Work Psychology Group
Thinking differently

Analysis of the Situational Judgement Test for Selection to the Foundation Programme 2013

Technical Report

Prof Fiona Patterson
Vicki Ashworth
Helena Murray
Laura Empey
Amy Aitkenhead

May 2013

Table of Contents

1 Introduction

1.1 Purpose and Structure of the Report

1.1.1 The Foundation Programme (FP) Situational Judgement Test (SJT) was delivered for entry to FP2013 between December 2012 and January 2013, over three administration sessions. The SJT, in combination with the Educational Performance Measure (EPM), was used to rank applicants applying for Foundation Year One training. The operational usage of the SJT follows a period of successful piloting and recommendations that were made to the Department of Health.

1.1.2 This technical report provides an overview of the results from the operational delivery of the FP2013 SJT. The report is divided into three main parts:

- Part One describes the development process of items that were trialled alongside the operational SJT.
- Part Two describes the results and analysis of the operational SJT, as well as initial analysis of the trial items.
- Part Three provides a summary and recommendations going forward.

1.2 Background

1.2.1 In 2009, the Department of Health in England (DH), on behalf of the four UK health departments, commissioned the Medical Schools Council (MSC) to lead a cross-stakeholder steering group to design, develop and pilot new arrangements for the selection of medical students into the FP. The FP is a two-year generic training programme, which forms the bridge between medical school and specialist/general practice training.

1.2.2 This steering group recommended the pilot of an SJT and an EPM, and that these two assessments in combination should be used for selecting applicants and allocating them to foundation schools. The SJT must therefore be developed and validated in accordance with accepted best practice, so that it provides an effective, rigorous and legally defensible method of selection.

1.2.3 In August 2011, a report was produced on the design, analysis and evaluation of an SJT for Selection to the FP. This included the specification of the domains to be targeted in the SJT[1]. Recommendations were for the implementation of an SJT, alongside the EPM for entry to FP2013.

[1] Please see FY1 Job Analysis report 2011 for full details of how domains were derived and what comprises each domain.

1.2.4 In March 2012, a report was produced on the analysis and evaluation of a full scale pilot SJT for selection to the FP titled the Parallel Recruitment Exercise. This report summarised the SJT as an appropriate methodology to be used within this context.

Part One: Item Development

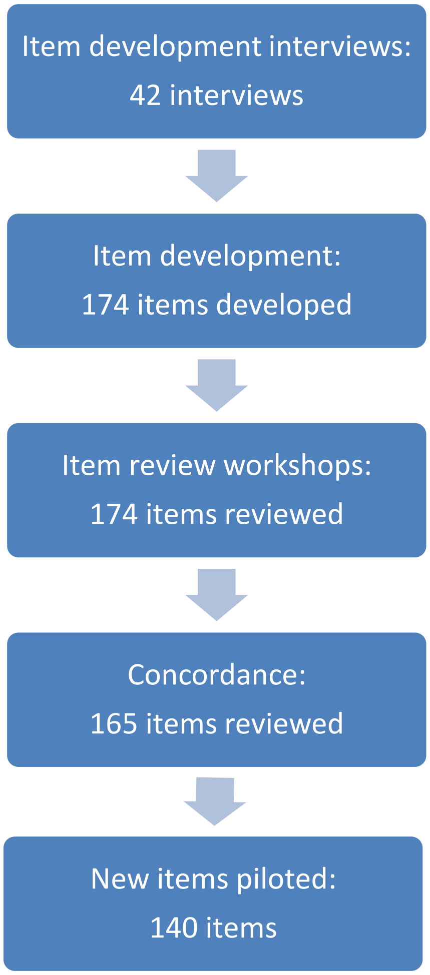
2 Development of Trial Items

2.1 Process Overview

2.1.1 To ensure that there are sufficient items within the operational item bank to support operational delivery, and to continually refresh and replenish the bank with a wide range of relevant and current scenarios, trialling of new items takes place alongside the operational SJT each year.

2.1.2 Figure 1 summarises the development and review process undertaken for the new items that were trialled alongside the FP2013 operational delivery.

Figure 1: Item development and review process



2.2 Item Development Interviews

2.2.1 Item Development Interviews (IDI's), using the Critical Incident Technique (CIT), were held to develop SJT items. CIT interviews aim to elicit, from Subject Matter Experts (SMEs), scenarios or incidents involving Foundation Year One (FY1) doctors that demonstrate particularly effective or ineffective behaviour and that reflect the SJT target domains.

2.2.2 Using interviews such as these has a number of benefits, including that a broad range of individuals can be involved in the design process from across the country, without the need for a significant commitment in terms of time and effort.

2.2.3 Prior to the interviews taking place, a review of the current operational bank (n=153), containing items that were established through the previous trials between 2010 and 2012, was carried out. This included a review of the spread of target domains and topic areas. The aim of this review was to focus item writing on under-represented domains or topic areas and to identify topic areas to be avoided, due to over-representation.

2.2.4 An invitation was sent out via the UKFPO Foundation School Managers network, which cascaded information to those working closely with FY1 doctors. Individuals who had previously been involved, or had expressed an interest in being involved, in the process were also invited to take part in the interviews.

2.2.5 In total, 42 interviews were conducted by trained interviewers. Table 1 shows the range of job roles held by the SMEs. FY1/FY2s were deliberately targeted; as these individuals are closest to the role therefore they are well placed to provide relevant and realistic scenarios.

Table 1: Subject Matter Experts' job roles

Job Role	Number
Clinical/Educational Supervisor	1
Clinical Tutor	3
Foundation School Director	3
Medical School Director	1
FY1/FY2	14
ST1/ST2	2
Not stated	18

2.2.6 Table 2 shows the range of the primary specialties of SMEs. Where possible, SMEs were selected to ensure a broad range of specialties were involved in generating items. This helps to ensure that the bank of items represents a wide spread of topic areas and medical settings.

Table 2: Subject Matter Experts' primary specialties

Specialty	Number
Anaesthetics	3
Emergency Medicine	3
Endocrinology	1
Gastroenterology	1
General Practice	2
Haematology	1
Neurosurgery	1
Obstetrics & Gynaecology	1
Paediatrics	4
Physiology	2
Psychiatry	2
Respiratory Medicine	1
Renal medicine	1
Rheumatology	1
Surgery	5
Urology	1
Not stated/non-clinical	12

2.2.7 The telephone interviews lasted between 30 and 45 minutes. During the interview, a trained interviewer asked the interviewee to describe a number of scenarios, providing as much information as possible. This included the pre-cursor to the incident, who was involved, what the outcome was and other possible ways that the scenario could have been dealt with (to enable alternative responses to be developed). The trained interviewer then used this information to develop the SJT items.

2.2.8 A total of 174 items were written. This equals an average of 4.2 items per interview, which is consistent with previous item development processes.

2.3 Item Review

2.3.1 All 174 items that were submitted were logged on a spreadsheet, which detailed who the item writer was, the date the item was submitted, the type of item, the target domain, the answer key and a short one line summary of the item.

2.3.2 As mentioned above, prior to the development of new items, a review of the operational item bank was undertaken. This revealed that the bank contained a larger number of items in the target domain 'Commitment to Professionalism'. Therefore the development of new items was deliberately targeted within the other areas. The breakdown of items developed relevant to each of the target domains was as follows:

- Commitment to Professionalism - 22
- Coping with Pressure - 48
- Effective Communication - 31
- Patient Focus - 41
- Working Effectively as Part of a Team – 32

2.3.3 Within each domain a number of different item topics are identified, broadly categorised into three main topic areas; colleague-related, patient-related and self/task-related. These are also further categorised under sub-topics which describe in more detail the sort of scenario presented in each item. This enables test construction to ensure an appropriate representation of item types across each paper.

2.3.4 The split of the item formats was designed to reflect the construction of the operational test (two thirds Ranking questions, one third Multiple Choice). The breakdown of items, regarding item format was as follows:

- Ranking - 122
- Multiple Choice - 52

2.3.5 Items were reviewed by the core team of item reviewers from Work Psychology Group. Each scenario was reviewed in accordance with SJT item writing principles and the specification of the FY1 SJT. In particular, scenarios were reviewed to ensure that they addressed one of the target criteria and were at an appropriate level for an FY1.

2.4 Review workshops

2.4.1 The aim of the review workshops was for SMES to review SJT items for relevance and fairness, as well as providing input into the scoring key. The benefit of holding these full-day review workshops is that it enables input from a range of clinicians in developing items that have face validity and SME agreement on the answer key. An added benefit is that the review workshops serve as an opportunity to build awareness of SJTs amongst the medical community and improve expertise in SJT design principles. All those who attend the review workshops are awarded 6 CPD points.

2.4.2 A small number of FY2 doctors attended the workshops to provide additional input in terms of the items' relevance and realism to the FY1 role.

2.4.3 Three review workshops were held. One workshop was held in London, one workshop was held in Manchester and one workshop was held in Edinburgh. A total of 21 individuals attended the four workshops, including five FY2s.

2.4.4 All applicants who volunteered to take part were sent briefing material that outlined the purpose of the review workshop and their role on the day. All applicants also completed a confidentiality and code of conduct agreement.

2.4.5 During the workshop, delegates were split into two groups. As a group, with the aid of a facilitator, delegates reviewed approximately 30 items. Delegates were asked to consider the scenario content and the response. They were also asked to discuss a possible answer key, which was compared with the answer key proposed by the author of the scenario. Their comments and suggestions were recorded by the facilitator and updates were made to items.

2.4.6 A total of 174 items were reviewed during the review workshops. During the course of the review workshops, it was agreed that 9 items should be rejected due to issues with either relevance or fairness.

2.5 Concordance Panel

2.5.1 Concordance panels were held following the review workshop stage. Concordance panels involve Subject Matter Experts, in this case clinicians working closely with FY1s, completing an SJT consisting of trial items. Following best practice in SJT design, the aim of a concordance stage is to identify a high level of consensus between experts on the item keys. Those items that exhibit high levels of consensus go forward to be trialled. Those items exhibiting low levels of consensus are put to one side for further review, with changes made if necessary. Concordance panels also provide the opportunity for additional feedback regarding fairness and relevance to be received about the items.

2.5.2 The criteria for Subject Matter Expert involvement in the concordance panel was that the clinicians worked closely with FY1 doctors and were very familiar with the responsibilities and tasks, as well as the necessary skills and abilities, that are required for the role.

2.5.3 Two concordance panels were held, each in two sessions, with one paper reviewed at each panel. Paper One consisted of 82 items and Paper Two consisted of 83 items. Therefore, a total of 165 items went to concordance. At this stage, the tests were not constructed as final tests, i.e. no consideration was given as to the spread of item topics or domains, as the aim of the concordance panels was to analyse individual items. This was made clear to those attending the panels.

2.5.4 A total of 18 individuals attended the concordance stage. One panel consisted of 10 individuals and the other panel consisted of 8 individuals. A minimum of 10 individuals is required for robust concordance analysis, with ideally 12 or more undertaking each paper. Where there are less than the ideal numbers, the results have to be interpreted with caution.

2.5.5 After completing a confidentiality form, the panel was asked to complete the SJT items under test conditions. There was no set time limit, although the panels were told that the test should take no more than two hours and 30 minutes to complete.

2.5.6 Feedback on the item content was provided by the panel, and this resulted in some minor alterations to a small number of items in order to provide clarification. No item was altered sufficiently enough to affect the interpretation of the question or the answer key.

2.5.7 Following the concordance panel meeting, concordance analysis was undertaken to analyse the experts' level of agreement over the keyed response for each trial item. Using established criteria of acceptance levels[2], items were deemed either to have satisfactory levels of concordance (144) or unsatisfactory levels of concordance (21). As well as taking into account concordance data, a further review of all items was undertaken, including feedback from SMEs to identify further items that may not be suitable for piloting.

2.5.8 Following concordance analysis, 140 items were chosen for piloting. To identify these 140 items, as well as taking into account results from the concordance stage as outlined above, an assessment of an item's similarity to those already in the item bank was also undertaken. The answer key provided by the concordance panel was used in combination with information from item writers and review workshops to determine a scoring key for the trial data. However, it must be noted that this does not necessarily reflect the final key, as information is used from the trial to develop the items and their keys further. For example, if highly performing applicants consistently provide a different key from the established key, then the key will be reviewed with the assistance of Subject Matter Experts.

[2] The main criteria for selecting an item for use in the pilot was a significant Kendall's W. Following best practice, any item that produces a low and non-significant Kendall's W is reviewed for removal from the item bank for further review. An inclusion criterion of approx 0.60+ is also used to assist in selecting items. However, there is a 'tolerance' around this figure and the decision will depend on a number of factors, including how many people have taken the concordance. Consideration of the raw statistics must be combined with consideration of the concordance keys versus item writer and focus group keys, as well as further feedback gained from the concordance panel. In this context, a Kendall's W of 0.60 or above indicates a good level of concordance, although any value above 0.50 can be described as having satisfactory levels of concordance.

Part Two: Scoring, Analysis and Evaluation

3 Operational Test Structure and Construction

3.1 All SJT items used operationally have been part of an item bank that has been developed over the period of trialling between 2010 and 2012 (n=153). Every item within the operational item bank has been deemed to have sufficient psychometric properties to be used operationally. However, good practice dictates that a yearly review of the operational bank should be conducted to ensure that the scenarios are still relevant and fair. The intention of this review is not to amend the content of the items, but purely to ensure that they are still appropriate to be used in this context. A review was undertaken by two clinicians in August 2012 and all items were deemed still appropriate to be used.

3.2 Three administrations of the SJT were undertaken, which required the production of three versions of the test paper which were subsequently equated. Version three was included as a 'mop up' paper and includes items from both version one and version two; version three did not contain any unique items as the number of applicants was expected to be small.

3.3 For each version, 70 items were administered. Of these, 60 were operational items and 10 were trial items (see 3.8). There were 40 ranking operational items and 20 multiple choice operational items for each paper. The test versions were designed with specific overlaps ('anchor' questions), which could be used to compare populations and link the different versions.

3.4 The three papers were developed to be as similar as possible in terms of content, psychometric properties and difficulty. The process for maximising the equivalency of the tests includes:

- Same number of items of each item type (i.e. ranking and multiple choice)
- Similar number of items for each target domain
- Similar spread of item topics/content areas
- Similar mean and range of item facility/difficulty
- Similar mean and range of item quality/partial

3.5 In addition to ensuring that the spread of target domains was similar *across* papers, as far as possible, an equal spread of the five target domains and topic categories was selected for each paper. However, the proportion of items from each target domain and topic category is also a reflection of the make-up of items within the operational item bank and has to be balanced with the two item response formats. Of the 99 items that were used across the three papers the spread of target domains was as follows: 26 are categorised under the 'Commitment to Professionalism' domain, 18 'Coping with Pressure', 16 'Effective Communication, 17 'Patient Focus' and 22 'Teamworking'. An outline of the spread of domains for each of the papers is illustrated in Table 3 below. Information concerning the topic categories is not provided as this provides insight into the item content and therefore could compromise the security of the test.

Table 3: Spread of target domains within each paper

Paper	Commitment to Professionalism	Coping with Pressure	Effective Communication	Patient Focus	Working Effectively as Part of a Team
1	16	12	10	10	12
2	16	11	9	10	14
3	16	12	9	11	12

3.6 As mentioned above, as well as selecting items based on their domains and topics, attention was also paid to ensuring that the range and mean item facility and difficulty were broadly similar across the three papers. Table 4 shows the mean item facility for ranking items and multiple choice items, as well as the mean item partials for all three papers. This demonstrates that all three papers were broadly equivalent, based on known psychometric properties.

Table 4: Spread of item facility and item quality within each paper

Paper	Item Facility (Ranking items)			Item Facility (Multiple Choice Questions)			Item Partial		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
1	14.07	18.67	16.52	7.19	10.20	8.54	.166	.458	.243
2	14.78	18.31	16.47	6.55	10.19	8.59	.169	.458	.244
3	14.07	18.67	16.52	6.55	10.19	8.59	.166	.458	.247

3.7 There were a total of 1037-1038 marks available for each operational version of the SJT paper, with a maximum of 20 marks available for each of the 40 ranking items and 12 marks for each of the 20 multiple choice items. For a small number of items, based on psychometric evidence and expert review, two of the options may be tied; that is an applicant may provide these options in either order. Due to the way the scoring convention is designed, this results in the maximum possible score being 19 rather than 20. This is the same for all applicants.

3.8 A total of 140 items were trialled. To enable this number of items to be trialled, 15 papers were developed with 10 trial items in each. Papers 1-10 incorporated operational version one, Papers 11-14 incorporated operational version two and Paper 15 was operational version three. The number of items to be trialled was based on a minimum sample size to obtain robust item level results and to ensure confidence in the findings. Best practice dictates a minimum sample size of 400. The trial items in Paper 13 and Paper 15 were the same.

4 Scoring & Test Equating

4.1 Following the scanning of all responses and a quality check undertaken by MSC to ensure that the data had been received from all applicants, the raw responses were received by WPG for scoring.

4.2 The scoring quality assurance procedure follows the process outlined below:

- **Scoring syntax QA:** this includes a check for typographical/SPSS errors, item type, key, number of options and tied scores. At this stage, dummy data are also run to check the syntax is working correctly.
- **Data cleaning (Excel):** this includes a check for unexpected characters as well as the checking of variable names and number of cases.
- **Data cleaning (SPSS):** this includes ensuring that data has converted to the correct format from Excel, the running of frequencies to identify potential errors and impossible data scores and ensuring that all applicants have a reasonable number of responses.
- **Scoring QA:** this includes initial analysis to ensure that the mean, reliability and test statistics are in the expected range and the running of frequencies of scored data to ensure they are in the expected range and that there are no anomalies.

4.3 During the scoring process, in some circumstances and according to best practice, the removal of items that are not differentiating as expected and that are significantly reducing the reliability of the test would be required. For the FP2013 operational test, no items were excluded from final scoring.

4.4 While the papers are developed to be as equivalent as possible, test equating also takes place so that the results from each of the different papers are comparable and fair to all applicants. Statistical equating procedures place all scores from different papers on the same scale. Without this, it is not possible to determine whether small differences in scores between papers relate to real differences in ability in the populations assigned to a paper, or differences in the difficulty of the papers themselves. In reality, observed differences will be a function of both sample and test differences. A minor statistical adjustment can be used to ensure that the scores are fully equivalent.

4.5 There are a number of approaches to equating. For this SJT, the most suitable approach is a chained linear equating process. The test papers were designed with specific overlaps (‘anchor’ items), which could be used to compare populations and link the different papers.

4.6 The raw equated SJT scores were transformed into points on a 50 point scale. The lowest scoring applicant was assigned a score of 0 and the highest scoring applicant a score of 50. A linear transformation was used to convert all other scores onto intermediate values on the scale. This is a linear transformation so it has no impact on the relative position of any applicant. The maximum number of applicants with a single score was 147. Scores were rounded to 1 decimal place during standardisation; using 2 decimal places would have reduced the number of applicants with a single score to 125. Differences in scores in the second decimal place are not meaningful and candidates that only differ to this degree should be considered to have the same level of performance. However, if the process requires the ordering of candidates with equivalent scores then using the second decimal place is more appropriate than an arbitrary or random method for selecting which of equal candidates will be given preference.

5 Analysis

5.1 Purpose

5.1.1 Following any operational delivery of an SJT, it is important that the test is evaluated with regards to reliability, group differences and the ability for the test to discriminate between applicants. Item level analysis of operational items also takes place. This is because, although previous trials have demonstrated that the items had sufficient psychometric properties to be used operationally, items can perform differently over time. It is therefore important to continually monitor all operational items.

5.1.2 Evaluation of trial items is also undertaken to analyse whether they exhibit sufficient psychometric properties to enter the operational item bank.

5.2 Evaluation Overview

5.2.1 This section outlines the psychometric analysis for the SJT. Any high stakes, high profile test needs to meet exacting psychometric standards in terms of the quality of individual items and of the test as a whole, including reliability, validity and fairness. The main analysis and evaluation activities reported here include:

- Test level statistics, including reliability and scoring distributions
- Item level statistics, including item facility and effectiveness
- Analysis of group differences at a test and item level to explore fairness
- Evaluation of applicant reactions
- Relationships between the EPM and the SJT

5.3 Sample

5.3.1 There were a total of 8,162 applicants who took the FP2013 SJT. They were final year medical students, including students who had been pre-allocated to a Defence Deanery Foundation Programme, UK students who had taken time out post-graduation and international medical students/graduates applying through the Eligibility Office.

5.3.2 A breakdown of the number of applicants who sat each of the three papers can be seen in Table 5 below. One version of a paper was undertaken at each school for logistical reasons, and to minimise security risk to the items.

5.3.3 Schools were given the choice as to which testing administration date they preferred, and as such the samples for the papers are not randomly allocated. Caution should be taken when interpreting the data from Paper Three, as the number of applicants is extremely low. The sample sizes for Paper One and Paper Two are well above the minimum requirement for psychometric robustness (n=400) and as such, confidence can be placed in the outcomes of the psychometric analysis.

Table 5: Number of applicants taking each paper

	No. of applicants	Percentage of Overall Sample
Paper One	6126	75.1%
Paper Two	2020	24.7%
Paper Three	16	0.2%

5.3.4 Applicant demographic data were collected from the FPAS application.

5.3.5 Table 6 outlines the breakdown of applicants by sex. Overall, more females completed the test (4555, 55.8%) than males (3515, 43.1%), reflecting the male/female split of applicants to the Foundation Programme.

Table 6: Applicant sex by paper

		Male	**Female**	**Not declared**
Overall	No. of applicants	3515	4555	92
	% of applicants	43.1	55.8	1.1
Paper One	No. of applicants	2686	3368	72
	% of applicants	43.8	55.0	1.2
Paper Two	No. of applicants	821	1179	20
	% of applicants	40.6	58.4	1.0
Paper Three	No. of applicants	8	8	0
	% of applicants	50	50	0

5.3.6 Table 7 outlines the breakdown of applicants by ethnicity. Overall, the majority of applicants reported their ethnicity as 'White' (5180, 63.5%), with the smallest proportion of applicants (241, 3.0%) reporting themselves as being from 'Black backgrounds'. The proportion of individuals in each ethnic group was roughly equivalent in Paper One (n=6,126) and Paper Two (n=2,020). Paper Three had a very small sample size and reflected just three ethnic backgrounds; 'White' (n=5), 'Asian' (n=10), and 'Other' (n=1).

Table 7: Applicant ethnicity by paper

		White	Asian	Black	Chinese	Mixed	Other	Not declared
Overall	No. of applicants	5180	1556	241	364	313	264	244
	% of applicants	63.5	19.1	3.0	4.5	3.8	3.2	3.0
Paper One	No. of applicants	3872	1159	183	270	247	202	193
	% of applicants	63.2	18.9	3.0	4.4	4.0	3.3	3.2
Paper Two	No. of applicants	1303	387	58	94	66	61	51
	% of applicants	64.5	19.2	2.9	4.7	3.3	3.0	2.5
Paper Three	No. of applicants	5	10	0	0	0	1	0
	% of applicants	31.3	62.5	0.0	0.0	0.0	6.3	0.0

5.3.7 Table 8 outlines the breakdown of applicants' ethnicity (White and Black and Minority Ethnic (BME) group). 5,180 (63.5%) applicants were classified as White and 2,738 (33.5%) applicants were classified as being from BME groups. 244 (3.0%) applicants did not declare their ethnicity. Paper One had 33.6% BME applicants, Paper Two had 33.1% BME applicants and Paper Three had 68.8% BME applicants.

Table 8: Applicants' ethnicity by paper

		White	BME	Not declared
Overall	No. of applicants	5180	2738	244
	% of applicants	63.5	33.5	3.0
Paper One	No. of applicants	3872	2061	193
	% of applicants	63.2	33.6	3.2
Paper Two	No. of applicants	1303	666	51
	% of applicants	64.5	33.1	2.5
Paper Three	No. of applicants	5	11	0
	% of applicants	31.3	68.7	0

5.3.8 Table 9 outlines the breakdown of applicants by their country of medical education (UK and Non-UK medical schools). 7,822 (95.8%) applicants were from UK medical schools and 340 (4.2%) applicants were from non-UK medical schools. Paper One had 2.9% non-UK applicants, Paper Two had 7.7% non-UK applicants and Paper Three had 50% non-UK applicants.

Table 9: Applicants' country of medical education by paper

		UK	Non-UK
Overall	No. of applicants	7822	340
	% of applicants	95.8%	4.2%
Paper One	No. of applicants	5949	177
	% of applicants	97.1%	2.9%
Paper Two	No. of applicants	1865	155
	% of applicants	92.3%	7.7%
Paper Three	No. of applicants	8	8
	% of applicants	50%	50%

5.3.9 The mean age of the sample was 25.5 years, with a range of 21 – 58 years.

5.4 Test Completion Analysis

5.4.1 The time allowed for the SJT (including trial items) was 140 minutes for 70 items. Table 10 provides an overview of test completion across all of the test versions. Across all test versions, 99.3% of applicants attempted the last item on the test. 97.9% answered all items and 0.5% failed to answer four or more items.

5.4.2 Test completion was also examined by paper. 0.6% (37) of applicants did not finish Paper One and 1.1% (22) of applicants did not finish Paper Two. 97.9% of applicants in Paper One answered all items and 99.7% of applicants in Paper Two answered all items. 0.3% of applicants in Paper One and 0.6% of applicants in Paper Two failed to answer four or more items. Therefore, it seems that there is a slightly higher completion rate for Paper One than Paper Two. All applicants in Paper Three completed the test.

5.4.3 These results are comparable with previous trials (97.2% completion rate in 2011 pilot, 96% completion rate in the 2012 Parallel Recruitment Exercise) and confirm that the SJT is a power test, rather than a speeded test. This indicates that 140 minutes is an appropriate length of time to complete 70 items.

Table 10: Analysis of Test Completion

	Attempted last item	Answered all items	Failure to answer four or more items
All Papers	99.3%	97.9%	0.5%
Paper 1	99.4%	97.9%	0.3%
Paper 2	98.9%	99.7%	0.6%
Paper 3	100.0%	100.0%	0.0%

5.5 Operational Test Level Analysis

5.5.1 Test level analysis was carried out for all three papers separately before the scores were equated. Table 11 illustrates the test level descriptive statistics.

Table 11: Operational test level descriptive statistics by paper

	N	Reliability (α)	SEM	Mean	Mean % correct	Skew	Kurtosis	SD	Min	Max	N items
Paper One	6126	0.67	16.8	861.5	83.1	-1.47	8.06	29.2	546	935	60
Paper Two	2020	0.76	16.7	856.6	82.5	-1.91	8.95	34.0	577	927	60
Paper Three	16	0.85	18.8	814.2	78.5	-0.50	-0.98	48.5	723.5	871	60

5.5.2 Mean scores are broadly similar between Paper One and Paper Two, with Paper One exhibiting a slightly higher mean score and a slightly higher percentage correct. However, this is comparable with Paper Two and indicates that the two papers have comparable levels of difficulty. Paper Three has a lower mean score and percentage correct, but this result is not reliable because of the very small numbers of applicants taking the test. The equating strategy that follows scoring takes into account any differences between the papers in terms of difficulty.

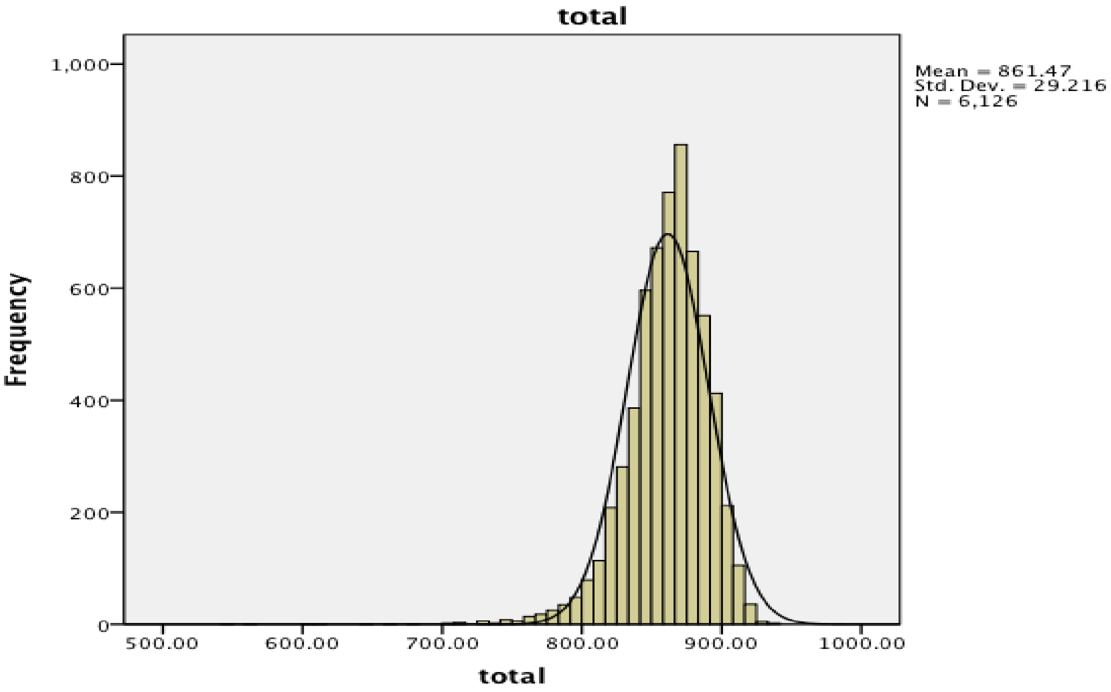
5.5.3 In terms of scoring distribution, the score range is highest for Paper One. Paper Three has the lowest distribution of the three papers. However, the standard deviation (SD) is a much better indicator of the spread of scores than the range, as the range can be strongly affected by a single score.

5.5.4 The SD is a measure of the distribution of scores and indicates how much variation there is from the mean. A low SD indicates that the data points tend to be very close to the mean, whereas a higher SD indicates that the data are spread out over a large range of values. The SD for Paper One (SD =29.2) is a little lower than that for Paper Two (SD=34.0). This suggests greater variation in the applicants sitting the second paper. The actual variance observed will depend on the variance within the applicant pool. Applicants are not randomly assigned to the two papers and different universities sit different papers, which may account for this difference in variance. The SD for Paper Three (SD=48.5) is substantially higher, but any measure of distribution will be unstable in such a small sample. These SDs are expected and are comparable with previous pilot results.

5.5.5 The reliability for all three papers is α=0.67 and above; sufficient for the use of an operational SJT (Paper One α=0.67, Paper Two α=0.76, Paper Three α=0.85). In summary, the average of the reliability across the three papers is 0.76. Paper Two appears to show substantially higher reliability than Paper One, however inspection of the standard error of measurement (SEM[3]) indicates that the underlying accuracy of scores on the two papers is highly comparable. It is important to note when interpreting the results that reliability coefficients vary according to the sample. Where there is a greater spread of scores, reliability coefficients are higher. In this case, Paper Two applicants exhibit a varied and greater spread of scores (indicated by the higher SD) and therefore the reliability coefficient is higher.

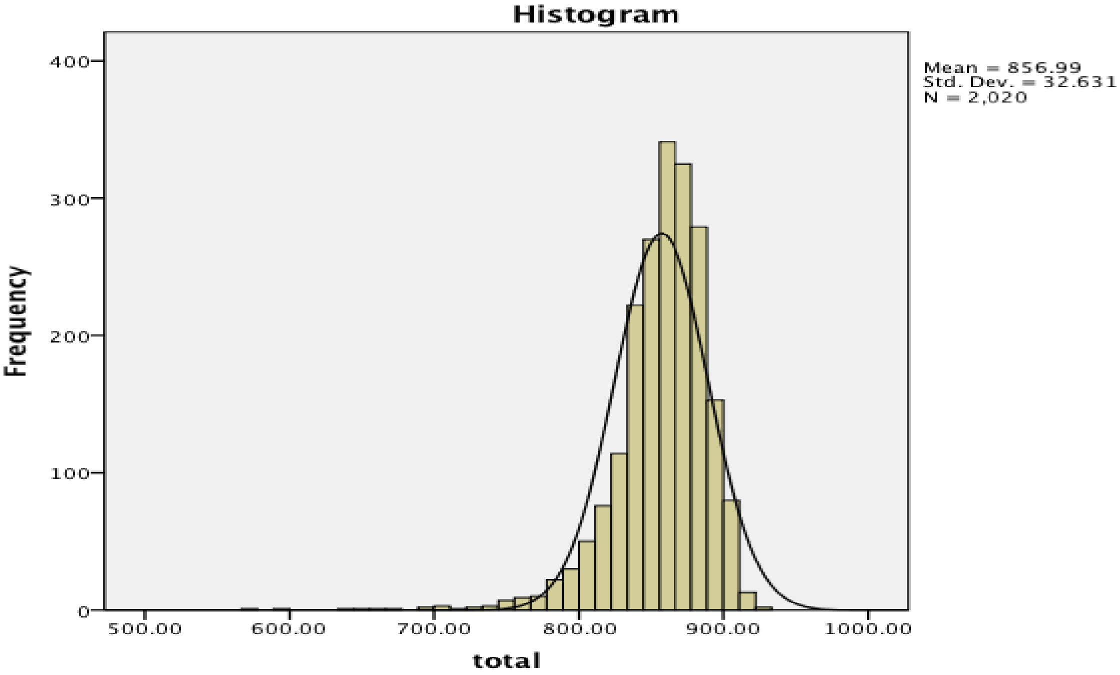
5.5.6 Figures 2 and 3 illustrate the distribution of scores for Paper One and Two, which are slightly negatively skewed. A negative skew indicates that the tail on the left side is longer than the right side/bulk of the values. However, the results do show a close to normal distribution, as would be desired for a selection test for employment.

Figure 2: Distribution Statistics for Paper One



[3] The Standard Error of Measurement (SEM) is an estimate of error that is used to interpret an individual's test score. A test score is an estimate of a person's 'true' test performance. SEM estimates how repeated measures of an individual on the same test have a tendency to be distributed around the individual's 'true' score. It is an indicator of the reliability of a test; the larger the SEM, the lower the reliability of the test and the less precision in the scores obtained.

Figure 3: Distribution Statistics for Paper Two



5.6 Operational Item Level Analysis

5.6.1 Item analysis was used to look at the difficulty and quality of individual SJT items within the operational test. Although the psychometric properties of the operational items are known beforehand, it is important that these continue to be monitored. As the number of individuals who have completed the items increases, the potential for error in the item partial decreases, therefore it is possible that in comparison to earlier pilots (when sample sizes were smaller) the psychometric properties of some items will change. This may result in a need to remove poorly performing items from the operational bank.

5.6.2 Item facility (difficulty) is shown by the mean score for each item (out of a maximum of 20 for ranking items and 12 for multiple response items). If the facility value is very low, then the item may be too difficult and may not yield useful information. If the facility value is very high, then the item may be too easy and may not provide useful information or differentiate between applicants. A range of item facilities is sought for an operational test, with few very easy (characterised by a mean score of greater than 90% of the total available score) or very difficult (characterised by a mean score of less than 30% of the total available score) items. Prior psychometric analysis of the items indicated that the operational items fell within these parameters, however these are reviewed again as part of the item level analysis.

5.6.3 The SD of an item should also be considered. If an item's SD is very small, it is likely to not be differentiating between applicants. The SD for an item should be at least 1.0 and no more than 3.0. If the SD is very large, it may mean that the item is potentially ambiguous and there is not a clear 'correct' answer, especially if this is coupled with a relatively low mean. Again, prior to operational delivery, all operational items fell within these parameters, based on their psychometric properties from the piloting stages.

5.6.4 Table 12 outlines the item level statistics for Papers One and Two. Paper Three has been excluded, as the small sample size will skew the overall results. The mean item facility for ranking items is 16.9. The mean item facility for multiple choice items in 9.1. The facility ranges and SDs for both ranking and multiple choice items are in line with expectations, however for both ranking and multiple choice questions the facility value has increased with operational use compared to when originally trialled. This increase in facility value is likely to be due, in part, to the sample population. The operational population has a greater motivation to perform well compared to a trial population and, as such, this may result in higher facility values for some items. This highlights why it is important for an operational item analysis to be completed, especially after a period of extended trialling. There are some items at the 'easy' end of the scale (more than 90% of the total available score) for both ranking and multiple choice questions. These items will be reviewed to ensure that they are sufficiently differentiating between applicants (and are therefore 'useful'). If this is not the case, then they will be removed from the operational bank. Those items with low SDs will also be reviewed.

Table 12: Item level statistics: Facility values

		Ranking			**Multiple Choice**		
	N	**Mean**	**Facility Range**	**SD Range**	**Mean**	**Facility Range**	**SD Range**
Overall (Paper One and Two)	8,146	16.9	14.5-19.0	1.5-2.8	9.1	5.8-11.3	1.6-2.7

5.6.5 Item quality (or discrimination) was determined by the correlation of the item with the overall operational SJT score, not including the item itself (item partial)[4]. This analysis compares how individuals perform on a given item with how they perform on the test overall and attempts to identify whether an item discriminates between good and poor applicants. One would expect that high scoring applicants overall would select the correct answer for an item more often than low scoring applicants; this would show a good to moderate correlation/partial. A poor correlation would indicate that

[4] With regards to acceptable levels of correlations for item partials, guidelines suggest in general 0.2 or 0.3 as identifying a good item (Everitt, B.S.,2002 *The Cambridge Dictionary of Statistics*, 2nd Edition, CUP). In this process we have used heuristics based on these guidelines and based on identifying items with sufficient level of correlation to be contributing to the reliability of the test.

performance on the individual item does not reflect performance on the test as a whole. Table 13 below outlines how items performed for each of the three papers and overall.

Table 13: Operational Item level statistics: Item partial

	Range of Item Partials	Mean Item Partial	Good (>0.17)	Moderate (0.13-0.17)	Unsatisfactory for operational bank (<0.13)
Overall (Papers One and Two)	.04-.33	.17	45 (45.5%)	36 (36.4%)	18 (18.2%)

5.6.6 45 of the 99 (45.5%) items are deemed as having good psychometric properties with regards to item quality and 36 (36.4%) of the items are deemed as moderate. 18 of the 99 (18.2%) items were deemed as unsatisfactory to remain in the operational bank. One of the reasons for this is that as the number of individuals completing the items increases, the potential for error in the item partials decreases (fewer false positives). In addition, over-exposure of items or other factors can cause item partials to decrease with additional use. This level of redundancy is to be expected each year and is in line with SJTs used in other contexts and has been built into the test development process with regards to the building of the item bank.

5.6.7 Following review, it is likely that these 18 unsatisfactory items will be removed from the operational bank to ensure that only those items that sufficiently differentiate between applicants are used operationally. It is anticipated that those items with poor item partials are also those that exhibit high facility values and low SDs, therefore it is not expected that more than 18 will be removed from the operational bank. Each of these items will be reviewed to try and identify if there is any specific reason as to why the partial has decreased. This may help to identify a pattern that will assist with future item development.

5.7 Group Differences

5.7.1 In order to examine fairness issues regarding the use of an SJT for selection into the FP, group differences in performance at a test level (equated scores) were analysed on the basis of sex, ethnicity, country of medical education and age.

5.7.2 Table 14 shows group differences in performance on the SJT based on sex. Overall, female applicants scored higher than male applicants by less than .25 of an SD. However, based on T-test[5] results in combination with analysis of effect-size using

[5] Independent sample T-tests are used to compare the mean scores of two different groups to assess if there is a statistically significant difference. The significance of the difference is assessed using the probability value (p-value). The closer the p-value is to 0, the more likely it is that the null hypothesis is false and hence, a difference is likely to exist.

Cohen's d[6], it is determined that the difference in the mean SJT scores for males and females was not significant, indicating that performance on the SJT does not appear to be influenced by sex differences.

Table 14: Group differences by sex

	Sex	N	Mean	SD	Sig Difference
Equated Data	Male	3,515	855.6	29.5	ns
	Female	4,555	862.1	30.6	

5.7.3 Table 15 shows group differences in performance on the SJT based on ethnicity by White and Black and Minority Ethnic (BME) groups. White applicants scored higher than BME applicants by approximately.5 of an SD and a T-test revealed that the difference was statistically significant (p<.001, d =.55).

5.7.4 Whilst items are designed to avoid group differences (i.e. avoiding the use of colloquial words/phrases, which might disadvantage particular groups), a richer understanding of the implications of the observed group differences in practice (for sex and ethnicity) is needed as an impetus for future research. Without detailed systematic research in this area, causal factors cannot be reliably identified.

Table 15: Group differences by ethnicity

	Ethnicity	N	Mean	SD	T-test Sig.	Cohen's D
Equated Data	White	5,180	865.1	26.4	p < .001	0.55
	BME	2,738	848.4	33.9		

5.7.5 Table 16 shows group differences in performance on the SJT based on the country of medical education (UK or Non-UK). Applicants from UK-based medical schools performed significantly better than those from non-UK medical schools by approximately 1.3 SDs and a T-test revealed that the difference was statistically significant (p<.001, d = 1.30).

Table 16: Group differences by country of medical education

	Country	N	Mean	SD	T-test Sig.	Cohen's D
Equated Data	UK	7822	861.4	27.3	p < .001	1.30
	Non-UK	340	810.4	48.2		

[6] Cohen's d is an effect size statistic used to estimate the magnitude of the difference between the two groups. When comparing differences between large sample sizes, it is best practice to include a measure of effect size. Cohen's d is defined as the difference between two means divided by the standard deviation for the data. The guidelines (proposed by Cohen, 1988) for interpreting the d value are:. 2 = small effect, .5 = medium effect and .8= large effect.

5.7.6 In terms of age, there was a negative correlation between age and scores on the SJT (r=-.075 - p<.001), with younger applicants scoring significantly higher on the SJT than older applicants. This finding is in line with previous findings from the Main Pilot (July 2011) but in contrast to findings from the Parallel Recruitment Exercise (March 2012). The effects of age on SJT performance should therefore continue to be monitored.

5.7.7 Differential Item Functioning (DIF) was used to examine group differences at an item level. The DIF analysis is a procedure used to determine if test items are fair and appropriate for assessing the ability of various demographic groups. It is based on the assumption that test takers who have similar ability (based on total test scores) should perform in similar ways on individual test items, regardless of their sex or ethnicity. DIF is a necessary, but not sufficient condition, for bias: bias only exists if the difference is illegitimate, i.e. if both groups should be performing equally well on the item. An item may show DIF but not be biased if the difference is due to actual differences in the groups' ability to answer the item, e.g. if one group is of high proficiency and the other of low proficiency, the low proficiency group would necessarily score much lower.

5.7.8 DIF, undertaken using a multiple regression analysis[7], was used to examine whether sex or ethnicity significantly predicts performance on each item once overall test performance has been controlled for (i.e. to determine if there is a difference in item performance beyond that which was expected due to differences between groups on the test overall).

5.7.9 30 (30.3%) items were flagged for sex differences (males performed better on 19 items and females on 11). Of the items flagged for sex differences, nine (9.1%) had previously been flagged as exhibiting the same sex differences when previously piloted.

5.7.10 22 (22.2%) items were flagged for ethnicity differences (White applicants performed better on 11 items and Black and Minority Ethnic applicants on 11 items). As items on which White applicants and Black and Minority Ethnic applicants performed better on are present in equal proportions, this suggests that the test is not biased with regards to ethnicity. Of these items flagged for ethnicity differences, three (3.1%) had previously been flagged as exhibiting the same ethnicity differences when previously piloted.

5.7.11 Items which have been flagged for gender and ethnicity differences will be reviewed in light of these results to identify whether there appears to be any bias in the item content. Once reviewed, if the items do appear to demonstrate bias (as outlined above, DIF is a necessary but not sufficient condition for bias), items will be amended or removed from the item bank.

[7] To account for the large sample sizes (Paper 1: n=3872 White, n=2061 BME, n=2686 male, n=3368 female. Paper 2: n=1303 White, n=666 BME, n=821 male, n=1179 female) items were flagged as exhibiting a significant difference between groups at the p < 0.01 threshold. For previous pilots, with smaller sample sizes, the p < .05 threshold was used.

5.8 Correlations with Educational Performance Measure

5.8.1 The relationship between SJT total scores and the Educational Performance Measure (EPM), which was the other FP selection method used in 2013, was assessed using correlations. The EPM is a measure of the clinical and non-clinical skills, performance and knowledge of applicants up to the point of their application. It takes into account medical school performance, additional degrees and other educational achievements.

5.8.2 At the p<.01 level, significant correlations were found between SJT scores and EPM decile scores for Papers One and Two and between SJT scores and total EPM score for Papers One and Two. Although these correlations are significant, indicating some shared variance/commonality between the assessment methods, there is also a large amount of variance that is not explained, therefore the SJT appears to be assessing somewhat different constructs from the other methods.

Table 17: Correlations between SJT total scores and the Educational Performance Measure (EPM)

	Current selection methods	SJT total scores[8]
Overall	Total Score	r=.30**
	Decile	r_s=.30**
Paper One	Total Score	r=.31**
	Decile	r_s=.31**
Paper Two	Total Score	r=.29**
	Decile	r_s=.27**
Paper Three	Total Score	NS
	Decile	NS

** Significant at the p<.01 level; NS = not significant

5.9 Item Level Analysis – Trial Items

5.9.1 Fourteen sets of items were trialled in the FP2013 SJT, with each set consisting of seven ranking and three multiple choice items. Ten sets were trialled alongside operational Paper One, and four alongside Paper Two. One set from Paper One was also trialled alongside Paper Three. However, as the analysis involves correlating the trial items against the operational total, due to the very small number of applicants that sat Paper Three, these data were not included in the item analysis.

[8] Correlation coefficients provide information about the direction and strength of the relationship between two variables. Correlation coefficients can range from -1 to +1. A positive value indicates that there is a positive correlation (i.e. as one variable increases so does the other) a negative value indicates that there is a negative correlation (i.e. as one variable increases, the other decreases). The size of the value provides information on the strength of the relationship. For normally distributed data (i.e. the SJT total score), the Pearson product-moment correlation coefficient is used (r). For non-normally distributed data (i.e. the SJT decile), the Spearman's rank correlation coefficient is used (rs).

5.9.2 Item analysis was used to look at the difficulty (item facility) and quality (item partial) of trial SJT items. Together these can help to identify how well the items differentiate between applicants and the results are used to identify which items can enter the operational item bank and which items may need further refinement. The same criteria are applied as per the operational item level analysis.

5.9.3 Table 18 outlines the item level statistics for all fourteen sets of trial items. For the ranking items, the mean facility value was broadly similar across papers, with Papers 10 and 14 displaying a slightly lower mean facility (15.78 and 15.72 respectively) and Papers 11 and 12 demonstrating a slightly higher mean facility (17.36 and 17.45 respectively). For the multiple choice items, the mean facility was broadly similar across papers, with Paper 9 having the lowest mean facility value (7.7) and Paper 6 having the highest mean facility value (9.8). These mean scores are comparable with previous pilots and are in line with expectations.

5.9.4 The range of facility values across papers is broadly similar. Papers 11 and 12 had a small range (16.19-18.98 and 16.03-18.94 respectively) indicating that these sets of trial items did not contain any particularly ‘difficult’ ranking items. The standard deviation range is also broadly similar across papers. The range of facility values for multiple choice items differed across the different papers. Those items that are at the ‘easier’ end of the scale will be reviewed alongside other psychometric evidence (i.e. SD and partial) in relation to their inclusion in the item bank. This is to ensure that the operational item bank does not contain too many ‘easy’ items.

Table 18: Trial Item level statistics: Facility values

		Ranking			Multiple Choice		
	N	Mean	Facility Range	SD Range	Mean	Facility Range	SD Range
Overall	8156	16.68	13.56-19.1	1.36-3.30	8.74	6.61-10.67	2.11-3.05
Paper One	596	16.91	15.15-18.20	1.60-2.96	8.41	7.99-9.24	2.61-2.85
Paper Two	572	16.25	13.56-18.64	1.93-2.55	8.39	8.04-8.76	2.59-2.99
Paper Three	569	16.66	15.54-19.1	1.72-2.35	9.77	8.86-10.34	2.28-2.78
Paper Four	593	16.77	15.33-17.94	2.01-2.50	8.78	8.26-9.36	2.11-2.85
Paper Five	729	17.19	15.70-18.13	1.72-2.20	8.63	7.52-10.04	2.58-3.04
Paper Six	577	16.86	14.78-18.23	1.72-2.78	9.81	8.49-10.67	2.19-2.67
Paper Seven	720	16.84	15.47-18.13	1.81-2.30	8.57	8.11-9.01	2.47-2.65
Paper Eight	549	16.59	14.92-17.93	1.90-2.62	8.26	6.61-9.14	2.33-2.54
Paper Nine	573	16.20	14.05-17.74	2.22-2.72	7.69	6.79-8.77	2.32-2.85
Paper Ten	648	15.78	14.71-16.98	1.99-2.60	9.40	8.80-10.44	2.49-3.05
Paper Eleven	515	17.36	16.19-18.98	1.86-2.52	9.57	9.34-9.76	2.51-2.69
Paper Twelve	515	17.45	16.03-18.94	2.20-3.05	8.90	7.96-10.1	2.28-2.79
Paper Thirteen	519	16.97	15.79-17.95	1.36-2.37	8.16	7.08-9.50	2.29-2.72
Paper Fourteen	471	15.72	14.49-17.62	2.12-3.30	8.09	7.24-8.82	2.33-2.56

5.9.5 Item quality is determined by the correlation of the trial item with the total score on the **operational** items of the test. This analysis compares how individuals perform on a given trial item with how they performed on the operational items of the test overall. Although the item partial provides vital information in terms of how well an item is performing and helps to decide whether to enter it into the operational item bank, this needs to be taken into consideration with a number of other statistics (item facility, SD) as well as how the best performing applicants performed (i.e. if the best performing applicants have a very different key to that of the SMEs then it suggests that there may be problems with the item). It is also recommended that item partials are balanced with other considerations, e.g. the need to provide coverage of all target domains.

5.9.6 Table 19 outlines how items performed for each of the 14 papers and overall. Trial Papers 1, 3, 11 and 13 have the fewest items with partials above .17. Papers 3 and 13 have a large proportion of items below .13.

5.9.7 61 of the 140 (43.6%) items were deemed as having good psychometric properties with regards to item quality. 39 of the 140 (27.9%) items were deemed as having moderate psychometric properties. 40 of the 140 (27.9%) items were deemed as having poor psychometric properties.

Table 19: Trial item level statistics: Item partials

	Range of Item Partials	Mean Item Partial	Good (>0.17)	Moderate (0.13-0.17)	Item requires further review (<0.13)
Overall	-.146-.470	.169	61 (43.6%)	39 (27.9%)	40 (28.6%)
Paper One	.055-.211	.124	2 (20.0%)	4 (40.0%)	4 (40.0%)
Paper Two	.084-.271	.169	4 (40.0%)	4 (40.0%)	2 (20.0%)
Paper Three	.026-.212	.100	1 (10.0%)	2 (20.0%)	7 (70.0%)
Paper Four	.104-.269	.201	6 (60.0%)	3 (30.0%)	1 (10.0%)
Paper Five	-.074-.250	.149	5 (50.0%)	3 (30.0%)	2 (20.0%)
Paper Six	.130-.285	.187	5 (50.0%)	5 (50.0%)	0 (0.0%)
Paper Seven	.045-.194	.131	3 (30.0%)	4 (40.0%)	3 (30.0%)
Paper Eight	.090-.223	.152	3 (30.0%)	4 (40.0%)	3 (30.0%)
Paper Nine	.163-.369	.250	9 (90.0%)	1 (10.0%)	0 (0.0%)
0aper Ten	.018-.255	.138	3 (30.0%)	2 (20.0%)	5 (50.0%)
Paper Eleven	.054-.203	.122	1 (10.0%)	5 (50.0%)	4 (40.0%)
Paper Twelve	.242-.470	.350	10 (100.0%)	0 (0.0%)	0 (0.0%)
Paper Thirteen	-.142-.221	.084	2 (20.0%)	2 (20.0%)	6 (60.0%)
Paper Fourteen	-.146-.350	.203	7 (70.0%)	0 (0.0%)	3 (30.0%)

5.9.8 Following further review by the WPG team considering the available data (e.g. item facility, SD and the key of the best performing applicants), 59 (42.1%) of the items are deemed to be appropriate to enter the operational item bank and 39 (27.9%) are deemed to be inappropriate and un-amendable and will therefore not enter the item bank. The remaining 42 (30%) are deemed to be appropriate to be reviewed in collaboration with a clinician to potentially re-enter the 2014 trial item development process. These figures are broadly in line with expectations for SJT development and sufficient for the on-going refresh of the item bank.

5.10 Applicant Reactions

5.10.1 All applicants who participated in the SJT were asked to complete an evaluation questionnaire regarding their perceptions of the SJT. A total of 7770 applicants (95.2% of applicants) completed the questionnaire.

5.10.2 Applicants were asked to indicate their level of agreement with several statements regarding the content of the SJT paper, the results of which are shown in Table 20.

Table 20: Applicant Evaluation Responses

% candidates (N=7770)	Strongly Disagree 🟥 %	Disagree 🟥 %	Neither 🟨 %	Agree 🟩 %	Strongly Agree 🟩 %
The test seemed well-run and invigilated	2.0	4.5	11.1	61.4	21.2
The content seemed relevant to what I think the role of a foundation doctor will be	5.4	15.2	27.0	45.6	6.9
The content of the SJT appeared to be fair for selection to the Foundation Programme	10.0	21.5	29.9	34.5	4.1
The results of the SJT should help to differentiate between weaker and stronger applicants	16.4	28.9	30.3	21.2	3.2
I felt I had sufficient time to complete the SJT	12.3	26.0	13.2	41.0	7.6

5.10.3 82.6% of applicants agreed or strongly agreed that the SJT seemed to have been well run and invigilated; 52.5% thought that the content seemed relevant to what they thought the role of a foundation doctor would be; and 58.6% agreed or strongly agreed that they were given a sufficient amount of time to complete the SJT.

5.10.4 Only 38.6% agreed or strongly agreed that the content of the SJT appeared to be fair for selection to the Foundation Programme, however 29.9% of applicants neither agreed nor disagreed with this statement. Only 25.4% of applicants agreed or strongly agreed that the results of the SJT should help selectors to differentiate between weaker and stronger applicants, although 30.3% neither agreed nor disagreed with this statement.

5.10.5 75.9% of applicants thought that the SJT was of an appropriate level of difficulty, with 1.0% indicating that they believed the test was too easy, and 23.1% indicating that they thought the test was too difficult.

5.10.6 Applicants were invited to provide qualitative comments on the SJT. 1680 (20.6%) applicants took the opportunity to provide a comment. Whilst, due to the number of responses, WPG have not conducted a qualitative analysis of these data, a student representative undertook an analysis to identify common themes in responses, these can be summarised as follows.

5.10.7 In general, most applicants welcomed the introduction of the SJT and indicated that they thought it was a fairer assessment than the white space answers that were previously used, as it was not possible for individuals to obtain help from third parties and it also relied less on creative writing skills. However, students viewed it to be less fair than the EPM at present, and were concerned about the statistical validity of ranking students on a scale of 0-50 for their SJT score. They were also less convinced that it would pick out stronger applicants than the EPM and were concerned about the relative points difference in comparison to the EPM.

5.10.8 Applicants thought that the preparation they were offered for the SJT was helpful, despite the fact that it was 'unrevisable'. They suggested that it would be useful if MSC could release more practice questions so that applicants could get a better idea of the content before sitting the test. They also indicated that it was not fair for some universities to offer practice sessions and others not to; and that question books from alternative sources were helpful, but expensive.

5.10.9 Most applicants indicated that they thought that the SJT was generally well run and invigilated.

5.10.10 Many students felt that they were pressured for time and some found it difficult to fill in the answer sheet.

5.10.11 With regards to the content of the SJT, some applicants indicated that they felt the questions were repetitive and that the relevance to the working life of an FY1 was not always clear. For the ranking items, some applicants found it difficult to put several 'correct' options into a ranking order. Applicants also indicated that some items were too long and wordy, and relied heavily on British idiosyncrasies, making these items difficult for individuals who did not speak English as their first language, as well as those with specific disabilities.

Part Three: Summary & Recommendations

6 Summary

6.1 This report details the operational use of the SJT for selection to FP2013 as well as the development of new items to be trialled alongside selection to FP2014.

6.2 A test completion analysis revealed that the majority of applicants (97.9%) answered all items within the test, indicating that 140 minutes is an appropriate length of time to complete 70 items.

6.3 The psychometric analysis presented in this report supports all preceding evidence that the SJT is a reliable tool that is able to sufficiently differentiate between applicants (with a mean reliability across all papers of 0.76). Test-level analysis demonstrates that the three different versions of the test are broadly similar, however test equating techniques are undertaken to ensure equivalence in scores across the different versions.

6.4 Operational item level analysis revealed that the majority of operational items (81.8%) can be classified as good or moderate in terms of their psychometric properties. A more in depth review of these items, including analysis of facility values and DIF analysis, and a comparison to how they have previously performed, will take place and it is expected that virtually all of these items will remain the item bank. The remaining items (18.2%) are likely to be removed from the bank in accordance with best practice.

6.5 Group differences analysis reveals significant differences between performance in the test based on ethnicity and age. Although females scored higher than males, these differences were not deemed to be significant. Further research is needed to explore these differences.

6.6 Significant correlations were found between SJT scores and EPM decile scores and between SJT scores and total EPM scores. Although these correlations are significant, indicating some shared variance/commonality between the assessment methods, there is also a large amount of variance that is not explained, therefore the SJT appears to be assessing somewhat different constructs from the other methods.

6.7 140 items were trialled alongside the operational items during FP2013. 42.1% of the items are deemed to be appropriate to enter the operational item bank and 27.9% are deemed to be inappropriate and un-amendable and will therefore not enter the item bank. The remaining 30% are deemed to be appropriate to be reviewed in collaboration with a clinician to potentially re-enter the 2014 trial item development process. These figures are broadly in line with expectations for SJT development and sufficient for the on-going refresh of item bank.

7 Item Writing Methodology

7.1 The item development process should continue as outlined in this report. The inclusion of a diverse range of individuals from a range of specialties is encouraged, as is the involvement of FY1/FY2s to ensure that the items developed are relevant and realistic.

7.2 As the item bank continues to grow, the potential for duplicate items to emerge increases. Therefore, in developing new items to be trialled during FP2014, attention should be paid to generating a broader range of scenarios through utilising a wider range of question wording (for example; *'Rank in order* ***the importance of the following considerations*** *in the management of this situation', 'Rank in order the* ***extent to which you agree with the following statements*** *in this situation'* and *'Rank the* ***order in which you should carry out the following tasks'****).* In doing so, this will directly address specific feedback from applicants about the perceived similarity of items within the existing test. With the introduction of this different wording, amendments to applicant communications and the practice paper will need to be made.

7.3 The findings from item level analysis should be used to inform ongoing item development. For example, existing items should be reviewed to determine if there are particular characteristics of the content of poorly performing items, which can be avoided in future test development. Additionally, items flagged for either gender or ethnicity differences should also be examined to identify if there are particular characteristics of the content of these items which can be avoided in future test development.

8 Scoring and Scaling

8.1 For FP2013, the applicants' total score was created by summing the total EPM score and the scaled SJT score. The transformation of the raw SJT score to a scaled score was determined by mapping the lowest attained score to zero and the highest score to 50 and then mapping the remaining scores linearly between these two points.

8.2 It is recommended that in future a standardised scale should be created using the mean score and the SD of the scores to determine the linear transformation. The mean score is mapped to the desired point on the new scale and the SD is used to control the spread of scores on the new scale. The advantages of this process are:

- The scaling is stable because it is based on parameters from all applicants' scores (mean and SD). Because of the large population size, even large outliers will have minimal impact on these parameters, although good practice is to ignore very extreme outliers in determining the scale.
- The weighting of the SJT score when combined with the EPM can be controlled since this is determined by the SD, and the SD can be fixed to any desired value, for example to mirror the EPM.

- The scores are likely to remain consistent from year to year. While there may be minor changes, the average level of performance for such large groups is likely to remain fairly constant. Outliers, which can change from year to year have little, if any, impact on the scale. This means that those using the scores can start to understand the level of performance implied by a particular score.
- Changes to the scores of a few applicants after scaling will have little, if any, impact on the scaling for the majority of applicants. The scaling is robust against changes to some scores.
- The scale can be described in terms of its mean and SD. Scores therefore have a meaning when compared with the whole applicant group. This allows applicants to better understand how their performance has been rated, relative to the group, and so makes the scale more transparent. This helps to improve applicant perceptions of the appropriateness of the weighting process.

The disadvantage of this approach is that occasional very extreme scores can exceed the set range. The extreme scores can be left as they are or can be rounded to the desired extreme scores.

9 Equality and Diversity

9.1 Differences were found for ethnicity at both a test and item level and for place of medical education at a test level. This will continue to be monitored. To assist with ensuring that the test is not adversely discriminating against some minority groups, two recommendations are provided below.

9.2 Efforts should continue to be made to ensure that there is a diverse range of individuals involved in the development of the SJT in the future with regards to sex, age and ethnicity. This eliminates any unintended bias as a factor of those individuals involved in item writing and quality assurance.

9.3 Although during test development item writers were trained to ensure that the language used did not unfairly discriminate against applicants who were not educated in the UK, an additional independent equality and diversity check is recommended to be carried out as part of the design methodology to help ensure that the content of the items does not unfairly discriminate against any particular group due to a factor unrelated to the job specification. This is in line with UK employment law.

10 Applicant Communications

10.1 Comments from applicants undertaking the test as part of FP2013 suggest that there are a number of improvements which could be made to the communications provided to applicants prior to taking the test.

10.2 Several applicants reported that they felt that the items provided in the practice paper were not entirely reflective of the content of the operational papers. This is not unexpected, as the practice paper was not constructed in order to be an accurate reflection of the operational papers, but instead was an early pilot paper which was made available to applicants because the security of the paper had been compromised during the piloting process. As such, the pilot paper was a reflection of the items that were required to be piloted at that time and may not have covered all potential topic areas. In addition, the paper was also only 30 items in length and therefore would also not have covered the breadth of the operational item bank, as an operational paper would be designed to. Therefore, it is recommended that the content of the practice paper should be reviewed, amended and the number of items increased to ensure that these are reflective of the item content, domains and level of difficulty of the operational test. This will be especially important as the newly worded question types enter the papers in FP2014.

10.3 Another concern reported by applicants regarded the perceived weighting of the SJT against the EPM, with many individuals feeling that the restricted range on the EPM scale meant that the SJT accounted for more than 50% of the total score. Whilst, as discussed above, the weighting and scaling process can account for any discrepancy in scales across the two measures, it is recommended that the way in which this is communicated to applicants is reviewed to ensure that there is an accurate understanding of this process.

11 Computer-Delivered Test Administration

11.1 MSC is currently assessing the feasibility of administering the SJT electronically (i.e. using a computer, laptop or tablet device) from FP2015 onwards. Whilst recognising that there are a number of practical considerations to be taken into account, especially considering the large number of applicants required to take the test, WPG recommend this form of test administration over traditional pencil and paper forms for a number of reasons:

- It offers the potential to reduce the lack of adherence to instructions, for example by not allowing applicants to use the same rank twice or choose more than three options for the MCQs.
- It provides an easier interface in which to respond to questions and therefore offers improved applicant reactions.

- Applicants can be informed of how much time and how many questions they have left.
- Responses are collated and can be scored electronically, which avoids the need for responses to be scanned. This both reduces the risk of error and results in quicker turnaround for the response data.
- Response time can be measured more accurately and it allows exact control of the time that applicants have to complete the test.
- The screen resolution can be adapted to meet individual requirements, for example it should be possible to increase the text size.