ACADEMY OF MEDICAL ROYAL COLLEGES

The Academy of Medical Royal Colleges

in collaboration with

The Division of Medical Education at UCL,

The Acute Specialties Pilot team

and

Work Psychology Group

ST1 Selection Pilot 2010

Project Report

August 2010

Summary
This report describes a pilot project for selection into Specialist Training 1 (ST1) undertaken by a consortium led by the Academy of Medical Royal Colleges (AoMRC) on behalf of the Department of Health.

The aim of the project was to determine whether a common computer delivered Clinical Problem Solving (CPS) test could contribute to a selection process for ST1 in a range of specialties and deaneries. Six specialties – Anaesthesia, Acute Care, General Practice, Histopathology, Medicine and Paediatrics agreed to participate. The work was undertaken in collaboration with a separate investigation into the use of Situational Judgement Tests (SJTs) in selection to Anaesthetics and Acute Care Common Stem (ACCS). All the English deaneries participated in the investigation while many of the remaining specialties and the Scottish deaneries were involved with the project Steering Group and the workshops.

The CPS test was specially devised for this project and consisted of 120 single best answer questions, to be completed in two hour test slots at invigilated test centres around the country. All applicants to the six participating specialties were invited to take this test in one of a choice of 24 test centres. A 45 item, 60 minute SJT was constructed by the Anaesthesia/AC Pilot team from a range of questions validated in previous pilots. All applicants to Anaesthesia and ACCS training were invited to take this test immediately prior to taking the CPS test.

The tests were staged over a two day period in early January 2010, to coordinate with the process the specialties and deaneries were using to select ST1 candidates. Candidates were advised that the results would not influence their chances of being selected for a training post More than 1300 doctors registered to take the test. A number of those who did book a test failed to attend, in the main due to adverse weather conditions, but 666 of the ST1 candidates were successful in completing one or both tests.

Candidates who completed the tests were invited to provide feedback on the experience. Two stakeholder workshops were held to solicit views on the usefulness of computer delivered tests in selection and to share expertise in the development of SJTs.

The SJT and CPS test results were analysed in a variety of ways including comparison of the rankings they produced with those generated by the processes actually being used by the specialties and deaneries to select candidates for ST1. Both tests had good predictive validity for the eventual outcome of the recruitment process, regardless of specialty. A combination of CPS and SJT would produce the most reliable result in a live selection process.

Other data analyses included an investigation of gender, age, ethnicity and country of primary qualification. There was little difference in terms of age or gender, however doctors that qualified in the UK scored significantly higher in both tests.

Conclusions

a. It is feasible to hold SJT and CPS tests as part of the ST1 selection process – it would be time and cost efficient and would accurately discriminate between candidates.
b. These tests are acceptable to candidates.
c. The specialties and deaneries worked well together.
d. Use of either an SJT or a CPS test in an early stage of the selection process would increase predictive validity and the combination of both types of test would enhance this.
e. The cost of staging such tests would be in the region of £55-65/candidate, with lower costs, but greater effort, being incurred if local test venues were used instead of managed test centres.
f. Computer delivered tests can be used to replace shortlisting processes currently employed by many specialties and deaneries.
g. There is currently wide variation in the performance of shortlisting methods in specialties that are not part of a national standardised process.

Recommendations

a. An SJT and CPS test, probably common to a group of related specialties, should be an integral part of the National Health Service's ST1 selection procedures in future years, most probably as a component of candidates' overall Selection Process Scores.
b. This overall score should also take account of performance in a Selection Centre and qualifications, experience and research interests, as set out in their (simplified) portfolios.
c. Elimination of the shortlisting process, so that every candidate received at least one interview, would enable the tests to be staged at the same time as interviews or Selection Centres, thereby streamlining the process.
d. The combination of these machine markable selection instruments would provide a fair test, and reduce the expense and workload associated with the current processes.
e. Machine markable tests can potentially eliminate the variation in shortlisting methods between regions and provide a test which is objective, reducing the subjective nature of shortlisting by multiple raters in Deaneries across the country.

Acknowledgements
We are grateful to the Department of Health for funding this pilot, and to the many collaborators who participated in the test development, delivery and analysis of results. The main members of the collaborating teams are:

AoMRC core working group team: Professor Jane Dacre (lead), Dr Alison Sturrock, Dr Luci Etheridge, Dr Hilary Spencer, Dr Katherine Woolf.

Acute Specialties Pilot team: Dr Tom Gale (lead), Dr Alison Carr, Mr Martin Roberts, Dr Ian Anderson

Work Psychology Group: Professor Fiona Patterson (lead), Ms Victoria Carr.

Pearson Vue: Mr Thierry Berthou, Mr Ivan Coutinho.

Steering group members:

Bill Irish (Chair)	GP NRO
Maggie Blott	RCOG
Steve Buggle	DoH
Colin Campbell	RCPCH
Alison Carr	DoH
Victoria Carr	WPG
Jane Dacre	UCL
Nicola Dagnall	London Deanery – Histopath. recruitment
Tom Dolphin	BMA JDC
Neil Douglas	AoMRC
Gai Evans	GP NRO
Siobhan Fitzpatrick	MSC
Ashley Fraser	NHS Employers
Tom Gale	RCoA Acute Specialties pilot
Sue Heenan	RCR
Sarah Hill	RCPath
Carole Jones	RCOpht
Malcolm Lewis	COGPED
Damien Longson	RCPsy
Graham Muir	RCPCH
Chris Munsch	JCST
Jonathan Myers	ATDG
Fiona Patterson	WPG
Simon Plint	DoH
Vicky Ridley-Pearson	KSS Deanery
David Sowden	COPMeD
Hilary Spencer	Project Manager
Alison Sturrock	UCL
Gary Waltham	London Deanery – Histopath. recruitment
Tom Waterman	RCP
Clare Wedderburn	GP Stage 2 item writing group
Kevin West	RCPath

Participating Colleges:
College of Emergency Medicine
Royal College of Anaesthetists
Royal College of Paediatrics and Child Health
Royal College of Pathologists
Royal College of Physicians
Royal College of General Practitioners

Participating Deaneries:
East Midlands Healthcare Workforce Deanery (North & South)
East of England Deanery
KSS Deanery
London Deanery
Mersey Deanery
Northern Deanery
North Western Deanery
Oxford Deanery
Severn Deanery
South West Peninsula Deanery
West Midlands Deanery
Wessex Deanery
Yorkshire & The Humber Deanery

In particular, we would like to thank all those in the deaneries and specialty recruitment organisations for posting notices about the pilot on their web sites, sending out numerous e-mails encouraging candidates to volunteer, and responding to the onerous requests for data about the tested candidates' results in the real process

We would also like to thank all the technical and administrative staff at Sheffield Hallam University and University College London who organised our university hall venues and helped to set up and run the technical facilities there.

Contents

1. Introduction to AoMRC pilot
Professor Jane Dacre

This pilot was commissioned by the Department of Health to contribute to the developing evidence base around the use of computer delivered tests in selection to postgraduate training programmes for doctors. The consortium responsible for this pilot is drawn from clinicians, clinical academics and psychologists with experience of the development and implementation of computer delivered tests for selection across a number of specialties.

The purpose of this pilot was to prove the concept and add to the evidence base by the development of a generic CPS test blueprinted on to the Foundation Programme curriculum. We were also evaluating whether it was feasible to host a computer based test in invigilated centres. We incorporated a new situational judgment test (SJT) component, aimed at candidates applying for Acute Care Common Stem (ACCS). The original project proposal was to develop a generic CPS with families of SJT questions for groups of specialties. Delays in securing the funding made it impossible to deliver this within the timescale so a compromise position was reached where we developed a generic CPS and worked with an existing SJT developed and already piloted for ACCS.

The tests were administered in parallel with the existing application process to ST1 in participating specialties, and in all English deaneries.

The design of this pilot was influenced by consortium members who have previously developed computer delivered tests and also by the National Recruitment Office for General Practice (NRO) who have successfully implemented computer delivered tests (CPS and SJT) for GP selection.

The introduction of a reliable computer delivered test blueprinted on to the Foundation curriculum will increase the robustness of the selection process overall. Early results from recent CPS and SJT pilots suggest that candidates will think that it is fairer. Entry to many specialties is competitive, with a large volume of potential applicants. This is a significant administrative challenge to the deaneries. The addition of computer marked tests will make the process more efficient as the use of these tests alongside a streamlined national process will reduce the current duplication of effort between individual deaneries.

The use of a generic test as a selection instrument is applicable to all units of application. It has already been use successfully in General Practice but could however be extended to cover all specialties and applications to posts at all levels.

A well validated generic test has the potential to be applied to cohorts of candidates at any level of application. This pilot was designed for selection at the end of F2, but has the potential to be applied to other levels.

The issues being explored by this pilot are:

- The collaborative approach of the participants in the development and staging of the test
- Its multispecialty nature, and the introduction of a common test
- The addition of a SJT component for ACCS
- The introduction of computer delivered testing at a series of test centres throughout England
- The aims of the pilot were:
- To develop and implement a generic clinical problem solving test for evaluation as a shortlisting tool for specialty training
- To pilot a computer delivered test in a number of venues around the UK
- To compare computer delivery with paper and pencil delivery of the test
- To build on the collaboration with stakeholders necessary to deliver a national project across deaneries and specialties.
- To evaluate the test as a selection instrument, and to consider its acceptability to potential candidates and other stakeholders
- To stage a training workshop on SJT as a selection method as part of the project.
- The project was hosted by the Academy of Medical Royal Colleges. A project steering group was set up to oversee the running of the project. This had representation from all interested specialties. Weekly telephone conference calls were held to implement the project plans.

2. Previous research on SJT and CPS Tests
Professor Fiona Patterson, Ms Victoria Carr

Introduction to CPS and SJT
Clinical Problem Solving (CPS) and Situational Judgement Tests (SJT) are examples of machine markable tests (MMTs, i.e. tests that could potentially be marked by machine) used in selection. MMTs offer standardised, reliable tools which can substantially increase the utility of a selection process, for example by reducing assessor time. MMTs developed according to established test development principles can provide valid and robust selection instruments.

Clinical Problem Solving tests are designed to assess the application of clinical knowledge. Typically, CPS tests use multiple choice questions (MCQs) which might involve solving a problem, reflecting a diagnostic process or developing a management strategy for a patient.

Situational Judgement Tests are a measurement method designed to assess individuals' judgement regarding situations encountered in the workplace. Candidates are presented with a set of hypothetical work based scenarios and make judgements about possible responses, which are evaluated against a predetermined scoring key. SJT scenarios are based on detailed analysis of the job role, to ensure that test content reflects the most important situations in which to test candidates' judgement. SJTs can be designed to assess a variety of job-relevant professional attributes.

SJTs have become increasingly popular over the last 20 years in large scale selection in many occupations. The research literature indicates that SJTs have significant validity in predicting job performance and can offer incremental validity over other methods, such as ability tests and personality questionnaires[1, 2]. SJTs tend to show smaller differences between candidate groups (based, for example, on race) than cognitive ability tests and favourable candidate reactions because they appear directly relevant to the job role[3, 4].

Development of CPS and SJT for General Practice
In 2005, COGPED (Committee of General Practice Education Directors) began to develop and pilot a machine markable test for shortlisting into GP training, comprising a CPS test and SJT. The aim was to provide a reliable, valid, standardised and efficient way to assess over 6000 candidates per year against key person specification criteria such as clinical expertise, empathy, integrity and resilience.

The GP CPS was initially developed from an existing clinical MCQ paper developed in North Western Deanery. It assesses the application of clinical knowledge to solve patient problems, based on 12 clinical topic areas relevant to GP. A group of subject matter experts (senior GP educators) worked with three psychologists to generate, review and refine a bank of questions according to an agreed test specification. Following an initial pilot in several deaneries in 2005, the CPS went live nationally in spring 2006, initially alongside invigilated structured application form questions.

The GP SJT ('Professional Dilemmas') was developed in 2006 to provide a reliable and valid assessment of key non clinical domains (empathy, integrity and resilience) and replace structured application form questions. As with the CPS, a group of subject matter experts worked with psychologists to develop test content according to an agreed specification. Following a series of pilots in 2006, the SJT was implemented nationally in spring 2007. This was the first known application of an SJT in postgraduate medical selection, although they had previously been used in medical school admissions[5]. There is an ongoing cycle of development and piloting to expand the item bank for both tests and create new test papers each year.

Research conducted in 2006[6] demonstrated that the CPS, SJT and structured Application Form Questions were all effective predictors of candidate performance at the GP selection centre. The SJT served as the best single predictor; however the optimum shortlisting methodology in terms of effectiveness and efficiency was a combination of SJT and CPS.

The development and successful implementation of the GP shortlisting process has been a significant innovation in UK postgraduate specialty selection. The shortlisting process has enabled all candidates to be assessed on the same day across the UK, nationally ranked and allocated for selection centre to their highest preference deanery. Evaluation of the process has shown increased reliability of selection and increased predictive validity of performance during GP training[7]. Evaluation of the national GP selection system from the candidate's perspective has also consistently demonstrated that candidates have confidence in the relevance and fairness of the process[8].

A retrospective evaluation and national pilot of the GP CPS and SJT with candidates applying for Core Medical Training (CMT) in 2008 and 2009 demonstrated that both tests showed good reliability and predictive validity in a CMT sample, with incremental validity over the current CMT shortlisting process[9].

3. Project logistics
Dr Hilary Spencer

The DoH confirmed the aims and funding of the pilot in July 09. This meant that there were six months to arrange and stage the main portion of the pilot and a further six months to analyse the data. Below is a list of the logistics that were encountered over the course of this period.

1. Between the start of the project and the first day of the test, the team undertook a number of parallel streams of activities including:
   - Negotiating with specialties and deaneries to participate in the pilot and designing a process which would be acceptable to them all
   - Obtaining a suitable software platform
   - Choosing and setting up test venues
   - Contacting all potential applicants and informing them how to take the test
   - Constructing and validating the test (as described in Section 3 and 4 of this report)
   - Implementing the communication strategy
   - Manning help lines (telephone and e-mail) for up to 12 hours/day to deal with queries from the candidates who were interested in taking the test and from the IT service suppliers who were completing their work on the platform.
2. Two test events were staged in university halls in Sheffield and London.
3. Results were assembled and sent to candidates along with a certificate and, in three randomly selected cases, small prizes following completion of the live process.
4. Analysis of the results included:
   - Question performance
   - Comparison with scores
   - Demographics of candidates
   - Feedback obtained from the candidates.
5. A workshop was held on the construction and use of situational judgement tests
6. A stakeholder workshop was held to describe the process and emerging results

Participating deaneries and specialties
A power calculation was used to estimate the number of test candidates needed to achieve statistically significant correlations between scores and potential selection outcomes (Table 1).

Actual correlation	Lower confidence interval desired		
	-0.2	-0.1	-0.05
0.5	*not calculated*	67	243
0.6	17	52	182
0.7	13	36	122
0.8	10	22	67
0.9	6	11	25

Table 1: **Power calculation**

Six specialties participated: General Practice (GP), Medicine (CMT), Anaesthetics and ACCS (in collaboration with the Acute Care Specialties Pilot), Paediatrics and Histopathology. All applicants to all participating specialties were invited, from all of the English deaneries.

Designing the pilot process
The CPS test was constructed to reflect core skills expected of all ST1 applicants and based on the Foundation Programme curriculum. SJT questions were provided from the bank of questions developed and validated for the Acute Specialties Pilot Project.

All applicants were asked to affirm their consent by tick box as part of the on line registration process for the tests. Ethics approval was granted by the UCL research ethics committee.

Candidates were invited to undertake the test **in parallel with** their actual applications to the deaneries and specialties concerned. Rankings for the two different processes were compared. The test had to precede the participating specialties' shortlisting results; they were therefore held on Friday January 8th and Saturday January 9th 2010.

Selecting the platform
Four suppliers of the test platform were considered: Moodle, PrimalPictures, Questionmark and Pearson Vue. All met our criteria regarding functionality and timeframe but only Pearson Vue had a well established network of test centres.

Inviting the candidates
Applicants to the participating specialties' ST1 programme were invited, either by e-mail or via a web link, to take the CPS test. Those applying for Acute Care (Anaesthesia or one of the ACCS specialties) were invited to take an SJT in addition to the CPS. Candidates could choose test day, time and location (Figure 1).

Figure 1: Chosen Pearson Vue locations

Help lines
Help lines (telephone and e-mail) dealt with candidates' queries and problems. Lines were open for 12 hours each day until two weeks after the test was staged. Queries included: problems with the e-mails which we sent out, request for further details about the tests, cancellations because of snow and requests for feedback on their performance.

Candidate booking
At initial booking, candidates were asked to provide information on their country of primary medical qualification (PMQ), sex, ethnicity (using UK Census 2001 codes) and age at time of testing. They were also asked which speciality/specialties they had applied to, and if this was more than one specialty, they were asked to name their preferred specialty.

Managing the test centres
Tests taken in Pearson Vue centres were managed by Pearson Vue staff. Tests held in the university halls required technical support and invigilation by the project team.

Economic considerations
There were seats for 400 candidates in the university halls but only 22 candidates chose to book at SHU and 70 at UCL. University halls generated a fixed cost whereas Pearson Vue charged per candidate.

Test day challenges
Over 1400 candidates registered and almost 1,000 tests were booked. However, on the week of the test, England experienced its worst snow for many years so many candidates cancelled or did not turn up. At the test centres, a few candidates brought the wrong (or no) identification, arrived late (often because of the weather) or turned up without previously booking.

4. The CPS test

Dr Alison Sturrock, Dr Kath Woolf, Ms Victoria Carr

Purpose, format and content

A Clinical Problem Solving Test (CPS) is an assessment method which measures a candidate's ability to apply knowledge in a clinical scenario.[10] Specific features that it is designed to test include:

- Gathering and using data required for clinical judgement
- Choosing examinations, investigations and interpretation of clinical findings
- Applying knowledge
- Demonstrating diagnostic skills
- Ability to evaluate undifferentiated material
- Ability to prioritise
- Making decisions and demonstrating a structured approach to decision making
- Ability to apply ethical guidance to clinical scenarios

This chapter reports on the development of the CPS test and the test specification, the statistical analyses relating to the CPS test and item analysis and the comparison with results achieved at shortlisting. Candidates' perceptions of the face validity of the tests and of the administration of the tests are presented in Chapter 8.

Test format

The CPS test used in this pilot was based on similar assessments used for this purpose including the 2008 selection project[11] and National Recruitment Office for General Practice. The test was 120 single best answer items (best of 5), computer delivered and time limited to two hours.

Content relevance and selection of items

The Academy of Medical Royal Colleges' assessment working group identified named content/subject experts from participating Colleges to each select 30 appropriate and validated questions. The questions were requested to be at the appropriate level of training and needed to have individual item statistics including facility and discrimination indices.

We received questions from the Acute Specialties selection project, the Royal College of Psychiatrists, the Royal College of Physicians, The Royal College of Paediatrics and Child Health, the National Recruitment Office for General Practice and the Royal College of Pathologists. The remainder of the items were selected from a bank of questions used in a previous selection pilot.

All items were selected by the assessment working group based on their relevance to foundation year doctors. This group then blueprinted all items on to the foundation program curriculum to ensure an adequate breadth of coverage across all areas and specialties.

Computer delivered CPS
The test was delivered either in Pearson Vue assessment centres in 22 venues across England and in 2 computer-enabled university halls in London and Sheffield. All of these centres used the same Pearson Vue platform to deliver the test electronically. In the Pearson Vue centres, candidates could book time slots over a 2 day period. In the university halls, the test could be taken either at 11:00 or 14:00 on January 9th 2010.

Candidates
The candidates were all volunteer doctors who had applied for at least one ST1 post in 2010. Applicants in ACCS, Anaesthetics, Medicine, Paediatrics, Psychiatry, Histopathology or General Practice in the UK were all eligible.

Candidate demographics
667 candidates took the CPS test. Data relating to scores were excluded for three candidates: one took the test twice (only the first result was used) and two spoiled their papers (one candidate skipped 62 items; the other took only 27 minutes to complete the test). The demographic data of the candidates with spoiled papers was included in the analysis.

Descriptive data for candidate factors and outcome measures were conducted, followed by inferential statistical analyses of the relationships between candidate factors (demographics, preferred specialty) and outcome measures using t-tests, ANOVA, Pearson's r and chi-squared tests as appropriate. Multivariate analyses (multiple regression and multivariate ANOVA) were conducted to determine the relationships between candidate factors and outcome measures i.e. to determine the relationships between sex, ethnicity, PMQ, age, preferred specialty, test scores, and survey results.

Parametric tests were used except when assumptions of Normality were violated, in which case non parametric equivalents were used. Statistical significance was set at alpha=0.05.

Region of primary medical qualification (PMQ)
Seventy percent (466/665) of candidates qualified in the UK. Four percent (27/665) qualified in the EEC and the remaining 26% (172/665) qualified outside the EEC.

Due to small numbers in the EEC group, in many subsequent analyses the EEC and other categories were amalgamated. When appropriate, analyses were repeated with both variables to check for possible differences.

Sex
There were 263 (40%) males and 402 (60%) females.

Ethnicity
655 candidates reported their ethnic group. Due to small numbers in some groups, in subsequent analyses the ethnicity variable was converted into : Ethn4 ('white', 'black', 'Asian', 'Chinese', 'mixed' 'other'), Ethn2 ('white'; 'Asian excluding Chinese' and 'else'), Ethn3 ('white' and 'non-white'). When appropriate, analyses were repeated using each variable to check for discrepancies.

Thirty seven percent (246/655) of candidates were white British or white Irish, and 31% (209/655) were Asian (Indian, Pakistani, Bangladeshi or Asian other). Overall there were 301 white candidates and 354 non white candidates.

Within UK qualified doctors, 57% (260/460) were white and 25% (115/460) were Asian. Of the doctors who qualified outside of the UK, 48% (94/195) were Asian, and 5% (41/173) were white, 23 of whom qualified in the EEC but outside the UK.

Ethnic group	UK	Non UK	Total
white British	239	4	243
white Irish	3	0	3
white other	18	37	55
black British	8	5	13
black African	6	31	37
Indian	61	37	98
Pakistani	25	39	64
Bangladeshi	2	5	7
Chinese	30	1	31
Asian other	27	13	40
Mixed	22	4	26
Other	19	19	38
Total	460	195	655

Table 2: Ethnic group shown by region of primary medical qualification.

Age
614 candidates reported their age. The median age was 27 (range 23-63). Due to the skewed nature of the data, non parametric statistics were used to calculate age differences between different groups.

Male candidates were on average older [median male=28 years; median female=26 years; Mann Whitney U z=3.7; p<0.001], as were non UK trained candidates [median UK=26 years; median EEC=31 years; median non EEC=34; Kruskal-Wallis Chi squared=221.0; p<0.001].

Preferred specialty
General practice was the most popular specialty (350/665; 53%), followed by Core Medical Training (128/665; 19%). Around four percent (28/665) had no preference. Only four candidates chose psychiatry as their preferred specialty.

A chi-squared test showed no significant sex differences in preferred specialty. However, there were differences by PMQ with non UK qualified doctors being less likely to prefer ACCS and more likely to prefer GP (chi-squared= 26.7 df=7 p=0.002).

Due to small numbers, ACCS and Anaesthetics were combined, as were Paediatrics, Histopathology, Psychiatry and no preference, giving a four category variable (CMT, GP, ACCS & Anaesthetics, Else). Repeating the analyses with this variable produced similar results, with non UK graduates being less likely to prefer ACCS + Anaesthetics (p<0.001) (Table 3).

Preferred specialty		UK	Non UK	Total
ACCS & Anaesthetics	Observed	88	14	102
	Expected	*71.5*	*30.5*	*102.0*
General Practice	Observed	224	126	350
	Expected	*245.3*	*104.7*	*350.0*
Core Medical Training	Observed	96	32	128
	Expected	*89.7*	*38.3*	*128.0*
Else	Observed	58	27	85
	Expected	*59.6*	*25.4*	*85.0*
Total	Observed	466	199	665
	Expected	*466.0*	*199.0*	*665.0*

Table 3: Preferred specialty by PMQ (UK versus non UK)

Evaluation of CPS
The descriptive statistics for the clinical problem solving test are shown in Table 4 below.

Number	Minimum score (%)	Maximum score (%)	Mean (%)	Std. Deviation (%)
665	*20*	*90*	*65.7*	*11.6*

Table 4: CPS test descriptive statistics

Scores were roughly Normally distributed (mean=65.7%; SD=11.6%) (Figure 2).

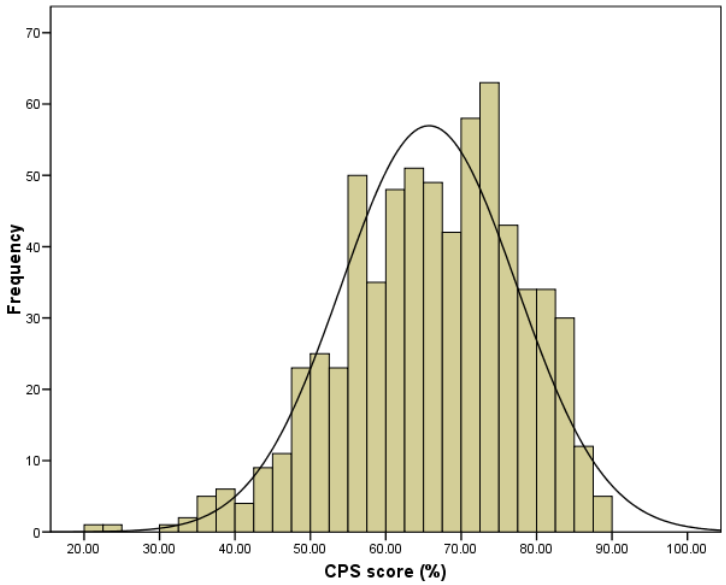


Figure 2: Distribution of CPS scores

The CPS test had a Cronbach alpha of 0.89 (95% confidence interval 0.88 – 0.90) and a SEM of 4.71. Thus it demonstrated a good degree of reliability.

Table 5: Summary item statistics

	Minimum	Maximum	Mean	Std deviation
Facility index	0.24	0.97	0.66	0.17
Point biserial	-0.05	0.46	0.27	0.10

The frequency of p values (facility indices) and point biserials are shown in Figures 3 and 4.

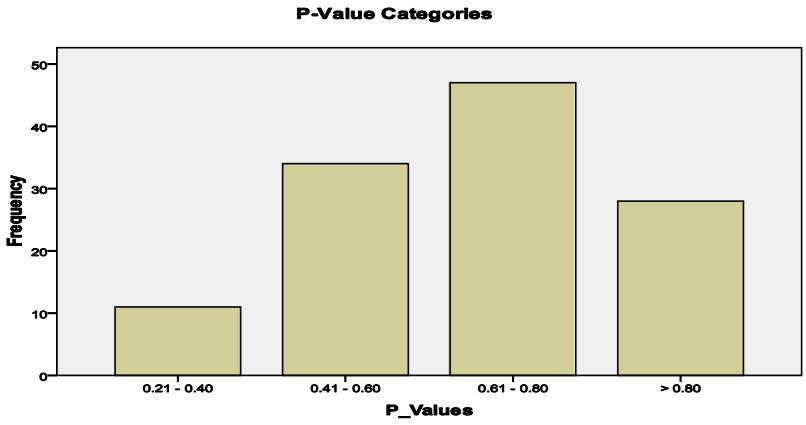


Figure 3: Distribution of facility indices for CPS

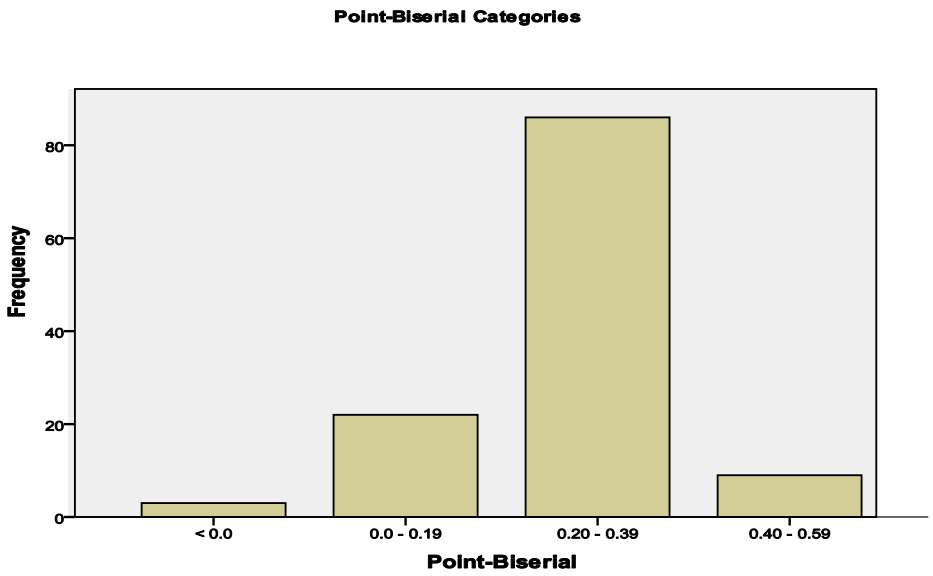


Figure 4: Distribution of point biserials for CPS

Items with a facility index (p-value) less than 0.5 or a negative point biserial were identified and checked to ensure there were no keying errors.

Time taken to complete the CPS test
Candidates took approximately 100 minutes to complete the CPS test (mean=99.0 minutes; median=103.7 minutes; SD=18.4), with a range of 45 to 120 minutes. As expected, the time data were very negatively skewed – see Figure 5.

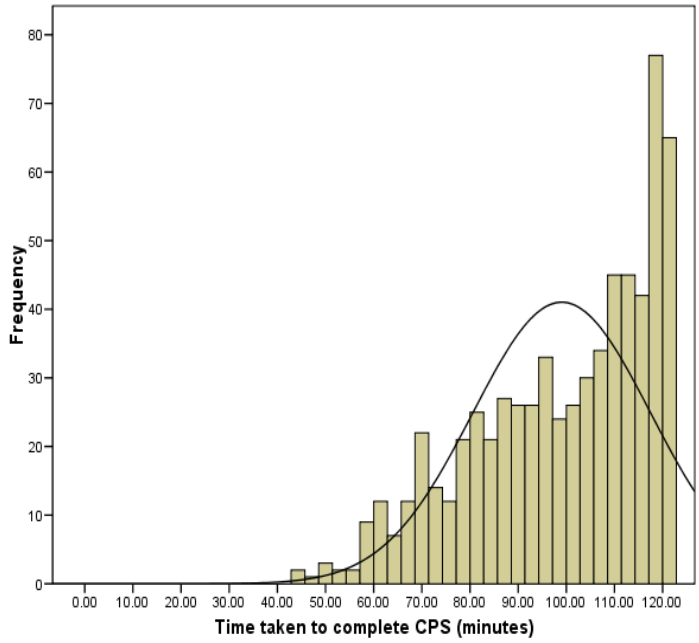


Figure 5: Time taken to complete the CPS test

Men (median=110 minutes) took on average twelve minutes longer to complete the CPS than women (median=99 minutes), a difference which was statistically significantly different [Mann Whitney U z=-5.0; p<0.001].

Doctors who qualified outside of the UK also took significantly longer to do the CPS (median UK=97 minutes; median EEC=111 minutes; median non-EEC=114 minutes); [Kruskal-Wallis chi-squared=79.3; df=2; p<0.001].

Non white doctors took significantly longer to complete the CPS than white doctors [Mann Whitney U z=-7.4; p<0.001]. When the analysis was restricted to UK qualified doctors, this difference remained statistically significant, with non white UK doctors taking 13 minutes longer on average than white UK doctors [Mann Whitney U z=-5.4; p<0.001]

Number of CPS items attempted
Nearly 90% of candidates (592/665) took all test items. Six candidates did not answer more than a fifth of the questions.

Table 6: No. items unanswered and mean test scores for that number unanswered

No. items not answered	No. of candidates	Mean score (SD)
0	592	70.0 (11.1)
1	16	61.3 (11.8)
2	5	61.2 (6.0)
3	4	63.3 (13.9)
4	4	56.3 (8.5)
5	5	55.8 (6.3)
6	3	65.6 (4.2)
9	6	60.3 (5.5)
10	4	55 (1.4)
12	3	60.8 (11.2)
13	1	43.3 (0)
14	2	47.1 (12.4)
15	2	48.0 (8.8)
16	4	51.2 (8.1)
17	2	56.7 (7.1)
18	3	37.8 (7.9)
19	2	58.8 (2.9)
23	1	41.7 (0)
24	1	35.8 (0)
26	2	49.6 (7.7)
29	1	33.3 (0)
35	1	51.7 (0)
41	1	20.0 (0)
Total	665	65.6 (11.9)

Region of PMQ
A one way ANOVA showed that UK qualified doctors achieved significantly higher scores than doctors who qualified either in the EEC or elsewhere in the world [$F(5,664)=120.5$; $p<0.001$]. Post hoc testing showed that there was no significant difference in the performance of doctors who qualified in the EEC or elsewhere outside of the UK (Table 7).

Table 7: Ryan-Einot-Gabriel-Welsch F post hoc test results: mean scores by PMQ region

Region of PMQ	Number	Mean CPS score	
EEC	27	55.25	
Other	172	56.74	
UK	466		69.65
p value (for differences within subsets)		0.47	1.00

Sex
Although women achieved slightly higher CPS scores, the sex difference was not statistically significant [mean female score=66.3; mean male score=64.9; t=1.6; p=0.117].

Ethnicity
A 1 x 6 ANOVA showed significant differences by ethnic group [F(5,654)=24.2; p<0.001] with white candidates achieving the highest scores. Post hoc testing using the Ryan-Einot-Gabriel-Welsch method showed that white and mixed candidates scored significantly higher than other groups. The black group scored the lowest, followed by the Asian group – see Table 8.

Table 8: Ryan-Einot-Gabriel-Welsch F post hoc test results: mean scores by ethnicity

Ethnic group	n	Mean CPS score			
		Subset 1	Subset 2	Subset 3	Subset 4
Black	50	56.98			
Asian	209		61.57		
Other	38		61.86	61.86	
Chinese	31			67.47	67.47
Mixed	26				69.04
White	301				70.05
p value (for differences within subsets)		1.00	1.00	0.09	0.67

NB: Scores in the same subset are not statistically different from one another, as shown by the p value.

Interaction between ethnicity and PMQ
A 2 x 3 ANOVA using the Ethn2 variable was calculated to determine whether ethnic differences in the data were due to differences in PMQ. The results showed a significant ethnic group by PMQ interaction on the CPS [F(2,654)=5.3; p=0.005], as well as a main effect of ethnic group [F(2,654)=5.0; p=0.007], and a larger main effect of PMQ [F(1,654)=178.9; p<0.001].

Plotting the results (Figure 5) shows that the reason for this interaction effect is the under performance of the UK Asian group, whereas the non UK Asian group achieve similarly to the non UK white group. The graph also shows that part of the reason for ethnic minority underperformance on the test is the poorer performance of non white UK graduates.

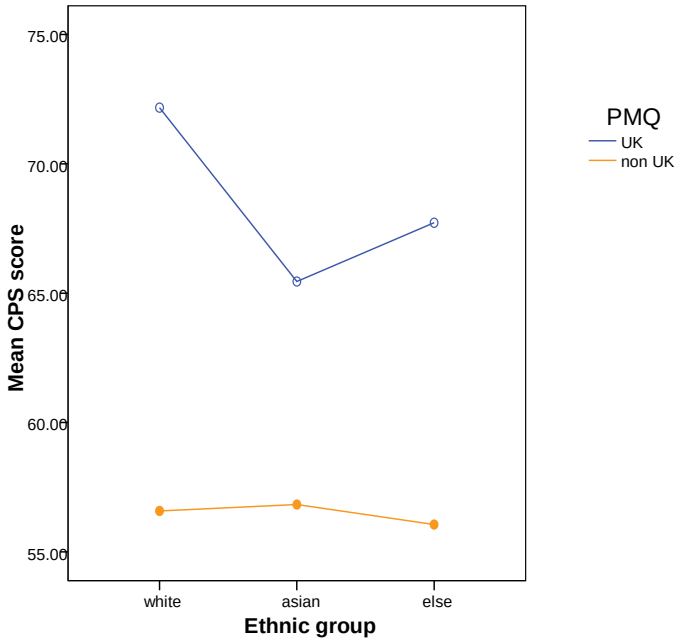


Figure 5: Interaction between ethnic group (Ethn2) and PMQ on CPS score. Lines shown to aid interpretation and do not represent continuous variables

A 1 x 3 ANOVA [$F(2,459)=25.4$; $p<0.001$] and a 1 x 6 ANOVA [$F(5,459)=11.1$; $p<0.001$] on the UK data using the Ethn2 and Ethn4 variables respectively, confirmed the relative under-performance of the UK non white candidates on the CPS (Table 9).

Table 9: Post hoc Ryan-Einot-Gabriel-Welsch F test results showing white UK candid-ates achieved higher scores than any other UK ethnic group except the mixed group

Ethnic group	Number	Subset 1	Subset 2
Back	14	63.93	
Asian	115	65.45	
Other	19	66.97	
Chinese	30	67.44	
Mixed	22	70.27	70.27
White	260		72.17
p value (for differences in subsets)		0.14	0.71

Age

Older candidates had significantly lower CPS results (Spearman's Rho=-0.39; $p<0.001$)

Preferred specialty

A one way ANOVA showed that candidates whose preferred specialty was CMT, ACCS, Anaesthetics, Paediatrics, or those who had no preferred choice of specialty achieved higher CPS scores [$F(7,664)=10.7$; $p<0.001$]. However, the numbers, particularly in the psychiatry group, were very small, so the analysis was repeated with the four category variable. This produced very similar results (Table 10) with the GP group scoring lowest and the CMT and ACCS+ Anaesthetics groups scoring highest [$F(3,664)=19.0$; $p<0.001$].

Table 10: Post hoc Ryan-Einot-Gabriel-Welsch test of mean CPS scores by specialty

Preferred specialty	Number	Subset 1	Subset 2	Subset 3
GP	350	62.83		
Else	85		66.12	
ACCS+ Anaesthetics	102		69.74	69.74
CMT	128			70.19
p value (for differences within subsets)		1.00	0.06	0.94

Regression of demographics on CPS scores

A regression of CPS score on to age, sex, ethnic group (white, non white) PMQ (UK vs non UK) and preferred specialty showed that ethnic group, PMQ and preferred specialty each had independent and significant effects on CPS score. The largest effect was of PMQ. Once other demographic factors were taken into account, age was no longer a significant predictor of CPS score. Candidates wanting to go into ACCS+ Anaesthetics or CMT achieved significantly higher scores than those wanting to go into GP, even after taking demographic factors into account (Table 11).

Table 11: Regression output showing the independent effects of ethnic group, PMQ, preferred specialty and age on CPS scores.

CPS					
Model	*B*	*Standard Error*	*Beta*	*t*	*p value*
(Constant)	91.94	2.20		41.24	<0.001
Ethnic group	-4.59	0.80	-0.20	-5.76	<0.001
Sex	-0.021	0.77	-0.01	-0.27	0.790
PMQ	-10.71	0.89	-0.42	-12.09	<0.001
Age (years)	-0.001	0.01	-0.02	-0.60	0.550
ACCS+ Anaesthetics	3.81	1.08	0.12	3.54	<0.001
CMT	5.87	0.98	0.21	5.99	<0.001
Else	1.87	1.16	0.06	1.61	0.108

The scatterplot below shows the correlation between age and CPS score in candidates who qualified in the UK (blue) and outside the UK (orange lines are Lowess curves for each subgroup. One candidate (age=63) removed from CPS results for maintenance of confidentiality)

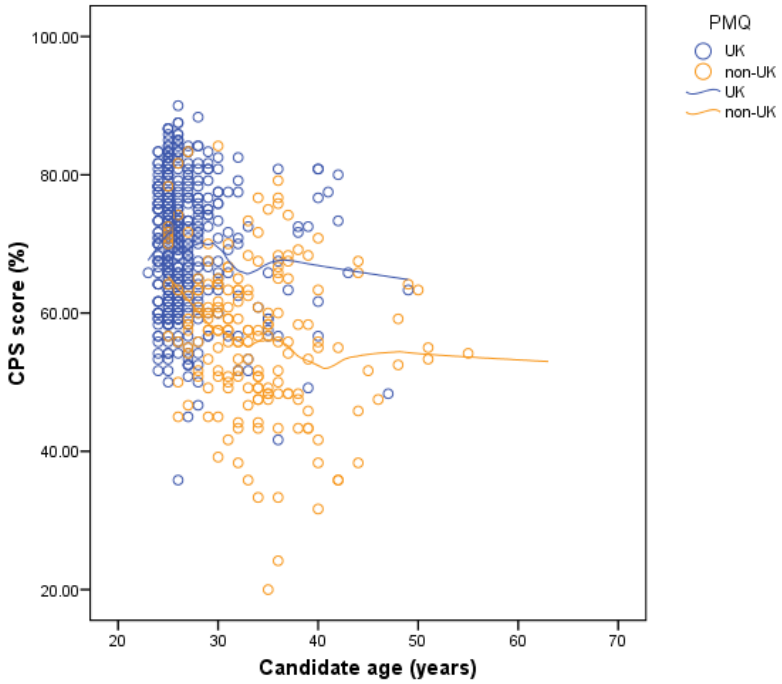


Figure 6: Correlation between age/PMQ and CPS score

Comparison with performance in live selection process

Correlations between CPS pilot scores and live selection scores (shortlisting and interview) were calculated for participating specialties. Selection data were available for GP, CMT, Paediatrics, Histopathology, Anaesthesia and ACCS. Separate analyses were conducted for each specialty. Analyses were conducted by deanery for specialties where the selection process differed across locations (anaesthesia and ACCS).

Table 12 shows descriptive statistics and correlations for CPS, shortlisting and interview scores by specialty/deanery for those candidates who participated in the pilot[1]. Scores for CPS, shortlisting and interview showed close to normal distributions except where indicated. It should be noted that sample sizes are very small in almost all cases and results should therefore be interpreted with caution.

[1] Anaesthesia/ACCS selection descriptives include all candidates; Paediatrics selection data includes several candidates whose data was submitted but who did not complete the pilot

Table 12: CPS, shortlisting & interview descriptive statistics & correlations by specialty & deanery

Specialty	CPS			Shortlisting			Interview			Shortlist correlation		Interview correlations			
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	CPS & S/L	N	CPS & I/V	N	S/L & I/V
GP															
National	424	76.7	13.8	425	70.5[2] 648.1 510.4	9.3 33.7 67.9	375	39.0	5.7	424	.75**Error! ookmark not defined.** .51** .71**	375	.43**	375	.45** .56** .57**
CMT															
National	204	82.1	13.5	204	22.3	9.6	204	45.2 (65.6[3])	8.0 (12.2)	204	.46**	204	.61** (.64**)	204	.49** (.64**)
Histopathology															
National	27	70.9	13.9	24	59.2	19.3	18	370.1	103.9	24	.37	18	.70**	18	.80**
Paediatrics															
Deanery1	50	78.2	12.2	55	24.8	10.8	47	109.6 (26.8[4])	20.9 (5.9)	47	.28	39	.55** (.28)	47	.29* (-.02)
Deanery2				55	24.1	9.5	40	109.3 (26.2)	19.6 (6.2)	48	.27	35	.48** (.37*)	40	.32* (.32*)
Anaesthetics															
KSS[5]	42	80.0	16.0	31	25.1	5.6	16	107.8	10.4	31	.62**	16	.42	16	.33
London	46	87.5	11.1	290	41.2	7.4	174	74.7	8.5	45	.53**	29	.09	174	.11
Oxford	31	83.2	11.8	261	39.2	11.7	47	107.3	48.4	31	.21	4	n/a	47	.03
Severn	16	82.9	11.7	160	42.3	12.2	31	83.9	18.5	15	.43	2	n/a	31	.41*
WMids	18	81.3	12.1	149	30.8	7.7	70	38.6	5.2	15	.40	2	n/a	61	.22
Y&H	26	82.1	10.7	165	38.4	10.4	87	50.1	6.2	24	.62**	11	n/a	87	.50**
ACCS															

[2] Figures refer to GP CPS, SJT and overall shortlisting total respectively
[3] Weighted interview score
[4] Communication component score
[5] Non parametric correlations due to non-normal data

EoE	36	80.3	16.3	160	21.2	8.2	80	45.6	6.4	24	.49*	14	n/a	80	.38**
LondonAM	16	85.5	18.6	87	54.4	12.3	29	84.4	10.0	16	.73**	7	n/a	29	.32
LondonAn	42	88.0	10.9	243	44.3	8.4	108	100.8	12.6	41	.43**	19	.00	108	.20*
NW	21	77.3	13.3	192	96.0	22.2	91	108.8	15.0	19	.51*	6	n/a	91	.34**
OxfordAn	25	84.6	10.6	153	42.2	11.3	6	71.0	15.0	25	.25	2	n/a	6	n/a
SevernAn	20	84.4	12.4	162	42.7	12.9	51	79.4	23.2	18	.55*	6	n/a	51	.06
Y&HAn	20	83.5	11.8	115	37.5	9.1	21	39.6	4.2	18	.77**	5	n/a	21	.05

**Correlation is significant at the 0.01 level (2-tailed).

The CPS showed significant positive correlations with GP live shortlisting components (CPS, SJT & shortlisting total; r=.51-.75, p<.01) and significantly predicted performance at the GP selection centre (r=.43, p<.01). The CPS did not demonstrate incremental validity over GP shortlisting total in predicting selection centre performance however.

In CMT, the CPS showed a significant positive correlation with live shortlisting (r=.46, p<.01) and was a good predictor of interview performance (r=.61-.64, p<.01). The CPS and CMT shortlisting both offered incremental validity over each other in predicting CMT interview scores (raw and weighted).

In Histopathology, the CPS was not significantly correlated with live shortlisting scores but was a good predictor of interview performance (r=.70, p<.01).

In Paediatrics, correlations between CPS and shortlisting at candidates' first and second choice deaneries were not significant. However, the CPS was a significant predictor of interview performance at first and second choice deaneries (r=.55 & r=.48, p<.01), offering incremental validity over shortlisting alone.

In Anaesthesia, correlations between CPS and live selection were only possible for 6 deaneries (KSS, London, Oxford, Severn, West Midlands, Yorks & Humber) due to sample size[6]. The CPS correlated positively with shortlisting at 3 of these deaneries (r=.53-.62, p<.01). Correlations between CPS and interview scores were calculated for 2 deaneries but were not significant.

In ACCS, correlations were calculated by programme due to different selection processes and only 7 cases had sufficient sample sizes (East of England & North Western ACCS; London ACCS AM; London, Oxford, Severn and Yorks & Humber ACCS Anaesthesia). The CPS correlated positively with shortlisting in 6 programmes (r=.43-.77, p<.05). Correlation between CPS and interview was possible for 1 programme but no relationship was found.

Conclusions

It is feasible to deliver a CPS test on line in a number of invigilated test centres.

- Overall test statistics show that the CPS had good reliability (Cronbach α 0.89).
- Over 97% of items were successful with a point biserial of < 0.
- The participants were representative of the ability range of applicants (as measured by the shortlisting scores).
- UK qualified doctors scored significantly higher than non UK doctors.
- There was no significant difference in scores between sexes.
- White candidates scored highest.
- Older doctors scored lowest.
- Candidates who applied for CMT, ACCS, Anaesthetics and Paediatrics scored higher.

The CPS was a strong predictor of shortlisting score in GP and CMT. It also correlated with the interview scores for GP, CMT, histopathology and paediatrics.

[6] Only deaneries with sample sizes of 15+ for correlations were included

5. The SJT

Dr Tom Gale, Dr Alison Carr, Ms Victoria Carr, Mr Martin Roberts, Dr Ian Anderson

Background

A Situational Judgment Test (SJT) is an assessment method which measures how an applicant may behave when posed with difficult professional dilemmas. Features include:

- The ability to target particular attributes for assessment
- Questions can be tailored to specific professional groups in order to increase authenticity of scenarios used
- A pre-defined answer key is determined using extensive pre-piloting of items
- An SJT is NOT a test of knowledge or problem solving ability

The predictive validity and incremental validity of SJTs over other methods of selection is well established and there is recent evidence that performance in SJTs used for selection to General Practice training programmes correlates well with clinical performance, once appointed.[6] In this regard SJTs seem to perform differently to Clinical Problem Solving Tests and knowledge based tests which are known to predict performance at knowledge based professional examinations.[13,14]

Acute Specialties Selection Pilot

The Acute Specialties Selection Pilot has been run through the South West Peninsula Deanery since 2008 with the support of Department of Health funding. This has involved the development and evaluation of both shortlisting and interview methods for recruitment to training in the Acute Specialties (Anaesthesia, Acute Care Common Stem training). The project has focused on 3 main areas:

- Clinical Problem Solving test
- Situational Judgment Test
- Multi-station selection centre utilising simulation and work sample tasks.

Acute Specialties SJT

In 2009 this test was piloted in three English deaneries utilising a paper based format at the point of interview for Acute Specialties. Pre-piloting was performed prior to the interviews utilising existing trainees and Consultants in the Acute Specialties. The question bank was developed by subject matter experts who were trained to write SJTs from Acute Medicine, Anaesthesia, Emergency Medicine and Intensive Care Medicine.

During 2010 the SJT was piloted in two ways:

1. At the point of interview utilising a paper based format in 7 participating deaneries
2. At the point of application utilising a computer delivered format through the AoMRC pilot.

The remainder of this chapter discusses the demographics and results obtained for the computer-delivered version of the SJT. Chapter 6 discusses the results of the paper-delivered SJT.

Computer based SJTs
Test specification and implementation

- Blueprinted onto four attributes / non-technical skills identified by thorough job analysis of Anaesthesia and ACCS:[15,16]
- Empathy and Sensitivity
- Vigilance and Situational Awareness
- Professional Integrity
- Coping with Pressure
- 45 items
- 20 ranking questions, requiring candidates to rank 5 items in order of appropriateness from 1 (most appropriate) to 5 (least appropriate).
- 25 multiple response questions, requiring candidates to choose the 3 best answers from a selection of 8 possible responses.
- 1 hour test duration

The SJT was offered to all applicants applying to Anaesthesia / ACCS at CT1 entry through the AoMRC pilot website. Anaesthesia / ACCS applicants could opt to take the AoMRC CPS test (2 hours duration) or the AoMRC combined test which included the SJT (1 hour duration) followed by the CPS (2 hours). All applicants sitting the combined test did so using a computer based format in Pearson VUE centres.

Candidate demographics
Although the SJT was incorporated into the AoMRC pilot as a test for applicants to the Acute Specialties, many candidates applying to other specialties applied to take the test. A total of 382 candidates took the SJT of whom a slight majority (57%) were female. The distribution of candidate ages was strongly skewed (Figure 7) with a median of 27 years or less.

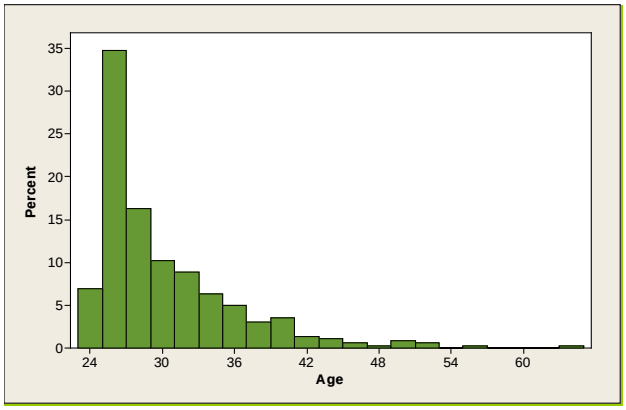


Figure 7: Age distribution of candidates (N=382)

The majority (64%) of candidates were UK-trained and the largest ethnic groups were Asian (38%) and White British (31%). Table 13 shows the distribution of candidates by ethnicity and region of primary medical qualification.

Table 13: Distribution of candidates by ethnicity and region of primary medical qualification.

	Region of Primary Medical Qualification				
Ethnicity	**UK**	**EU**	**Other**	**All**	**%**
Asian / Asian British	*78*	*4*	*64*	146	38%
White - British	*106*	*2*	*10*	118	31%
Black	*11*	-	*26*	37	10%
Any other Ethnic group	*12*	*1*	*11*	24	6%
White - Other	*10*	*7*	*6*	23	6%
Other Ethnic Groups - Chinese	*18*	-	*1*	19	5%
Mixed	*8*	*1*	*2*	11	3%
Not declared	*2*	-	*2*	4	1%
All	245	15	122	*382*	*100%*
%	64%	4%	32%	*100%*	

Applications to more than one specialty were made by 169 (44%) candidates and General Practice was the most frequently preferred specialty (Table 14).

Table 14: SJT candidates by sex and preferred specialty

Preferred Specialty	ALL	%
ACCS	60	16%
Anaesthetics	32	8%
General Practice	219	57%
Histopathology	4	1%
Medicine (CMT)	39	10%
Paediatrics	10	3%
Psychiatry	3	1%
No preference	15	4%
ALL	382	100%

Applicants to the Acute Specialties of ACCS and Anaesthesia (the target population for the SJT) comprised 115 (30%) of the 382 test candidates.

Representativeness of the Acute Specialties subsample

Data on all applications to the Acute Specialties (ACCS and Anaesthesia) was obtained from 13 of 14 English deaneries allowing us to put together an almost complete picture of applicants and applications to these specialties across the country. A total of 1498 doctors made 5118 applications for Acute Specialty training posts. The number of applications per doctor ranged from 1 to 33 with a median of 2 (Table 15).

Table 15: Frequency of **Acute Specialty** applications per doctor

Applications	Frequency	%
1	485	*32%*
2	309	*21%*
3	184	*12%*
4	166	*11%*
5	96	*6%*
6-10	201	*13%*
11-20	50	*3%*
21-33	7	*0.5%*
Total	1498	*100%*

Just 115 (8%) of these applicants to the Acute Specialties sat the computer based version of the SJT (compared to 261 (17%) who sat the paper based version). Demographic and recruitment data provided by the deaneries allowed us to examine the extent to which the volunteer sample of Acute Specialties applicants was representative of the wider population of those applicants. We report here on possible sampling bias in the computer based test sample only.

Logistic regression analysis was conducted using the variables gender, age, time since registration, ethnic group, language and region of medical training as predictors of SJT participation. Only ethnic group and time since registration were significant predictors (Table 16).

Table 16: Coefficients in logistic regression model

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
Age	-0.039	0.031	1.647	1	0.199	0.961
PMQ region			1.240	3	0.743	
PMQ region(1)	0.644	0.612	1.109	1	0.292	1.904
PMQ region(2)	0.071	0.468	0.023	1	0.879	1.074
PMQ region(3)	-0.082	0.403	0.042	1	0.838	0.921
Months registered	-0.020	0.009	4.856	1	**0.028**	0.980
Ethnic			20.918	3	**0.000**	
Ethnic(1)	0.915	0.569	2.581	1	0.108	2.496
Ethnic(2)	0.921	0.579	2.530	1	0.112	2.513
Ethnic(3)	-0.106	0.568	0.035	1	0.852	0.900
Sex(1)	-0.177	0.204	0.757	1	0.384	0.838
English(1)	-0.729	0.748	0.951	1	0.330	0.482
Constant	-1.169	0.960	1.483	1	0.223	0.311

These results show that the odds of participation in the SJT pilot reduced as time since registration increased (mean time =10.5 months for participants vs. 14.5 months for non-participants) and that Asian and other non white applicants were significantly over-represented in the computer-based SJT sample (Table 17).

Table 17: Acute Specialty applicants by ethnicity & computer-based SJT participation

	Participation		
Ethnic background	***No***	***Yes***	**Rate**
Asian / Asian British	*310*	*39*	11%
Other	*172*	*24*	12%
White - British	*729*	*46*	6%
White – non-British	*112*	*6*	5%
Unknown	*60*	*0*	0%
ALL	***1383***	***115***	**8%**

We also examined whether participants in the computer based SJT pilot differed from non participants in terms of their shortlisting scores. There were 47 deanery × post subsamples, 45 of which allowed testing for a difference in average shortlisting scores between those who sat the SJT and those who didn't. After allowing for multiple testing no significant differences were found amongst the 45 subsamples [Mann-Whitney tests, family wise error rate = 0.05]

Conclusions

The participants in the computer based SJT pilot appeared to be representative of the ability range of Acute Specialties applicants (as measured by shortlisting scores) and were demographically similar to the non-participant group in terms of their gender and region and language of medical training. Younger, less experienced doctors were slightly over-represented in the participant group as were doctors from non-white ethnic groups.

Evaluation of the SJT

The objective of the evaluation was to explore the effectiveness of the computer delivered SJT pilot in terms of its psychometric properties, as follows:

- **Reliability** of the Acute Specialties SJT.
- **Psychometric review** of SJT item properties.
- **Fairness** of the SJT in terms of group differences.
- **Criterion related validity** of the SJT in terms of correlations with other assessments and with performance in AS selection processes.

The SJT was made up of 45 items (25 multiple response, 20 rank), with 60 minutes testing time. There was a maximum of 12 points available for each item. Table 18 gives descriptive statistics for the SJT[7] and Figure 8 shows the distribution of the scores.

Table 18: SJT pilot results

	SJT overall (N=351)
No. of items	45
Mean raw score (%)	386 (71%)
Std deviation	45.88
Range	256-474
Reliability	0.8

Overall test statistics show that the SJT had good reliability (Cronbach alpha=0.84). Scores were approximately normally distributed (very slight negative skew), indicating that the test was capable of differentiating between candidates.

[7] 31 cases were removed from the SJT analysis due to a high number of unanswered items (13 or more).

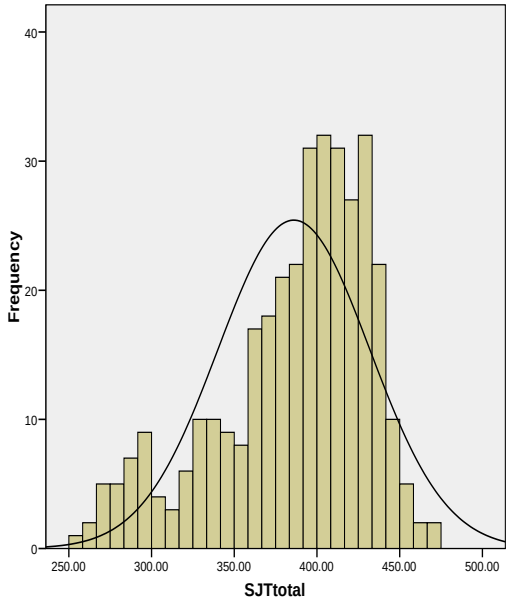


Figure 8: Distribution of SJT scores

Descriptive statistics were explored on the basis of candidates' preferred specialty. Table 19 shows SJT results for each preferred specialty. One way ANOVAs and post hoc tests showed that candidates whose preferred specialty was Anaesthesia/ACCS scored significantly better than candidates whose preferred specialty was GP ($p<.001$) or Other specialties ($p<.05$).

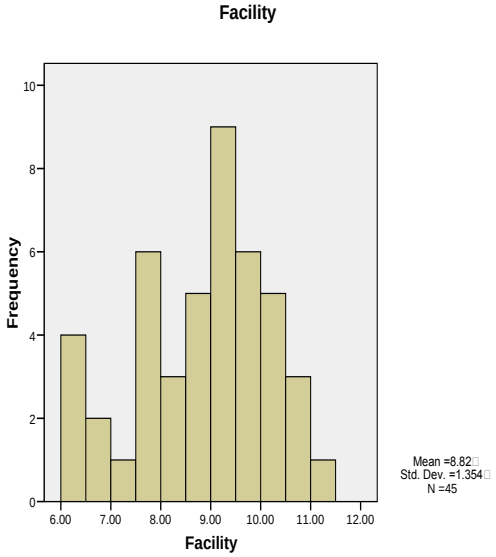
Table 19: SJT results by preferred specialty

Preferred Specialty	Number	Mean Score	SD	Range
Anaes/ACCS	90	401	42.22	266-474
GP	196	380	43.35	272-462
CMT	35	388	49.73	260-456
Other[8]	30	376	58.01	256-460

Item analysis was used to look at the difficulty and quality of individual SJT items. Full results of the item analysis are shown in Appendix E. Item facility (out of a maximum of 12) ranged from 5.14 to 10.69, with a mean of 8.58.

In terms of item quality, item partials ranged from 0.05 to 0.52, with a mean of 0.30. Eighty two percent of items were successful (item partial >0.2) and are potentially suitable for inclusion in an operational test. Figures 9 & 10 shows the distribution of item facility and item partials respectively.

[8] Candidates with a preferred specialty of Paediatrics, Psychiatry, Histopathology, No Preference or no response were combined into Other due to small sample sizes



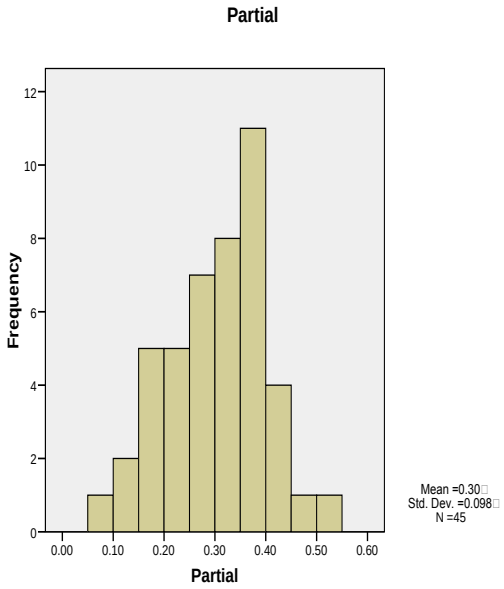


Figure 9: Distribution of SJT item facility **Figure 10:** Distribution of SJT item partials

In order to examine fairness issues regarding the use of the SJT with an Acute Specialties sample, group differences in performance were analysed on the basis of gender, age and ethnic group.

In terms of gender, there was no significant difference (t-test) in SJT scores between male and female candidates (male mean 384.2, female 388.2). In terms of age, there was a significant negative correlation between age and SJT scores (rho=-.43, p<.001).

In terms of place of medical qualification, UK trained candidates performed significantly better on the SJT than candidates trained outside the UK (UK trained mean 403.8, non UK trained 349.2, p<.001). This pattern of findings is similar to other assessments in this context but indicates that regular monitoring of group differences is recommended.

The correlation between SJT pilot scores and CPS pilot scores within the Acute Specialties SJT sample was r=.52.This is similar to correlations found in other settings (e.g. GP) and indicates that the two tests have both shared and independent variance.

Comparison with performance in live selection process

Correlations between SJT scores and live selection scores (shortlisting and interview) were calculated for participating specialties. Selection data were available for GP, CMT, Paediatrics, Histopathology, Anaesthesia and ACCS. Separate analyses were conducted for each specialty, and by deanery for specialties where processes differed across locations (Anaesthesia and ACCS).

Table 20 shows descriptive statistics and correlations for SJT, shortlisting and interview scores by specialty/deanery for those candidates who participated in the national MMT pilot[9]. Scores for SJT, live shortlisting and interview showed close to normal distributions except where indicated.

[9] NB Descriptive statistics for live selection include candidates who completed the CPS pilot but not the SJT

Table 20: SJT, shortlisting & interview descriptive statistics & correlations by specialty & deanery

	SJT Pilot			Shortlisting			Interview			Shortlisting correlations		Interview correlations			
Specialty	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	SJT & S/L	N	SJT& I/V	N	S/L & I/V
GP															
National	257	366.3	60.7	425	70.5[10] 648.1 510.4	9.3 33.7 67.9	375	39.0	5.7	257	.43** .58** .57**	229	.58**	375	.45** .56** .57**
CMT															
National	92	371.5	63.0	204	22.3	9.6	204	45.2 (65.6[11])	8.0 (12.2)	92	.32**	92	.69** (.68**)	204	.49** (.64**)
Histo															
National	8	312.3	56.0	24	59.2	19.3	18	370.1	103.9	6	n/a	3	n/a	18	.80**
Paeds															
Deanery1	26	366.3	58.3	55	24.8	10.8	47	109.6 (26.8[12])	20.9 (5.9)	24	.35	17	.71** (.58*)	47	.29* (-.02)
Deanery2				55	24.1	9.5	40	109.3 (26.2)	19.6 (6.2)	25	.31	17	.48 (.38)	40	.32* (.32*)
Anaes															
KSS[13]	38	390.7	50.0	31	25.1	5.6	16	107.8	10.4	29	.36	15	.19	16	.33
London	41	410.9	28.9	290	41.2	7.4	174	74.7	8.5	40	.01	26	.15	174	.11
Oxford	26	407.2	33.0	261	39.2	11.7	47	107.3	48.4	26	.10	4	n/a	47	.03
Severn	16	417.3	24.9	160	42.3	12.2	31	83.9	18.5	15	.64*	2	n/a	31	.41*
Y&H	21	400.7	53.3	165	38.4	10.4	87	50.1	6.2	20	.45*	7	n/a	87	.50**
ACCS															
EoE	27	374.6	56.4	160	21.2	8.2	80	45.6	6.4	19	.39	11	n/a	80	.38**
LondonAn	36	402.8	34.3	243	44.3	8.4	108	100.8	12.6	35	.20	17	.32	108	.20*
NW	18	383.8	53.8	192	96.0	22.2	91	108.8	15.0	15	.66**	5	n/a	91	.34**
OxfordAn	23	408.8	32.3	153	42.2	11.3	6	71.0	15.0	23	-.05	2	n/a	6	n/a

[10] Figures refer to GP CPS, SJT and overall shortlisting total respectively
[11] Weighted interview score
[12] Communication component score
[13] Non-parametric correlations due to non-normal data

	SJT Pilot			Shortlisting			Interview			Shortlisting correlations		Interview correlations			
Specialty	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	SJT & S/L	N	SJT& I/V	N	S/L & I/V
SevernAn	20	407.0	35.0	162	42.7	12.9	51	79.4	23.2	18	.44	6	n/a	51	.06

**Correlation is significant at the 0.01 level (2-tailed). *Correlation is significant at the 0.05 level (2-tailed).

The Acute Specialties SJT showed substantial positive correlations with GP live shortlisting (CPS, SJT & overall shortlisting score; r=.43-.58) and was a good predictor of performance at the GP selection centre (r=.58), showing a similar level of prediction to the GP SJT.

In CMT, the SJT correlated positively with live shortlisting (r=.32) and was a strong predictor of interview performance (r=.69). The SJT offered incremental validity over CMT shortlisting alone, explaining a further 33% of the variance in CMT interview scores.

In Paediatrics, correlations between the SJT and shortlisting at candidates' first and second choice deaneries were not significant. The SJT was a significant predictor of interview performance at first choice deanery (r=.71) and correlated with the communication interview score, although results should be interpreted with caution due to sample size.

In Anaesthesia, correlations between SJT and live selection were only possible for 5 deaneries (KSS, London, Oxford, Severn, Yorks & Humber) due to sample size. The SJT correlated positively with shortlisting at 2 deaneries (mean correlation r=.31 across 5 deaneries). Correlations between SJT and interview were calculated for 2 deaneries but results were not significant.

In ACCS, correlations were calculated by programme where these used different selection processes and only 5 cases are included due to sample size (East of England & North Western ACCS; London, Oxford & Severn ACCS Anaesthesia). The SJT correlated positively with shortlisting at 1 deanery (mean correlation r=.33 across 5 deaneries). Correlation between SJT and interview was possible for 1 deanery but was not significant.

Conclusions

- The Acute Specialties SJT was a popular choice amongst candidates applying to many specialties within the AoMRC pilot.
- Amongst applicants to the Acute Specialties, participants in the computer based SJT pilot were representative of the ability range (as measured by shortlisting scores) and were demographically similar to the non participant group in terms of their gender, region and language of medical training. Younger, less experienced doctors were however slightly over represented in the participant group, as were doctors from non-white ethnic groups.
- Overall test statistics show that the SJT had good reliability (Cronbach α=0.84) which is over the level of α=.80 that is desirable for high stakes assessment.
- 82% of items were successful (item partial >0.2) and are potentially suitable for inclusion in an operational test.
- Candidates applying for the Acute Specialties performed significantly better than other specialty groups in the SJT.
- The SJT showed good criterion related validity with strong correlations with live shortlisting for GP and CMT applicants.
- The SJT was a strong predictor of performance at interview / selection centres for GP, CMT and paediatric applicants.
- In Anaesthesia and ACCS, correlations between the SJT and current selection processes are encouraging but the strength of any associations are limited due to the fact that these specialties are not currently part of a nationally standardised shortlisting and interview process.

6 The paper-based SJT

Dr Tom Gale, Dr Ian Anderson, Ms Victoria Carr, Mr Martin Roberts, Dr Peter Davies

Background

The SJT was also administered by the Acute Specialties Pilot team across seven English deaneries during the 2010 recruitment round using a *paper based* format. All CT1 applicants who were interviewed for Anaesthesia and ACCS training posts were invited to sit the SJT in participating deaneries at the point of interview.

Test specification was identical to the computer based version described in Chapter 5 with a 1-hour test duration comprising 45 items. 33 items were identical to the computer based test with 12 items unique to the paper based format. Applicants were NOT asked to sit the test if they had already taken part in the computer based pilot.

Sample Demographics

Across the seven participating deaneries, 424 applicants attended 663 interviews for CT1 training posts in Anaesthesia and ACCS. Numbers of interviews conducted in each deanery and specialty are shown in Table 21 below.

Table 21: Interviews conducted by deanery and type of post, 2010

	POST				
Deanery	*ACCS-AM*	*ACCS-AN*	*ACCS-EM*	*ANAES*	ALL
EMidN	6	30	17	7	*60*
EMidS	6	15	10	19	*50*
Oxford	10	6	12	47	*75*
Severn	29	48	28	30	*135*
SW Pen	5	26	13	61	*105*
W Mid	5	22	32	70	*129*
Wessex	9	8	29	63	*109*
ALL	*70*	*155*	*141*	*297*	*663*

Twenty nine interviewees turned down the chance to sit the paper-based SJT because they had already taken the computer-based version offered via the AoMRC pilot. The paper-based SJT was sat by 261 of the 395 eligible candidates, an overall response rate of 66%. Response rates are broken down by post and deanery in Table 22.

Table 22: SJT response rates by deanery and post, 2010.

	POST				
Deanery	*ACCS-AM*	*ACCS-AN*	*ACCS-EM*	*ANAES*	ALL
EMidN	*40%*	*64%*	*50%*	*71%*	54%
EMidS	*40%*	*62%*	*33%*	*40%*	46%
Oxford	*25%*	*100%*	*75%*	*65%*	63%
Severn	*61%*	*62%*	*56%*	*68%*	56%
SW Pen	*40%*	*83%*	*100%*	*90%*	84%
W Mid	*100%*	*89%*	*69%*	*72%*	70%
Wessex	*50%*	*83%*	*72%*	*93%*	82%

Just over half (55%) of the paper-based SJT candidates were male and, as with the computer based test, the age distribution was positively skewed with a median age of 27 years (range 23-50). The overwhelming majority (93%) of candidates were UK-trained and the largest ethnic groups were White British (72%) and Asian (14%). These proportions differ from the corresponding figures in the computer-based SJT sample (64% UK-trained, 31% White British and 38% Asian). The differences may be attributed to regional variation in the two samples and the fact that candidates in the paper-based test were drawn from interviewed applicants only.

Sampling Bias

The shortlisting process will have filtered out poorer quality applicants so the paper-based SJT sample is unlikely to be fully representative of the Acute Specialties applicant population. We examined self-selection bias in this sample by investigating possible demographic and application outcome differences between those eligible interviewees who volunteered to sit the SJT and those who did not. Logistic regression analysis was conducted using the variables gender, age, time since registration, ethnic group, language and region of medical training as predictors of SJT participation.

Table 23: Coefficients in logistic regression model

Variables	**B**	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
Age	0.014	0.033	0.181	1	0.670	1.014
PMQreg			3.912	3	0.271	
PMQreg(1)	-1.005	1.055	0.906	1	0.341	0.366
PMQreg(2)	-0.800	0.949	0.712	1	0.399	0.449
PMQreg(3)	-1.287	0.768	2.806	1	0.094	0.276
MonthsRegistered	-0.013	0.008	2.515	1	0.113	0.987
Ethnicgroup			1.902	3	0.593	
Ethnicgroup(1)	-0.124	0.668	0.035	1	0.852	0.883
Ethnicgroup(2)	0.142	0.720	0.039	1	0.843	1.153
Ethnicgroup(3)	0.310	0.626	0.245	1	0.621	1.363
Male	0.555	0.233	5.679	1	**0.017**	1.741
Eng	-1.597	1.063	2.256	1	0.133	0.203
Constant	1.602	1.559	1.056	1	0.304	4.965

These results show that the odds of volunteering to sit the pilot SJT were not significantly related to age, time since registration, ethnic group, language or region of medical training. Males however, were over-represented in the test sample: 55% of the 261 participants were male compared to 45% of the 134 non-participants.

Evaluation of the SJT

The objective of the evaluation was to explore the effectiveness of the SJT paper-based pilot in terms of its psychometric properties, as follows:

- **Reliability** of the SJT.
- **Psychometric review** of SJT item properties.
- **Fairness** of the SJT in terms of group differences.
- **Criterion-related validity** of the SJT in terms of correlations with other assessments and with performance in AS selection processes.

The AS SJT was made up of 45 items (25 multiple response, 20 rank), with 60 minutes testing time. There was a maximum of 12 points available for each item. Table 24 gives descriptive statistics for the AS SJT for the paper-based pilot sample[14]. Figure 12 shows the distribution of SJT scores.

Table 24: SJT paper-based pilot results

	SJT overall (N=249)
No. of items	45
Mean raw score (%)	396.9 (74%)
Std dev	37.00
Range	250-458
Reliability	0.78

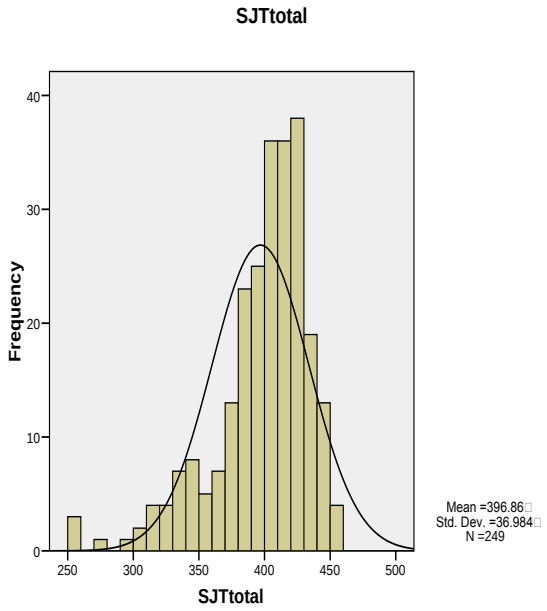


Figure 12: Distribution of SJT paper-based scores

- Overall test statistics show that the paper-based pilot SJT had good reliability (α=0.78), although reliability of α=.80 is desirable for high stakes assessment. Scores showed a small negative skew due to some very low scores but generally indicated that the test was capable of differentiating between candidates.
- Item analysis was used to look at the difficulty and quality of individual SJT items. Full results of the item analysis are shown in Appendix F. Item facility (out of a maximum of 12) ranged from 6.02 to 11.18, with a mean of 8.82.

[14] 12 cases were removed from the SJT analysis due to a high number of unanswered items (13 or more).

- In terms of item quality, item partials ranged from -0.01 to 0.61, with a mean of 0.23. 49% of items were successful with an item partial >0.2; 58% showed item partials >.18 and are potentially suitable for inclusion in an operational test. Figures 13 & 14 show the distributions of item facility and item partials respectively.

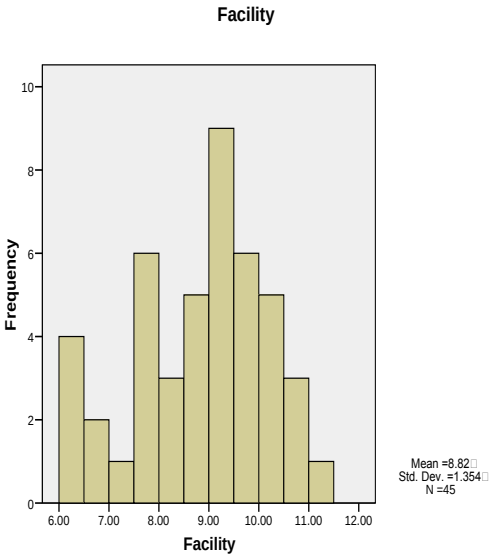


Figure 13: Distribution of SJT item facility

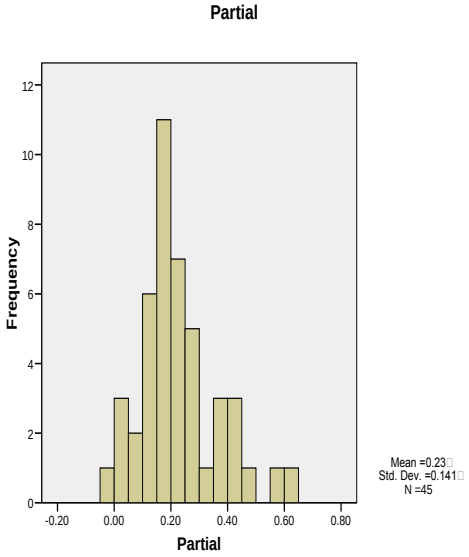


Figure 14: Distribution of SJT item partials

- In order to examine fairness issues regarding the use of the SJT with an Acute Specialties sample, group differences in performance on the paper-based SJT were analysed on the basis of gender, age and ethnic group. The number of non-UK trained candidates was too small to consider group differences on the basis of place of medical qualification.
- In terms of gender, there was a small but significant difference in SJT total score between male and female candidates (male mean 392.3, female 402.7, Mann-Whitney p<.01). In terms of age, there was no significant correlation between age and SJT scores.
- In terms of ethnicity, due to sample size candidates were classified into White UK/Irish, Asian or Other ethnic groups. There was a significant difference in SJT total scores between ethnic groups (Kruskal Wallis p<.01) and post-hoc tests showed that White UK/Irish candidates performed significantly better than both Asian and Other candidates (White UK/Irish mean 402.7, Asian 384.1, Other 380.1, Mann-Whitney p<.01). This pattern of group differences findings is similar to other assessments in this context but indicates that regular monitoring of group differences is recommended.

Comparison with performance in live selection process

- In order to explore the criterion-related validity of the SJT, correlations between SJT paper-based scores and live selection scores (shortlisting and interview) were calculated for participating deaneries and programmes. Selection data were provided by 14 deaneries for a total of 50 programmes. Separate analyses were conducted for each deanery and programme where selection processes differed. NB These analyses are based on applications rather than unique applicants, as most applicants applied to multiple locations/programmes.
- Table 25 shows descriptive statistics and correlations for SJT, shortlisting and interview scores by programme/deanery for candidates who completed the SJT paper-based pilot; programmes are only included if 15 or more candidates completed the SJT. Scores for SJT, live shortlisting and/or interview were somewhat skewed in many cases therefore non-parametric correlations (Spearman's rho) are shown for all cases for ease of comparison.
- In Anaesthesia, correlations between SJT and live selection were possible for all 14 deaneries[15]. The SJT showed a significant positive correlation with shortlisting at the majority of deaneries (8 out of 13; mean correlation r=.27). The SJT correlated positively with interview at only 3 out of 11 deaneries however (mean correlation r=.12).
- In ACCS (encompassing ACCS, ACCS-Anaesthesia, ACCS-AM, ACCS-EM), correlations between SJT and live selection were possible for a total of 16 programmes, covering 10 deaneries. The SJT showed a significant positive correlation with shortlisting in 5 out of 16 programmes (mean correlation r=.27) but correlated positively with interview in only 1 out of 8 programmes (mean correlation r=.27).

[15] Correlations are included for sample sizes of 15+; therefore in some locations correlations were calculated for shortlisting or interview only

Table 25: SJT, shortlisting & interview descriptive statistics & correlations by programme & deanery

	SJT Pilot			Shortlisting			Interview			Shortlisting correlations		Interview correlations			
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	SJT & S/L	N	SJT & I/V	N	S/L & I/V
Anaes															
E/Mids N[16]	23	385.5	31.1	23	26.8	6.8	5	56.4	4.5	23	.02	5	-	5	-
E/Mids S	29	376.6	51.0	29	27.4	6.9	6	39.2	6.4	29	.48**	6	-	6	-
EoE	31	382.0	40.0	0	-	-	18	72.4	10.6	0	-	18	.01	0	-
KSS	53	382.5	54.6	40	25.0	5.6	24	105.1	9.9	40	.44**	24	-.01	24	.10
London[16]	45	400.0	38.0	45	41.7	6.1	33	72.6	8.6	45	.32*	33	-.13	33	.07
Mersey	32	376.2	45.2	29	43.0	9.1	17	122.3	17.5	29	.39*	17	.67**	17	.40
N/Western[16]	23	386.5	47.2	23	88.9	18.8	20	89.9	11.5	23	.62**	20	.45*	20	.78**
Northern	23	383.5	54.4	23	49.6	8.2	5	38.9	4.3	23	.48*	5	-	5	-
Oxford	86	397.7	39.8	86	43.0	9.6	28	130.4	32.0	86	.14	28	-.22	28	.23
Severn	76	389.2	47.1	75	45.0	11.9	19	86.8	11.4	75	.32**	19	-.13	19	.69**
SWP	64	382.0	56.4	62	12.8	4.1	53	168.5	23.9	62	.26*	53	.36**	53	.41**
W/Mids	73	378.6	65.5	63	33.1	6.8	49	39.6	5.0	63	.08	49	.27	40	.11
Wessex	62	400.1	39.9	62	50.6	6.8	53	71.5	11.1	62	-.05	53	.06	53	.30*
Y&H[16]	26	378.0	41.9	21	43.9	9.1	15	47.9	6.1	21	-.02	15	-.06	15	.75**
ACCS															
EoE	33	373.6	54.9	26	22.6	7.1	10	46.7	4.5	26	.19	10	-	10	-
KSS	30	378.4	51.5	20	15.9	3.5	8	70.4	3.9	20	-.01	8	-	8	-
N/Western[16]	19	382.5	58.9	18	102.8	17.6	7	103.0	20.7	18	.72**	7	-	7	-
ACCS-Anaes															
E/Mids N	28	365.1	59.6	28	19.9	4.5	16	76.8	9.3	28	.47*	16	.33	16	.25
London[16]	39	399.3	35.3	39	45.2	5.7	19	98.3	10.9	39	.28	19	-.02	19	.14
Oxford	59	395.0	44.8	59	45.1	7.7	4	76.3	10.1	59	.31*	4	-	4	-
Severn	64	387.0	48.3	64	46.7	11.4	28	77.3	24.4	64	.24	28	.21	28	-.01
SWP[16]	42	386.1	52.7	40	12.8	4.3	20	139.2	15.2	40	.24	20	.13	20	.28
W/Mids	43	374.2	61.4	43	34.3	6.2	17	36.5	4.5	43	.02	17	.63**	17	.06
ACCS-AM															
Severn[16]	18	394.7	29.9	18	46.6	11.1	17	83.5	7.9	18	.47*	17	.30	17	.78**

[16] Data approximately normally distributed

	SJT Pilot			Shortlisting			Interview			Shortlisting correlations		Interview correlations			
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	SJT & S/L	N	SJT & I/V	N	S/L & I/V
ACCS-EM															
London	18	376.9	61.1	18	47.3	10.8	11	51.8	7.3	18	.67**	11	-	11	-
Oxford	22	382.7	63.2	22	42.1	8.9	9	68.1	15.0	22	.03	9	-	9	-
Severn	28	393.7	44.1	28	46.3	10.3	14	85.4	8.5	28	.11	14	-	14	-
SWP	21	387.8	47.8	20	13.5	5.2	11	140.7	17.4	20	.24	11	-	11	-
W/Mids[16]	28	372.0	65.4	20	32.2	5.8	20	33.4	4.6	20	.37	20	.42	12	-
Wessex	21	403.8	44.1	21	44.3	5.7	21	69.0	10.0	21	-.02	21	.16	21	.51*

**Correlation is significant at the 0.01 level (2-tailed).

*Correlation is significant at the 0.05 level (2-tailed).

Conclusions

- Interviewees who participated in the paper-based SJT pilot were demographically similar to the non-participant group in terms of their age, length of medical experience, ethnic origin, language and region of medical training but males were over-represented in the test sample.
- The test was able to differentiate between candidates with a wide variation in scores in a near normal distribution.
- Overall test statistics show that the paper based SJT had good reliability (Cronbach α=0.78) which is close to the level of α=.80 that is desirable for high stakes assessment but slightly lower than the reliability of the computer delivered version of the test (α=0.84).
- 58% of items showed item partials >.18 and are potentially suitable for inclusion in an operational test. Items that were common to both the paper based and computer based test performed generally less well in the paper based version.
- The lack of a nationally standardised shortlisting and interview format resulted in low sample numbers for testing criterion related validity of the test.
- There is a wide variation in correlations between shortlisting and interview scores across deaneries which may be attributable to the fact that shortlisting and interview strategies vary.
- In Anaesthesia and ACCS, correlations between the SJT and current selection processes are encouraging but the strength of any associations are limited due to the fact that these specialties are not currently part of a nationally standardised shortlisting and interview process.

7. Correlation of CPS and SJT scores
Dr Kath Woolf, Dr Alison Sturrock

Average and distribution of test scores
On the CPS, the scores ranged from 20% to 90%. Scores were very slightly negatively skewed with a mean of 65.7% and standard deviation (SD) of 11.6%.Two candidates scored lower than 3 SD below the mean, neither of whom took the SJT. No candidates scored 100%, which was just at 3 SD above the mean.

Minimum score (percent correct) on the CPS was 32% and the maximum 88%. SJT scores were more negatively skewed than the CPS scores. The mean and SD of the SJT were very similar to the CPS at 69.4% and 11.1%, respectively. Two candidates scored lower than 3 SD below the mean (those candidates both scored within 3 SD of the mean on the CPS).

Correlation between test scores
CPS and SJT scores were statistically significantly correlated (r=0.57; p<0.001).

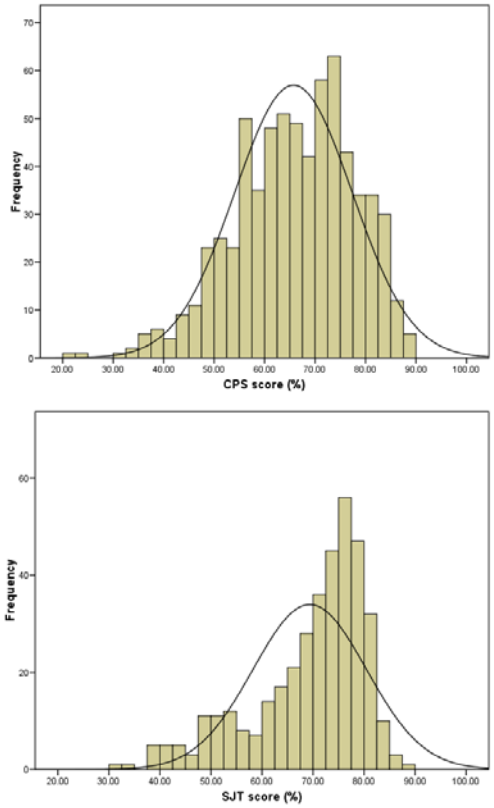


Figure 11: Distribution of CPS scores (n=665) was roughly Normal. The distribution of SJT scores (n=379) was negatively skewed**.**

Time taken to complete both tests
Candidates who took both CPS and SJT tests took about one minute less to complete the CPS on average (mean=97.9 minutes; median=101.7 minutes; SD=18.6), with an identical range of 45 to 120 minutes.

Sex
Although women achieved slightly higher CPS and SJT scores, the sex difference was not statistically significant [CPS p=0.117, SJT p=0.368].

Regression of demographics on CPS and SJT scores

Two regressions of CPS and SJT scores, respectively, on to age, sex, ethnic group (white, Asian not Chinese, else) and PMQ (UK vs non UK) showed that ethnic group and PMQ both had independent and significant effects on CPS and on SJT scores. The largest effect was of PMQ. Once other demographic factors were taken into account, age was no longer a significant predictor of CPS score; whereas it remained significant on the SJT. Differences in preferred specialty choice were no longer significant on the SJT though, once other factors were taken into account.

Table 23: Regression output showing the independent effects of ethnic group, PMQ, preferred specialty and age on CPS and SJT scores.

CPS					
Model	*B*	*Standard Error*	*Beta*	*t*	*p value*
(Constant)	91.94	2.20		41.24	<0.001
Ethnic group	-4.73	.81	-.21	-5.85	<0.001
Sex	-.18	.78	-.01	-.22	0.82
PMQ	-10.82	.90	-.43	-12.00	<0.001
Age (years)	-.01	.01	-.03	-.95	0.35
Preferred specialty	-1.66	.42	-.13	-3.93	<0.001
SJT					
(Constant)	104.0	3.28		31.75	<0.001
Ethnic group	-4.39	1.01	-.19	-4.33	<0.001
Sex	-.58	.94	-.03	-.62	0.54
PMQ	-7.52	1.26	-0.32	-5.96	<0.001
Age (years)	-0.49	0.10	-0.26	-5.07	<0.001
Preferred specialty	-0.77	0.50	-0.07	-1.55	0.12

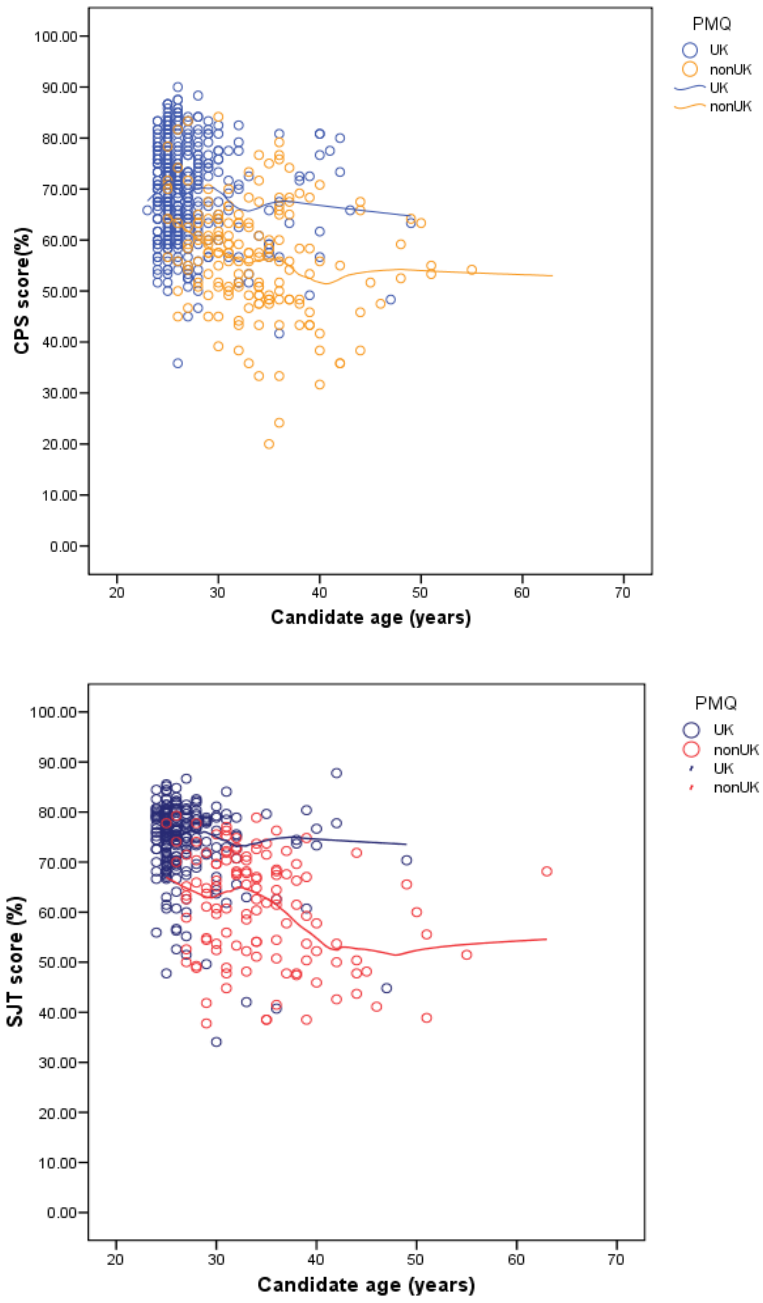


Figure 11: Scatterplot showing the significant correlations between age and CPS/SJT score in candidates who qualified in the UK (blue/dark blue) and outside the UK (orange/red). Lines are Lowess curves for each sub group. One candidate (age=63) removed from CPS results for maintenance of confidentiality.

8. Candidate perceptions of the computer delivered tests
Dr Luci Etheridge, Dr Kath Woolf, Mr Martin Roberts

Overall ratings, all candidates

All but six of the candidates completed the post-test survey. The results for those who took just the CPS and those who took both tests are given in Table 24, except for perceptions of the test length, which are given later. It is not possible to determine how participants felt about the SJT alone, we can only see whether their experience of doing the SJT coloured their experience overall (which it appeared not to do).

Table 24: Candidates' ratings of the test's usefulness, validity and practicality

		Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Total
The test gave me a chance to demonstrate my knowledge and skills	CPS only	3	39	55	156	28	281
	CPS & SJT	3	41	64	216	48	372
I had sufficient information about the purpose of the assessment	CPS only	0	2	5	157	117	281
	CPS & SJT	2	6	15	213	136	372
The test is a fair way to help decide who should be shortlisted for training posts	CPS only	8	49	54	148	22	281
	CPS & SJT	8	41	79	200	45	373
The test is a useful addition in helping decide who should be shortlisted for training posts	CPS only	8	39	60	149	26	282
	CPS & SJT	6	37	64	219	47	373
The questions were at a level appropriate for doctors at the end of foundation training	CPS only	3	30	30	189	29	281
	CPS & SJT	6	41	63	222	40	372
The test was easy to arrange and book	CPS only	3	18	22	178	60	281
	CPS & SJT	3	20	34	238	77	372
It was easy to find a test centre that was convenient for me	CPS only	3	20	4	150	104	281
	CPS & SJT	3	17	14	206	133	373
I found it easy to understand the instructions for the computerised test	CPS only	1	19	17	147	97	281
	CPS & SJT	9	39	31	196	98	373
I am confident that my data will be stored securely	CPS only	1	9	64	155	51	280
	CPS & SJT	2	9	78	203	80	372
Mean ratings	CPS only	*3*	*25*	*35*	*159*	*59*	*281*
	CPS & SJT	*5*	*28*	*49*	*213*	*78*	*372*

NB Statements with >50 ratings are highlighted.

Relationship between candidate factors and test perceptions
Generally the test was well received. Candidates felt it was appropriate and practical. They agreed most strongly with the statements concerning the practicalities of taking the test; and were least convinced by the statements regarding the validity of the test and with the data security.

CPS-only candidates generally felt the test was about the right length, with 70% (198/287) of candidates stating that the amount of testing time was “just right”. By comparison, only 52% (194/372) of those who also took the SJT felt the time was “just right”. Those who took the CPS and SJT were more likely to thing the testing time was too long (23% of CPS + SJT vs 6% of CPS only). Exactly the same proportion (23%) of those who took the CPS (65/287) and the CPS-SJT (87/372) felt the testing time was too short.

It seems that candidates were partly basing their ratings on the amount of time it took them to complete the test. Those who felt the testing time was too short completed the CPS in a median of 120 minutes (CPS only) or 112 minutes (CPS+SJT); those who felt the testing time was too long completed the test in a median of 76 minutes (CPS only) or 97 minutes (CPS+SJT); and those who felt the testing time was about right completed the test in a median of 100 minutes (CPS only) or 94 minutes (CPS+SJT).

Test perceptions related to test scores
For those who took the CPS only, there were weak positive correlations of approximately r=0.15 between performance on the CPS and perceptions of its validity (fairness, usefulness, chance to display knowledge). No correlations were found between perceptions of the practicalities of taking the test (e.g. ease of finding a test centre; understanding the items) and performance; however candidates who felt the testing time was too short performed statistically significantly worse than candidates who felt they were given the right amount or too much time [$F(2,280)=12.0$; $p<0.001$].

For those who took both tests, there were weak positive correlations between perceptions of test usefulness for shortlisting and CPS scores ($r=0.13$; $p<0.05$) and perceptions that the test was set at the right level ($r=1.3$; $p<0.05$). There was a negative correlation between perceptions of data security and CPS scores ($r=-0.15$; $p<0.01$).

In terms of SJT scores, there was a weak negative correlations between SJT score and perceptions of the fairness of the test to select candidates ($r=-0.12$; $p<0.05$) and data security ($r=-0.11$; $p<0.05$) and performance – i.e. who were less likely to think the test was valid and their data were stored correctly performed better on the SJT. There was a weak positive correlation between SJT score and feeling that there was enough information about the purpose of the test ($r=0.12$; $p<0.05$). SJT candidates who felt the testing time was too short performed statistically significantly worse than candidates who felt they were given the right amount or too much time [$F(2,371)=25.4$; $p<0.001$].

These correlations were statistically significant but only at the $p<0.05$ level, except for data security and CPS score in those who took both tests. Bearing in mind the number of statistical tests calculated, interpretation of these results requires some caution.

PMQ
In candidates who took the CPS only, non UK graduates were more likely to feel that the testing time was too short (chi-squared=20.2, df=2, p<0.001)

In those who took both tests, non UK graduates were more likely to report feeling that the test was fair to use in selection (t=3.0; p=0.003) and that it gave them a chance to display their knowledge and skills (t=2.3; p=0.023). However non UK graduates were less clear about the purpose of the test (t=-2.6; p=0.01) and more likely to feel that the testing time was too short (chi-squared=18.3, df=2, p<0.001).

Ethnicity
In those who took the CPS only, non white candidates were more likely to feel that the testing time was too short (chi-squared=9.3, df=2, p=0.01).

In candidates who took the CPS and SJT, non white candidates were less likely to report that they were clear about the purpose of the test (t=4.1; p<0.001), but more likely to state that they felt the test was fair to use in selection (2.3; p=0.025). Non white candidates were slightly more likely to feel that testing time was too short, but this was only bordering statistical significance (chi-squared=5.9, df=2, p=0.05).

Sex
In all candidates, there were no sex differences in reported perceptions of the validity or practicalities of the test.

Age
Older candidates who took the CPS only were less likely to perceive the test as fair to use in selection (Rho=-0.15; p=0.017), and that the test gave them the opportunity to display their knowledge and skills (Rho=-0.14; p=0.023).In addition, candidates who perceived they were given too much time (median age=25) were on average one year younger than those who perceived they were given the right amount time (median age=26); and they in turn were a year younger than those who felt they were given not enough time (median age=27) [Kruskal-Wallis chi-squared=13.1, df=2, p=0.001].

Contrary to those who took the CPS only, older candidates who took both tests were more likely to perceive the pilot was a fair way to shortlist candidates (Rho=0.15; p<0.01) but, as with those who took the CPS only, old candidates were more likely to think that the testing time too short [Kruskal-Wallis chi-squared=10.9, df=2, p=0.004]. Those who perceived the testing time was too short were on average 3 years older (27 years vs 30 years) than those who perceived it to be just right or too long.

Conclusions
- The CPS and SJT tests showed good face validity and were well received by candidates.
- Both tests showed an acceptable spread of scores, although there was evidence that candidates were more likely to achieve high scores on the SJT.
- There was weak evidence that those who felt most positive about the test achieved higher scores on it
- Candidates who felt that the testing time was too short scored lower on the tests.
- It is credible that those differences in perceptions reflect true variation in candidates' knowledge and skills and could be interpreted as supporting the test validity.

Candidate feedback on paper-based SJT

Candidate evaluations of the paper based SJT were collected via the same short questionnaire that had been used in the two previous years of the paper based pilot. The questionnaire was kept very brief because candidates took the test immediately before or after their high stakes training post interview and we felt it unreasonable to ask for more of their time on what was probably a very stressful day. The questionnaire simply asked candidates to rate the SJT, using the 5 point rating scale *1= Poor; 2= Borderline; 3= Satisfactory; 4= Good; 5= Excellent*, for the qualities of:

- Fairness
- Opportunity to demonstrate ability.
- Relevance to selection

242 questionnaires were returned, a response rate of 93%. The percentage of candidate ratings falling in each category of the scale, together with the mean and standard deviation of the ratings, are shown in below. The high proportion of negative ratings for *Opportunity to demonstrate ability* (30%) may be due to the fact that the SJT does not touch on the candidate's clinical skills and knowledge but focuses only on certain non technical aspects of their ability.

Table 25: Distribution of candidate ratings of SJT for Fairness, Opportunity to Demonstrate Ability and Relevance. N=242.

Attribute	1 Poor	2 Border line	3 Satisfactory	4 Good	5 Excellent	Mean	SD
Fairness	0.4%	7%	30%	48%	15%	3.70	0.82
Opportunity to demonstrate ability	7%	23%	32%	35%	4%	3.06	1.00
Relevance to selection	2%	12%	40%	37%	8%	3.35	0.89

Effect of candidate variables on feedback ratings for paper based SJT

We examined whether feedback ratings were related to any of these demographic variables. None of the three ratings (*Fairness, Opportunity, Relevance*) differed between the sexes [Mann-Whitney test, p = 0.897, 0.074, 0.306] or between the ethnic groupings *White* (N=176), *Asian* (N=30) and *Other* (N=17) [Kruskal-Wallis test, p = 0.624, 0.103, 0.639] but all three ratings were significantly higher from non UK trained doctors [Mann-Whitney test, p = 0.001, <0.001, 0.007], although there were only 15 such doctors in the sample.

None of the three ratings (*Fairness, Opportunity, Relevance*) was correlated with the candidate's score on the SJT but there was a slight positive correlation between ratings for *Opportunity to Demonstrate Ability* and age [Spearman's rho = 0.16, p=0.013]. This suggests that older candidates, who presumably have more experience of the sort of scenarios presented in the SJT, may feel more confident about their ability to tackle the test. This hypothetical confidence was not however borne out by the actual test scores which were slightly negatively correlated with age [Spearman's rho = –0.18, p=0.005].

Free text comments about both computer delivered tests
The candidate survey included the following three free response items:

- The thing I liked most about the test was…
- To improve the test I would…
- To improve the selection process I would…

Candidates were able to type responses to these items in text boxes with no word limit. The typed responses were collated verbatim, along with the other survey information, by the Pearson Vue computer system. In total, survey data was available for 659 candidates. Each candidate's responses were converted into an individual text document and imported into NVivo 8 data analysis software (QSR International). The responses to each question were coded and organised into themes and simple statistical analysis was done on the coded groups.

As stated previously, the same survey instrument was used whether candidates took the CPS alone or the combined CPS-SJT test, so it is not always possible to determine whether a candidate is referring to a particular portion of the test in their response, unless they specifically indicate this.

Results
Codes are presented as tables showing the main emergent themes. Illustrative quotes and further relevant analysis are then presented.

The thing I liked most about the test was…

Main theme	Sub themes	Frequency
Set up	Ease and organisation	103
	Suitability of location	47
	Choice of locations	10
Functionality of computers	Navigation	85
	Review of questions	39
	Progress timing	13
	Images	3
Test specification	Breadth of coverage	102
	Question type and quality	82
	Relevance	75
	SJT specifically	14
	CPS specifically	13
	Test length	3
Overall issues	Preparation for real selection	75
	Perceived fairness	26
	Did not like	9

Table 26: Distribution of themes regarding positive aspects of test

Set up
The most frequent group of responses to this question concerned issues around organisation prior to the test itself. The booking system with a choice of locations near to the candidate was popular:
"Easy access to test centres and booking opportunity." [candidate 385]
"Ease of organisation, if it was to be an additional way of selecting candidates would seem to be easy/flexible to organise." *[candidate 802]*

Computerised testing
The use of computerised testing was also popular and candidates commented on most of the special features of the computerised system, including the layout, navigation, review feature and progress timer:
"It was done on the computer so very easy to click on chosen answer and to change answers. It was also nice to be able to go back to previous answers using the review button." [candidate 938]
"Easy format with clear 'timer countdown' to see how much time you could / could not afford per question." [candidate 234]

Test specification
The test specification was also commented on favourably. As candidates did not always specifically indicate which section of the test they were commenting on, it was not always possible to tell whether comments related to the CPS, the SJT or both. However, an analysis of comments by type of test taken (CPS only or CPS-SJT) showed that candidates who took the CPS only were more likely to comment favourably on the breadth of coverage [65/102 comments, Fischer's exact test p = 0.0015] than candidates who took the combined test. There was no difference between the groups for other test factors, although some candidates did choose to comment on a particular section of the test specifically. Illustrative comments included:
"Wide variety of topics – medical, paediatric, palliative care, surgical topics." [candidate 221]
"(questions) clear and not too long instructions." [candidate 139]
"CPS was fairly well organised and thought out with reasonable questions." [candidate 395]
"The Situational Judgement Test allowed you to demonstrate prioritisation and certainly see the benefit of this as part of the shortlisting process or the process thereafter." [candidate 303]
"It was relevant to some of the things I have come across in my training." [candidate 232]

As expected, several candidates commented that they found the test useful as they felt it gave them some preparation for the live selection process.

Some candidates chose to use this free response section to comment on the fact that they did not like this type of test. These were, however, very much in the minority:
"There was not a great deal I liked about the test." [candidate 331]

To improve the test I would…

Main theme	Sub themes	Frequency
Before the test	Preparatory information	43
	More choice location and/or time	19
Computerised testing	Improve image quality and/or size	42
	Improve facilities	40
	Allow breaks	24
	Improve computer capabilities	12
	Clarify instructions on computer features	8
	Prefer paper test	4
Test spec. & construction	Questions ambiguous	63
	Use other topic areas	33
	Use alternative question types	26
	Make more specialty specific	25
	Concerns about level	21
	Ensure equal specialty representation	14
	Avoid duplication of topics	13
	More judgement testing	6
	Less cognitive testing	5
	Divide into specialty areas	3
	Make more real life	3
	Less SJT	1
	Make more challenging	1
	Add negative marking	1
Test length	Decrease number questions &/or increase time	126
	Reduce item length	17
	Allow less time	1
After the test	Instant feedback	12

Table 27: Distribution of themes of candidates comments on how to improve test

Before the test

A number of candidates would have liked more information before the test:

"Give more information to candidates about what to expect and what sorts of questions will be asked." [candidate 74]

Some of the issues commented on reflect decisions taken by the test administration team, such as having an email only contact point. A number of candidates would have liked more choice of locations to take the test and more choice of testing times:

"Make it easier to contact the testing centre via phone directly, rather than having to wait for email responses or being passed around various departments on the phone." [candidate 387]

"Provide the test on more dates, to enable more people to attend. Give more notice of the date of the test, to allow time to be booked off work." [candidate 141]

All these factors will need to be considered in further live test administration.

Computerised testing

Several suggestions were made to improve the computer capabilities and the comfort of taking a computerised test. An important factor that hadn't been fully considered in test administration was the prolonged use of VDU screens:

"Maybe give a 5 to 10 minute break away from the computer screen between the 2 tests. Looking at a computer screen for 3 hours is testing." [candidate 1257]

Other suggestions included:

"Have the facility to enlarge images used in the test, e.g very difficult to view ECG image as so small." [candidate 493]

"Give a warning, about 5 minutes before the end of the test, rather than it just ending suddenly." [candidate 1173]

"Include a question stem for each question in the review page." [candidate 1352]

These were all generally minor points but would help improve candidates' perceptions of computerised testing. A small minority of candidates specifically commented that they prefer paper and pencil tests.

Test specification

There was a significant variety of comments on the test construction and specification again, although there was little clear consensus of opinion here. The most commonly reported issue was that of potential item ambiguity, within both the CPS and the SJT test, and some candidates felt very strongly about this:

"Make some of the questions clearer- sometimes it seemed to me there were several options that were correct." [candidate 617]

"Avoid having SJT as a paper test, unless it was conducted as a oral exam where you can justify your reasoning for a particular order or ranking in particular situations. Many of the questions were opinionated and debatable and it would be unfair to calculate a doctor's empathy and professional skills and allow that to influence shortlisting when one hasn't even meet the doctor or interacted with the doctor face to face." [candidate 472]

"Possibly make the answers less vague as for some questions it felt like there were several answers which you would do in the situation described, but knowing which was the correct or 'most appropriate' was difficult. A clear right or wrong answer would have been better." [candidate 610]

There was no significant difference in the frequency of these comments between candidates who took the CPS only and those who took the combined test. This is a factor that can be difficult to unpick, as a well constructed Single Best Answer (SBA) question will have more than one plausible distracter in the list of options to avoid simple factual recall and encourage clinical problem solving[16]. This may be perceived as ambiguity by the candidate. In addition, SJT questions explicitly deal with professional 'dilemmas' where careful weighing up of options is required[17].

Test length
Several candidates also commented on the length of the test and would have preferred a combination of more time per question, less testing time overall and less questions to answer: "Reduce the length of the question. Reduce the number of questions - I felt the test was too long (especially if doing combined SJT + CPS). Give more time for the SJT - it was difficult to complete this particular test in the time allocated." [candidate 628]

To improve the selection process I would …

Main theme	Sub themes	Frequency
Initial application	National coordination	18
	Give more info about expectations	15
	Modify application form	11
	Timing	8
	Apply locally	6
	Apply to single specialty	1
	Two start dates	1
Format of selection	Include a written test	121
	Face to face structured interview	63
	Continue current process	35
	Do not include testing	19
	Test clinical skills	16
	Use selection centre methods	15
	Use references	15
	Use existing PG assessments	11
	Consider foundation progress and/or assessments	5
	Test English language ability	1
Overall aims of selection	Opportunity to demonstrate achievements	20
	Less weight on non clinical achievements	10
	Concern about disadvantage to some groups	8

Table 28 : Distribution of themes on how to improve selection process

Again, there was little clear consensus amongst candidates in response to this question.

Initial application
There were some comments on the application process; some candidates commented that they would like to see national selection procedures (but many specialties have now moved to this procedure, which may explain why this comment did not occur often). A small number of candidates specifically commented that they would prefer a local application process, but these were in a minority.

Written testing

There were several comments recommending the inclusion of a written test of some format as part of the procedures:

"Take into consideration clinical knowledge and situational judgements assessed in this test alongside other information as given on application form." [candidate 676]

However, there were also several candidates who felt this type of testing was included elsewhere in a doctor's career and should not be added at this stage:

"I would not rely on a test, there are plenty of tests to test knowledge and safety of doctors throughout their career progression." [candidate 940]

In particular, applicants for core medical training commented that they were rewarded for gaining their MRCP Part 1 as part of the application process and that this test duplicated much of that. This is clearly part of a wider issue about the suitability of using postgraduate examinations as part of selection criteria:

"Most candidates will have either passed or attempted part 1 of MRCP by the time they apply for CMT and so will already have evidence of their core knowledge." [candidate 1367]

Interviews

Candidates were generally in favour of keeping a structured and/or face-to-face interview as part of the process:

"Have an interview for all specialities as it is impossible to assess someone via online/ computer assessments only." [candidate 205]

"Base decision making for jobs on interview and questions face to face too." [candidate 247]

There were also a number of comments regarding keeping the current process. 24/35 (69%) of these came from candidates applying to GP as their first choice.

"I feel that the current selection process for GP is very structured and sounds fair although I have not been through the process yet." [candidate 55]

Overall aims

A small number of candidates commented on the overall aims of selection and were concerned about potential disadvantage to some groups:

"I'm not sure - at the moment you get the impression that the people who do well are the people who spend more time doing audit etc instead of spending time actually doing their jobs whilst other people pull up the slack but don't have good cv's. Doesn't seem fair really!" [candidate 863]

"Doctor's apply from a wide ranging background, some have been working in different areas and the CPS would not reflect this clearly. The CPS is ok for Doctors who have only just finished their foundation year training but I am not sure it is a fair way to assess Doctors in other situations." [candidate 118]

Conclusions
Overall, the pilot test, both CPS and SJT, were well received by candidates, as has been highlighted in previous sections of this report.

Computerised testing was popular with candidates, particularly some of the capabilities allowed by using a computerised format. However, there are particular issues around computerised testing that may need to be more carefully considered if this is to be adopted in the future, such as safe use of VDU screens and careful instructions to candidates.

The breadth of topic coverage in a generic CPS is popular, as well as the use of questions in an SJT that test outside the cognitive domain. However, candidates remain concerned about potential ambiguity in these types of questions and in a high stakes process worry that questions with no straightforward right or wrong answer are unfair. Although the inclusion of a written test of some format is generally acceptable, they would like clarification on the purpose of the test in a selection process and reassurances that this will not be used as the sole method of selection.

8. Conclusions and Recommendations
Professor Jane Dacre, Professor Fiona Patterson

Conclusions
There has been a significant increase in interest in the development of robust selection methods for entry into higher specialist training in medicine and its specialties. It is clear that the old system of individual Deaneries or Trusts short listing and interviewing in a poorly coordinated and non standardised fashion is no longer desirable. In addition, there has been considerable advance in the development of robust evidence based techniques for selection in medicine, learning from the experience of occupational psychologists and recruitment specialists.

Following MTAS, colleagues have learned about the importance of a thorough evaluation and piloting phase of any selection instruments which are to be used in live selection processes.

This pilot project has demonstrated that it is feasible and practical for a consortium to work together to develop a generic MMT (including SJT and CPS items) that adds incremental reliability to the overall selection process. The tests are reliable and have been shown to improve the rigour of the selection process. Feedback from volunteers suggests that a MMT is an acceptable component of a selection process and is seen to improve the fairness of the approach. This pilot was administered successfully as an on-line test. It opens up possibilities for such a test being made available out of hours and over a stipulated time period, which would reduce the amount of time that candidates would need to be away from the workplace.

Since the initiation of this project, the economic downturn has begun to affect the sector and has resulted in a need to re-evaluate what is proportionate and achievable in the short and medium term. Investment in the development and validation of a bespoke suite of selection instruments for individual specialties is unlikely to be met with enthusiasm in the current climate. The results of this project have shown that collaboration between specialties may be the key to continuing the momentum of this work in the most cost effective way. It is clear that there is more commonality than difference in the characteristics of the specialities included in this pilot. One way to exploit this commonality is to explore the possibility of the aggregation of specialties into ‘Job Families’ which share attributes and which could also share a common bank of assessment instruments.

Currently, competition ratios are variable between specialties. A pragmatic approach may be to consolidate the pilot work that has already been completed, to ensure that the data sets have been carefully examined and to focus on the implementation of an additional MMT to support the short listing process in the Job Families of the high competition specialties, where there is a clear need to discriminate reliably and fairly between candidates.

In summary:
- It is feasible to hold SJT and CPS tests as part of the ST1 selection process – it would be time and cost efficient, would accurately discriminate between candidates.
- These tests are acceptable to candidates and perceived to be fair.
- The specialties and deaneries worked well together in this collaborative exercise.
- Use of either an SJT or a CPS test in an early stage of the selection process would increase predictive validity and the combination of both types of test would enhance this.
- The cost of staging such tests would be in the region of £55-65/candidate, with lower costs, but greater effort, being incurred if local test venues were used instead of managed test centres.
- Computer delivered tests can be used to replace shortlisting processes currently employed by many specialties and deaneries.

Recommendations

- An SJT and CPS test, probably common to a group of related specialties, should be an integral part of the National Health Service's ST1 selection procedures in future years, most probably as a component of candidates' overall Selection Process Scores.
- This overall score should also take into account of performance in a Selection Centre and qualifications, experience and research interests, as set out in their (simplified) portfolios.
- Elimination of the shortlisting process, so that every candidate receives at least one interview, would enable the tests to be staged at the same time as interviews or Selection Centres, thereby streamlining the process.
- The combination of these machine markable selection instruments would provide a fair test, and reduce, the expense and workload associated with the current processes.

8. References

1. McDaniel MA, Hartman NS, Whetzel DL, Grubb WL. Situational judgement tests, response instructions and validity: A meta analysis. *Pers Psych* 2007; 60: 63-91.
2. Chan D, Schmitt N. Situational judgement and job performance. *Hum Perform* 2002; 15: 233–54.
3. Weekley JA, Ployhart RE. Situational judgment: Antecedents and relationships with performance. *Hum Perform* 2005; 18: 81-104.
4. Clevenger J, Pereira GM, Wiechmann D, Schmitt N, Harvey VS. Incremental validity of situational judgment tests. *J Appl Psychol* 2001; 86; 410–417.
5. Lievens F, Buyse T, Sackett PR. The operational validity of a video-based situational judgement test for medical college admissions: illustrating the importance of matching predictor and criterion construct domains. *J Appl Psychol* 2005; 90: 442–52.
6. Patterson F, Baron H, Carr V, Plint S, Lane P. Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. *Med Educ* 2009; 43(1): 50-7.
7. Patterson F, Carr V, Irish B, Price R. A predictive validity study to evaluate selection methods for training in general practice. *Proceedings of Ottawa Conference* 2010 May; Miami, USA.
8. Patterson F, Zibarras L, Carr V, Irish B, Gregory S. Evaluating procedural justice in postgraduate medical selection. *Med Educ* 2010 *(accepted)*.
9. Patterson F, Carr V, Zibarras L, Burr B, Berkin L, Plint S, Irish B, Gregory S. New machine-marked tests for selection into core medical training: evidence from two validation studies. *Clinical Medicine* 2009; 9(5): 417-20.
10. British Medical Association *Examining equality: A survey of royal college examinations.* BMA, May 2006
11. Dacre J, Spencer H, Collett A, Morris R, Sales D. *Clinical Problem Solving Test Pilot Report*. Royal College of Physicians, November 2008.
12. Dewhurst NG, McManus IC, Mollon J, Dacre JE, Vale AJ. Performance in the MRCP(UK) Examination 2003–4: analysis of pass rates of UK graduates in relation to self-declared ethnicity and gender. *BMC Medicine* 2007, 5:8
13. Matveevskii AS, Gravenstein N: Role of simulators, educational programs, and nontechnical skills in anesthesia resident selection, education, and competency assessment. *J Crit Care* 2008; 23: 167-72
14. Patterson F, Ferguson E. Selection for medical education and training. *Understanding Medical Education; ASME Publications* 2007.
15. Kearney RA: Defining professionalism in anaesthesiology. *Med Educ* 2005; 39: 769-76
16. Patterson F, Ferguson E, Thomas S. Using job analysis to identify core and specific competencies: implications for selection and recruitment. *Med Educ* 2008; 42:1195–1204
17. NBME, 2002. *Constructing Written Test Questions For the Basic and Clinical Sciences* (3rd edition). Philadelphia, PA: National Board of Medical Examiners

Annex A Project aims and communications strategy

A.1 Test Purpose and specification

Fiona Patterson, Jane Dacre

1. Background & Rationale

1.1. Research evidence shows that well-designed computer delivered tests can provide **valid & efficient shortlisting tools**, with the potential to reduce selector shortlisting time whilst increasing the reliability and validity of the shortlisting process. Computer delivered tests used for shortlisting do not normally replace an interview/selection centre; rather they are used in combination as a component of a selection process.

1.2. The Academy of Medical Royal Colleges is piloting a **multi specialty machine markable Test** (MMT) in January 2010 building on previous work from the GP MMT used for shortlisting purposes.

1.3. The 2010 test is a **pilot** and results will not contribute to the live selection process for any of the participating deaneries or specialties. Participation in the pilot is voluntary. In 2010, the pilot MMT is being delivered via computer at Pearson VUE testing centres and in University examination halls. The test will last for 2 hours and contains over 120 items.

1.4. The **rationale** for the pilot is to explore whether a reliable & valid multi-specialty Clinical Problem Solving test can be developed, i.e. a single test that will provide reliable and valid evidence regarding candidate aptitude for training in a range of postgraduate specialties.

2. Purpose of the Test

2.1. The purpose of the MMT is **selection** into specialty training, i.e. to test **aptitude** for specialty training. It is **NOT** designed as a test of attainment of knowledge or achievement of Foundation training competence (i.e. unlike an examination).

2.2. The function of the MMT is for **shortlisting** candidates into specialty training, potentially alongside or as a replacement for other shortlisting tools (e.g. application forms). The tests will not replace the interview/selection centre stage.

3. Test Specification

3.1. The MMT being piloted in 2010 is a **Clinical Problem Solving** test. This aims to assess the application of clinical knowledge and expertise in content areas that are relevant to specialty training and are mapped to the Foundation curriculum.

3.2. The MMT is **designed for selection into ST1/CT1 only**. The test therefore does not assume any knowledge or experience of the target specialties beyond that expected to be gained during Foundation training or equivalent.

3.3. The MMT test specification uses the **Foundation curriculum** to define appropriate areas of clinical practice to be covered in the test, to ensure it is pitched at a level commensurate with entry into specialty training. Items have been submitted by subject matter experts to cover all areas of the Foundation curriculum. All items have established psychometric properties demonstrating good performance.

4. Evaluation

4.1. The evaluation of the MMT will include test and item level **psychometric analysis**, reliability & validity (face, content, criterion related, including predictive validity), **fairness, candidate reactions, logistics/utility,** and **stakeholder acceptance**.

A.2 Communications strategy
Max Pragnell

This communications plan is set out to ensure the successful introduction of a pilot study into a new selection process for doctors entering speciality training. It is based on a messaging workshop held on 10-Nov-09 between members of the AOMRC pilot team) and Rosie Carlow and Max Prangnell of AoMRC and *Milbank Media*.

Overarching objective
To successfully carry out a nationwide set of tests on Jan 9th 2010 for doctors entering the speciality selection process.

Communications objectives:
1. To ensure that the pilot study is implemented with as little negative comment or references to previous online testing systems and selection process as possible.
2. To ensure sufficient take up of the study (more than 1100 attendees) so that sufficient data can be gathered to judge the efficacy of both the format and process of online tasting.

Communications strategy
Our ambition must be for a 'fuss-free' approach. While we may have views about the ultimate direction of travel for the speciality selection process, this pilot study is not the place to air them. Our job is to get the pilot study done so that we can make informed decisions later based on the evidence gathered now. Without the evidence in the first instance there can be no meaningful conversations later.

Our approach must therefore always be to be low key, inclusive and collegiate. This is about taking some gentle first steps towards finding a fairer selection process – it's about getting it right for doctors.

We should aim to keep this item off the radar of the more vocal groups for as long as possible, but some discussions – as outlined below and particularly with the JDC should be embarked upon now.

Third party advocacy from respected figures will be invaluable.

Our overall approach is that this is a pilot study and we've got nothing to hide.

In pure communications terms, this pilot study is achievable with minimum fuss provided we keep it in perspective. If we do not want some doctors jumping to conclusions about the 'inevitable' direction of the study then we should do all we can to lessen that exposure.

Stakeholders and influencers
The table overleaf is a snapshot now of influencers. It will be further populated as the project progresses and does not take account of groups with no direct input such as BAPIO or Fidelio.
It is important to introduce a regular feedback and monitoring mechanisms. We need to know what the other side is saying. This should happen naturally at weekly project meetings. RC will monitor the various websites above for news and chatter.

Our primary channel of communication will be our website and one-to-one advocacy. Any calls to the Dept. of Health press office should be referred to MP in the first instance.

Key Stakeholders & Opinion Formers

Organisation	Influence	Risk	Approach	Delegate
BMA/ JDC	High	Med	Inclusive. We should establish a formal reason to involve the JDC in the evaluation process. They should feel enfranchised if possible, but if not then at least not excluded from the process.	JD and 3rd party advocacy (LE?)
REMEDY	Medium	High	None – let them come to us when and if they have a problem. We will not engage with single issue pressure groups	None
Doctors. Net	Medium	High	Take out paid-for advertising for the pilot study on this group's website?	HS
Medical Royal Colleges	Medium	Low	Inclusive. This is a pilot study carried out in part by the AoMRC to establish whether a fairer selection process can be implemented in future	Direct request to MRC presid-ents to cascade to specific c'tee chairs (ND).
BMA/ CCSC	High	??	Unknown	TBA
Deaneries	High	Low	Inclusive and underway	JD et al
Specialist Press	Medium	Med	None – responsive rebuttal only	MP
General Press	Medium	Low	None – responsive rebuttal only	MP
DH	Potential for -ive influence	Low	The DH should not have any fingerprints on this study.	JD/MP

Key message
This pilot study will enable deaneries, specialties and applicants to design the best method of selecting doctors to training posts. It's about testing a system which we hope will be more effective, more open and provide more support to doctors during the selection process. There's no pass or fail and it won't replace interviews, but it is about finding a fairer way of ensuring the right people get the right jobs.

Supporting messages
The tests themselves are of the same type that GP's, lawyers and accountants have been taking for years. They'll be marked centrally which will be fairer because everyone will be marked in the same way – so there'll be consistency across the country. The results will help doctors see where their strengths are and which specialties are most appropriate for them as they progress their career. The focus is less on establishing what doctors know and more on establishing what they're good at.

If you take part in the study and give us feedback you can have a real say in the shape of selection testing in the future. For most applicants the sessions will be held locally and will last just over two hours. You'll get valuable experience of a test system you are very likely to encounter later in your career and we'll make sure you're not out of pocket. You'll also get a certificate to say you took part.

Final selection will only be made after interviews have taken place. This pilot study is to help develop an aid to the selection process that offers clarity, consistency and takes a common sense approach when it comes to finding out who is good at what.

Annex B Notes from the two workshops
Both these events were working sessions, designed to augment data already gathered by the pilot. Summary notes from the two workshops are given below.

B1 SJT workshop
Attendees: Ian Anderson, Vicky Archer, Colin Campbell, Alison Carr, Victoria Carr, Jane Dacre, Luci Etheridge, Tom Gale, Sue Heenan, Nora Pashayan, Fiona Patterson, David Rowley, Hilary Spencer, Mark Stott, Alison Sturrock, Kath Wolf.

Objectives
- To summarise current research evidence regarding the use of SJTs in postgraduate medical selection.
- To share evidence and learning from SJT pilots conducted to date in individual specialties.
- To review & summarise learning to date regarding the use of SJTs to assess non-cognitive domains in postgraduate specialty selection.
- To explore areas of commonality and divergence between specialties with regard to SJTs.
- To discuss potential further work to develop SJT(s) for use in PG specialty selection

Desired outcomes
- A policy statement to summarise current thinking regarding the use of SJTs in PG specialty selection.
- A proposal for further work to develop SJTs for use in PG specialty selection.

Presentations
- Introduction & objectives JD
- SJTs in postgraduate medical selection: research evidence FP
- Summary of SJT development in individual specialties
  - GP VC
  - ACCS/Anaesthesia TG/IA
  - Surgery DR
  - Public Health NP
  - CMT JD
- Considerations for future development of specialty SJTs FP/JD
- Next steps FP/JD

Discussion

All discussions were held in open session, during the presentations. Points made were as follows.

- Selection tests are very different from tests of competency; SJTs may be suitable for one but not for the other
- There is considerable research evidence about the use of SJTs in selection; Warwick, Durham & Newcastle Universities have done literature reviews on selection methods
- It is important to be clear whether selection into ST1 should be made on knowledge or personality or potential aptitude or what – the same type of test is not necessarily suitable for all of these
- It could be questionable whether there is any difference at all between applicants for different specialties at ST1 level – is there any point trying to test for it?
- It's unclear whether current shortlisting methods might be replaced or just augmented by instruments such as an SJT
- Some conclusions from this pilot regarding SJTs may be that:
  - it's both possible and useful to include an SJT in the ST1selection process
  - writing SJTs is difficult, and takes considerable time and skill
  - selection using SJTs is not yet evidence-based (but neither is CV-based selection)
  - the evidence in favour of the current ST1 selection process is no longer sustainable
- It would be interesting to compare results of the SJT with those of the CPS test for each candidate who took both
- We should look for clusters of applicants to particular deaneries/specialties, to see which cohorts might be big enough to analyse individually
- Currently, every specialty is doing ST1 selection differently from the others; will this be considered acceptable in the future?
- Every specialty should do a job analysis; there may a set of generic traits, common to all specialties, plus some specialty-specific ones. If so, this could be reflected in any SJT to be used in ST1 selection, with perhaps some 'contextualisation' of the common questions to each specialty, to give them face validity
- Specialties should share SJT questions they have developed, to see if they're in fact the same but contextualised, or if not, whether they could be used by other specialties after contextualisation.
- It would be useful to derive, from the presentations at this workshop, a guide on "Steps required when designing and validating an SJT" as they all seem to include common elements
- SJTs should be developed to a common standard – the cut-off level can be varied according to how popular a specialty is, but the test standard should be constant
- It is important to avoid SCs drifting into becoming OSCEs – the two are different
- Someone needs to write a Good Practice guide into the use of SJTs for selection – there are too many examples around of their misuse, which could devalue their reputation as a reliable tool
- The trainees' views need to be taken into account in our reporting – the JDC invitees to the workshop could not attend but JD is meeting them next week to get feedback, and the test participants' feedback will also be analysed
- The objective of all these pilots is to find a process which is "more rigorous but not more onerous" (to quote from one of the presentations)
- It appears that the SJT might look at different things from those assessed by other elements of current selection processes – if so, it should augment existing methods, not replace them
- It's very important not to lose the value of the portfolio in the selection process but it may be better to consider it within a SC rather than at the shortlisting stage
- A good output from this workshop (and this pilot) would be a recommendation on how best to go forward with analysis, design and implementation new ST1 selection process(es)

Next Steps

- It was generally agreed that:
- the ‘volunteer’ model is no longer useful – any further analysis needs to be on the total ST1 candidate population, so any further pilot tests must be compulsory; all our reporting must make this absolutely clear
- we need to canvass all our stakeholders’ views (eg our SG) before writing a consensus report from this workshop for next SG meeting
- The following need to be considered further:
- whether an SJT could be used within a Selection Centre, or whether it is just a short-listing tool
- establishing a ‘collaborative’ selection process (across all deaneries & specialties) with all ST1 candidates going through all stages, so the results can be comprehensively analysed
- if AoMRC should lead another pilot in 2011, involving all MRCs, with common CPS test & varied SJTs (and/or common but contextualised questions)
- who might fund further work, when the DH stops doing so (in 2011?) MRCs? Medical Schools? Deaneries? Who benefits? Who would be responsible & own the results?
- if it might be worth doing a small project to work out exactly what should be done the following year, instead of leaving it up to individual deaneries/specialties/pilots
- if we should do a generic/multi-specialty SJT, on a compulsory basis, next year
- The recommended process would be:
- Review results of all pilots to determine which jigsaw parts are still missing
- Check our data for each non-GP application to see if SJT works equally for other specialties
- Hold a (mandatory) SJT + CPS test for all ST1 applicants 2011?
- Repeat with extra SJT questions, specific to specialty applied for 2012?
- Repeat, but with all deaneries & specialties doing everything 2013?

B2 stakeholder workshop

Attendees: John Adams, Gavin Anderson, Graham Buckley, Stuart Carney, Alison Carr, Victoria Carr, Helen Cugnioni, Jane Dacre, Nicola Dagnell, Neil Dewhurst, Paul Dilworth, Ian Doughty, Luci Etheridge, Siobhan Fitzpatrick, Ashley Fraser Gordon French Tom Gale Stephen Harding Sue Heenan Kim Hinshaw Humphrey Hodgson Matthew Huggins Bill Irish Peter Lamont Susan McCArthy Angus McGregor Graeme Muir Fiona Patterson Katie Petty-Saphon Marcia Reid Martin Roberts Hilary Spencer Mark Stott Alison Sturrock Kath Wolf Adrian Woodthorpe

Objectives

- Share results from the pilots
- Discuss the implications
- Recommend a way forward

Presentations

• Introduction	JD
• Logistics of CPS test	HS
• Test performance-CPS	AS
• Candidate perceptions	LE
• Comparison with live selection data	TG
• Report on the Anaesthesia/ACCS Pilot	FP
• DH perspective	AC
• Discussion	All

All discussions were held in open session, during the presentations. Points were made as follows.

• Local or national selection? Most specialties still recruiting at deanery level are planning to go national; they've concluded that local selection:

- leads to a lot of process variation, even if national standards have been agreed
- is too hard to evaluate (e.g. that the process was fair & gave the best outcomes)
- can lead to excessive applications (e.g. one candidate made 31 applications this year)

However, moves to a national process should be taken 1-step-at-a-time, to get acceptability.

- MMT or not? Delegates agreed that the 'ideal' selection process was probably an eclectic mix of the various available instruments, might vary according to circumstances (e.g. the applicants : posts ratio) and was still unknown, despite the considerable amount of research which has been done.
- CPS or SJT or both? All the evidence so far indicates that each test type has incremental validity over current methods & that candidates don't mind doing both. The consensus of the meeting was that the best way to improve selection methods would be to scrap shortlisting– it's universally unpopular, adds almost nothing to the process, can be onerous & can lead to acceptable candidates failing to get interviews. Checks currently made in shortlisting could be done at interview with little loss in efficiency. Then an MMT, if used, could be set at the interview (as some specialties do now) which would be cheaper than setting it in advance.

NB: It was noted that information provided by candidates at shortlisting was not always true, so had to be verified at interview, which was another reason why shortlisting adds little value

- Multiple interviews? It was generally felt that 2nd interviews added nothing to the process, it was better to offer everybody one interview & move towards a process where deaneries would accept other deaneries' interview results.

• Acceptability of using an MMT (post MTAS)? This question generated a number of points*:*
- The pilot teams reported that it had still been a serious issue (eg with the JDC) at the start of the pilot but, as in other pilots, the predicted hostility never actually materialised.
- Feedback from candidates (admittedly, a self-selected group) had been generally good.
- Specialties already using an MMT (e.g. GP) reported no complaints from candidates.
- Specialties who'd tried using an MMT (eg Medicine) said they'd taken pains to 'sell' it to candidates, and thereby avoided acceptability problems
- 'Acceptability' begged the question 'for what?': shortlisting/overall score/increasing standardisation; the question couldn't easily be answered without clarification on this
- The meeting concluded that, with care, an MMT would be acceptable to most stake-holders (candidates and recruiters) as some part of the selection process.

• Correlation of MMT scores with actual outcomes. It was observed that the correlations aren't in fact outstanding (but 'test + interview' scores correlate better with outcomes with than interview scores alone do)

• Fairness of process: An MMT can be useful to level the playing field eg between candidates who've recently qualified & those with more experience. But the process should still take into account all the other things (publications, audit work, etc) currently covered by shortlisting.

• Need for specialty-specific tests: This was felt to be an important question, as yet unresolved. It was suggested that a single test might be seen as biased towards some specialties and delegates wondered about tailoring a standard MMT by specialty rather than adopting a one-size-fits-all approach. It was pointed out that most ST1 candidates had similar know-ledge & experience and that even SJTs, which aimed to test aptitude for a specialty, ended up fairly similar for different specialties. However, face validity was thought to be important: candidates for e.g. radiology may not like questions phrased in GP terminology. To this end, work was ongoing on 'job families' but it was not yet known if this would be the answer (and some delegates doubted the necessity for separate tests, when their similarities far exceeded their differences).

• Sample bias? Contrary to some delegates' fears that the test group might have been untypical (e.g. just the "keener and better" candidates), analysis had shown that the sample was in fact representative on ability as well as demographics (that is, their outcomes from the real process were distributed proportionately across the range of possible outcomes).

• Terminology: It was agreed that "aptitude" was a better term to use than "trainability" in discussions of what SJTs might be able to measure

• Why use an MMT when everyone gets an interview? It might still be useful for:
- calibration & standardisation of nationally-coordinated, locally delivered recruitment;
- fairness, esp. for junior doctors with less experience but enough aptitude for ST1
- as part of a selection centre, to give a more complete picture of a candidate's abilities (GP is currently looking at including an MMT (including a SJT) in its overall SC score, both to benchmark the SC and to introduce more standardisation into the process;
- oversubscribed first-choice specialties and deaneries, to balance applications and posts;

The important thing is to have a tool kit of well-understood selection instruments available for use as appropriate to each specialty, deanery and circumstances.

(Note of caution: there might be ethical objections to requiring candidates to take an MMT whose main function was to assist research and, conversely, to using them to score candidates when they are still insufficiently researched)

• Predicting long-term performance: Very little research has been found on how SJTs compare with SCs in predicting candidates' performance over the long-term but what does exist shows SJTs to be the better single independent predictor for this.

• Funding: It seems unclear who should be funding work such as these pilots when it is the deaneries and candidates who stand to gain from them. In recent years, the deal has been that the deaneries fund the current (live) process and the DH funds research into improved processes but now that no more DH funds are available, money will have to come from the elsewhere (e.g. deaneries, specialties or the candidates themselves) and/or something else will have to be dropped to release the necessary funds. Suggestions included scrapping shortlisting and using SJTs instead of, rather than as well as, some SC stations.
NB: Need to remember that the cost of a failed ST trainee or failed doctor in their later career far outweighs the cost of researching & using accurate selection methods.
• *Same-day testing*: the issues associated with wider use of MMTs in specialty selection included that of volume; it would not be possible to release every ST1 candidate to take an MMT simultaneously but multi-day testing required test-equating and a much bigger pool of questions to be available.

Summary
The following summary of the proceedings was agreed at the end of the workshop:

The pilot was a success
• It worked – the tests were staged successfully, the participating specialties and deaneries worked together harmoniously, and a wealth of useful information was generated as a result.
• Both the CPS and SJT provided incremental validity to the current selection process, with a combination of the two adding the most value as a predictor of eventual outcomes from the process.
• All indications from the pilot results are that a common test could be used for selection into some, and possibly all, specialties at ST1.
• The process was generally acceptable to candidates, and other stakeholders (specialties, deaneries, BMA and the DH).

But
• Its costs were not insignificant and were extra to those of the current process. In the current financial climate, alternative funding would need to be found or (more likely) something else would have to be dropped if MMTs were to be introduced as part of the ST1 selection process.
• The consensus of opinion was that the various shortlisting processes currently used by many specialties and deaneries add the least value to the overall selection process and could be dropped without significant detriment to selection outcomes.

Further work
• A detailed report of the pilot and its findings, including a record of this workshop, will be written and published, to inform the design of future selection processes
• Some further work within current budgets may be done into job families and whether a single test for each would be better than a common test across all specialties.
• It would be very useful to have longitudinal data (on how well performance at ST1 selection correlates with subsequent performance in the job) but it is unclear how this might be obtained or who would fund such work.

Annex C Process design, confidentiality and ethics
The pilot's process design was strongly influenced by data confidentiality and ethical considerations. The resultant design and the measures taken to ensure informed consent was obtained are shown in this Annex

C1 Process design
The following process description was agreed by all participants in December 2009.

Application Process Description
The objective of this paper is to clarify the process through which applicants to the participating deaneries & specialties will be invited to take the CPS test & (Anaesthetics / ACCS applicants only) the SJT

Specialty	Recruitment method	Contact	IT System	Approx # posts	No. of App's	Candidate :post ratio
ACCS	Local (via deaneries)	Local co-ordinators	ICAMs / London system	350	1260	3.5:1
Anaes-thetics	Local (via deaneries)	Local co-ordinators	ICAMs / Lon system	400	800	2:1
CMT	National (via RCP)	SH	Konetic	1100	2,400	2:1
GP	National (via GPNRO)	GE	Konetic	2,700	5,000	2:1
Histopa thology	Central (via London)	ND	Local system	120	200	8:1
Paedi atrics	Central (via RCPaed)	TR	ZMR	500	760	1.5:1

Process
a. Dec 12th-18th – Each contact (see table) e-mails all applicants to the specialty concerned, asking them to participate in the pilot and to follow a link to the pilot's website straight away
b. Candidates follow the link to the PV-AoMRC web-page; this gives more information and asks them to complete a registration form
c. Candidate completes the form (including signing a consent form and stating their preferred specialty) then click on a link to book a test
d. They choose a location (out of 23) & a time slot and the booking is complete
e. If no offered location is suitable, they may contact the help number to discuss other options
f. Regular review of bookings throughout the booking period will enable adjustments to be made to locations on offer, if necessary
g. When the real application process is over, each contact supplies ranking data to enable comparisons to be made with the candidates' results in our test

---

C2 Memorandum of Understanding
The following was agreed between all parties before they shared data and test questions with other participating organisations for the purposes of this pilot.
Memorandum of Understanding between ("the Organisation") and UCL ("the AoMRC pilot team")

Ownership:
All test items (CPS and SJT) shall remain the intellectual property of the Organisation which owned them on the date of this MOU.

Test items may only be used with the express permission of the Designated Individual of the item's owner-Organisation.

Any items that are used for test purposes may be used for one diet only with further permission being sought before any additional use. Currently there is an agreement to use them in test preparation, staging and evaluation for the AoMRC pilot for 2010 only; this would constitute use for one diet.

Security:
Live pilot test items will be stored securely by the AoMRC pilot team in a dedicated item bank accessible only by their Designated Co-ordinator or their Designated Item Bankers.
All test items shall be encrypted for access.

Test items shall be transmitted by e-mail only in an encrypted fashion.

All test items belonging to the Organisation that have been used in the live AoMRC pilot test shall be stored separately in a secure item bank and only accessed for purposes of evaluation of the project. They shall not be used again in any live test without the express permission of their owner-Organisation's Designated Individual.

All items that are not used in the AoMRC pilot test will be returned to the owner-Organisation in an encrypted format and securely deleted from all computer and other storage records; written confirmation that this has taken place will be provided if required.

Evaluation:
Any of the Organisation's test items used in the AoMRC pilot test must be seen in their delivery format for signing off by the Organisation's Designated Individual before use; approval (or otherwise) for their use will be given within 48 hours of it being requested by the AoMRC pilot team.

The use of the Organisation's test items for teaching and training purposes is strictly prohibited. Any evaluation of the specific test items belonging to the Organisation can only be undertaken with the prior approval of the Organisation's Designated Individual. In particular, WPG (on behalf of the NRO) will be invited to collaborate in the academic evaluation of the AoMRC pilot project carried out on behalf of the AoMRC, as an integral part of its pilot, and this will result in co-authorship of any resulting publications.

Any additional evaluation carried out by the Organisation using data collected as part of the AoMRC pilot will only be done with the agreement of the AoMRC pilot team's Designated Individual and will result in co-authorship of any resulting publications.

Designations
The Organisations "Designated Individual" shall be ……..
The AoMRC pilot team's "Designated Individual" shall be ……..
The AoMRC pilot team's "Designated Co-ordinator" shall be ……..
The AoMRC pilot team's "Designated Item Bankers" shall be …….

Signed

…… on ../../…. for the Organisation
…… on ../../…. for the A0MRC pilot team

C3 Candidates' consent form

The following consent agreement was signed by every candidate at registration for a test.

I agree to data collected about me during this test and subsequent recruitment processes being used anonymously for research purposes by the Academy of Royal Colleges selection pilot group. All data will be confidential and stored in accordance with the Data Protection Act 1998. I understand that I am free to withdraw from this pilot at any time. ☐

Annex D Detailed statistical results

CPS item statistics

Item	p-value	pt-biserial	time
I1	0.81	0.28	**37.70**
I10	0.64	0.37	**41.01**
I100	0.66	0.17	**70.23**
I101	0.55	0.19	**43.97**
I102	0.53	0.3	**45.67**
I103	0.8	0.29	**33.55**
I104	0.88	0.27	**38.18**
I105	0.7	0.29	**63.27**
I106	0.88	0.27	**24.04**
I107	0.73	0.38	**34.50**
I108	0.92	0.33	**17.63**
I109	0.77	0.42	**49.99**
I11	0.47	0.18	**56.46**
I110	0.63	0.36	**50.69**
I111	0.13	-0.1	**45.04**
I112	0.87	0.34	**63.95**
I113	0.81	0.32	**55.92**
I114	0.03	-0.05	**38.86**
I115	0.52	0.25	**103.88**
I116	0.9	0.25	**34.66**
I117	0.46	0.34	**132.82**
I118	0.84	0.08	**33.43**
I119	0.66	0.12	**62.56**
I12	0.43	-0.01	**57.84**
I120	0.58	0.36	**37.52**
I121	0.32	-0.04	**54.11**
I122	0.65	0.26	**40.97**
I123	0.56	0.11	**51.30**
I124	0.5	0.32	**49.87**
I125	0.85	0.27	**51.72**
I126	0.57	0.19	**45.87**
I127	0.85	0.26	**36.40**
I13	0.84	0.33	**42.15**
I15	0.6	0.24	**43.71**
I16	0.59	0.17	**38.04**
I17	0.66	0.28	**49.97**
I18	0.38	0.26	**43.76**
I2	0.78	0.28	**48.01**
I20	0.11	-0.15	**78.58**
I21	0.84	0.34	**35.43**
I22	0.79	0.31	**38.10**
I23	0.96	0.31	**23.95**
I24	0.34	0.24	**72.50**
I25	0.65	0.35	**62.54**

Difficulty	Discrimination
Easy (.8-1.0)	Poor (<0.0)
Medium (.41-.79)	Fair (0.0 - 0.20)
Hard (.00-.40)	Good (> 0.20)

Item	p-value	pt-biserial	time
I26	0.31	0.1	**46.21**
I28	0.76	0.29	**43.20**
I29	0.71	0.14	**46.11**
I3	0.85	0.21	**59.66**
I30	0.6	0.28	**57.12**
I31	0.78	0.23	**29.68**
I32	0.73	0.38	**44.36**
I33	0.38	0.16	**65.40**
I34	0.92	0.25	**30.57**
I35	0.87	0.2	**46.90**
I36	0.29	0.21	**53.29**
I37	0.57	0.31	**68.85**
I38	0.4	0.07	**45.97**
I39	0.46	0.27	**66.07**
I4	0.52	0.2	**52.92**
I40	0.6	0.3	**57.41**
I41	0.58	0.32	**34.84**
I42	0.76	0.23	**61.66**
I43	0.65	0.38	**65.33**
I44	0.51	0.1	**60.63**
I45	0.75	0.38	**42.30**
I46	0.65	0.35	**88.35**
I47	0.78	0.21	**45.08**
I48	0.13	-0.09	**28.33**
I49	0.79	0.28	**56.40**
I50	0.66	0.09	**52.63**
I51	0.71	0.25	**44.25**
I52	0.71	0.4	**36.59**
I53	0.59	-0.08	**114.88**
I54	0.73	0.17	**57.26**
I55	0.41	0.34	**95.10**
I56	0.38	0.22	**42.28**
I57	0.36	0.27	**41.80**
I58	0.71	0.26	**62.78**
I6	0.59	0.28	**53.09**
I60	0.47	0.35	**68.56**
I61	0.82	0.41	**82.71**
I62	0.49	0.25	**90.51**
I63	0.85	0.34	**30.06**
I64	0.69	0.34	**63.24**
I65	0.86	0.32	**34.19**
I66	0.66	0.16	**61.69**
I67	0.37	0.16	**44.34**
I68	0.97	0.2	**23.70**
I69	0.8	0.36	**38.65**
I7	0.73	0.3	**62.46**
I70	0.69	0.15	**47.87**
I71	0.86	0.17	**29.16**
I72	0.54	0.42	**47.87**
I73	0.69	-0.05	**49.98**

Item	p-value	pt-biserial	time
I74	0.53	0.28	**67.25**
I75	0.75	0.09	**36.89**
I76	0.94	0.06	**45.69**
I77	0.91	0.2	**24.25**
I78	0.75	0.22	**43.39**
I8	0.59	0.32	**59.84**
I81	0.58	0.13	**49.45**
I82	0.83	0.4	**41.00**
I83	0.78	0.3	**46.23**
I84	0.63	0.34	**115.36**
I85	0.71	0.23	**33.62**
I86	0.61	0.21	**43.87**
I87	0.75	0.31	**48.53**
I88	0.74	0.12	**58.94**
I89	0.8	0.24	**30.29**
I9	0.43	0.22	**70.86**
I90	0.76	-0.02	**29.08**
I91	0.81	0.33	**25.57**
I92	0.24	0.04	**49.61**
I93	0.53	0.17	**62.28**
I94	0.39	0.21	**50.28**
I95	0.41	0.22	**72.76**
I96	0.61	0.26	**58.37**
I97	0.52	0.27	**52.19**
I98	0.47	0.21	**76.68**
I99	0.81	0.27	**41.53**

Annex E SJT computer based item statistics

Item	Type	Title	Domain	Mean	SD	Partial	Quality
1	MRQ	Difficult Colleague	ES	8.05	3.11	0.31	Good
2	MRQ	Cancelled Patient	PI	10.69	2.43	0.35	Good
3	MRQ	Self Discharge	PI	9.66	2.81	0.24	Mod.
4	MRQ	Canulation	PI	9.66	2.49	0.27	Good
5	MRQ	Central Line	CP	9.69	3.10	0.28	Good
6	MRQ	Working under Pressure	CP	9.89	3.18	0.26	Good
7	MRQ	Breaking Bad News	ES	10.27	2.72	0.30	Good
8	MRQ	Legal Guardian	ES	10.44	2.72	0.36	Good
9	MRQ	Obstetric Trauma	ES	8.24	3.08	0.36	Good
10	MRQ	Respect	PI	7.81	2.50	0.13	Poor
11	MRQ	Dementia	ES	9.85	2.83	0.18	Mod.
12	MRQ	Paediatric Resus	ES	9.38	2.89	0.25	Mod.
13	MRQ	Self Harm	ES	9.69	2.77	0.20	Mod.
14	MRQ	Paediatric Consent	ES	8.02	2.69	0.05	Poor
15	MRQ	Patient Complaint	PI	9.52	2.73	0.20	Mod
16	MRQ	Pacing	PI	7.15	2.64	0.11	Poor
17	MRQ	Elder Abuse	ES	9.39	2.77	0.26	Good
18	MRQ	On Call Cover	PI	8.83	3.01	0.41	Good
19	MRQ	ICU Admission	CP	9.09	3.17	0.27	Good
20	MRQ	Safe Prescibing	PI	8.74	2.88	0.36	Good
21	MRQ	Abdominal Pain	CP	8.79	3.27	0.32	Good
22	MRQ	Jehovah's Witness	ES	6.01	2.70	0.31	Good
23	MRQ	IV Opioids	ES	5.14	2.40	0.15	Poor
24	MRQ	Dental Phobia	ES	7.54	2.97	0.17	Poor
25	MRQ	Consultation	PI	8.21	2.88	0.20	Mod.
26	RQ	Gifts	PI	8.57	3.47	0.26	Good
27	RQ	Anxious Patient	ES	9.78	2.71	0.41	Good
28	RQ	Trauma	SA	7.88	3.37	0.41	Good
29	RQ	Dealing with Relatives	ES	6.70	2.65	0.38	Good
30	RQ	Problem Colleagues	PI	8.95	2.85	0.31	Good
31	RQ	Insomnia	ES	8.28	2.58	0.21	Mod.
32	RQ	Cardiac Arrest	SA	8.07	2.73	0.37	Good
33	RQ	Child Protection	ES	8.95	2.46	0.34	Good
34	RQ	Pleuritic Chest Pain	PI	8.58	2.84	0.35	Good
35	RQ	Appraisal	CP	9.37	3.33	0.52	Good
36	RQ	CT Scanner	SA	7.09	3.23	0.44	Good
37	RQ	Patient	PI	8.69	3.13	0.35	Good

Item	Type	Title	Domain	Mean	SD	Partial	Quality
		Relations					
38	RQ	Silent Treatment	ES	6.97	2.82	0.25	Good
39	RQ	Surgical Difficulty	SA	9.49	2.88	0.37	Good
40	RQ	Missing Cross-Match	CP	8.41	2.92	0.33	Good
41	RQ	Cancer Pain	PI	7.50	2.88	0.38	Good
42	RQ	Confidentiality	ES	9.23	3.06	0.46	Good
43	RQ	Dental Injury	PI	9.38	2.87	0.36	Good
44	RQ	Surgical Risk	ES	6.82	2.77	0.21	Mod.
45	RQ	Death on Intensive Care	ES	7.58	2.62	0.32	Good

RQ=ranking question, MRQ=multiple response question, ES=Empathy & Sensitivity, SA=Situational Awareness,
PI=Professional Integrity, CP=Coping with Pressure

Annex F SJT paper-based pilot item level statistics

Item	Type	Title	Domain	Mean	SD	Partial	Quality
1	RQ	Gifts	PI	10.17	2.20	0.26	Good
2	RQ	Anxious Patient	ES	9.57	1.91	0.25	Good
3	RQ	Trauma	SA	9.10	2.54	0.21	Mod.
4	RQ	Dealing with Relatives	ES	7.94	2.30	0.23	Mod.
5	RQ	Problem Colleagues	PI	9.17	2.41	0.16	Poor
6	RQ	Insomnia	ES	8.58	2.28	0.18	Mod.
7	RQ	Cardiac Arrest	SA	8.45	1.92	0.24	Mod.
8	RQ	Child Protection	ES	9.30	2.11	0.20	Mod.
9	RQ	Pleuritic Chest Pain	PI	9.04	2.27	0.23	Mod.
10	RQ	Appraisal	CP	10.36	2.20	0.16	Poor
11	RQ	CT Scanner	SA	7.56	3.11	0.18	Mod.
12	RQ	Patient Relations	PI	9.56	2.09	0.29	Good
13	RQ	Silent Treatment	ES	7.57	2.47	0.17	Mod.
14	RQ	Surgical Difficulty	SA	10.11	2.04	0.21	Mod.
15	RQ	Missing Cross-Match	CP	9.24	2.13	0.17	Poor
16	RQ	Cancer Pain	PI	8.33	2.31	0.05	Poor
17	RQ	Pre-Operative Checklist	PI	6.52	2.49	0.18	Mod.
18	RQ	Penicillin Allergy	PI	7.82	3.02	0.18	Mod.
19	RQ	Second Opinion	SA	8.99	2.51	0.12	Poor
20	RQ	Late Arrival	CP	9.35	2.17	0.22	Mod.
21	MRQ	A Difficult Colleague	ES	7.82	2.65	0.14	Poor
22	MRQ	Cancelled Patient	PI	10.99	1.92	0.34	Good
23	MRQ	Self Discharge	PI	10.02	2.07	0.02	Poor
24	MRQ	Canulation	PI	9.65	2.13	0.10	Poor
25	MRQ	Central Line	CP	10.76	2.26	0.11	Poor
26	MRQ	Working under Pressure	CP	9.19	3.31	-0.01	Poor
27	MRQ	Breaking Bad News	ES	11.18	1.77	0.07	Poor
28	MRQ	Legal Guardian	ES	10.91	1.96	0.19	Mod.
29	MRQ	Obstetric Trauma	ES	8.98	2.38	0.27	Good
30	MRQ	Respect	PI	8.19	2.21	0.03	Poor
31	MRQ	Dementia	ES	9.73	2.68	0.05	Poor
32	MRQ	Paediatric Resus	ES	10.38	2.10	0.13	Poor
33	MRQ	Self Harm	ES	9.77	2.40	0.16	Poor
34	MRQ	Paediatric Consent	ES	7.78	2.32	0.11	Poor

Item	Type	Title	Domain	Mean	SD	Partial	Quality
35	MRQ	Patient Complaint	PI	9.51	2.62	0.30	Good
36	MRQ	Pacing	PI	7.45	2.63	0.22	Mod.
37	MRQ	Elder Abuse	ES	9.00	2.97	0.36	Good
38	MRQ	Drug Overdose	LEA	8.55	3.10	0.42	Good
39	MRQ	Confusion	ES	6.28	3.44	0.44	Good
40	MRQ	Domestic Violence	ES	9.45	4.03	0.61	Good
41	MRQ	Fundoscopy	PI	9.20	4.02	0.57	Good
42	MRQ	Endoscopy	PI	6.88	3.90	0.43	Good
43	MRQ	Telephone Protocol	ES	6.02	3.95	0.40	Good
44	MRQ	Mastectomy	ES	6.06	3.54	0.48	Good
45	MRQ	HIV Infection	PI	6.38	4.08	0.39	Good

RQ=ranking question, MRQ=multiple response question, ES=Empathy & Sensitivity, SA=Situational Awareness,
PI=Professional Integrity, CP=Coping with Pressure, LEA=Legal & Ethical Awareness