standard.

The report of *Selection to Foundation: An Option Appraisal* deemed that a measure of performance at medical school could only be taken in relation to the cohort from that medical school. Owing to the timing of the application to the Foundation Programme, it is not possible to use Finals. There is no mechanism by which to standardise assessment scores from UK and non-UK medical schools to take into account the number of assessments, pass mark, level of difficulty – or to take into account that performance might be banded by individual assessments. Performance within cohort is the only fair, feasible and defensible measure of performance at medical school.

In the long term, development of assessment items in collaboration with all UK medical schools, through the Medical Schools Council Assessment Alliance (MSC-AA), will address issues of comparability between graduates from different medical schools. By sharing expertise in assessments, and by sharing assessment items themselves, it is anticipated that the collaboration will i) reassure the General Medical Council, NHS employers and the public that medical schools are looking in detail at the equivalency in standards between medical schools, ii) ensure confidence in the quality of medical school graduates, and iii) improve the student learning experience.

Distance from the mean vs distance from the median

One of the concerns about quartile bandings is regarding differentiation at the boundaries. By taking into account the distance from the mean, rather than median, it may be possible to more fairly reflect the performance of individual students in relation to their cohort, with the result that different bandings have a different number of students within. For illustration, the distribution of written EPM scores from Keele Medical School is illustrated in Figure 11, with a mean of 877.28 and a standard deviation of 77.72.

Figure 11: Distance from the mean, Keele written EPM scores

	Boundary score	Number of applicants
Mean -4 St.Dev	566.39	
		1
Mean -3 St.Dev	644.11	
		3
Mean -2 St.Dev	721.83	
		12
Mean -1 St.Dev	799.55	
		36
Mean	877.28	
		36
Mean +1 St.Dev	955.00	
		17
Mean+2 St.Dev	1032.72	
		0
Mean +3 St.Dev	1110.44	

Figure 11 illustrates differences in the number of individuals in a septile band, when taking into account difference from the mean as opposed to distance from the median. There are four applicants from Keele Medical School who scored significantly lower than other students in the cohort, but were allocated into the same quartile as students who were within two standard deviations of the mean. This would suggest that banding by standard deviations from the mean would achieve two advantages over quartiles, in that it would provide greater differentiation at the extremes, and could provide additional granularity – in this instance, six bandings compared with four quartiles. However, assuming a bell-curve of student performance, it would result in more tied ranks around the middle and thus less differentiation. Further, the calculations to devise rankings of performance in cohort would be more complex, with a greater risk of error, and less transparent to the non-expert

In order to compare distance from the mean in terms of standard deviations, the difficulty of the assessments would need to be known. Furthermore, it would need to be known that the distribution of scores followed a normal distribution. A goodness of fit test for a normal distribution, using the Chi-squared distribution, was performed on practical EPM scores from one school - Aberdeen - for illustration. However there was insufficient evidence to either accept or reject that the data (combined number of assessments) follow a normal distribution. Further tests would need to be undertaken to determine whether the total dataset, as well as its composites, follow a normal distribution.

It is extremely difficult to compare distance from the mean between cohorts, unless the data follows a normal distribution, and that scores have been standardised to take into account difficulty of assessment and pass mark. Whilst in one cohort, students may all perform within 1 or 2 standard deviations of the mean, in another there may be much greater variation. Similarly, in one school the mean may not be close to the median, such that a greater number of students may score above than the mean than below it.

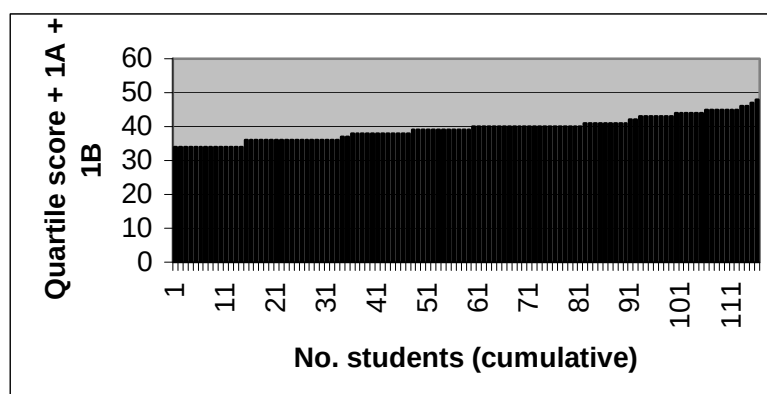
Statistically, distance from the median is a more robust measure, as unlike the mean, the median is unaffected by relatively high or low values. On the basis of this analysis, it is not possible to compare individuals in different cohorts using distance from the mean.

Increasing granularity (deciles, centiles)

One method of increasing granularity achieved by EPM scores is to divide scores into a higher number of bandings than quartiles, for example deciles or centiles.

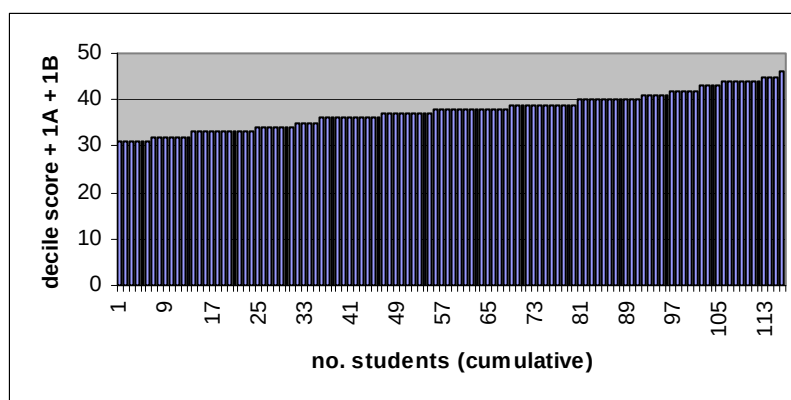
Figures 12 and 13 illustrate the effect of increasing the range of EPM scores, illustrated using data for practical EPM scores from Hull York Medical School. Points for quartiles (34, 36, 38, 40) and deciles (31, 32...39, 40) were combined with points for Question 1A and Question 1B (maximum of 10 points). Applicants under each scoring methodology could achieve a maximum of 50 points (1st quartile plus 10 points for Q1). 64 of 116 applicants scored 0 for both parts of Question 1.

Figure 12: EPM quartiles plus Question 1A and 1B, illustrative example (HYMS practical EPM quartiles)



Score achieved	34	36	37	38	39	40	41	42	43	44	45	46	47	48
Number of applicants	14	19	2	12	12	22	9	2	7	6	7	2	1	1

Figure 13: EPM quartiles plus Question 1A and 1B, illustrative example (HYMS practical EPM quartiles)



Score achieved	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
Number of applicants	6	7	11	7	4	11	9	13	12	11	5	5	4	7	3	1

Increasing the range of possible EPM scores achieved by applicants inevitably increases the range of possible total scores when combined with other performance measures, in this instance, educational achievements as evidenced on Questions 1A+B. It is interesting to note that although the full range for the deciles is 31-50 points, applicants scored between 31-46 points when their decile points score was combined with their score on Question 1 (range of 17). Similarly, when divided into quartiles, applicants could achieve 34-50 points, yet the scores achieved ranged from 34-48 (range of 15). With both scoring systems, there continue to be a number of applicants achieving tied scores across the range of scores.

With the example illustrated in Figures 12 and 13, the distribution of scores with quartiles produced a higher number of applicants with tied scores towards the lower end and middle of the scale (median 38 points), and a higher number of applicants with tied scores (e.g. 19, 22). In contrast, the distribution of scores with deciles produced a greater spread of applicants with different scores, with the most tied scores clustered in the middle of the scale (36-40 points, median 39 points) and fewer applicants with tied scores (11, 12, 13).

Earlier analysis (Figure 3) indicated the small positive correlation between performance on EPM scores and performance on Question 1A and Question 1B. With both scoring systems, there is greater granularity at the top end of the scale, supporting the assumption that the scoring scheme identifies and rewards the highest achieving applicants.

In summary, increasing the number of bandings, such as from quartiles to deciles, can help to differentiate applicants by i) providing a greater spread of scores and ii) reducing the number of applicants with tied scores, without greatly affecting the values of scores achieved (mean, median, mode). Issues of feasibility and desirability need to be addressed.

Increasing granularity (changing the difference in points awarded to quartiles)

A second method of increasing granularity achieved by EPM scores is to allocate a greater difference in points for existing quartile – or other – bandings.

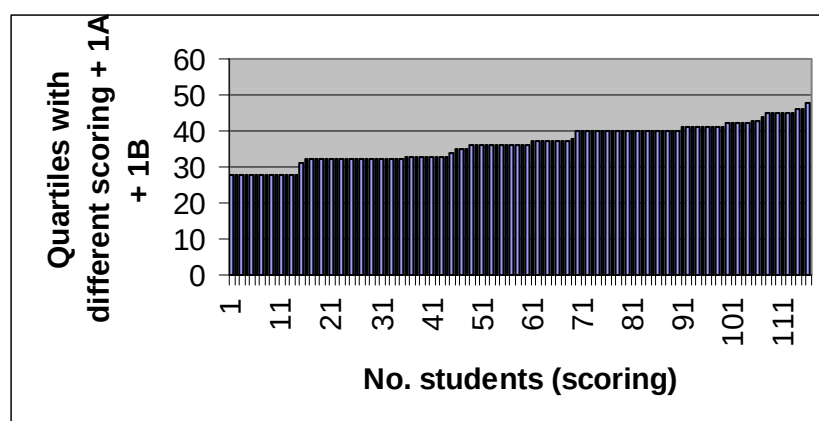
Figure 14 illustrates the effect of increasing the range of points available to quartiles, illustrated using data for practical EPM scores from Hull York Medical School. Points for quartiles (28, 32, 36, 40) were combined with points for Question 1A and Question 1B (maximum of 10 points).

It is interesting to note that although the full range for the deciles is 28-50 points, applicants scored between 28-48 points when their quartile points score was combined with their score on Question 1 (range of 21). A range of 17 possible scores were achieved by applicants within this range (not 29, 30, 39, 47).

Within this scoring system, there continues to be a number of applicants achieving tied scores across the range of scores, widely mirroring the range of scores illustrated in Figure 12 (combining quartiles plus Question 1A and 1B), with up to 20 applicants with the same tied score. The median of 36 reflects the lower range of points available to applicants.

In summary, increasing the difference in points awarded to bandings, such as quartiles, can help to differentiate applicants by i) providing a greater spread of scores and ii) reducing the number of applicants with tied scores. However the additional level of granularity is comparatively less than that achieved by increasing the number of possible bandings. The desirability of changing the range of points allocated to quartiles need to be addressed.

Figure 14: Quartiles (alternative scoring) plus Question 1A and 1B, illustrative example (HYMS practical EPM quartiles)



Score achieved	28	31	32	33	34	35	36	37	38	40	41	42	43	44	45	46	48
Number of applicants	14	1	20	9	1	3	12	8	1	21	9	5	2	1	6	2	1

7. Other findings

The pilot generated a considerable amount of helpful feedback from medical schools. The feedback, summarised below, is set out in greater detail in Appendix B.

Practical versus written assessments. It is not possible for all schools to provide separate scores or rankings for practical and written work, as some schools have assessments that combine both competencies.

Timing of course delivery and assessments. There is considerable variation in the timing of courses and assessments. Different cohorts – often within the same school – may be at medical school for four, five or six years. It is challenging for schools to standardise the assessment data for students who have not taken the same assessments, or the same combination of assessments. Given this, the EPM framework should not try to specify that certain assessments should be taken from certain years to produce an EPM.

Grading and banding. There is considerable variation in the way in which the outcomes of assessments are quantified and recorded by universities. Some universities have policies that require schools to keep raw scores from assessments; others require schools to retain only broad gradings. Some assessments yield very granular results; at the other extreme some assessments result only in a pass or a fail. There are also more complicated scoring schemes including merit and demerit points, negative marking and where the number of passes or fails of individual questions determine an overall pass or fail. The EPM framework needs to cater for this degree of variation.

Transfers, Student Selected Components etc. There is a particular difficulty in finding a satisfactory way to deal with students who transfer between schools, as such students will not have undertaken the same mix of assessments as the others in the school to which they have transferred. Schools currently take one of two approaches: a complex process of standardisation, or developing quartiles according to assessments that all students have taken in common. There is a similar problem in dealing with other cases in which not all the students at a school undertake the same assessments, for example where the students have some freedom to select modules of work.

Re-sits. There needs to be an agreed standard way of allocating marks in relation to re-sits, taking account of any extenuating circumstances, and the varying opportunities to re-sit failed finals in the same academic year.

Authorised absence. There needs to be an agreement about how to treat cases in which a student is authorised to miss an assessment that is taken by the rest of the cohort.

Comparability across schools. The basis of the EPM (and the original quartile score) is an assumption that the cohorts are broadly comparable across medical schools. Some schools have expressed concerns that this assumption may not be valid.

8. The way forward

The EPM pilot has demonstrated the importance – and the benefits – of a nationally agreed framework for EPM scores in terms of transparency and consistency. The 25 medical schools which participated fully in the EPM pilot were able to return raw data relating for both written EPM scores and practical EPM scores. Feedback from the remaining five schools indicated that they too would be able to provide data on curriculum knowledge and clinical skills, although they were unable to adhere to the framework within the timeframe.

Whilst many, if not all, schools consult on and publish the composition of quartiles, it is evident that there is wide variation between medical schools in the way they produce quartiles. The pilot revealed the different compositions of current measures of performance at medical school – which include assessments in all years in some schools, and a single year in others. It is not practical to specify the exact composition of the EPM in terms of specific assessments, as the different approaches to assessment and learning should be encouraged. Nationally agreed EPM standards would be beneficial in avoiding unnecessary variation across schools, for example by including a requirement for a standard way of allocating scores to students who re-sit assessments. The movement of students between medical school quartiles and EPM scores further highlights the importance of a defensible and consistent approach to developing EPM scores.

The EPM pilot, and the issues raised by schools in adhering to the EPM framework, highlighted that a framework which specifies X% of assessments from one year and Y% from another is inappropriate, since it does not recognise the variation across schools in the timing of courses and assessments. Similarly a framework that calls for a clear split between assessments of curriculum knowledge and clinical skills is not feasible, as one competency underpins the other, and assessments frequently assess both competencies.

In principle there is scope to make the EPM more granular than quartiles, provided that the underlying data enable sufficient differentiation between students. This granularity is separate to the weighting that might be used for the EPM with the SJT. It may be prudent for there to be a requirement for a minimum number of assessments to be taken into account, to ensure that the desired granularity can be reached.

Averaging a student's performance over a 'basket' of assessments should be more representative of the student's capability than taking any one of the assessments individually. The pilot has shown that the rules for constructing such a basket cannot be specific about the nature or timing of the assessments to be taken into account, or about the split between written and practical assessments, as these factors vary by school. The EPM framework needs to be defined in a sufficiently generic way to take this variation into account.

Given the above, a pragmatic approach would be as follows:

Each medical school should calculate an overall average score for each of its students, based on a basket of assessments of clinical skills and curriculum knowledge. Each school will be responsible for constructing the most appropriate basket of assessments for the purpose of providing the EPM, taking into account local factors. The basket should not be artificially skewed. For transparency, all schools should develop the composition of EPM scores in consultation with its students, and would be required to publish the details of the basket to its students.

The students in a cohort should be competition-ranked according to their score, then given an overall EPM at an agreed level of granularity (e.g. quartiles, deciles, etc).

To ensure that EPM scores are determined in a consistent and fair way across schools, there must be standards to address:

- General principles to be applied in calculating the weighted averages of assessments
- A minimum number of assessments
- Re-sits (either of single assessments or of an entire year)
- Missing assessment scores (e.g. where an applicant has been on sick leave during an assessment)
- Treatment of pass/fail assessments and other forms of banding
- Transfer students
- Cases in which different students within a cohort take different assessments
- Cases in which the strict boundary of a class (e.g. quartile) falls among a number of equally-ranked students
- Whether and how additional points may be awarded for other achievements (e.g. prizes, publications, presentations and extra degrees)

A draft of these standards should be worked-up by an EPM Task & Finish Group, reporting to the ISFP Project Group, through liaison with those schools who are particularly affected by the issues. The revised EPM standards should then be issued for consideration by all UK medical schools and the ISFP Project Group. Legal opinion will be sought.

Appendix A: Draft Framework for Educational Performance Measurement (EPM)

Component 1: Written assessments – 50% of EPM

This component uses the results of summative written assessments throughout the student's undergraduate career, with the assessments in the later years more heavily weighted. Please determine the total mark for each student using the weighting below.

- Early years (40% of Component 1)
 - o Year 1 of 4 year graduate entry courses
 - o Years 1 and 2 of 5 year courses and
 - o Years 1-3 of 6 year courses*
- Later years (60% of Component 1)
 - o Years 2 and 3 of 4 year graduate entry courses
 - o Years 3 and 4 of 5 year courses and
 - o Years 4 and 5 of 6 year courses*

Component 2: Practical assessments (e.g. OSCEs) - 50% of EPM

This component uses the results of summative practical assessments, such as OSCEs. Please determine the total mark for each student using the weighting below.

- o Years 2 and 3 of 4 year graduate entry courses
- o Years 3 and 4 of 5 year courses and
- o Years 4 and 5 of 6 year courses*

1. Data will be drawn from summative rather than formative assessments.
2. Schools will need to devise their own weightings for modules and assessments – which may include summative SSCs - within the three components above. These weightings will be made transparent.
3. Where re-sits have been taken, the marks awarded will be capped at the pass mark. An exception to this will be in the case of extenuating circumstances, as approved by an examinations board or committee, for which the marks from the repeat assessment will then be used.
4. Schools whose students move after 3 years will provide the receiving School with the necessary data to allow early years' results to be included in the calculation.
5. The calculations will be made using actual marks (standard-set) awarded for modules and not the grade points, to keep as much discrimination as possible.

*Six year courses including a Foundation or pre-clinical year should count the foundation year as Year 0, with the following five years as Years 1-5.

Please note that the length of the course excludes time out for intercalation.

Appendix B: Issues raised by medical schools in adhering to the EPM framework

Difficulties in accessing data

- Difficult to separate scores for written and practical competencies, as assessments usually include both and it is difficult to assess the appropriate weightings to be applied retrospectively (Barts, Birmingham, Cambridge, Edinburgh)
- Data are held by another school within the university (Cardiff – years 1&2 are managed by the School of Biosciences)
- Data for early years assessments is not held for transfer students (Aberdeen, Barts, Cardiff, Edinburgh, Imperial, KCL, Manchester, Oxford, Imperial, UCL)
- Data is not stored in a format that enables the framework to be followed easily – many schools reported that it was a manual process, for example UCL referred to 36 different spreadsheets plus more to track students who didn't progress normally, and checking individual cases of extenuating circumstances.

Granularity

- University-wide policy to record assessment bandings or gradings, not raw scores (Aberdeen, Leeds, Liverpool, Manchester, UEA, Warwick)
- Some assessments are pass/fail (Liverpool, Southampton, UEA)

Ranking within a cohort

- Students from different intakes (i.e. graduate entry, five year and six year courses) ranked separately (Imperial, St George's)
- Students from different intakes ranked together (KCL, Southampton)

Normalisation

- Students from the same intake do not take the same combination of assessments (Manchester)
- Students from different intakes (four, five and six year courses, plus direct entrants to Year 2 in Aberdeen, to Year 3 for BDS graduates in KCL) do not take the same combination of assessments (Aberdeen, Birmingham, HYMS, Imperial, KCL, Manchester, Southampton, UCL). This is currently addressed by two methods – using only data common to all intakes (KCL – 3 years; Manchester – 2 years; Southampton – 1 year), or using a complex process of standardisation for the pass mark, level of difficulty and scale (e.g. Aberdeen)

Transfer students

- Transfers of whole cohorts (St Andrews to Manchester, Swansea to Cardiff – both at year 3)
- Transfers of small groups of students – from Oxbridge, from the International Malaysia University (Aberdeen, Barts, Cardiff, Edinburgh, Imperial, KCL, Manchester, Oxford, Imperial, UCL)

Composition of EPM scores

- The framework limits the number of assessments to be included
 - o Clinical skills assessments in early years are excluded (BSMS, Brighton, Birmingham, Cardiff, Edinburgh, Oxford, Warwick, Sheffield)
 - o Assessments with pass/fail are excluded from the framework if resit marks are capped at the passmark (BSMS usually includes 49 assessments within quartiles, but the EPM framework is limited to 14)
 - o Cambridge places a greater emphasis on practical assessments – only two assessments are purely 'written'
- The framework requires additional assessments to be included

- Interpretation of definitions of 'practical' and 'written' – i.e. mode of assessment vs competencies being assessed
 - o Cardiff - assessments termed 'skills and competencies' include academic skills panel, group work, critical reading, essay under exam conditions
 - o St George's – includes Student Select Components
 - o Edinburgh – includes Professionalism/ PPD measures, psychiatry viva, interpretations of x-rays assessed online, mini-CEX
 - o Sheffield – moving towards greater emphasis on workplace-based assessments, only one OSCE (year 3)
 - o UEA – includes presentations

Resits

- Use of actual first mark (Aberdeen, Birmingham, Liverpool, Manchester, Sheffield, St George's, UCL) vs capped at pass mark (Imperial, Sheffield) vs compensated fail (Edinburgh, St George's, UCL)
- Verifying extenuating circumstances in retrospect is a time-consuming and manual process
- Capping resits at the passmark leads to clustering at the lower end of student performance

Other issues for agreement

- Is the divide between written and practical artificial? Definition of 'written' and 'practical' – in terms of competencies?
- Enable flexibility for Schools to include other assessments (e.g. SSMS, professionalism measures, portfolios) etc – but in a way that does not contract the goal of standardisation across schools
- Timeframe to introduce framework - not fair to impose a framework retrospectively – if early years data are to count, this must be openly agreed from the start of the course
- What of authorised absence and no mark returned? In Manchester, the mean mark for the module is used.
- Inclusion and weighting of points for presentations, publications, prizes or previous degrees

Additional comments

- Aberdeen - Putting in a pass mark of 9 for any students who resit a course means that the bottom quartile will be more clustered. This was something that our students felt strongly about; the students who have been at the lower end of the class but who have never failed an exam are then clustered in with students who have re-sat, sometimes across lots of exams, and they felt that this was not fair.
- Barts - Weighting of pre clinical and clinical years. As we do not differentiate significantly in terms of content this is a fairly artificial division for us and no doubt some other schools. Early and late years are a better distinction, but then the weighting is for the value of the contribution to the finishes product rather than the proportion of time spent in a clinical setting.
- Edinburgh - Having been through the process of allocating marks to either 'Written' or 'Practical' and faced the difficulties we have highlighted, we wonder if the effort of splitting marks in this way is worthwhile, especially in the absence of an underpinning principle or construct for the division. I understand UKCAT may also be moving to considering 'clinical' versus 'academic' abilities, another difficult distinction to make in practice since academic knowledge underpins most clinical activity and practical clinical activities such as interpreting data and x-rays, and even prescribing might be assessed by way of an online 'written' exam.

Perhaps competencies, such as history taking, information giving, and clinical examination, demonstrated through interactions with patients, are the important ones to distinguish, since in practice there appears to be a handful of students who, though able in other ways, struggle to master these competencies. However these

are most robustly tested in formal summative tests such as OSCEs to overcome the leniency of workplace-based assessments and this would result in very few assessments that could be collated for such a mark.

- Leicester – the (ranking) order will be very different in the middle, due to the different weightings, although the students at the top and bottom should be the same. Pass/fail is a set standard– students have to pass minimum number of questions, not a minimum percentage – not a compensatory system.
- Oxford - I am afraid that I continue to regard this exercise as a red herring. Increasing the 'granularity' used for compiling scores to discriminate between students at individual schools is not going to improve the fairness of the process one jot -the problem continues to be the underlying assumption that EPM scores are comparable between medical schools, which the evidence suggests they are not. I have no problem with following a common methodology, but I do hope people will stop pretending that doing so is going to make selection fairer. The assumption that quartiles are equal between medical schools is of course made in the existing rankings, but the impact is low because so few points are awarded for the quartile rankings.

If the EPM were used to increase the fraction of the ranking score allocated for medical school performance, we believe that our lower scoring students would be unfairly discriminated against, and consequently could not agree to it.

- Southampton – to ensure transparency, students are provided with their grades which they can then verify (based on single assessment). If the framework required complex normalisation this transparency would not be possible.
- St George's - We are in the midst of a change of curriculum and even aside from that, we had a big change to the structure of our pre clinical and clinical exams. So for the students on the list, anyone who did not go straight through the system and took a year out or intercalated etc will have done a different style of exam than other students.

There is also a problem when students don't sit the exams with the year they are being ranked with i.e. a student who failed a Year 3 exam in 06-07 gets the pass mark for that exam for that year, but they are ranked with other students who failed a Year 3 exam a year later in 07-08. This student would get the pass mark for that exam, but if the pass marks are different each year (as they are every year), this could skew the ranking. The only way to get round this would be to % each exam and pass mark and make everything out of 100. This was too much work for the purposes of this exercise and would have taken me weeks and weeks to work out individual pass mark % for each student across all the exams of all the years, especially breaking up a combined exam into OSCE and written components.

We have some exams that you can't pass or fail, and they just add into another exam which you can pass or fail overall. It is not formative per se as the overall exam, which includes the results from this previous exam as above, is summative. Therefore if someone gets below the pass mark for the first part, should I give them the pass mark or their actual mark? I gave them the pass mark, but they haven't actually failed this exam and do not resit it.

- Warwick - We do not use overall percentages in written examinations as a way of determining who has passed. Each question has a separate threshold that the student has to pass to be given a single point. These points are then added up and students have to reach a required level of points to pass. Points are also used to allocate grades. We therefore worked out our percentage for the written components based on

how many grade points out of the total available for example 36 points available 18 points achieved equalled 50%.

For our clinical examinations students achieve possible marks for certain grades whilst other are considered to be 'demerit' points. For these examinations we looked at whether students had been required to sit further tests or not and how they had done overall. 100% meant no demerit points, 75% demerit accrued but no further exam, 50% demerit points, further exam and passed, 0% demerit points and failed.

The results have moved students within quartiles and in one case a student has moved ranking from 131st to 27th! We are not happy with this as this student has now sat Finals and has been required to re-sit in November 2010. The additional information that we used in our original calculations and the instruction to put all re-sit candidates in the 4th quartile meant that this student was placed, in our opinion, in the correct quartile using the whole body of knowledge.

- UCL - The data requirements included many complexities that we considered and rejected at a series of working parties with representatives from Oxford, Cambridge and the other London Medical Schools when quartile rankings were first introduced and when we were first determining how to calculate quartiles for UCL students.

We have a complication at UCL relating to IBSc's, which is that we currently have a compulsory IBSc for non-graduates which may be taken between years 3 and 4 or between years 4 and 5. It makes compiling the marks for UCL students complicated because we have to draw on exams taken in a number of different sessions by different cohorts and with different exams and different standard set pass marks. For the EPM pilot, because of the different cohorts and the structure of our summative exams, we had to work with 36 spreadsheets to compile data for students who had progressed normally, and more to track students who did not progress normally. For MBPhD students and others with long period of interruption we had to track results back to spreadsheets dating over 8-9 years. Each year had different standard set pass marks with, sometimes, considerable variation.

The compilation of exam results dating back, including MBPhDs, for 8-9 years and matching it to the EPM cohort took perhaps a day of admin time. That's feasible and represents about twice the amount of time currently needed to compile our rankings based on a customised export of year 3 and year 4 exam results. However, in our current system for calculating rankings we sought deliberately to minimise manual manipulation of data because of the risk of error, and this is my main concern with the complexity of the pilot.

Tracking extenuating circumstances posed the biggest problem and we also have concerns about the fairness of using extenuating circumstances because at UCL we apply a tariff depending on the severity of the extenuating circumstances and the likely impact on student performance. For candidates graded 1 for extenuating circumstances, the mitigation is very weak and we wouldn't feel comfortable allowing an advantage to these candidates by using a re-sit mark achieved after an additional year's study. Also, 5 years ago information about extenuating circumstances and tariffs awarded wasn't recorded reliably in our exam mark sheets and so we had to go back to different spreadsheets and to student files and paper records to work this out. It's very unlikely that this type of information will ever be stored in the electronic student database against exam results and so there will always be a risk that some candidates with extenuating circumstances will be missed when tracking cohorts of 380-400 students over 8-9 years of a programme, and that some candidates with weak mitigation will be advantaged if their re-sit score is used.

Data manipulation increases the risk of error, and, even with easier compilation of exam marks, the amount of data manipulation in the pilot for re-sits and capping and extenuating circumstances is considerable. Even

with a proper database it would be impossible to put systems in place retrospectively to track capped/original marks and extenuating circumstances. In years 3 and 4, our compensation rules for calculating the final result mean that capping the written and clinical marks separately at the standard set pass mark for each components as requested for the EPM meant, in some cases, raising a re-sit mark for one component from below the pass mark and awarding a higher mark than was achieved at the re-sit attempt.

In summary, I think our main concern is that the amount of manual manipulation of data and the amount of cross-checking against different mark sheets with different standard examinations and different pass marks means that we could not be confident in using the data provided for the EPM pilot to calculate the real quartile rankings to be used for our students, and a simpler method would be preferable as it would reduce the risk of error and produce more reliable data for rankings.

Appendix C: Caveats for interpretation of EPM scores

Aberdeen

- Scores based on points awarded by band, not raw scores
- First attempt marks used, not capped at the pass mark

Barts and The London

- Transfer students – assumption that the mark achieved in years at that medical school used for early years marks

Birmingham

- First attempt marks used, not capped at the pass mark

Cambridge

- The 'early years' part of component 1 includes both written and practical – it is not possible to split the marks retrospectively

Edinburgh

- The 'early years' part of component 1 includes both written and practical – it is not possible to split the marks retrospectively
- Practical scores include Year 2 OSCE
- First attempt marks used, not capped at the pass mark
- Transfer students – assumption that the mark achieved in years at that medical school used for early years marks

Leeds

- Scores based on points awarded by band, not raw scores.
- Transfer students – assumption that the mark achieved in years at that medical school used for early years marks
- First attempt marks used, not capped at the pass mark
- Practical assessments – based on Year 3 OSCE only

Liverpool

- Scores based on points awarded by band, not raw scores
- First attempt marks used, not capped at the pass mark

Manchester

- Scores based on points awarded by band, not raw scores.
- First attempt marks used, not capped at the pass mark
- Only data from Years 3 and 4 is used

Oxford

- Clinical exams are scored 1-4 (can be decimal places) and converted to percentages.
- First attempt marks used, not capped at the pass mark
- Scores for preclinical years calculated using a merit system and not raw data (3 points for distinction, 1 point for merit, failed paper cancels out points for merits)

- Scores for preclinical years includes marks from Yrs 1&2 for the graduate entry cohort, as science based written exams are in both years not just year 1

Sheffield

- First attempt marks used, not capped at the pass mark

Southampton

- Practical assessments – based on Year 3 OSCE only
- Written assessments – based on two Year 3 assessments only

St George's, University of London

- Transfer students – assumption that the mark achieved in years at that medical school used for early years marks

UCL

- Transfer students – assumption that the mark achieved in years at that medical school used for early years marks

University of East Anglia

- Some modules are pass/fail (50/0 points) – however as the resit mark was capped at the pass mark, these modules were then excluded from the EPM pilot

Warwick

- Points are allocated to grades (up to 36 points)
- Clinical examinations are awarded 0%, 50%, 75%, 100% based on demerit points
- Only two clinical assessments could be used, as the framework did not allow for Year 1

Improving Selection to the Foundation Programme

Appendix I

Report of the EPM Task and Finish Group

Final report of the Educational Performance Measurement Task and Finish Group

April 2011

Contents

1. Introduction.....	420
Structure of the document.....	421
2. Background.....	421
3. Quality criteria	423
4. Issues and Recommendations	424
The number and type of assessments to be taken into account	424
Mathematical principles	425
Re-sits.....	426
Missing scores	427
Definition of a cohort.....	427
Transferring groups of applicants.....	427
Other special cases	428
Other achievements.....	428
Granularity	429
Ranking procedure.....	430
5. Final recommendations	431
6. Document History.....	434

1. Introduction

1.1 This is the report of the Educational Performance Measurement Task and Finish Group, hereafter referred to as the EPM Group. This report reflects the recommendations and agreements reached at the meeting of the EPM Group on 10 November 2010, and through subsequent work led by members of the EPM Group. The report is written for the ISFP Project Group and ISFP Programme Board for their consideration.

1.2 The membership of the EPM Group represents expertise in medical school assessments, business analysis and mathematical modelling as well as key stakeholder interests, with continuity from previous work of the ISFP project. The membership is as follows:

Professor Mike Greaves (Chair)	Head of Medical School, University of Aberdeen
Dr Frazer Anderson	Chair of BM Finals Working Group, University of Southampton
Dr Alison Carr	Dean Adviser, Department of Health
Dr Ian Curran	Associate Dean for Postgraduate Medicine, London Deanery
Dr Anne Edwards	Foundation School Director, Oxford Deanery
Ashley Fraser	Medical Director, NHS Employers
Dr Christine Kay	Medical School Curriculum Manager, University of Aberdeen
Roddy McDermid	Medical Students Society President, Sheffield
Dr Katie Petty-Saphon	Executive Director, Medical Schools Council
Karin Purshouse	Chair, BMA Medical Students Committee
Dr Susan F Smith	Deputy Director of Education, Imperial College Faculty of Medicine
Professor Anthony P Weetman	Chair, Medical Schools Council

Support for the EPM Group was provided by Siobhan Fitzpatrick and Denis Shaughnessy, ISFP Project Team.

1.3 The EPM Group was tasked with defining and recommending specific and workable standards for the EPM framework, in consultation with medical schools and stakeholders as required, and building upon the assumptions made in relation to the EPM, namely that:

- 1.3.1 The EPM should be a means of ranking applicants to an agreed level of granularity, by averaging their performance over a representative basket of assessments taken at medical school up to the point of application to the Foundation Programme.
- 1.3.2 Individual schools should be responsible for determining which combination of assessments is most representative for their applicants, and for agreeing this in consultation with their applicants.
- 1.3.3 The composition of the baskets of assessments should be transparent - each school should publish an unambiguous description of how its basket of assessments is formed.
- 1.3.4 There would continue to be a spread of 30-100 points available for selection to the Foundation Programme.
- 1.3.5 The EPM should satisfy the agreed quality criteria.

Structure of the document

- 1.4 This sets out the rationale and requirements of the EPM framework, the quality criteria, and the issues that the EPM Group was tasked to address. Section 5 summarises the recommendations for the EPM framework.

2. Background

- 2.1 Applicants to the Foundation Programme currently receive a score based on performance at medical school in relation to their cohort, as ranked into four quartiles worth 40, 38, 36 and 34 points. This score is then combined with their score from an online application form, white space questions (total of 60 points), which includes 10 points for degrees, prizes, publications and presentations.
- 2.2 Concerns about the use of academic quartiles, raised before and during an extensive and detailed Option Appraisal¹, include the comparability of applicants from different medical schools and discrimination of applicants at the margins between quartiles. Given issues around the lack of transparency and lack of consistency across medical schools, one of the main drivers to review the quartile system is to ensure defensibility in the event of legal challenge. Stakeholder feedback shows strong support for the use of some measure of academic performance as well as non-academic and possibly extra-curricular activities. An advisory international Expert Panel supported the principle of making greater use of information accumulated during medical school, and the development of a standardised measure of educational performance.
- 2.3 The recommendations of the ISFP Steering Group were to pilot and evaluate:
- 2.3.1 A Situational Judgement Test (SJT) to assess professional attributes, judgement and employability.
- In combination with*
- 2.3.2 An Educational Performance Measurement (EPM) of applicants at their medical school to assess clinical knowledge/ skills as well as wider personal achievement.
- 2.4 The EPM as a selection tool refers to a differential ranking score produced by the medical school to reflect the applicant's achievements or performance on a range of assessments in relation to their cohort up to the point of application to the Foundation Programme. It was envisaged that the EPM would be derived using a standardised and transparent framework of existing performance measures, with a weighting agreed with medical schools in consultation with applicants and other stakeholders. All UK and non-UK medical schools would be required to provide an EPM score to the UKFPO derived using the standardised framework.
- 2.5 The ISFP Project Group envisaged that a uniform transparent framework for an EPM would address some of the current concerns about comparability between applicants of the same quartile from

¹ Medical Schools Council (2009) *Selection to Foundation: An Option Appraisal*

different schools, and that it would enable greater granularity². Depending on the results of the 2010-11 pilots, it is anticipated that the EPM would be used in combination with a second selection tool, resulting in a combined higher level of granularity than could be achieved by the EPM alone.

- 2.6 Two in-depth consultations in autumn and winter 2009 involving all 30 UK medical schools whose applicants apply to the Foundation Programme³ informed the development of the draft EPM framework for piloting. The first consultation, in September and October 2009, gathered evidence around the information currently used to inform quartile rankings, and the assessment information on applicant performance currently collected and utilised by UK medical schools. The second consultation, in November and December 2009, consulted around the principles of a possible EPM framework, the weighting between different elements and how additional points for prizes, publications, presentations and previous degrees might be awarded.
- 2.7 A working draft EPM framework for piloting, informed by the feedback from the two consultations, was agreed by the ISFP Project Group and the Medical Schools Council in early summer 2010. The draft EPM framework called for applicants to be given two scores - one relating to written assessments (curriculum knowledge), and the other relating to practical assessments (clinical skills) – according to a specific prescription of weightings between earlier and later years of the course. The two scores would then be combined to provide an overall score, with additional granularity to the current quartile ranks.
- 2.8 The EPM pilot, and the issues raised by schools during the pilot, highlighted that a framework which specifies the proportion of assessments from different stages in medical school is inappropriate, since it does not recognise the variation across schools in the timing of courses and assessments. Further, a framework that calls for a clear split between assessments of curriculum knowledge and clinical skills is not feasible, as one competency underpins the other.
- 2.9 The EPM pilot demonstrated the importance – and the benefits – of a nationally agreed framework for EPMs in terms of transparency and consistency. The 25 medical schools which participated fully in the EPM pilot were able to return raw data relating separately to written assessments and practical assessments. Feedback from the remaining five schools indicated that they too would be able to provide data on curriculum knowledge and clinical skills, although they were unable to adhere to the framework within the timeframe.
- 2.10 Whilst many schools consult on and publish the composition of quartiles, there is wide variation between medical schools in the way they produce quartiles. The pilot revealed the different compositions of current measures of performance at medical school – which include assessments in all years in some schools to a single assessment in a single year in others. It is not practical to specify

² Granularity is taken to mean the differentiation between applicants (e.g. the number of different scores achieved)

³ At the University of St Andrews, applicants transfer to the University of Manchester or to other Scottish medical schools at the end of Year 3, ahead of their application to the Foundation Programme.

the exact composition of the EPM, and the different approaches to assessment and learning should be encouraged.

- 2.11 In principle, there is scope to make the EPM more granular than quartiles, provided that the underlying data enable sufficient differentiation between applicants.
- 2.12 Averaging an applicant's performance over a 'basket' of assessments should be more representative of the applicant's capability than taking any one of the assessments individually. The pilot has shown that the rules for constructing such a basket cannot be specific about the nature or timing of the assessments to be taken into account, nor about the split between written and practical assessments, as these factors vary by school. The EPM framework needs to be defined in a sufficiently generic way to take this variation into account.

3. Quality criteria

- 3.1 The recommendations of the EPM Group will be subject to:
 - 3.1.1 Legal opinion, which is to be sought on behalf of the ISFP Project Group by the Medical Schools Council.
 - 3.1.2 Quality review by the ISFP Project Group.
 - 3.1.3 Approval by the ISFP Programme Board.
- 3.2 The Medical Schools Council will also need to seek formal confirmation from its membership that the recommendations will be supported by all UK medical schools.
- 3.3 The following quality criteria are measures against which to evaluate the EPM:
 - 3.3.1 The EPM is a reliable and representative measure of the applicant's performance in their educational progression at medical school up to the point of application.
 - 3.3.2 The EPM is a valid measure in relation to selection to the Foundation Programme (e.g. it is measuring factors that have a bearing on suitability for the job).
 - 3.3.3 The EPM is a sufficiently granular measure.
 - 3.3.4 The EPM is a fair measure.
 - 3.3.5 The EPM is not overly expensive to administer and quality assure.
 - 3.3.6 Medical schools are able to adhere to the framework, and introduce it within suitable timescales.
 - 3.3.7 Where medical schools are not able to produce a reliable EPM, there is a legal and justifiable fall-back.
 - 3.3.8 There is no requirement on any medical school (UK or non-UK) to modify its curriculum in order to comply with the framework.
 - 3.3.9 The framework complies with all relevant legislation.

4. Issues and Recommendations

- 4.1 The EPM Group was asked to make recommendations in relation to the following topics identified through the initial EPM pilots:
 - 4.1.1 The type and minimum number of assessments to be taken into account in constructing EPMs.
 - 4.1.2 Mathematical principles to be applied in calculating an overall score from a basket of assessments, taking into account the need to deal with pass/fail assessments and other forms of banded marking schemes.
 - 4.1.3 The scores to be used in the case of re-sits.
 - 4.1.4 The treatment of missing assessment scores (e.g. where an applicant has been on sick leave during an assessment).
 - 4.1.5 The definition of the cohorts within which rankings are to be determined.
 - 4.1.6 The treatment of transferring applicants.
 - 4.1.7 Other special cases in which different applicants within a cohort take different assessments.
 - 4.1.8 How other achievements (e.g. prizes, publications, presentations and extra degrees) should be taken into account.
 - 4.1.9 The degree of granularity to be achieved by the EPM.
 - 4.1.10 The mathematical ranking procedure to be adopted (expected to be competition-ranking) and the rules for dealing with equally-scoring applicants around the boundaries between the granular divisions.
 - 4.1.11 Any other aspects that the EPM Group considers should be standardised in order to produce an effective EPM.
- 4.2 This section of the report considers each of these topics in turn, explaining the significance of the topic and its associated issues, describing any options considered for resolving the issues, and presenting the EPM Group's recommendations and associated rationale.

The number and type of assessments to be taken into account

- 4.3 The EPM is intended to summarise the performance of the applicant over a wide range of assessments undertaken at medical school. In principle this approach should provide a more representative picture than could be achieved by a single assessment. In practice, however, the range of assessments needs to be considered, for example what sort of assessments should be taken into account and how many assessments need to be included.
- 4.4 The EPM pilot has underlined the variety of assessments at medical schools. Although all UK medical degree courses teach to the GMC's curriculum framework *Tomorrow's Doctors*, the timing and format of assessments varies considerably between medical schools. For example, some schools will hold separate assessments for written and practical work, while other schools hold assessments that

combine both elements. Similarly, some schools spread their summative assessments across the years of a course, while others hold most of their assessments in one or two years of a course.

- 4.5 Given the above, it is neither desirable nor possible to define a standard of the sort that stipulates the types of assessment taken in particular years of the degree course. Instead, the standards for the production of EPMS need to allow for the variety of ways in which assessments are conducted, and must reflect the effects to be achieved rather than the means by which they are achieved.
- 4.6 Given this, the EPM Group recommends that the EPM standards should require each school to be responsible for constructing a representative basket of assessments, subject to the following general requirements that the selected assessments should:
 - 4.6.1 Be summative (and hence subject to formal controls).
 - 4.6.2 Cover clinical and non-clinical performance.
 - 4.6.3 Cover clinical knowledge and clinical skills.
 - 4.6.4 Cover all aspects of the curriculum assessed up to the end of the penultimate year at medical school.
 - 4.6.5 Represent the average performance of the applicants over time, rather than being limited to a snap-shot.
 - 4.6.6 Include written and practical forms of assessment.
- 4.7 The school should also be required to publish, in advance of the related recruitment round, the details of the basket of assessments and weightings to be used to create EPMS for that round. This will mean that the approach of a school will be subject to scrutiny by its applicants, by other schools, and by those responsible for recruitment to the Foundation Programme.
- 4.8 The EPM Group specifically considered whether the standard should specify a lower limit on the number of assessments that must be taken into account in forming an acceptable EPM. The pilot has shown that some schools might implement a large number of relatively short assessments, while others might implement a smaller number of lengthier tests. Given this, the number of assessments required to form a representative picture of an applicant's performance will vary from school to school, so there is no way to specify an appropriate minimum number. Instead the EPM Group recommends that each school will be expected to use the results of all of the summative assessments taken by the cohort⁴, subject to the general requirements listed earlier.

Mathematical principles

- 4.9 The EPM pilot has shown that there is a wide variation in the format of the results of assessments held by medical schools. For example, some assessments result in numerical raw scores on a wide scale, say from 0 to 100 or more, while others result in a small number of grades (A to E, say), or even

⁴ 'Cohort' broadly means all of the applicants in a given year of a degree course at a medical school. However, the term has a more precise meaning in this context, which is defined later in this report.

just a binary pass/fail outcome. The assessments can also vary in difficulty, with some resulting in a wider spread of attainment than others. The EPM Group considered the principles upon which such a range of results should be combined to provide an overall measure for each applicant in a cohort.

- 4.10 The EPM Group recommends that assessments with a pass/fail outcome should be excluded from the EPM, as there are a number of variables which would affect their inclusion, for example:
- 4.10.1 The number of summative pass/ fail assessments undertaken by an applicant at each school.
 - 4.10.2 The pass rate of each pass/ fail assessment.
 - 4.10.3 The weighting assigned to each pass/fail assessment.
 - 4.10.4 The granularity of other assessments contributing to the EPM at each school.
- 4.11 Statistical modelling found that in order for pass/ fail assessments to contribute meaningfully to the overall EPM without skew, an above-average applicant must be likely to fail at least one summative pass/fail assessment. Further, failing a single pass/ fail assessment (regardless of weighting) can have a significant impact on the EPM score, whereas the results of other major assessments can give a lot of tied scores.
- 4.12 Where possible, the raw scores from assessments should be used to inform the EPM. However, there are a number of schools whose university-wide policy is to award grades. Including assessments with three, four or five outcomes (for example A-E grades) alleviates the problem of the skew, provided that the average applicant is likely to achieve a few borderline passes.
- 4.13 Where a medical school particularly wants to include a pass/ fail assessment within the EPM, they may do so subject to their own modelling with expert statistical advice.

Re-sits

- 4.14 There will be cases in which an applicant fails a summative assessment and is required to re-sit. The EPM Group sought to make a recommendation whether, in such cases, the original or the re-sit score should be used for the purpose of determining that applicant's EPM.
- 4.15 The EPM Group recognises that medical schools are usually obligated to follow university-wide policy on re-sits. The EPM Group recommends that re-sit scores should either be the original first-attempt mark, or capped at the pass-mark only, in line with university policy. This rule should be published as part of the locally agreed EPM framework.
- 4.16 An exception may be made in cases in which there are mitigating circumstances for the applicant's performance in the original assessment, but only where the circumstances are formally recognised by the university's policy on re-sits.

Missing scores

- 4.17 The EPM Group was asked to consider what should be done in cases in which an applicant had been unable (for example through illness) to take a given assessment the results of which formed part of the EPM calculation. The EPM Group has concluded that such cases should not arise if the EPM is to be based solely on summative assessments, since arrangements should be made to ensure that applicants can and do sit such assessments regardless of the circumstances.
- 4.18 If there is an exceptional circumstance that an applicant has been unable to take a component of a summative assessment, for example a presentation as part of a module assessment, the medical school should adhere to its own procedures, for example allocating a representative mark as agreed by its Standards Committee.

Definition of a cohort

- 4.19 The EPM Group considered whether a cohort should be taken as meaning all of the applicants in a given year of a course, or whether it should exclude those applicants who had chosen not to apply for the Foundation Programme.
- 4.20 The EPM Group recommends that the cohort should be taken to mean all of the applicants starting the final year together, including those who have chosen not to apply but excluding those who have previously applied (see 4.28), since the EPM needs to be a measure of how each applicant has performed relative to all of the other applicants who have taken the same course. The reason for this can be seen most readily by considering the extreme case of a medical school from which just one applicant applies for the Foundation Programme (which is entirely possible for overseas medical schools), where the EPM would be meaningless if it were limited only to the applying applicant.

Transferring groups of applicants

- 4.21 There are cases where a group of applicants transfers into a school part way through the degree course, having completed the earlier years elsewhere. For example, some applicants who begin their medical studies at the University of St. Andrews transfer at the end of year 3 to Manchester. In these cases the early assessments taken by the applicants from one school might not be in a format comparable with those taken by the applicants from the other school, although in the later years of the course the applicants all take the same assessments, with scores determined by the graduating medical school. The EPM Group considered the three options for dealing with these cases, as follows:
- 4.21.1 Treat the transferring applicants as a separate cohort.
 - 4.21.2 Treat all the applicants as one cohort, but exclude the early assessments (e.g. those taken before the transfer) from the EPM basket.
 - 4.21.3 Treat all of the applicants as one cohort, and find a way to equate the scores from the early assessments across both schools.

- 4.22 It is at local discretion whether applicants on different entry routes, for example graduate-entry and standard-entry courses, are treated as a single or separate cohorts for the purpose of ranking, provided that the composition of the EPM meets the stipulations of paragraphs 4.6 (representative basket of assessments) and 4.7 (transparent and published framework).

Other special cases

- 4.23 There are several other special cases that need to be dealt with by the EPM standards. These cases are considered in turn below.
- 4.24 **Student Selected Components.** The decision about whether to include SSCs in the basket of assessments should be taken by individual schools on the following basis: if the assessments relating to SSCs are robustly designed so that the results of all applicants can be fairly compared regardless of which SSC they have chosen, then the assessments should be included; in other cases the results of such assessments should be excluded.
- 4.25 **Transfer of individual applicants.** The EPM Group recommends that only the assessments taken at the institution of graduation should count within the EPM. It is at local discretion whether only assessments in common with the graduating cohort are counted, or a representative basket of assessments for each route of entry be considered.
- 4.26 **Applicants who take a year out or re-sit a year.** Locally agreed rules should be published in advance of the application round; in the case of re-sits, the same policy should be in place as for the re-sit of individual assessments.
- 4.27 **Applicants who delay their application to the Foundation Programme.** The EPM Group recommends that the original decile score should be carried forward. Points for additional achievements are awarded at the point of application.
- 4.28 **Applicants who re-apply after failing finals.** The EPM Group recommends that the original decile score should be carried forward. Points for additional achievements are awarded at the point of application.

Other achievements

- 4.29 The EPM Group recommends that individual applicants could earn up to five points for additional degrees achieved at the point of application to the Foundation Programme. Where an intercalated degree does not add an additional year (or more) to the length of the Primary Medical Qualification, applicants would receive one point less than a degree which extends the length of the Primary Medical Qualification.

4.29.1 5 points for a Doctoral degree (PhD, DPhil, etc).

- 4.29.2 4 points for a Masters degree, Honours degree 1st class, Bachelor of Dental Science (BDS), B Vet Med.
- 4.29.3 3 points for an Honours degree 2.1 class or an intercalated Honours degree within the length of the degree programme 1st class.
- 4.29.4 2 points for an Honours degree 2.2 class or an intercalated Honours degree within the length of the degree programme 2.1 class.
- 4.29.5 1 point for an Honours degree 3rd class, unclassified or ordinary degree, or an intercalated Honours degree within the length of the degree programme 2.2 class.
- 4.29.6 0 points for Primary Medical Qualification only.
- 4.30 The EPM Group recommends that individual applicants could earn a maximum of two points for the following:
 - 4.30.1 National/international prize – 1st prize only (1 point); *and/or*
 - 4.30.2 Publication in a peer-reviewed journal (1 point); *and/or*
 - 4.30.3 National/ international presentation (1 point).
- 4.31 EPM Group members considered that there would be an element of double-counting if applicants were to earn points for each of these achievements, for example it could be expected that an applicant who presents at an international conference would also be an applicant with a peer-reviewed publication. For this reason a maximum of 2 points is proposed.
- 4.32 The EPM Group does not consider there to be a fair, reliable and transparent mechanism to reward extra-curricular achievements with additional points within the EPM.
- 4.33 The workings of the scheme should be monitored over time to ensure that the points awarded are fair and recognise excellence. Consideration should be given to means by which further ideas to recognise other achievements might cost-effectively and fairly be incorporated.

Granularity

- 4.34 At present, applicants from a single cohort are split into four bands by their academic quartile score. The EPM Group has considered whether it would be beneficial to split applicants into a larger number of bands by a more granular EPM.
- 4.35 The original decision to adopt four bands was partly influenced by the analogy with the four degree classifications (1st, 2.1, 2.2, 3rd). The analogy is a limited one however, as degree classifications do not map on to quartiles (for example, it is not the case that the top 25% of applicants get 1st class degrees).
- 4.36 The EPM Group noted that in principle increasing the granularity of the EPM should increase fairness and increase discrimination between applicants by ability. For example, consider the case of a cohort of 100 applicants. With quartiles the applicant placed at 25th in the cohort is ranked the same as the

applicant who is placed first, while the applicant who is placed 26th is ranked the same as the applicant who is placed 49th. This means that the applicants in 25th and 26th place are rewarded quite differently, even though their performances might be quite similar, while the applicants in 26th and 49th place are rewarded in the same way even though their performances might be quite different.

- 4.37 Statistical analysis of the summer 2010 pilot indicated that septiles (combining two sets of quartiles by addition) added an additional level of differentiation between applicants to the Foundation Programme, particularly when combined with scores from the second application tool (the SJT score). Deciles add a further level of differentiation, provided that this is supported by the underlying data. In recognition that a number of schools record grade points rather than raw scores, the EPM Group believe that a greater degree of granularity beyond deciles would be artificial.
- 4.38 It is important to recognise that the granularity cannot be meaningfully increased beyond the 'resolving power' of the underlying assessments. Given this, the scope for increasing granularity should be further investigated through consideration of the data available from medical schools.
- 4.39 The EPM Group recommends that the granularity of the points for performance in relation to cohort should be increased to deciles.
- 4.40 The current range of points available to applicants for quartiles is 34-40 points out of a total 100 points for their FP application. The EPM Group recommends that each individual applicant earns between 34 and 43 points, based on their decile.

Ranking procedure

- 4.41 The EPM Group has noted that there are various ways in which applicants could be ranked within a cohort and allocated to granular bands, and each of the variants might lead to a different outcome from the others. Given this, the EPM Group recommends that a specific unambiguous process needs to be agreed in order for the EPM to be robust and fair. For the time being, the following approach is suggested:
- 4.42 The N applicants within a cohort will be allocated into deciles according to the following rules. The applicants will be competition ranked according to their overall score. This means that:
 - 4.42.1 Each applicant will have a rank place between 1 (highest scoring) and N (lowest scoring).
 - 4.42.2 Applicants with the same score will share the same rank place.
 - 4.42.3 Where x applicants share the same rank place, the next x-1 rank places will remain empty.
 - 4.42.4 The rank places will be allocated in order (from 1 to N) to ten roughly equal-sized groups, so that each group contains N/10 rank places, rounded up or down to the nearest whole number.

5. Final recommendations

- 5.1 The EPM framework is based on the premise that the EPM will be worth 50 points, earned as follows:
 - 5.1.1 34-43 points for decile scores; *plus*
 - 5.1.2 0-5 points for previous degrees; *plus*
 - 5.1.3 0-2 points for prizes, publications and presentations.
- 5.2 Performance in relation to the graduating cohort should be reported in deciles of roughly equal size (see 5.10). The cohort is defined as all of the applicants in a given year at the point of application to the Foundation Programme, including those who have chosen not to apply but excluding those who have previously applied (see 5.8). It is at local discretion whether applicants on different entry routes, for example graduate-entry and standard-entry courses, are treated as a single cohort or separate cohorts for the purpose of ranking.
- 5.3 Each medical school should be required to publish, in advance of the related recruitment round, the details of the basket of assessments and weightings to be used to create deciles for that round. This will mean that the approach of a school will be subject to scrutiny by its applicants, by other schools, and by those responsible for recruitment to the Foundation Programme.
- 5.4 Deciles should be composed of a range of assessments which
 - 5.4.1 Are summative (and hence subject to formal controls).
 - 5.4.2 Cover clinical and non-clinical performance.
 - 5.4.3 Cover clinical knowledge and clinical skills.
 - 5.4.4 Cover all aspects of the curriculum assessed up to the end of the penultimate year at medical school.
 - 5.4.5 Represent the average performance of the applicants over time, rather than being limited to a snap-shot.
 - 5.4.6 Include written and practical forms of assessment.
- 5.5 There is no minimum number of assessments to be taken into account in constructing deciles. Only assessments which achieve a fair spread of scores or grades should be included. Pass/fail assessments should not count within the decile score, unless there is a sufficient number of pass/fail assessments that an above-average applicant is likely to fail at least a few.
- 5.6 Locally agreed rules around the construct of the deciles for applicants from different entry routes, for example applicants transferring partway through the course, applicants who take a year out to intercalate, or applicants required to re-sit a year, should be published in advance of the related recruitment round.
- 5.7 Re-sit scores should either be the original first-attempt mark, or capped at the pass-mark only, except in the event of mitigating circumstances as accepted by the university's or medical school's policy. This rule should be published as part of the local framework for deciles.

- 5.8 Applicants who delay their application to Foundation Programme should be given their original decile score. Points for additional achievements are awarded at the point of application.
- 5.9 Applicants who re-apply to the Foundation Programme after failing finals should be given their original decile score. Points for additional achievements are awarded at the point of application.
- 5.10 The N applicants within a cohort will be allocated into deciles according to the following rules. The applicants will be competition ranked according to their overall score. This means that:
- 5.10.1 Each applicant will have a rank place between 1 (highest scoring) and N (lowest scoring).
 - 5.10.2 Applicants with the same score will share the same rank place.
 - 5.10.3 Where x applicants share the same rank place, the next x-1 rank places will remain empty.
 - 5.10.4 The rank places will be allocated in order (from 1 to N) to ten roughly equal-sized groups, so that each group contains N/10 rank places, rounded up or down to the nearest whole number.
- 5.11 In addition to points for deciles in relation to the graduating cohort, each applicant can earn up to 5 points for evidence of a previous degree (either prior to medical school or an intercalated degree) and up to 2 points for evidence of a national/ international 1st prize (1 point); a national/ international poster or oral presentation (1 point); a publication in a peer-reviewed journal (1 point).
- 5.12 Applicants can earn up to five points for additional degrees, as follows:

Previous degree	Number of points
• Doctoral degree (PhD, DPhil, etc)	5
• Masters degree • 1st class honours degree • Bachelor of Dental Science (BDS) • B Vet Med	4
• 2.1 class honours degree • 1st class intercalated degree which does not extend the length of the Primary Medical Qualification	3
• 2.2 class honours degree • 2.1 class intercalated degree which does not extend the length of the Primary Medical Qualification	2
• 3rd class honours degree • Unclassified or ordinary degree • 2.2 class intercalated degree which does not extend the length of the Primary Medical Qualification	1
• Primary medical qualification only • 3rd class intercalated degree which does not extend the length of the Primary Medical Qualification	0

Notes for points for additional degrees

- Honours degrees include any type of Bachelors honours degree, e.g. BSc, BA, BEng, LIB, BMedSci, etc.
- A Masters degree is where it represents a further year of study taken in addition to a basic medical qualification. Some international medical schools (e.g. the USA) award an 'MD' or similar as part of their basic medical qualifications. This qualification does not attract any additional points in this section.
- For students who have undertaken an exchange programme of study as part of a degree course, you must take the grade point average (GPA) and calculate the equivalent degree level and select the most appropriate.
- For a 4 point scale, a GPA of 3.6 - 4 should be scored as equivalent to a 1st class degree, a GPA of 3 – 3.5 as 2.1, a GPA of 2 – 2.9 as 2.2 and a GPA of 1 – 1.9 as a 3rd class degree.
- For a 5 point scale, a GPA of 4.4 - 5 should be scored as equivalent to a 1st class, a GPA of 3.8 – 4.3 as 2.1, a GPA of 3 – 3.7 as 2.2 and a GPA of 2.9 or lower as a 3rd class degree.
- If an applicant holds more than one degree at the time of application to the Foundation Programme, they should provide evidence of the degree that will achieve the highest number of points.
- Additional points for degrees can only be given to degrees awarded by the time of application to the Foundation Programme. Official notification from the university must be provided. Where the applicant has received a pass result but has not received the degree certificate, a letter from their medical school Dean confirming that they have passed must be provided, on letter headed paper, signed and dated by the Dean.

5.13 Applicants can earn up to five points for prizes, presentations and publications, as follows:

Educational achievements	Number of points
Prizes • 1st prize – National/international educational prize	1
Presentations • Oral or poster presentation at a national or international conference • Poster presentation at a national or international conference	1
Publications • Educational research paper published in a peer-reviewed journal	1
Maximum number of points available	2

Notes for points for publications, prizes and presentations

Prizes

Bursaries and medical school prizes will not count in this category. The prize must be 1st prize and it must be a national or international educational prize. A letter of evidence from the awarding body must be provided by the applicant.

Presentations

The conference must be hosted by a recognised professional medical body in order for a student to receive a point. The conference must have taken place by the time of application to the Foundation Programme. A letter of evidence from the conference host must be provided by the applicant.

Publications

Applicants must supply a PubMed ID (PMID) at the point of application to the Foundation Programme or provide a letter of evidence that the work has been accepted for publication and is 'in press' for a publication which has a PMID. This includes papers, abstracts, book chapters, audits and in rare cases, letters. The front page of the article including the title and authors' names must be provided by the student and uploaded into FPAS.

If an applicant has more than one publication, prize or presentation, they will receive a maximum of one point for any of the three categories individually; a maximum of two points in total.

Improving Selection to the Foundation Programme

Appendix J

Report of aggregating SJT and EPM scores



Durham
University



Aggregating Scores in the Foundation Selection Process



Professor John C. McLachlan
School of Medicine and Health
T: 0191 3340323
E: j.c.mclachlan@durham.ac.uk

Aggregating Scores in the Foundation Selection Process

Introduction

The Foundation Programme, introduced in 2005, had the aspiration of providing a cost effective, valid and national solution to the task of selecting doctors for the two-year training programme which would lead to full registration for the GMC. The initial selection process relies on a two stage process, in which 'white space' questions are worth 60 points, including 10 points for degrees, prizes, publications and presentations, while 'quartile' rankings provided by medical schools are worth 34-40 points. There have been issues about the on-line administration of the application process, and the authenticity of some applications. However, possibly more seriously, there is a lack of evidence for the validity of these measures, which may reduce the acceptability of the process, and even conceivably lead to legal challenge. Scoring the white space questions is also cost intensive and of low reliability.

As a result, the project 'Improving Selection into the Foundation Programme' has been undertaken by the Medical Schools Council. Reviews were commissioned (Illing et al, 2009; McLachlan & Turnbull, 2009; Thistlethwaite, 2009), and proposals for an improved selection approach brought forward. This will feature a combination of an Educational Performance Measurement (EPM) based on decile rankings from medical schools plus an additional points measure for educational achievements, and the outcomes of a Situational Judgement Test (SJT) Measure. The current working assumptions are that out of 100 points, the Educational Performance Measure (EPM) will be worth 50 points, including 7 points for degrees, prizes, publications and presentations. Quartiles will become deciles, worth between 34 and 43 points. The Situational Judgement Test (SJT) will therefore also be worth 50 points.

The SJT has no pass mark, and will not feature negative marking. The range of scores and scores available will vary year to year, depending on the composition of the paper and the scoring key.

However, there is an additional complication in aggregating the SJT and the EPM. If the arithmetic mean of two measures with different variances is calculated, then the outcome is unintentionally weighted towards the measure with the largest variance (McLachlan & Whiten, 2000; Eva and Reiter, 2004).

Brief

The brief from the MSC invited me to address this issue and to bring forward proposals for combining the EPM and SJT to consider:

- The range of points for each element (for example should the SJT be scaled to score between 0-50, or would applicants scoring 680 score a proportion of points available?)
- Addition/multiplication/z-scores – and factors required (e.g. do we need to know the total points available or would scoring be relative?)
- The benefits and drawbacks of the different approaches
- a way of combining the scores from the EPM and the SJT that is fair and does not distort the value of one or the other; and which is acceptable to stakeholders and easy to understand.

The Proposed Aggregation Method

I have approached this through modelling of score combination techniques, but only the proposed model is described in detail. This approach was greatly aided by the availability of two further sources of evidence: the SJT Initial Pilot Report (Patterson et al, 2010) and an MBA Dissertation (Moore, 2010). These sources provided the basis for estimates of the means and standard deviations of some key parameters. The key principles in developing the aggregation method were that it should provide an appropriate and defensible outcome but also that it should look intuitive, thus gaining a high level of acceptability. In summary, the proposal is that the SJT results are recalibrated to a 50 point scale, and added or averaged with the EPM without further data modification.

SJTs

The selection process for Foundation should be entirely normative in nature, since it is assumed that competence criteria are met by the possession of a Primary Medical Qualification. Variations in cohort ability from year to year are not relevant, since each year has a separate application process, so no element of test equating is required. Since we have to draw distinctions between approximately 8000 candidates each year, a relatively fine level of discrimination is needed.

It is therefore proposed that the SJT Score range is recalibrated each year to 50 Points, without rounding. For any given $Score_x$, the Point Score is given by the equation:

$$Point\ Score = \frac{(Score_x - Score_{min}) * 50}{(Score_{max} - Score_{min})}$$

For the Empirical Scoring Approach in the pilot, $Score_{min} = 710$ and $Score_{max} = 992$, hence for a range of scores, the following point equivalents are obtained

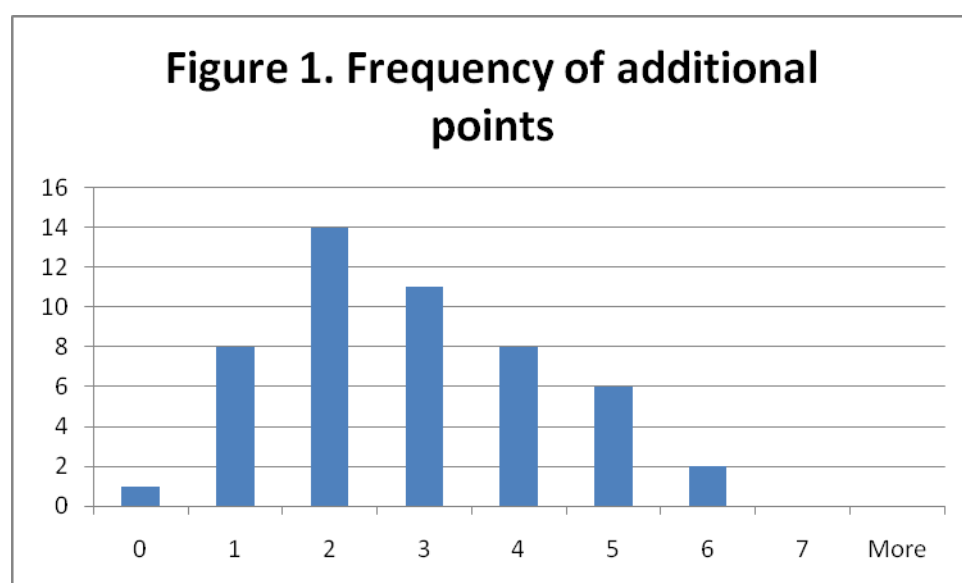
Score _x	Points
710	0
725	2.659574
750	7.092199
775	11.52482
800	15.95745
825	20.39007
850	24.8227
875	29.25532
900	33.68794
925	38.12057
950	42.55319
975	46.98582
992	50

EPMs, and aggregation with SJTs

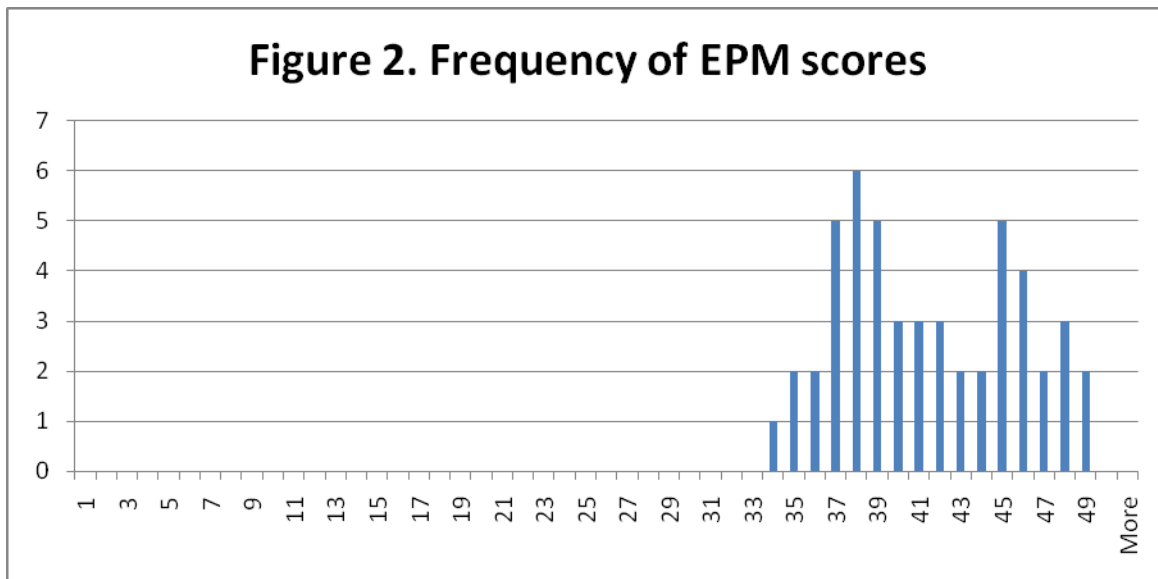
Considering the case in which the EPM score remains in the raw format proposed (i.e. 1 point per decile, starting at 34, plus a score of 0-7 points for other factors), it is necessary to estimate the distribution that will result from this approach.

I have considered 50 candidates who have undertaken the SJT, been ranked into deciles, and been awarded scores for performance indicators. Example means were obtained from Patterson et al, (2010) and Moore, (2010), and it was assumed that the distributions would be of skewed normal form, again based on these sources. It was also assumed that there would be a degree of coherence in the results: that applicants performing well on one measure may well perform reasonably well on the other, but the correspondence would not be perfect.

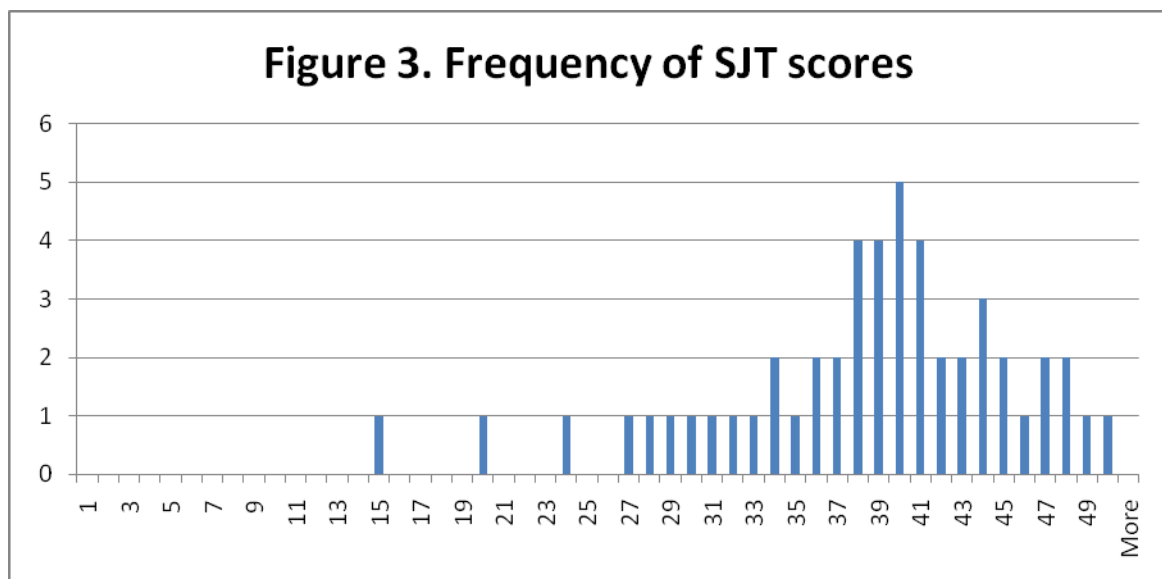
The range of model scores from the 'additional points' performance indicators is shown in Fig. 1.



Each candidate was also given a ranking, in the form of deciles (so obviously these have a flat distribution). A degree of coherence between the ranking and the 'additional points' score was assumed, and built into the model. A total EPM calculated for each individual by summing the 'additional points' components with the ranking which was scaled, as proposed, between 34 and 453 (Fig 2).



The model scores on the SJT were calculated from Patterson et al, (2010), 'empirical scoring' approach, page 35, and are shown in Fig. 3.



SJT Scores and EPM Rankings are available for each model candidate in an Excel Workbook, to allow further exploratory modelling if desired.

The EPM and SJT values for each candidate was then summed to give an uncorrected raw total score, and these were ranked by model candidate. The contribution of SJT Scores and EPM rankings to the variance of the total score can be calculated as the square of the correlation coefficients of each with the total score.

Carrying out this calculation, the SJTs represented 94% of the variance, and the EPMs 82% of the variance in the summed values. Obviously, there is considerable co-variance. However, these values seem similar enough to be acceptable.

To explore the effect of correcting for the different variances, the SJTs and EPMs were converted to the corresponding z-scores, and ranked. The two rankings are shown in Appendix 1. The rankings are essentially identical by statistical test. The correlation between the rankings obtained by raw scores or z scores is 0.91. The SJTs contribute more to the variance, and therefore carry a slightly higher weighting, but the SJTs already have a known reliability, and it is unlikely that the EPMs will have a similar or greater reliability. My view is that conversion to z-scores is unnecessary if the assumptions made here hold*, and that *not* converting to z-scores will seem more natural to stake holders, in that the data have not undergone a processing step.

***Assumptions**

The conclusion that the EPM and the SJT converted to a 50 point scale can be added without distorting the results relies in part on the assumptions made about the distribution of the 'additional points' measure. However, the outcome is stable to small differences, and if the distribution were mirror-symmetrical (i.e. negatively skewed rather than positively skewed) the same outcomes would emerge. Once the EPM process has been run, either as a pilot or as for real, this assumption can be tested against the data.

It is also a critical assumption that the variances of the SJT and the total EPM as described, are similar. In the version of the model presented here, the standard deviation of the EPM was 4.15 and that of the SJT was 7.3. Wide divergences from these values in practice will require reconsideration of the aggregation procedure.

It has also been assumed that the scores in the various categories have a degree of concordance, i.e. that a candidate who scores highly in one measure is more likely to score high in the others than to score low. This is plausible based on experience of the performance of medical graduates under many test settings, but again can only be tested with real data or following an authentic pilot project on the EPM comparable to those carried out with the SJT.

It has been assumed that the results of the Pilot SJT will be comparable with the results of the actual SJT in terms of descriptive statistics. This can be tested following the much larger pilot study envisaged by the MSC in the coming year. It would be surprising, however, if the conclusions were radically different.

The 'Empirical' scoring data were used in the model presented. However, this is very similar to the 'Rational' scoring data, and essentially the same outcome results if the Rational data are used as the template for the modelling process.

References

Boursicot KAM, Roberts T, Pell G. (2007) Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical Education*. **41**: 1024.

Boursicot, K. A. M., Roberts, T. E., & Pell, G. (2006), "Standard Setting for Clinical Competence at Graduation from Medical School: A Comparison of Passing Scores Across Five Medical Schools". *Advances in Health Science Education* **11**: 173-183.

Eva KW, Reiter HI. (2004) Where judgement fails: pitfalls in the selection process for medical personnel. *Adv Health Sci Educ Theory Pract* **9**: 161-174

Illing J, Campbell M, Kergon K, Thompson, N, Burford B, Morrow G, et al. (2009). *Selection methods for the Foundation Programme: A Literature review*. Newcastle University, Northern Deanery. London: Medical Schools Council.

McLachlan, JC & Whiten, S.C. (2000). Marks, scores and grades: scaling and aggregating student assessment outcomes. *Medical Education* **34**: 788-797.

McLachlan JC, Turnbull L. (2010) A review of available methods of selection for Foundation. Report for Medical Schools Council.

McManus JE, Elder AT, de Champlain A, Dacre JE, Mollon J, Chis L. (2008). Graduates of different UK medical schools show substantial differences in performance on MRCP (UK) Part 1, Part 2 and PACES examinations. *BMC Medicine* **14**: 6:5

Moore CG (2010) Can we predict which doctors will fail? Cardiff University MBA Dissertation.

Patterson F, Archer V, Kerrin M, Carr V, Faulkes L, Stoker H, Good D (2010) Initial Pilot Report: Improving Selection to the Foundation Programme. London: Medical Schools Council.

Thistlewaite J, Doug M, & Clay D. (2009). *A review of options for the selection into Foundations training*, London: Medical Schools Council.

John C. McLachlan

6th April 2011

Appendix 1

Low to High ranking by model candidate number	
Ranked by raw scores	Ranked by z-scores
4	4
13	13
1	1
2	2
15	15
20	20
19	7
7	19
18	10
10	18
6	5
21	3
5	6
12	12
3	21
16	16
9	9
14	14
17	17
27	8
8	27
11	11
22	22
23	25
25	23
29	29
24	24
26	26
28	28
33	33
32	32
30	30
31	31
34	34
41	41
35	35
36	37
37	36
40	40
42	39
43	42
44	44
38	38
39	43
45	45
46	46
47	47
48	48
49	50
50	49

Improving Selection to the Foundation Programme

Appendix K

ISFP Communications Plan

Improving Selection to the Foundation Programme Communications Plan Executive Summary

Introduction

A communications plan for the 'Improving Selection to the Foundation Programme' (ISFP) Project has been produced outlining the key messages, communications mechanisms, resources available and the budget required. The communications plan only relates to the work carried out by the MSC in relation to the ISFP project between January 2010 and July 2011.

Purpose

The purpose of the plan is to define:

- What information needs to be communicated with whom and when
- How the information will be communicated
- Who is to be responsible for which aspects of communication
- What resources will be needed
- How risks and issues relating to communications will be managed

Stakeholders

A number of stakeholders has been identified and grouped into order of importance. Those who have been deemed the most important include: medical students, medical schools, UKFPO, UK Scrutiny Group - Health Departments from the 4 countries, COPMeD, COGPED, UKFPO Rules Group. A full list of stakeholders and how they will be engaged with can be found in the detailed communications plan.

Branding/Website

The website – www.isfp.org.uk - and any other communications will be neutrally branded. Logos for those involved will only be used on relevant documents and presentations. Headed paper and a PowerPoint presentation have been designed to keep consistency and feature the main image from the ISFP website. A generic email address – admin@isfp.org.uk – has also been set up.

Communication Mechanisms

Various communications mechanisms that will be used throughout the project have been identified. These have been divided into 4 main headings; electronic media, events, publications and press. Activities include; ISFP website, Facebook, discussion forum, podcast, e-newsletters, student workshops, meetings with careers advisors, a large stakeholder event, project handbook, key facts document and leaflet and FAQ document.

Improving Selection to the Foundation Programme Communications Plan

Introduction

This document is the communications plan for the 'Improving Selection to the Foundation Programme' (ISFP) project and has been produced by the Medical Schools Council. The communications plan only relates to the work carried out by the MSC in relation to the ISFP project between January 2010 and July 2011 and it assumes that the reader is familiar with the ISFP project.

Purpose

The purpose of the plan is to define:

- What information needs to be communicated with whom and when
- How the information will be communicated
- Who is to be responsible for which aspects of communication
- What resources will be needed
- How risks and issues relating to communications will be managed

Principles

The plan is based on the following general principles:

- Communications must be clear and objective
- Communications should be restricted to topics that fall within the scope of the project
- Communications must be carried out within a set time frame and budget
- The responsibilities for communications must be clear
- Stakeholders should be given the information they need, when the time is right, through the correct networks and in the right way

Communications Objectives

The communications objectives are to:

1. Ensure that stakeholders - particularly medical students - understand the nature of the proposed changes to the FP selection process, and appreciate the rationale for those changes
2. Minimise the impact of misleading publicity or rumour about the project, and ensure that confidence in the current recruitment process is not undermined
3. Motivate sufficient participation in the project (e.g. recruiting volunteers for SJT item writing)
4. Ensure that all those participating in the project understand what they have to do in order to ensure its success
5. Ensure that the ideas or concerns of stakeholders are feedback to the Project Group so that they may be properly taken into account

Key Messages

The key messages that need to be reinforced throughout the communications activity are:

1. The proposed changes are evolution, not revolution
2. The proposed changes will be implemented only if they are proven to work and to be an improvement over the current approach
3. Extensive consultation will be carried out with all stakeholders, including students
4. Evidence showed that interviews along with other selection methods were not the best option
5. The pilots will not affect the outcome of current recruitment rounds in any way
6. The project is working collaboratively with UKFPO
7. SJTs are used successfully for GP short-listing
8. The changes will not affect the percentage of applicants who get their first choice of foundation school - as that depends on the algorithm, not the assessment method

Stakeholders

The stakeholder groups considered by the plan are shown below. These are grouped by importance.

Primary

- Medical students, including the BMA MSC and UKFPO Medical Student Board
- Medical schools
- UKFPO
- UK Scrutiny Group - Health Departments from the 4 countries
- COPMeD
- COGPED
- UKFPO Rules Group

Secondary

- Medical Education England
- NES
- Foundation School Directors, Managers and Deaneries
- Undergraduate Careers Advisors
- SJT Item Writers

Tertiary

- General Medical Council Education team
- Junior Doctors, including the BMA Junior Doctors Committee
- NHS Employers
- Employing Healthcare Organisations
- The Academy of Royal Colleges
- Patients
- Trusts
- SHA Communication Teams
- Educational Supervisors
- International Students

How these groups will be engaged with can be seen in appendix one to four.

Branding/Website

The website and any other communications will be neutrally branded. This will not feature any of logos of the organisations involved but logos may be used on relevant presentations and documents. The ISFP website has been updated to reflect this as it originally had the Medical Schools Council's logo and branding on it and the domain name has been changed to www.isfp.org.uk. A generic email address – admin@isfp.org.uk – is being set up to ensure consistency when sending out information. Headed paper and a PowerPoint presentation have been designed to keep consistency between communications.

Communication Mechanisms

Below is information on general communication mechanisms that could be used throughout the project.

Electronic Media

Website

A website has been created specifically for the project. All information on the project will be available here including briefing documents, information on SJTs and EPM, all final reports, a news page, a discussion forum, a newsletter sign up function and updates for students. The website will be the main communications method for reaching all stakeholders and the website address – www.isfp.org.uk – will be advertised on all communications that are sent out.

Facebook

Facebook continues to be a very popular networking and information site and provides users a platform to share information on topics of interest. An ISFP Facebook group has been set up, to provide 'friends' with regular updates as well as linking them to useful documents and websites for information. As at 19.7.10 there were 925 members. In total, 48 students groups have been contacting asking them to send information out to members.

Discussion Forum

An anonymous discussion forum has been set up on the ISFP website to allow people to post their views on the project. All comments will be communicated to the Project Group and included as part of a stakeholder feedback summary. Awareness of the discussion forum will be raised via letters to students, Facebook posts, Medical Schools Council website, press coverage in the Student BMA News and the Student BMJ and via the BMA Medical Students Committee (BMA MSC) and UKFPO Medical Students Board. Other blogs and discussion pages will be monitored to see what people are saying about the proposed options. This will help to ensure that any information sent out addresses the issues that the different stakeholders are concerned with.

RSS Feeds

RSS feeds will be added to specific pages on the ISFP and Medical Schools Council websites to allow those interested to receive notification when new information is posted. The RSS facility will be highlighted in correspondence to encourage people to sign up.

Podcast

In addition to the FAQ document it would be beneficial to have a student interviewing the Paul O'Neill, ISFP Chair, which would address the common questions and dispel up any rumours that may be circulating for example, that the new system will be implemented this year. It would be most appropriate to feature Nick Deakin, Chair of the BMA MSC and member of the ISFP Project Group. This would not involve the student having to confirm his/her support for the project but merely to highlight the main points. This would provide an added interactive dimension to the communications of the project and would be available in video podcast format via the website and publicised via Facebook. This will be available for the 2010/2011 academic year.

E-newsletters

Short and concise e-newsletters/bulletins will be sent to all stakeholders when there is a need to communicate. An option to sign up for newsletters will be available on the ISFP website. If a large number of people sign up for these, an organisation called iContact will send out the newsletter on your behalf to reduce the risk of blocking the server. If it is decided that this is the best way ahead iContact charge £70 per month to do this. Information on how to sign up will be added at the bottom of all correspondence. The newsletters/bulletins will hopefully reassure stakeholders that each stage is being carefully looked at and analysed, thus providing them with faith in the proposed system.

Medical School Newsletters and Intranet

The communications teams at each medical school will be contacted with a summary document to post on the intranet and in newsletters to help raise awareness of the project with the wider medical school network. It is hoped that an update will be posted once every term. The contacts for the newsletters and intranet will be added to the e-bulletin contacts list to ensure they get up to date information for students.

Events

Student events

During the project it will be important to provide students with information updates and ask them for their views. There are a number of different ways in which this can be done, including:

Workshops: Student workshops and meetings will provide an opportunity to discuss the project in more detail and talk to student representatives about concerns with the project. It is proposed that student workshops are carried out at each pilot medical school in the first instance. This will help to raise awareness of the pilots, encourage students to take part and gather feedback on the proposed system. Workshops may then be held at the remaining schools during 2011 once the pilots have been carried.

Student meetings: Meetings held by other organisations will provide opportunities to liaise with students. The UKFPO Medical Students Board will provide access to a number of medical school representatives and so some of the costs incurred by the UKFPO for running the meeting will be covered by the ISFP in exchange for a section of the day being dedicated to the project. The BMA MSC will also provide an opportunity to liaise with students and close liaison with the BMA MSC Secretariat will provide a good opportunity for joint initiatives.

Careers Advisors

In addition to the possibility of attending careers fairs it may be useful to meet with all careers advisors from each of the medical schools. The National Education Advisors Forum (NEAF) meets throughout the year. It may be useful to attend the meeting and inform them of the changes that are happening so that they can correctly relate information back to their students. A presentation will be produced for careers advisors for them to give to students when necessary.

Large stakeholder event

A large stakeholder event focusing on the outcomes of the pilots and the next steps and involving a large number of stakeholders will be held in the first week of June 2011 - i.e. after the pilots and before submitting the final report- and will focus on a number of different areas including; demonstrating that the evidence and work was done correctly; encouraging people to be more enthusiastic about the project, to announce the new governance arrangements and to critique the final report in terms of usability. A range of technology used to engage attendees, for example; voting buttons and group notepads to record discussions in seminars, could be used to add an interactive nature to the event. Presenters will include various members of the Project Group, someone from the Project Board – most likely Pat Hamilton, and the UKFPO.

Publications

FAQ document

To provide clarification on the project and to answer common/likely questions an FAQ document will be produced and will be available on the ISFP website. This will also be sent to members of the Facebook group and to stakeholders when necessary.

Project Handbook

A 4-page A5 'Project Handbook' outlining the main facts of the project will be produced and sent round to careers advisors to hand to their students. The information for this will be taken from the FAQ document and will also be available on the ISFP website. This will be developed and designed in-house using InDesign software to reduce costs.

Leaflets

A5 leaflets with the ISFP website address and some key facts will be printed and sent out with the Student BMA News and Student BMJ News in June 2010. This will reach 20,000 students. This information will also be available via the ISFP website.

'Key Facts' document for students

A one page A4 document outlining five key facts about the project will be produced and sent round to all medical students via the Dean or Head of the school. This will help the clear up any rumours that may be circulating about the implementation date of any new system. This will also be available on the ISFP website and will be sent round to Facebook members, the BMA MSC and the UKFPO to circulate.

Briefing documents

Briefing documents which highlight all relevant information will be sent out to stakeholders and will be posted on the ISFP website. A PowerPoint presentation which summarises the main points will also be produced. As the project progresses a number of briefing documents may need to be produced, for

example; information on the outcomes of the pilots or information for schools participating in the pilots outlining what they need to provide.

Press

No formal press engagement will take place until there is a significant update to highlight. It is likely that there will be minimal press interest until the pilots have been analysed and a decision is made on whether to implement SJTs and EPMs. However, it will be useful to keep a number of relevant publications informed of the progress including; Student BMA News, Student BMJ, BMJ Careers, NHS Employers Weekly Bulletin, Healthcare Workforce Bulletin and SHA communications teams. It may also be useful to invite them to any significant workshops or meetings.

Feedback

A large part of the communications strategy will involve talking to key stakeholders and asking for their feedback on the proposed process and it is extremely important that these views are taken into account. If we are not seen to be listening to these views this could have a negative impact on the project. Feedback will be gathered in a variety of ways; at meetings, via Facebook and the website discussion forum, through workshops at medical schools, etc, and these will need to be fed back to the Project Group for consideration. To ensure this is done in a consistent manner, all feedback will be consolidated and included as an update paper as part of the communications update, at Project Group meetings. If any comments are fed back that are particularly concerning, these comments will be sent to all members for information. A communications distribution list will be set up and will include a named contact at each of the Project Group member's organisation. This will ensure that there are clear communication streams and will allow any feedback from their organisations to be fed back to the Project Group. Relevant external websites and blogs will also be monitored for feedback and any comments will also be fed back to the Project Group.

Resources

In terms of staff time, there will only be one member of staff implementing the communications plan. Therefore, communications will be concise and only sent out when necessary. Activities will try to reach as many people as possible and have been designed with this in mind. The support of the Project Group will be essential and it will be important to utilise any existing networks when sending out any communications.

It may be useful to set up a Communications contacts list to ensure that key messages are reaching the intended audience and for some additional expert advice and input and to ensure the key messages are reaching the intended audiences.

Budget

Costs have been identified for the various communication mechanisms outlined above, which will be tightly monitored to ensure there is no overspend. Separate funding has been allocated to cover staff time and so where activities only require staff time they will be costed at £0. As at 19.7.10 these are estimated costs based on quotes and previous events and include VAT.

Activity	Estimated cost (inc VAT)
Electronic Media	
Website	
Re-branding website, new email address and IT support	3876.77
Google analytics	0
Discussion Forum	0
RSS Feeds	0
Facebook Profile	0
Podcast	
Filming, editing, script writing	3837.5
Room hire	200
Travel expenses	600
e-Newsletters	
BMN to set up sign up page	528.75
iContact	910
Medical School Newsletters/intranet	0
Events	
Student events	
Workshops at each medical school (x31) including expenses and room hire	20000
Careers advisors	
Meeting with NEAF	35
Large stakeholder event	17500
Publications	
FAQ	0
Handbook	4610
Leaflets	2047.25
Key facts document	0
Briefing document	0
Press	

Press updates	0
Other	
Software	
Buying software and training for InDesign and Photoshop to design leaflets, handbook etc	1886.85
Miscellaneous / contingency e.g. Extra events, travel, press event, medical staff group	15,000
Total	71032.12

Risk Management

The communication risks will be included in the risk register and managed along with the rest of the project activities. Records of all conversations, comments, quotes and stakeholder feedback will be kept to ensure there are accurate accounts for all correspondence. The Project Manager will help to ensure all activities are within budget and reports on the progress of the communications activities will be submitted regularly. The UKFPO will also be involved throughout the project to ensure this can be easily handed over to them without any confusion.

Measuring Success

The success of the communications strategy will be measured throughout the project. This will be done in a number of ways including; measuring website hits, monitoring the comments via Facebook and the discussion forum, monitoring the number of recipients of the newsletter and subsequent hits to the website, monitoring blogs every month to look at the positivity of comments and carrying out a survey with students at workshops to understand what messages are reaching them and how they feel about the project. Communications with the other stakeholders will be fed through the Project Group and any existing and new communication networks.

Appendix one

Key stakeholders

There are a vast range of stakeholders that need to be kept informed on the progress of the project. In order to optimise use of the Improving Selection to the Foundation Programme Project Group's limited resources key audiences have been identified.

Stakeholder	Current understanding	Desired understanding/behaviour	Mechanism
Medical students	Variable – some students have received information sent out whilst others have not. Better communication links need to be developed to help reduce this variation There is an issue that they are likely not to have understood, or bought into, the rationale for SJT/EPM and will want interviews, CVs, etc. It is important that we break down these perceptions	Detailed understanding of the project and the outcomes of the pilots leading to trust and support of the system	Website Joint letters with UKFPO Online discussion forum Podcast Newsletters FAQs Facebook Events Medical school intranet Medical student press
BMA Medical Students Committee	High. Nick Deakin is a member of the Project Group and the Secretariat is involved in the project as well	High – the BMA MSC will be a key organisation in helping disseminate information to students about the project, dispelling any rumours and holding joint meetings/workshops	Project Group meetings Secretariat meetings about the communications plan Email updates Briefing documents Workshops/meetings Newsletters
Medical schools	High – all medical school deans are aware of the project and are kept informed and up-to-date via the Medical Schools Council. Medical school deans need to ensure this information is being cascaded to all medical school staff	Detailed understanding and agreement that this is the best way forward as all medical schools will need to implement and manage the new system Medical Schools will need to agree an EPM standard and understand how to produce EPMS to that standard	Medical Schools Council meetings Meetings with medical school staff Newsletters Briefing documents on SJT and EPM Medical school intranet Pilots

Stakeholder	Current understanding	Desired understanding/behaviour	Mechanism
		The medical schools involved in the pilots need to know exactly what they need to do in order to make the pilots a success	
UKFPO	High – Professor Derek Gallen is a member of the Project Group and is involved in the decision making. The UKFPO will also be taking forward the second stage of the project and has been working with the project manager on what needs to be done	Detailed understanding and supportive of the options. The UKFPO will be implementing and running the new system and so it is imperative that they support any decisions made and are happy with the final outcomes	Project Group meetings Liaison with Project Manager Regular meetings with the Medical Schools Council/Professor Weetman Newsletters UKFPO Rules Group
UKFPO Rules Group	High – regular updates are given at the meetings	Full support and understanding of the project and its aims and objectives. This group will be important during the piloting and implementation stage	Regular update at their meetings Emails Newsletters Briefing documents/reports
GMC Education team	Medium – Martin Hart from the education team attends Project Group meetings and feeds back to the rest of the team	To be supportive of the new system and understand the reasons for implementation and benefits over a national exam	Project Group meetings Newsletters Briefing documents
Health departments, including MEE and NES	High – they have funded the project and are represented on the Project Group and Project Board and so it has to meet DH requirements	To fully support and understand the proposed system and agree that the changes being made are suitable and cost effective	Regular reports Project Group meetings Project Board meetings
Junior Doctors	Unknown	To understand the project and to be supportive in helping ensure the test reflects what a junior doctor is expected to do. It is important to liaise with this group as they are currently doing the different activities that SJTs will be assessing medical students	QA activities - job specification and item writing

Stakeholder	Current understanding	Desired understanding/behaviour	Mechanism
		on The experts undertaking the job analysis will be working closely with a small number of FY1s and so their support is needed to make sure the analysis truly reflects what junior doctors have to do	
BMA Junior Doctors Committee	Have knowledge through communications with the BMA MSC	Basic understanding is needed by the group but as they will not be directly affected by any changes this is more in an information capacity. If the committee did want any information this could be fed through the BMA MSC	Communications with the BMA MSC Email updates if the committee express an interest in hearing more about the project
Foundation School Directors and Managers	High – they have been informed on the process through meetings and emails. They are also helping with the SJT item writing	To support and understand the new system as they will be employing those students who have gone through the system	UKFPO Rules Group FSD and FSM meetings Newsletters SJT item writing and workshops Briefing documents FAQs
Employing Healthcare Organisations, including NHS Employers	Unsure – Ashley Fraser represents NHS Employers on the Project Group and Terry McMurray for NIMDTA, but it is unclear how much the different employing healthcare organisations know at this stage	High level of support and understanding Some of them will play an important role in the job analysis and need to be supportive of this work	Meetings with organisations Briefing documents Newsletters
Undergraduate Careers Advisors	Unknown	It is important this group fully understands what the new system involves and students will be coming to them for advice	Careers fairs Leaflets Emails via the medical school secretaries network and careers contacts on UKFPO website
COPMeD	Professor Derek Gallen represents COPMeD on the Project Group and feeds back to the council	Full support is needed from this group	Project Group Newsletters Briefing documents
COGPED	Unknown	Need to know the project is	Emails

Stakeholder	Current understanding	Desired understanding/behaviour	Mechanism
		underway and how it is progressing. SJTs are already used in GP selection so it would be useful to be able to share findings	Newsletters Briefing documents Meetings
The Academy of Medical Royal Colleges	Sir Neil Douglas is on the Project Group so there is understanding of the project. They also use SJT for entry into speciality training so understand the concept	Would be useful to share information and so understanding of the project and the reasons it is being carried out would be beneficial in developing a relationship	Project Group (via Sir Neil Douglas) Newsletters Briefing documents
Trusts	Unknown	Understanding of the project and the work that will be carried out in the different Trusts. Need them to give permission to carry out work	Liaise through communications with NHS Employers AUKUH Liaise with Medical Directors
SHA Communication Teams	Unknown	General knowledge of the project would be advantageous. The SHA communication teams could have established networks to share information with staff and patients	
SJT Item Writers	Unknown	High. This group will be writing the SJT items and so will need to understand why the project is being done and the different elements within this, for example the job analysis. Detailed understanding of the project may have an impact on the quality of the items	SJT Item Workshops Emails via Medical Schools Secretaries and Deans Email updates Newsletter for writers Emails regarding QA process
Patients	Low	To understand that this will not undermine patient safety and is about selecting the best doctors to the Foundation Programme. Would like this group to help with the job analysis	Academy of Royal Colleges Patient Liaison Group and the National Association for Patient Participation Involving patients in the job shadowing at Trusts
International applicants	Low	Need to understand the process they will need to	Trinity Dublin to pilot SJT and EPM and provide feedback -

Stakeholder	Current understanding	Desired understanding/behaviour	Mechanism
		follow when applying to the Foundation Programme	liaise with them via emails and meetings Information on the UKFPO website – but only when necessary as this is likely to cause confusion if the information is provided when not relevant Liaise with AMEE to inform of changes
SJT Item Writers	Unknown	High	Mainly through email. This includes emails about future item writing workshops and also the QA process for items. All previous item writers will be invited back to additional workshops and new ones will be contacted through existing networks

Appendix three

Stakeholder meetings

The table below outlines some of the stakeholder meetings which may require a project update.

Meeting	Dates
Project Group meetings	24 March 2010
	23 April 2010
	25 June 2010
	15 July 2010
SJT Call Updates / Checkpoints	09 March 2010
	23 March 2010
	07 April 2010
	21 April 2010
Medical Schools Council - Executive	12 March 2010
	23 April 2010
	25 June 2010
	10 September 2010
	22 October 2010
Medical Schools Council - Council	13 & 14 May 2010
	01 October 2010
	24 - 26 November 2010
Foundation School Directors	15 April 2010
	16 June 2010
	15 September 2010
	18 November 2010
Foundation School Managers	25 February 2010
	05 May 2010
	30 June 2010
	01 September 2010
	13 October 2010
UKFPO Medical Students Board	11 March 2010
	17 September 2010
UKFPO Rules Group	08 April 2010
	13 May 2010
	17 June 2010

	22 July 2010
	16 September 2010
	21 October 2010
	25 November 2010
	16 December 2010
MEE	24 April 2010
	02 June 2010
	14 July 2010
	01 September 2010
	13 October 2010
	24 November 2010
BMA MSC	17 March 2010
	08 April 2010
	9 & 10 April
	27 April 2010
	30 May 2010
	05 June 2010
	25 June 2010
	15 July 2010
Medical Workforce Forum	19 May 2010
	18 August 2010
	10 November 2010
NACT UK	22 June 2010
	13 October 2010
GMC Undergraduate Board	19 July 2010
	14 October 2010
GMC Postgraduate Board	17 June 2010
	28 September 2010
UK Scrutiny Group	02 June 2010
	29 September 2010
	17 November 2010
Medical Programme Board	02 June 2010
	14 July 2010
	01 September 2010
	13-Oct
	24-Nov

Appendix four

Key milestones

The key milestones table outlines the most important activities and the communication mechanisms used to raise awareness of these. A more detailed timeline is available in appendix three.

Key milestone/activity	Date	What information needs communicating	Communication mechanisms
Launch			
General information on the project	March/April 2010	What is being piloted Who is involved Impact for stakeholders Information on pilots When the system will be implemented if proven successful	- Website - Facebook - Letter to students - Podcast with Nick Deakin interviewing Paul O'Neill - FAQ document - Newsletter - Meetings – UKFPO & BMA MSC - Information on medical school intranets/ in newsletters - Printed business cards/leaflets possibly included in the Student BMA News
Pilots			
Micro pilots	March/April 2010	Which medical schools are involved When they are What the pilots will involve Outcomes/next steps	- Website; including information on who involved, the outcomes and a student's experience of the pilot - Facebook - Newsletter - Updates at meetings including BMA MSC and Medical Schools Council
Mini pilots	Autumn 2010	Which medical schools are involved When they will be held Who to contact if any questions Why students should get involved Format of the pilots Outcomes/next steps	- Website; including information on the process, the outcomes and the way forward - Facebook - Letter to students outlining what - Briefing packs to schools involved - Newsletter - Press – BMA Student News and Student BMJ - Intranet update for schools

			involved in pilots - Workshops at medical schools in partnership the BMA at the beginning of term - Possible presence at a careers fair - UKFPO MSB and BMA MSC meetings
Large pilot	Spring 2011	Which medical schools are involved When they will be held Who to contact if any questions Why students should get involved Format of the pilots Outcomes/next steps	- Website: including who is involved, the outcomes and what will be recommended - Facebook - Letters to students - Briefing packs to schools involved - Newsletter - Press – BMA Student News/BMJ - Workshops - Intranet and newsletter updates - UKFPO MSB and BMA MSC meetings - Outcomes – decisions made: what is happening with next stage - Large stakeholder event outlining the results from the pilots and what will be recommended in the final report
Job analysis			
Published job analysis	July 2010	Final version of the job analysis and how work was carried out	Website update Highlight at any relevant meetings e.g. Foundation School Directors and UKFPO Rules Group
Item writing workshops			
Completion and outcomes of workshops e.g. question styles, number of questions produced etc	May 2010 and Autumn 2010	That the item writing workshops have finished for that particular session. Include information on how successful the item writing sessions were, what was involved, who was involved, how many items were written the style of the questions	Website Include update in the next newsletter
Outcomes from workshops e.g.	May 2010 and	How successful the item writing sessions were, what	

question styles, number of questions produced etc	Autumn 2010	was involved, who was involved, how many items were written and the style of the questions	
Number of questions written	When milestone reached	Information on how many questions have been written	Update on website
Final report			
Pre-announcement	June 2011	The pre-announcement and key information that will be included in the final report	Website Newsletter At stakeholder event Facebook
Final report	July 2011	Final report outcomes and recommendations	Website Newsletter At stakeholder event Facebook Stakeholder meetings Press event – conference?

Appendix five

Timelines

The timelines have been produced in conjunction with the key milestones and show when the different communication activities will take place (by calendar year). A detailed timeline outlining specific deadlines and those involved

2010

March	April	May	June	July	August	September	October	November	December
Website amendments including new web address and information updates Set up Facebook group Update at UKFPO MSB Update at FSD and FSM meetings	Sort out generic admin@isfp.org.uk email address Produce a podcast between Paul O'Neill and Nick Deakin in collaboration with the BMA MSC Email students for feedback on the micro-pilots and ask them to write something on their experiences for the website	Publicise podcast – BMA TV, UKFPO and BMA MSC e-bulletins, newsletters, Facebook and website Newsletter update focusing on micro pilots, podcast, Facebook page and discussion forum, next stages and information on the item writing	Liaise with patient groups – send them information on the project Contact Careers Advisors at each Medical School with information, including leaflets, ready for the 2010/2011 term Send project update to	DEPENDING ON WHEN ANNOUNCED - COULD BE AUGUST: Update website to confirm which schools are involved in the pilots Student press update? Send briefing packs to schools involved in pilots including	Contact communication teams at the medical schools involved in the pilots to get the information on the internet and in the first newsletter of the term Hold workshops with schools about the pilots to ensure everyone understands what needs to	Provide update for UKFPO MSB on 17 September to help gage feedback and opinions from the group Newsletter launching pilots and welcoming students to the new term Contact SHA Communications teams with information on the project	Workshops at medical schools to provide an update to students on the project in conjunction with BMA MSC to help gage feedback and opinions from the group (this could possibly include non-BMA students on this occasion). COULD BE SEPT – date	Update website to include the outcomes of the pilots, next steps, what was involved, what did work and what didn't work. Student press update on the pilots?	Send project update to COGPED Update Trusts via NHS Employers and AUKUH

	Contact different stakeholders to get link to the ISFP website on their websites Produce FAQ document and send to Project Group, UKFPO and BMA MSC for feedback and then disseminate to staff and students and put on websites Produce business cards, leaflets and posters and enquire about including them in BMA Student News and Student BMJ. Send to medical schools. Contact school communication teams about getting information on their intranet and in newsletters	workshop and to help gage feedback and opinions from group Post summary information on micro-pilots and student feedback on website (link in newsletter) Website update on next stage of pilots and when they will happen Letter to students after micro-pilots with update – if necessary	COGPED Liaise with Trusts via NHS Employers and AUKUH	information on what they will need to do, FAQs, resources needed, etc. Update Facebook on which schools will be taking part in pilots and help gage feedback and opinions from people about the pilots Post the outcomes of the job analysis and updates from the item writing workshops on the website	be done		tbc Attend careers fair?		

2011

January	February	March	April	May	June	July
Book date and find venue for large stakeholder event in June (check against dates of finals etc as need student representation)	Publicise large scale pilot (possibly in Jan or March – tbc) via intranets and newsletters of those involved in pilots and briefing pack Newsletter focusing on announcement of large pilot Update website and Facebook Publicise event in June – save the date email, including invitations to speakers	Update for the UKFPO MSB to help gauge feedback and opinions from the group Workshops with students (similar to ones in Autumn 2010) to provide an update on the project	Large scale pilot Survey of those involved in the pilot?	Newsletter – focusing on the outcomes of the pilots and the last stage of the project	Large stakeholder event in the first week of June. What the pilots concluded, summary of work throughout the project, feedback and comments, final recommendations Use technology e.g. voting buttons/notepads Speakers could include Pat Hamilton, Paul O'Neill, UKFPO on taking the project forward Pre-announcement of final report?	Submit final report Press engagement – this could also be done once the report has been approved Update website Letter to students outlining the next stages of the project and handing it over to the UKFPO Update Facebook Newsletter focusing on key issues

Additional activities

Some information/activities that will be carried out in addition to the activities outlined above are shown below:

Information on communication activities for EPM will be added once more information is provided.

Facebook will be updated when necessary but it may be beneficial to provide a small update on the group page so that it is active and encourages people to join.

Updates will be sent via UKFPO and BMA MSC e-bulletins when required.

If information needs to be communicated which is not mentioned above this will be done via newsletters, letters to students and updates via the website.

As a general rule the newsletters will be sent quarterly.

Updates will be provided, when required, to/at the meetings of the various stakeholders – as outlined in appendix three.

An international pilot will be held at Trinity College Dublin. Liaison with the international students will happen once pilot dates are confirmed and when more information is available. This will be updated on the website. The UKFPO will post a link and information on its website when the time is correct, so as not to confuse international applicants on the process.

Liaison with those groups who are not involved in the process but will benefit from an update will be carried out during those months where there are less key activities being carried out.

All feedback and comments received via the communications mechanisms will be fed back to the Project Group at the regular meetings.