



Analysis & Evaluation of a Situational Judgement Test for Selection to the Foundation Programme – Parallel Recruitment Exercise

Final Report

Improving Selection to the Foundation Programme

Work Psychology Group

Professor Fiona Patterson

Vicki Ashworth

Dr Alec Knight

Dr Máire Kerrin

March 2012

Contents

1	Introduction	3
1.1	Purpose and Structure of the Report	3
1.2	Background.....	3
Part One: Item Development & Review		4
2	Item Development & Review	4
2.1	Process Overview.....	4
2.2	Item Writing Workshops	4
2.3	Item development interviews	7
2.4	Item Review.....	8
2.5	Review workshops.....	9
2.6	Concordance Panel.....	9
Part Two: Pilot Analysis & Evaluation		12
3	Pilot Analysis.....	12
3.1	Purpose and Specifications of the pilot.....	12
3.2	Evaluation Overview.....	12
3.3	Pilot Test Specification	13
3.4	Sample	14
3.5	Overall Summary of Results	23
3.6	Test Level Analysis	24
3.9	Item Level Analysis	29
3.10	Group Differences	32
3.11	Correlations with Quartiles and Application Form.....	34
3.12	Participant Evaluation	35
Part Three: Summary & Recommendations		38
4	Summary.....	38
5	Item Writing Methodology.....	39
6	Clinician review.....	40
7	Item bank.....	40

1 Introduction

1.1 Purpose and Structure of the Report

1.1.1 A Parallel Recruitment Exercise (PRE) was carried out in 2011/2012 as part of the Improving Selection to the Foundation Programme project. The purpose of the PRE was twofold:

- To evaluate the logistics of running a Situational Judgement Test (SJT) in a large scale across a number of key dates
- The development and piloting of additional items to maximise the item bank

1.1.2 This report is concerned with only the second objective; to provide details of the analysis and evaluation of a SJT as part of a PRE.

1.1.3 The report is divided into three main parts:

- Part One describes the development process of additional items that were piloted as part of the PRE.
- Part Two describes the results and analysis of the PRE pilots that were carried out.
- 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 Foundation Programme (FP). The FP is a two-year generic training programme which forms the bridge between medical school and specialist/general practice training.

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

1.2.3 In August 2011, a report was produced on the design, analysis and evaluation of a SJT for Selection to the Foundation Programme. Recommendations were for the implementation of a SJT, alongside the Educational Performance Measurement (EPM) in 2013.

Part One: Item Development & Review

2 Item Development & Review

2.1 Process Overview

2.1.1 The SJT item development process was conducted using two methodologies in parallel. The purpose of developing items using two methodologies was to evaluate each methodology with regards to efficiency and productivity.

2.1.2 Figure 1 summarises the development and review process undertaken.

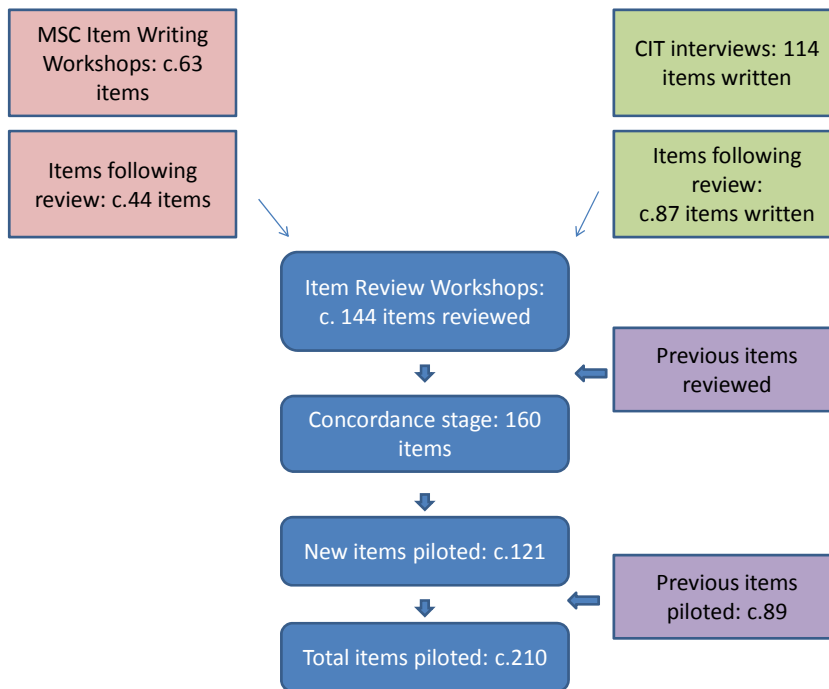


Figure 1: Item development and review process

2.2 Item Writing Workshops

2.2.1 A total of three item writing workshops were held in August 2011. Two workshops were held in London and one workshop was held in Birmingham.

2.2.2 A total of 11 item writers attended the workshops. All item writers were new to the process and had not been previously trained. Previously trained item writers were also invited to attend the second phase of workshops although none could attend.

2.2.3 The breakdown of the demographics of the item writers, their job role and primary speciality is provided in the tables below.

Table 1: Demographics of item writers

Sex	Male	4
	Female	7
	Not Stated	-
Age group	25 and under	-
	26-35	-
	36-45	4
	46-55	6
	56-65	1
	66 and over	-
	Not stated	-
Ethnicity	White	10
	Black	-
	Asian	1
	Mixed	-
	Chinese	-
	Other	-
	Not Stated	-

Table 2: Item writers' job role

Clinical Tutor	-
Clinical/Educational Supervisor	7
Foundation School Director	3
Lay Representative	-
Medical School Director	1
Other	-
Not stated	-

Table 3: Item writers' primary specialty

Specialty	No.
Anaesthetics	1
ENT	1
General Surgery	2
Geriatric Medicine	2
Obstetrics & Gynaecology	1
Radiology	1
Renal	2
Rheumatology	1

- 2.2.4 Information was sent to item writers prior to their attendance at item writing workshops, outlining the background to the project and the use of SJTs for selection, the purpose of the workshops, their role on the day and some example SJT items. Item writers were also provided with some initial information about item writing principles and were asked to complete some pre-work prior to attending the workshops. On the day, item writers were required to sign confidentiality and code of conduct agreements.
- 2.2.5 Prior to attending the workshop, each item writer was asked to think of at least three possible scenarios or critical incidents that could form the basis of SJT items. They were advised that these should be incidents that involved some dilemma and need for judgement, therefore avoiding straightforward incidents that would elicit text book responses. Item writers were also asked to think about possible responses to the scenarios that should be actions that could plausibly be taken by the applicant in response to the scenario.
- 2.2.6 Item writers were given guidance on the SJT target domains¹ and were advised that the scenarios should be relevant to one of these domains. This ensures that the scenario is relevant to the professional attributes that are expected of a FY1 doctor and that the content of the scenarios maps on to the person specification.
- 2.2.7 The SJT target domains which were presented in the pre-information sent to item writers and at the item writing workshops are outlined in the table below. Pre-information also advised that:
- Scenarios should ideally be fairly short (typically 30-60 words and no more than 80 words) but should provide enough detail to provide dilemma and complexity.
 - Scenarios and responses should be described clearly and unambiguously.
- 2.2.8 Item writers were requested to write two formats of items; ranking items and multiple choice items. Ranking items ask applicants to “rank in order the following actions in response to this situation (1= most appropriate; 5= least appropriate)” and multiple choice items ask applicants to “choose the three most appropriate actions to take in this situation.”
- 2.2.9 Item writers were then asked to work in pairs creating initial scenarios. Items were then group reviewed and more time was spent developing new items and reviewing items.
- 2.2.10 Unlike previous item writing workshops, item writers were not required to complete any follow on work. As such, item writers were not asked to review their items further and were not provided with any feedback on their items.
- 2.2.11 In total, over the three item writing workshops, 63 items were written. This equals an average of 5.7 items per item writer.

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

2.3 Item development interviews

- 2.3.1 Following recent development work in other contexts, Item Development Interviews using Critical Incident Technique (CIT) were held as an alternative methodology to write SJT items. CIT interviews aim to elicit from Subject Matter Experts scenarios or incidents involving Foundation Year One (FY1) doctors that demonstrate particularly effective or ineffective behaviour.
- 2.3.2 Using interviews such as these have 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.
- 2.3.3 An invitation was sent out via the UKFPO Foundation School Directors network who cascaded information to those working closely with FY1 doctors.
- 2.3.4 In total, 24 interviews were conducted by four trained interviewers. Details of interviewee role and specialty are provided below in table 4 and 5. Full demographic details of the interviewees were not recorded.

Table 4: Item writers job role

Clinical Tutor	-
Clinical/Educational Supervisor	19
Foundation School Director	1
Lay Representative	-
Medical School Director	1
FY1/FY2	3
Other	-
Not stated	-

Table 5: Item writers' primary specialty

Specialty	No.
Anaesthetics	2
Colorectal surgery	1
Emergency Medicine	1
Endocrinology	1
ENT	1
General Medicine	1
Geriatric Medicine	1
Infectious diseases	1
Obstetrics & Gynaecology	1
Orthopaedic Surgery	1
Oncology	1

Paediatrics	2
Respiratory	1
Not stated	9

2.3.5 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, including 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.

2.3.6 A total of 114 items were written. This equals an average of 4.6 items per 45 minute interview.

2.4 Item Review

2.4.1 All 177 items that were submitted were logged on a spreadsheet which indicated item writer, the date the item was submitted, type of item, target domain, answer key and a short one line summary of the item.

2.4.2 The breakdown of items relevant to each of the target domains written at or following the workshops was as follows:

- Commitment to Professionalism - 28
- Coping with Pressure - 49
- Effective Communication - 17
- Patient Focus - 50
- Working Effectively as Part of a Team - 33

2.4.3 The breakdown of items regarding item format was as follows:

- Ranking - 114
- Multiple Choice - 63

2.4.4 All items from the item writing workshops were reviewed by the core team of item reviewers from Work Psychology Group. Where necessary, items were passed to a Lead Clinician for further review, in particular where there were clinical based queries. The Lead Clinician is an individual expert in SJT design and review who has previously worked in SJTs for entry to specialty training.

2.4.5 Of the 63 items written in the item writing workshops, 19 were rejected due to not aligning with item writing principles. This is a 70% success rate. A Lead Clinician reviewed 17 (27%) of the items.

2.4.6 Of the 114 items written from the CIT interviews, 27 were rejected due to not aligning with item writing principles. This is a 76% success rate. A Lead Clinician reviewed 21 (24%) of the items.

- 2.4.7 In addition to new items written, a number of items that had not been successful at the previous pilot (79), and also some items that were unsuccessful at the previous concordance (16), were reviewed with the intention of potentially including them for the PRE SJT pilot.
- 2.4.8 With regards to those items that had not been successful at the previous pilot, following review, 45 had minor changes and were deemed suitable to be piloted. 34 items had more substantial changes and went either to the review workshop stage or the concordance stage. Ideally all items that had been amended would have gone to the concordance stage however minor the changes. However, this was not possible due to logistical constraints. If necessary, these items can go to a concordance stage following the pilot.

2.5 Review workshops

- 2.5.1 The aim of the review workshops was for SJT trained clinicians to review SJT items for relevance and fairness, as well as agreeing a scoring key. The benefit of holding these review workshops is that it enables input from a larger number of clinicians who are able to provide invaluable input into the development of the items.
- 2.5.2 A small number of Foundation Year Two (FY2) doctors also attended the workshops to provide additional input in terms of relevance and realism. This negated the need for separate FY1/2 focus groups.
- 2.5.3 Four review workshops were held; two in East Midlands and two in Peninsula. Attendees were invited based on existing networks within specialty selection. As such, the primary specialities of the attendees were general practice and anaesthesia.
- 2.5.4 A total of 19 individuals attended the four workshops, including 4 FY2s.
- 2.5.5 All participants who volunteered to take part were sent briefing material outlining the purpose of the review workshop and their role on the day. All participants also completed a confidentiality and code of conduct agreement.
- 2.5.6 During the workshop, delegates were split into two groups. As a group, with the aid of a facilitator, delegates reviewed no more than 20 items. Delegates were asked to consider the scenario content and the response. They were also asked to provide a possible answer key, which was compared with the answer key proposed by the item writers. Their comments and suggestions were recorded by the facilitator and updates were made to items.
- 2.5.7 A total of 144 items were reviewed during the focus groups. Following the review workshops, 5 items were rejected due to issues either with relevance or fairness.

2.6 Concordance Panel

- 2.6.1 In order to validate the SJT items further, concordance panels were conducted. 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 the pilot.

Those items exhibiting low levels of consensus are put to one side for further review with changes made if necessary.

- 2.6.2 The answer key provided by the concordance panel was used in combination with information from item writers and review workshops to determine the rational scoring key for the pilot data, however this may not reflect the final key as information will be used from the pilot to develop the items and their keys further. For example, if well good performing applicants consistently provide a different key to the established key, then the key will be reviewed with the assistance of Subject Matter Experts.
- 2.6.3 The criteria for Subject Matter Expert involvement in the concordance panel was that the individuals work closely with FY1 doctors and are very familiar with the responsibilities and tasks, as well as the necessary skills and abilities required for the role.
- 2.6.4 Two concordance panels were held with one paper reviewed at each panel; both papers consisting of 80 items. Therefore a total of 160 items went to concordance; 126 of these were new items. At this stage, the tests were not constructed as final tests i.e. no consideration was given as to spread of item topics or domains as the aim of the concordance panels was to analyse individual items.
- 2.6.5 A total of 23 individuals attended the concordance stage. One panel consisted of 11 individuals and one panel consisted of 12 individuals. (Ideally, for a concordance panel, a minimum of 10 individuals should be involved to ensure robust results). Details of interviewee role and specialty are provided below in table 6 and 7 below.

Table 6: Concordance participants' job role

Clinical Tutor	2
Clinical/Educational Supervisor	1
Foundation School Director	-
Foundation Programme Training Director	4
Lay Representative	-
Medical School Director	-
FY1/FY2	-
Other	-
Not stated	16

Table 7: Concordance participants' primary specialty

Specialty	No.
Anaesthetics	-
Colorectal surgery	-
Emergency Medicine	-
Endocrinology	3
ENT	-
Gastroenterology	1
General Medicine	1
General Surgery	1
Geriatric Medicine	-
Infectious diseases	-
Neurology	1
Obstetrics & Gynaecology	1
Orthopaedic Surgery	-
Oncology	-
Paediatrics	1
Respiratory	4
Not stated	10

- 2.6.6 After completing a confidentiality and code of conduct form, the panel was asked to complete the SJT items under test conditions. There was no time limit, although the panels were told the test should take no more than two hours 30 minutes to complete.
- 2.6.7 Feedback on the item content was provided by the panel, and this resulted in some minor alterations to a small number of items to provide clarification. No item was altered sufficiently to affect the interpretation of the question or the answer key.
- 2.6.8 Following the concordance panel meeting, a concordance analysis was undertaken to analyse the experts' level of agreement over the keyed response for each trial item. Using established criteria of acceptance levels², items were deemed either to have acceptable levels of concordance (149) or unacceptable levels of concordance (11).

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

Part Two: Pilot Analysis & Evaluation

3 Pilot Analysis

3.1 Purpose and Specifications of the pilot

- 3.1.1 The purpose of the initial pilot from an analysis and evaluation approach was to pilot a large number of items to maximise the item bank. The purpose was not to evaluate the use of the SJT for Selection to the Foundation Programme, although full evaluation of the tests is carried out.
- 3.1.2 The PRE was conducted at 30 medical schools and 2 centres for non UK students. All schools were asked to take part, although involvement in this process was not mandatory.
- 3.1.3 The proposed specification for the pilot, as defined by the authors, was a minimum of 400 participants for each paper. This would allow the psychometric properties of the SJT to be fully analysed and evaluated and would mean that greater confidence could be placed in the results.
- 3.1.4 Another requirement of both pilots was that the pool of participants was representative of the wider pool of applicants. For any voluntary pilot, this is likely to be a challenge as the individuals will be self selecting and may represent only a sample of the population. An associated issue relevant to the pilot population is that of test taking motivation. In any piloting of a selection test, the motivation of those taking the test may differ to that of applicants for a live selection test. Motivational issues may impact upon the time taken to complete the pilot test (with individuals completing the test in a quicker time than may be anticipated on a live test), the number of items completed (with fewer items being completed than on a live test) and the quality of the answers provided.
- 3.1.5 As the pilots were carried out in a number of schools, the possible influence of other factors is required to be controlled for as much as possible to ensure that any differences found are due to the test content/applicant ability rather than other factors. This includes ensuring that the tests are invigilated in a standardised way (using instructions provided) and are conducted in similar environments (i.e. test halls).

3.2 Evaluation Overview

- 3.2.1 This section outlines the psychometric analysis for the pilot. Any high stakes, high profile test needs to meet exacting psychometric standards in terms of the quality of individual items and of the test as a whole, including reliability, validity and fairness. The main analysis and evaluation activities reported here include:
- test level statistics, including reliability and scoring distributions
 - item level statistics, including item facility and effectiveness; only those items with sufficient psychometric properties will be used in an operational test
 - analysis of group differences at a test and item level to explore fairness
 - evaluation of participant reactions

- relationships between application form scores, quartiles and the SJT

3.2.2 All additional analysis carried out following both pilots used the entire test i.e. poorly performing items were not removed when analysing group differences, relationships with other assessments or criterion related validity.

3.3 Pilot Test Specification

3.3.1 A total of 210 items were piloted over 7 papers. This consisted of 121 new items and 89 reviewed and amended items that had been developed as previous item development processes.

3.3.2 Each pilot paper consisted of 30 items; this is less than the recommended 60 items for an operational test. As the 60 item test, and its suitability to be used as part of selection to the Foundation Programme, had been evaluated previously³ it was not deemed necessary at this stage to evaluate the full 60 item test. Instead, a shorter 30 item test was proposed. This still enables a large number of items to be piloted to maximise the bank, whilst also being logistically more straightforward. However, caution should be used when interpreting the test and item level analysis due to the shortened test length.

3.3.3 Each paper consisted of 19 ranking and 11 multiple choice items. All items were unique to each paper i.e. there were no anchor items between each paper. This was to assist with maximising the item bank.

3.3.4 As far as possible, an equal spread of target domains were selected for each paper, however the proportion of items from each target domain is also a reflection of the number of items written within each domain. This also had to be balanced with item response format and topic when constructing the papers. An outline of the spread of domains for each of the papers is outlined in table 8 below.

Table 8: 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
Total	57	37	30	48	38
1	7	4	4	8	7
2	8	8	4	5	5
3	9	5	2	9	5
4	7	4	6	8	5
5	8	6	3	7	6
6	10	5	5	5	5
7	8	5	6	6	5

³ FY1 SJT Final Report August 2011

3.3.5 There were a total of 512 marks available for each version of the SJT paper, with 20 marks available for each of the 19 ranking items and 12 marks for each of the 11 multiple choice items.

3.4 Sample

- 3.4.1 All applicants to FP 2012 were invited to take part in the PRE. In addition, the invitation to take part in the PRE was extended to all final year UK medical students, including those who did not complete an FP 2012 application. This included students who had been pre-allocated to a Defence Deanery Foundation Programme (separate recruitment – but in future, these applicants will also be required to complete the SJT); students who had chosen to take a year out post-graduation (usually for personal reasons); or international students who were returning overseas directly after graduation from medical school.
- 3.4.2 There was a total of 6842 participants in the PRE, of which 6706 were FP 2012 applicants (94% of all FP 2012 applicants) and a further 136 who did not complete an application to FP 2012 (i.e. pre-allocated to the Defence Deanery).
- 3.4.3 A breakdown of the number of participants who sat each of the seven papers can be seen in Table 9 below. As far as possible, only one version of a paper was undertaken at each school for logistical reasons, and to minimise security risk with the items. However, in 10 schools, participants undertook more than one paper as the medical school ran the SJT across more than one date.
- 3.4.4 The participant split between the seven papers aimed to be as similar as possible. Although attempts were made to ensure an equal as possible split between the papers, as the number of participants within each of the schools was unpredictable, this was not possible. The sample size for each paper is well above the requirements outlined above and as such confidence can be placed in the outcomes of the psychometric analysis.

Table 9: Number of participants taking each paper

	No. of participants	Percentage of Overall Sample
Paper One	1188	17.4%
Paper Two	881	12.9%
Paper Three	853	12.5%
Paper Four	1183	17.3%
Paper Five	889	13.0%
Paper Six	822	12.0%
Paper Seven	1026	15.0%

3.4.5 32 centres took part in the PRE. This included 30 medical schools and 2 centres that ran the pilot with applicants trained outside the UK (treated as a single school in Table 10). The breakdown of the number of participants from each centre is outlined in Table 10.

Table 10: Number of participants at pilot sites

	No. of participants	Percentage of Overall Sample	Paper Taken
School 1	149	2.2%	3&7
School 2	308	4.5%	1
School 3	237	%	5&6
School 4	335	4.9%	1
School 5	121	1.8%	1
School 6	242	3.5%	7
School 7	154	2.3%	7
School 8	316	4.6%	4
School 9	145	2.1%	1&6
School 10	163	2.4%	4&6
School 11	182	2.6%	7
School 12	208	3.0%	7
School 13	133	1.9%	5
School 14	221	3.2%	2
School 15	108	1.6%	3
School 16	429	6.3%	6
School 17	48	0.7%	3
School 18	214	3.1%	5
School 19	220	3.2%	5
School 20	298	4.4%	3
School 21	463	6.8%	2&6
School 22	70	%	3, 5 & 7
School 23	298	4.4%	4
School 24	326	4.8%	4
School 25	94	1.4%	6
School 26	163	2.4%	4&5
School 27	206	3.0%	2
School 28	266	3.9%	1&6
School 29	230	3.4%	1&7
School 30	331	4.8%	3&6
School 31 ⁴	164	2.3%	3&5

⁴ School 31 was a non-UK pilot site during which the UKFPO acted as a school for these participants

3.4.6 Participant demographic data were collected from the FPAS application. Demographic data were not collected for participants in the PRE who did not complete an FPAS application.

3.4.7 Table 11 outlines the breakdown of participants by sex. Overall, more females participated in the pilot (3724, 54.4%) than males (2657, 38.8%) (reflecting the male/female split of medical students) and the proportion of males and females is roughly equal across all seven papers.

Table 11: Participant sex by paper

		Male	Female	Not declared
Overall	No. of participants	2657	3724	461
	% of participants	38.8%	54.4%	6.7%
Paper One	No. of participants	451	668	69
	% of participants	38.0%	56.2%	5.8%
Paper Two	No. of participants	344	479	58
	% of participants	39.0%	54.4%	6.6%
Paper Three	No. of participants	330	426	97
	% of participants	38.7%	49.9%	11.4%
Paper Four	No. of participants	454	650	79
	% of participants	38.4%	54.9%	6.7%
Paper Five	No. of participants	355	489	45
	% of participants	39.9%	55.0%	5.1%
Paper Six	No. of participants	318	460	44
	% of participants	38.7%	56.0%	5.4%
Paper Seven	No. of participants	405	552	69
	% of participants	39.5%	53.8%	6.7%

3.4.8 Table 12 outlines the breakdown of participants by ethnicity. Overall, the majority of participants reported their ethnicity as 'white' (4159, 60.8%) with the smallest proportion of participants (145, 2.1%) reporting themselves as being from 'Other Ethnic Background'. This reflects the medical student profile in the UK. The proportion of ethnic breakdown varies across the seven papers, with Paper Seven having the highest proportion of white participants (70.1%) and Paper Six having the lowest proportion of white participants (46.8%).

Table 12: Participant ethnicity (5+1 groups) by paper

		White	Asian	Black	Chinese	Mixed	Other	Not declared
Overall	No. of participants	4159	1181	169	320	257	145	611
	% of participants	60.8%	17.3%	2.5%	4.7%	3.8%	2.1%	8.9%
Paper One	No. of participants	626	293	37	56	44	31	101
	% of participants	52.7%	24.7%	3.1%	4.7%	3.7%	2.6%	8.5%
Paper Two	No. of participants	491	158	17	69	41	26	79
	% of participants	55.7%	17.9%	1.9%	7.8%	4.7%	3.0%	9.0%
Paper Three	No. of participants	521	127	13	32	30	16	114
	% of participants	61.1%	14.9%	1.5%	3.8%	3.5%	1.9%	13.3%
Paper Four	No. of participants	814	146	22	29	49	18	105
	% of participants	68.8%	12.3%	1.9%	2.5%	4.1%	1.5%	8.8%
Paper Five	No. of participants	603	139	26	22	26	15	58
	% of participants	67.8%	15.6%	2.9%	2.5%	2.9%	1.7%	6.5%
Paper Six	No. of participants	385	212	39	51	39	27	69
	% of participants	46.8%	25.8%	4.7%	6.2%	4.7%	3.3%	8.4%
Paper Seven	No. of participants	719	106	15	61	28	12	85
	% of participants	70.1%	10.3%	1.5%	5.9%	2.7%	1.2%	8.2%

3.4.9 Table 13 outlines the breakdown of participants' ethnicity (White and Black and Minority Ethnic (BME) group). 4159 (60.8%) participants reported themselves as white and 2072 (30.3%) participants reported themselves as being from Black and Minority Ethnic groups. 611 (8.9%) participants did not declare their ethnicity. Paper Six has the highest proportion of BME participants (44.8%) followed by Paper One with 38.8%.

Table 13: Participants ethnicity (2 groups) by paper

		White	BME	Not declared
Overall	No. of participants	4159	2072	611
	% of participants	60.8%	30.3%	8.9%
Paper One	No. of participants	626	461	101
	% of participants	52.7%	38.8%	8.5%
Paper Two	No. of participants	491	311	79
	% of participants	55.7%	35.3%	9.0%
Paper Three	No. of participants	521	218	114
	% of participants	61.1%	25.6%	13.3%
Paper Four	No. of participants	814	264	105
	% of participants	68.8%	22.3%	8.8%
Paper Five	No. of participants	603	228	58
	% of participants	67.8%	25.6%	6.5%
Paper Six	No. of participants	385	368	69
	% of participants	46.8%	44.8%	8.4%
Paper Seven	No. of participants	719	222	85
	% of participants	70.1%	21.6%	8.2%

3.4.10 The mean age of the entire sample was 24.7 years and the median age was 23.5, with a range of 21 – 56 years.

3.4.11 To assist with establishing the equivalence of those taking the seven papers, the quartile scores and competency based application form (white space question) scores for the participants taking the seven papers were examined. Ideally, each population's scores should be normally distributed and should show a good and comparable spread of scores for both the application form and the quartile score as would be expected in a full operational sample. It should be noted that this information was not available for all participants.

3.4.12 Table 14 provides the descriptive data for the application form and quartiles by paper. The results show that the mean scores for the application forms vary between the papers, with Papers Two and Six having the highest mean score (40.4) and Papers Four and Five having the lowest mean score (38.0). T-test analyses showed that differences between the scores for all papers are significant at the $p < 0.01$ level. It may therefore be expected that the overall SJT scores for Papers Two and Six would be higher than for the remaining papers. This is based on the assumption that the application form is a reliable and valid assessment

method which correlates positively with the SJT. Histograms of application form scores for each of the 7 papers are provided below.

3.4.13 There were no significant differences in quartile scores for the seven papers.

Table 14: Application form and quartile scores by paper

	Application Form			Quartile		
	Mean	Range	SD	Mean	Range	SD
Paper One (1085)	39.2	10-54	6.7	37.0	34-40	2.2
Paper Two (757)	40.4	14-60	7.2	36.9	34-40	2.2
Paper Three (727)	39.1	14-54	7.5	37.1	34-40	2.2
Paper Four (1079)	38.0	15-54	7.0	37.1	34-40	2.2
Paper Five (788)	38.0	12-54	7.8	37.0	34-40	2.3
Paper Six (746)	40.4	15-55	7.0	36.8	34-40	2.2
Paper Seven (886)	39.0	16-56	7.2	37.0	34-40	2.2

Figure 2: Histogram for Application Form Total for Paper One

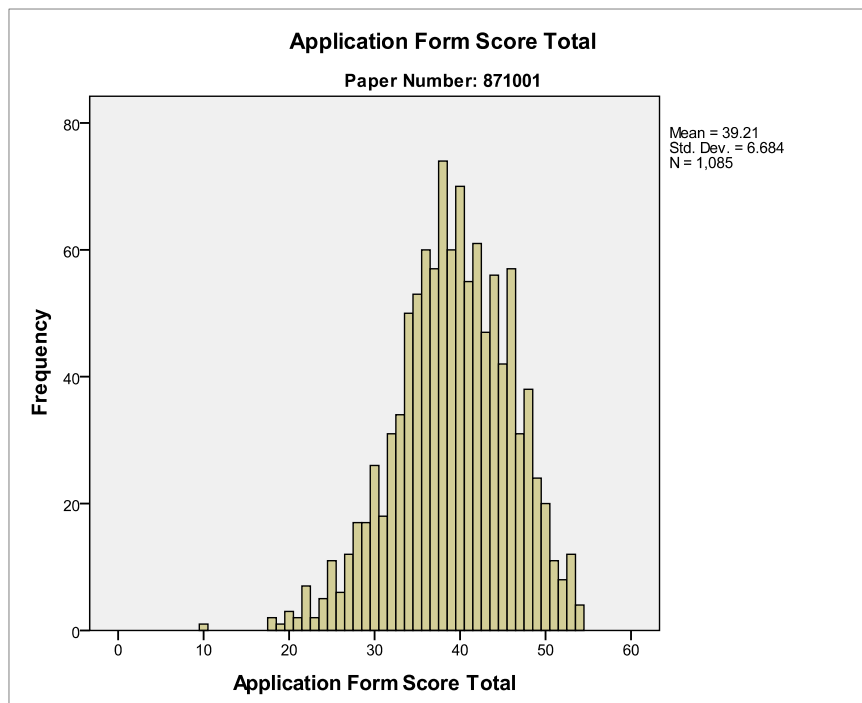


Figure 3: Histogram for Application Form Total for Paper Two

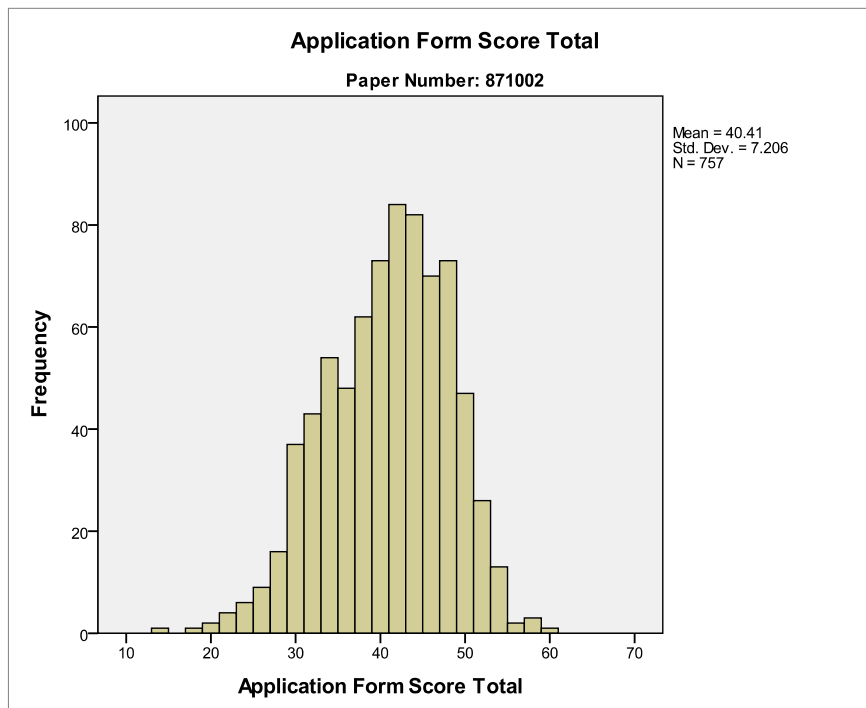


Figure 4: Histogram for Application Form Total for Paper Three

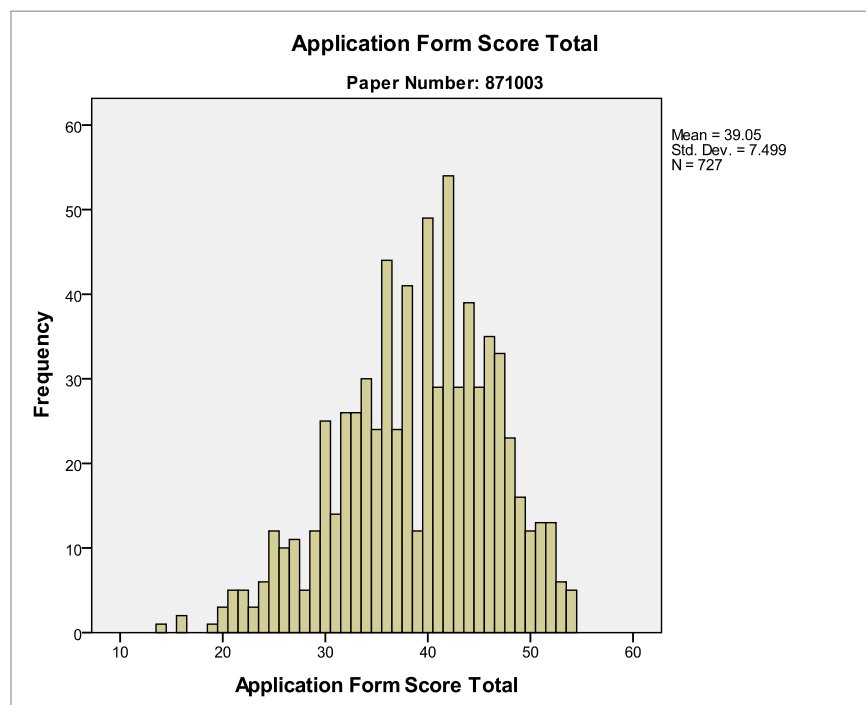


Figure 5: Histogram for Application Form Total for Paper Four

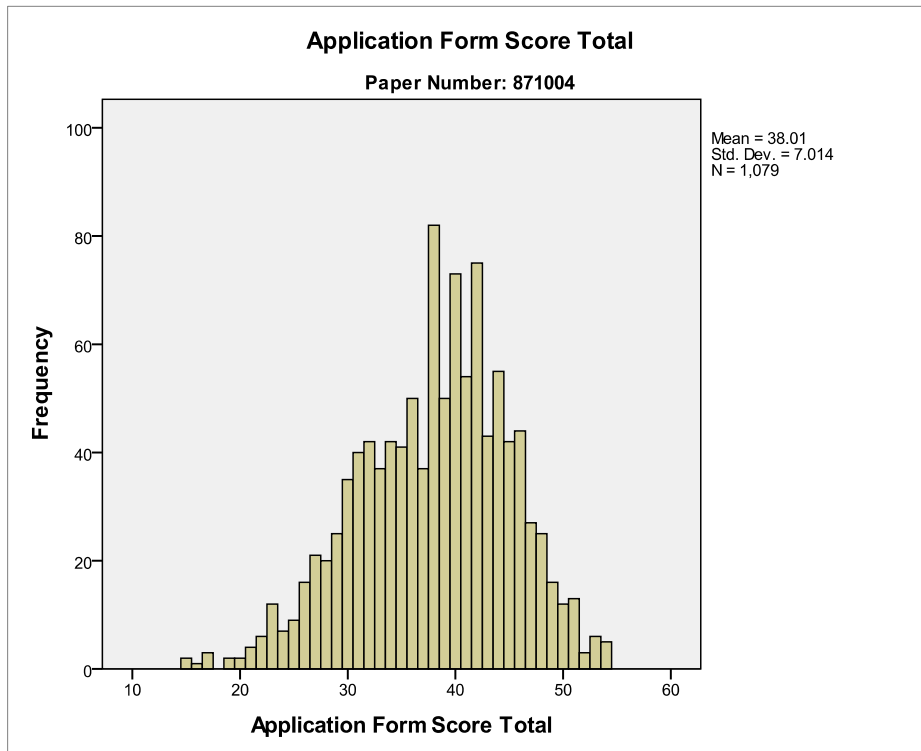


Figure 6: Histogram for Application Form Total for Paper Five

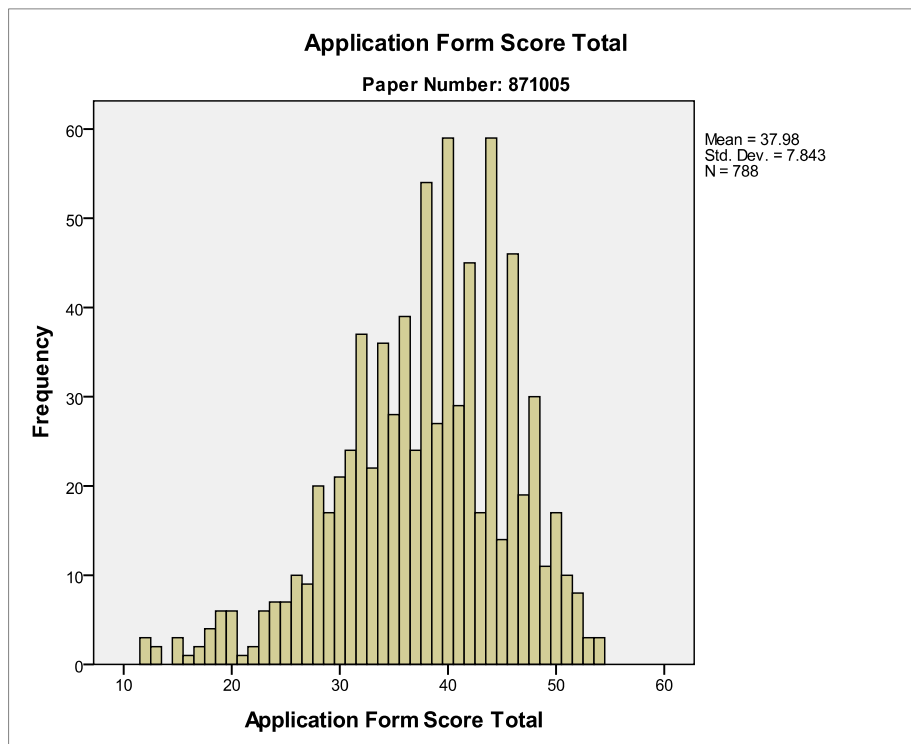


Figure 7: Histogram for Application Form Total for Paper Six

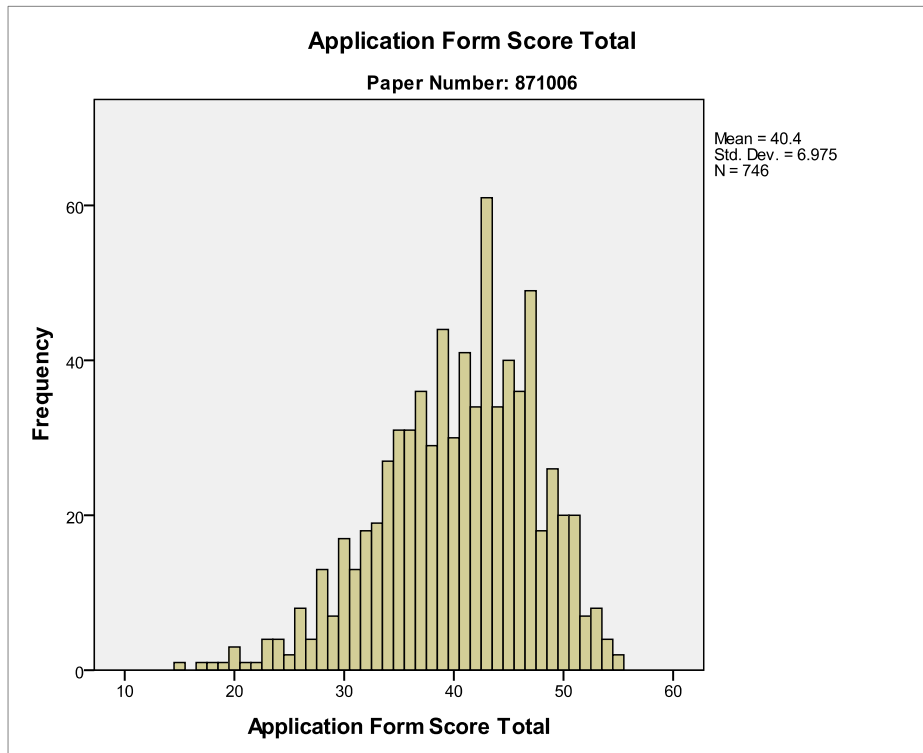
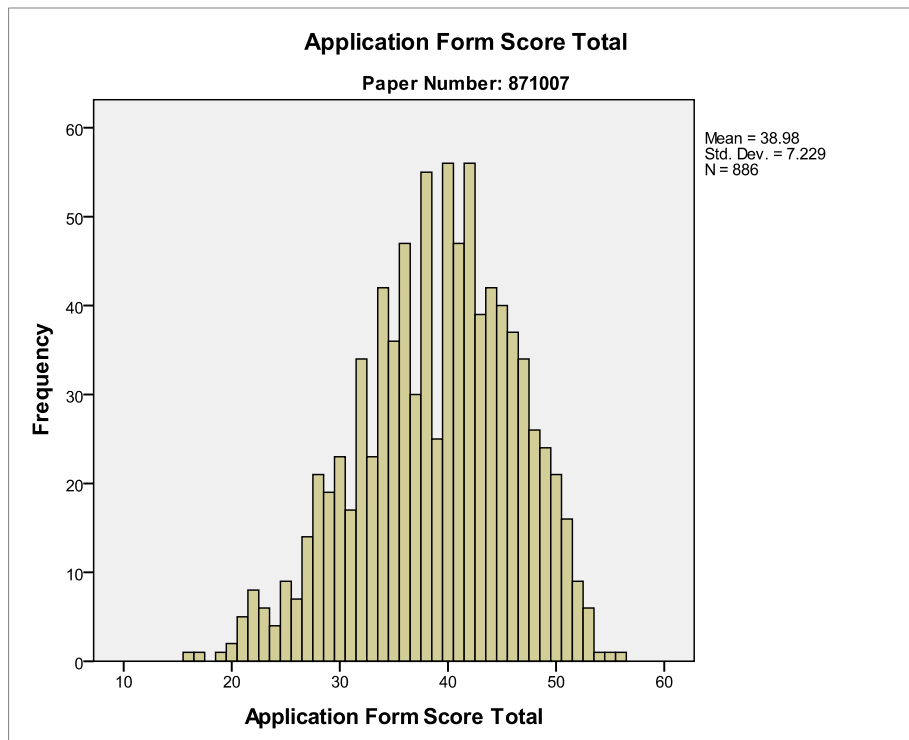


Figure 8: Histogram for Application Form Total for Paper Seven



3.5 Overall Summary of Results

3.5.1 Table 15 provides test completion data for all seven papers. Overall, 3.2% participants (216) did not finish the test, categorised by not completing item 30, 0.8% of participants missed more than 4 items and 96% of participants completed all 30 items within the paper. Paper Two had the lowest completion rate, with 94.2% of participants completing all items within the paper, and Paper Seven had the highest completion rate with 97.4% of participants completing all items within the paper. These results are comparable with previous pilots (97.2% completion rate in 2011 pilot) and confirms that the SJT is a power test, rather than a speeded test. This indicates that 120 minutes is an appropriate length of time to complete 60 questions.

Table 15: Test completion data

	Did not complete item 30		Completed fewer than 26 items		Completed all items	
	N	Percentage	N	Percentage	N	Percentage
Overall	216	3.2%	75	0.8%	6566	96.0%
Paper One	53	4.5%	17	1.3%	1125	94.7%
Paper Two	37	4.2%	14	1.4%	830	94.2%
Paper Three	18	2.1%	6	0.5%	826	96.8%
Paper Four	36	3.0%	12	0.5%	1139	96.3%
Paper Five	24	2.7%	9	0.6%	860	96.7%
Paper Six	27	3.3%	9	1.1%	787	95.7%
Paper Seven	21	2.0%	8	0.8%	999	97.4%

3.5.2 After initial review of the results, some participants were removed from the analysis for either high number of missing items or for erratic scoring patterns (e.g. tied ranks, only ranking best and worst)⁵.

- For Paper One, 11 participants were removed due to a large number of missing data and one very low scorer was removed
- Paper Two, 14 participants were removed from the analysis due to a large amount of missing data
- Paper Three, 4 participants were removed due to a large amount of missing data and 2 participants due to erratic answer patterns
- Paper Four, 3 participants were removed due to a large amount of missing data and 3 participants due to erratic answer patterns
- Paper Five, 5 participants were removed due to a large amount of missing data and 4 participants due to erratic answer patterns

⁵ These participants still received feedback on their performance.

- Paper Six, 6 participants were removed due to a large amount of missing data and 2 participants due to erratic answer patterns
- Paper Seven, 3 participants were removed due to a large amount of missing data and 2 participants due to erratic answer patterns

3.5.3 The sample figures we have therefore presented aim to give the most useful view of the pilot results by excluding only those participants with unusually high levels of missing data and/or erratic response patterns. In an operational test, analysis would be completed on the entire sample, and figures reported as such.

3.6 Test Level Analysis

- 3.6.1 Test level analysis was carried out for all seven papers separately. The figures below illustrate the test level descriptives along with further detail in Table 16. Data are not provided for all seven tests combined as the tests have not been equated and the data would therefore be meaningless. It is not possible to equate the tests until following piloting of all items when all psychometric properties of the items are known.
- 3.6.2 It is important to note that the papers undertaken as part of the PRE contained 30 items and therefore are only half the length of the full operational test. Based on this, test level analysis results should be interpreted with caution. Corrections have been made to the data to estimate based on a 60 item test using the Spearman Brown Formula, but this is not possible for all analyses.
- 3.6.3 The estimated internal reliability for a 60 item test (including those with poor psychometric properties) is provided in Table 16 below⁶. This is lower than may be expected, however this may be due to the composition of the items within the test. More detail is provided in section 4.9.11, however a relatively large number of previously poorly performing items that had been reviewed were included in the PRE. It therefore may be expected that a reasonable proportion of these items would still be classified as poor, and as such this will impact upon the reliability of the test.
- 3.6.4 One of the aims of the PRE was to pilot items to establish whether they have sufficient psychometric properties to enter the item bank. Only those items with sufficient psychometric properties⁷ and that have undergone final review based on all the available information will be used in a live test. Therefore the internal reliability for only those items which have sufficient psychometric properties with regards to item quality is also presented in column four⁸. The reliability for all seven papers is $\alpha=0.80$ and above; sufficient for the use of an operational SJT, and in one case (Paper 3) is $\alpha=0.87$.
- 3.6.5 The mean scores for the seven papers are similar and range from 399.2 to 413.1. The mean scores represent between 78.0% and 80.6% (maximum possible score of 512); this is

⁶ Reliability for a 30 item test is not provided as this would not provide useful information

⁷ Sufficient psychometric properties primarily refers to item quality established using item partials, but also takes into consideration item facility

⁸ Corrected using Spearman Brown formula to provide an estimation of the reliability of a 60 item test with similar quality of items

comparable with the mean score from the spring pilot (81.5%). The standard deviations range between 17.3 and 20.0. The standard deviation indicates how much variation there is from the mean. A low standard deviation indicates that the data points tend to be very close to the mean, whereas a higher standard deviation indicates that the data are spread out over a large range of values. As would be expected with a shorter test, the mean SD (18.6) is lower than in the previous spring pilot (mean SD=34.3).

- 3.6.6 The kurtosis figures are provided for each of the tests; the higher the kurtosis figure the more the variance is the result of infrequent extreme deviations (e.g. outliers), as opposed to frequent modestly sized deviations. All kurtosis figures are relatively low, with paper three having the highest kurtosis figure, indicating that more of the variance in this paper is due to infrequent extreme deviations.
- 3.6.7 Overall the distribution of scores is 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, results show a close to normal distribution and indicate that the SJT is capable of differentiating between participants.
- 3.6.8 In terms of scoring distribution, the scores range from 305 to 468 (a range of 163 scores) from 512 marks. Paper 7 has the lowest distribution of the seven papers, and paper 4 has the greatest distribution. The distributions of the seven papers are as expected based on the item number constraints if the PRE. As expected for a shorter test with a lower available maximum score, this is a smaller distribution than has been achieved in previous pilots (minimum 658 and maximum 922 from 1040 marks for the main pilot, minimum 682 and maximum 952 from 1116 marks for the initial pilot). For an operational test, once all the psychometric properties are known for a test, items with a range of difficulty will be used to assist with the distribution of scores.

Table 16: Test level descriptives by paper

	N	Reliability (α) ⁹	Reliability (α) ¹⁰	Mean	Mean %	Skew	Kurtosis	SD	Min	Max
Paper One	1176	0.69	0.84	399.2	78.0%	-0.63	0.85	20.0	317	452
Paper Two	867	0.65	0.85	399.5	78.0%	-0.44	0.23	18.5	322	444
Paper Three	847	0.71	0.87	414.3	80.1%	-1.14	2.79	18.9	305	454
Paper Four	1177	0.63	0.82	409.6	80.0%	-0.68	1.56	18.0	312	468
Paper Five	880	0.72	0.80	413.1	80.6%	-0.64	1.35	19.3	316	468
Paper Six	814	0.66	0.80	411.7	80.4%	-0.62	1.28	17.3	326.5	461
Paper Seven	1021	0.63	0.80	401.6	78.4%	-0.44	0.39	17.9	334	450

⁹ This is based on a 60 item test, including poorly performing items

¹⁰ Corrected using Spearman Brown formula for those items that were psychometrically robust

Figure 9: Distribution statistics for Paper One

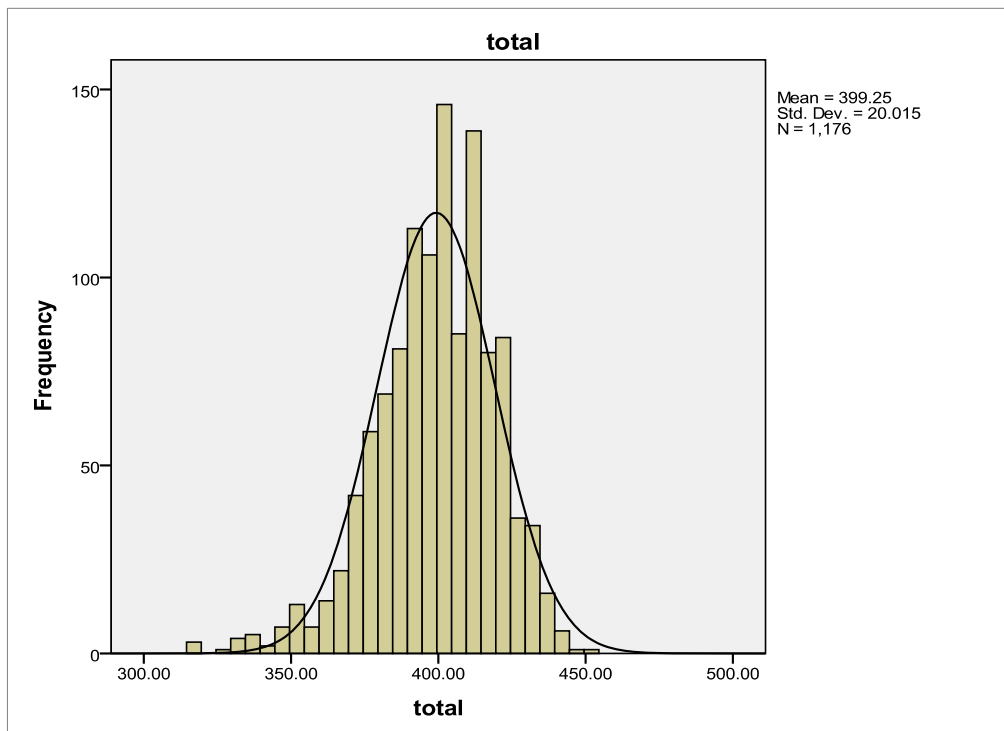


Figure 10: Distribution statistics for Paper Two

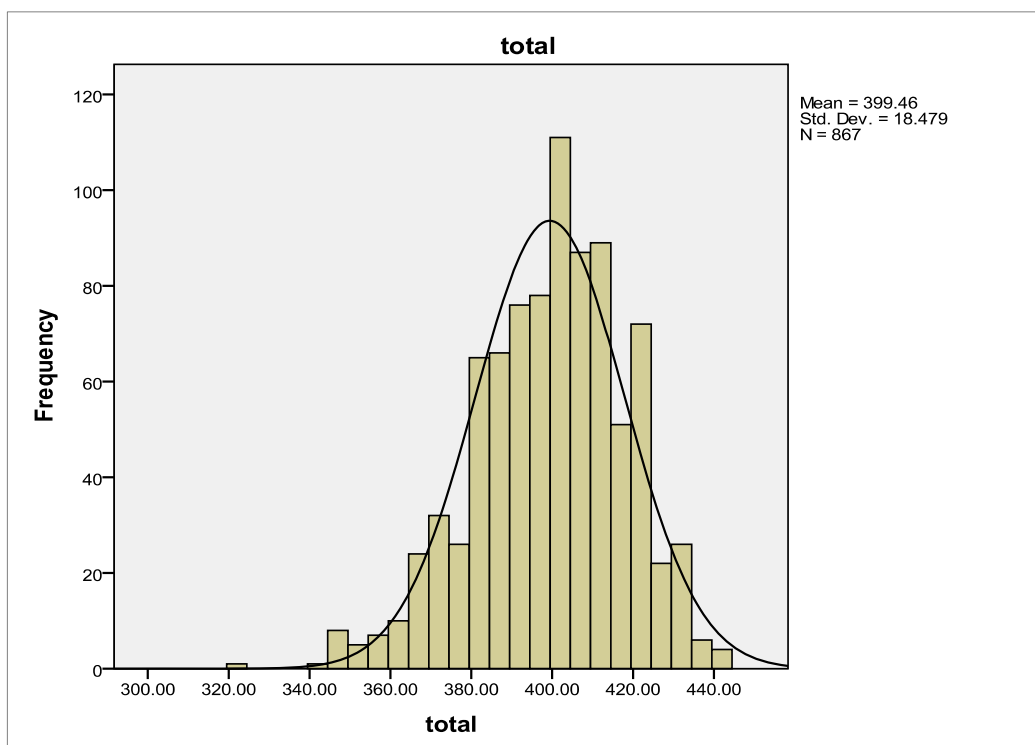


Figure 11: Distribution statistics for Paper Three

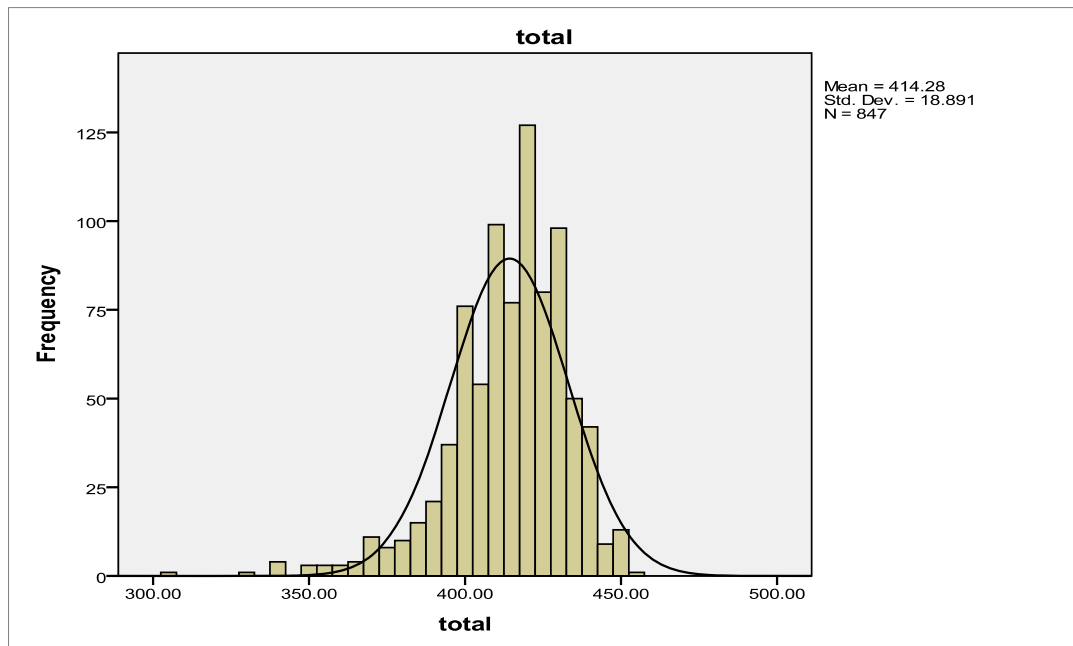


Figure 12: Distribution statistics for Paper Four

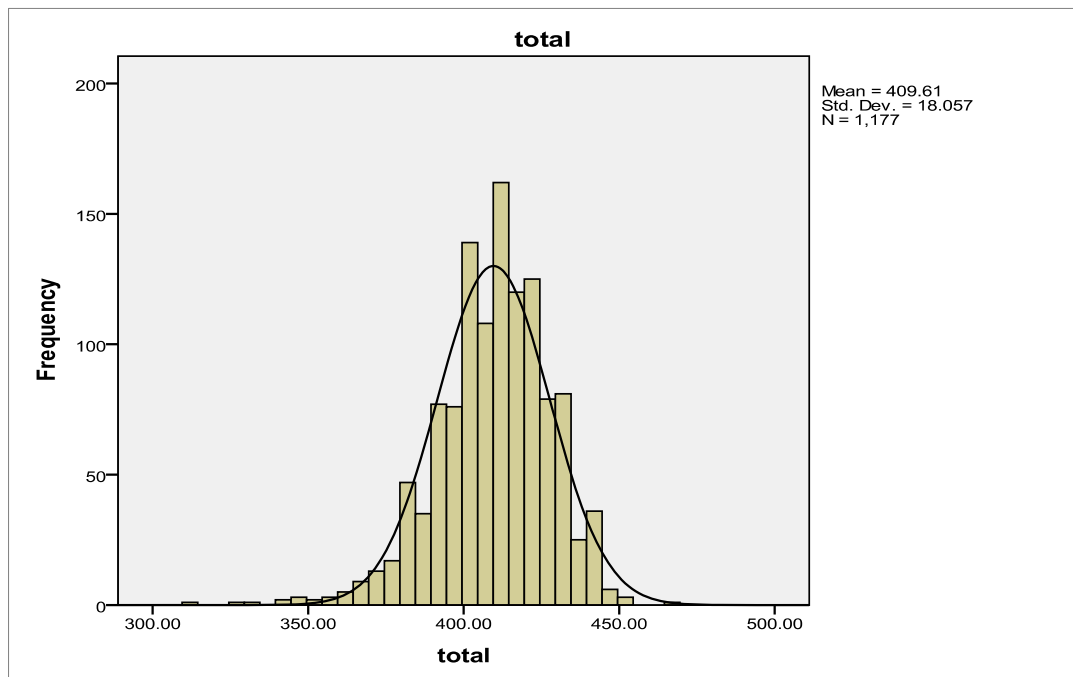


Figure 13: Distribution statistics for Paper Five

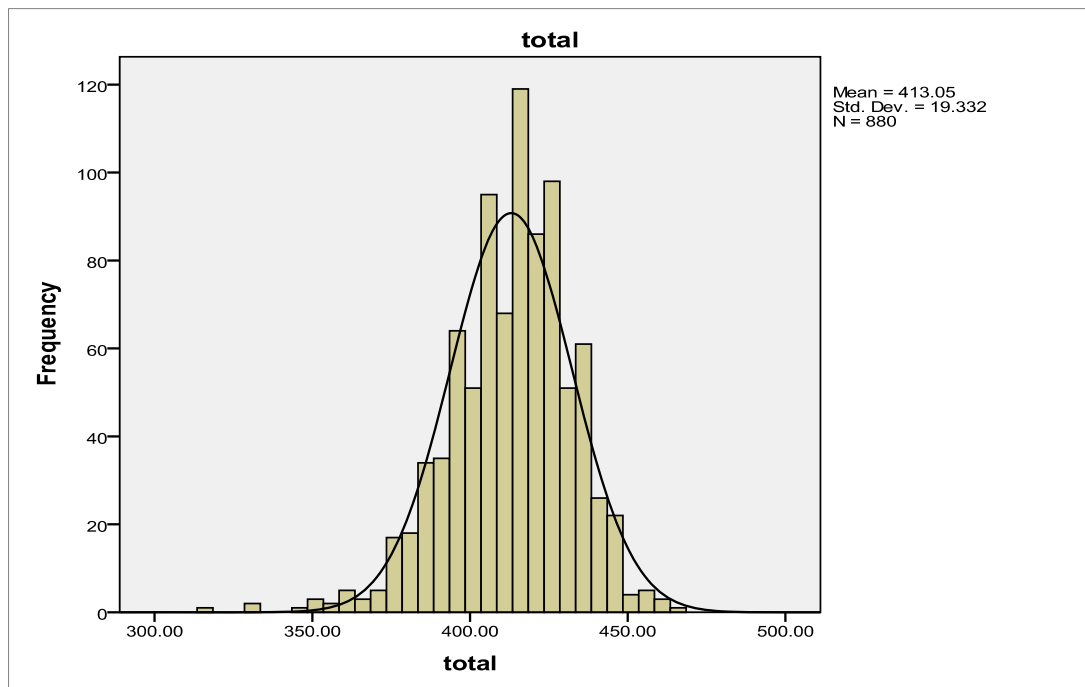


Figure 14: Distribution statistics for Paper Six

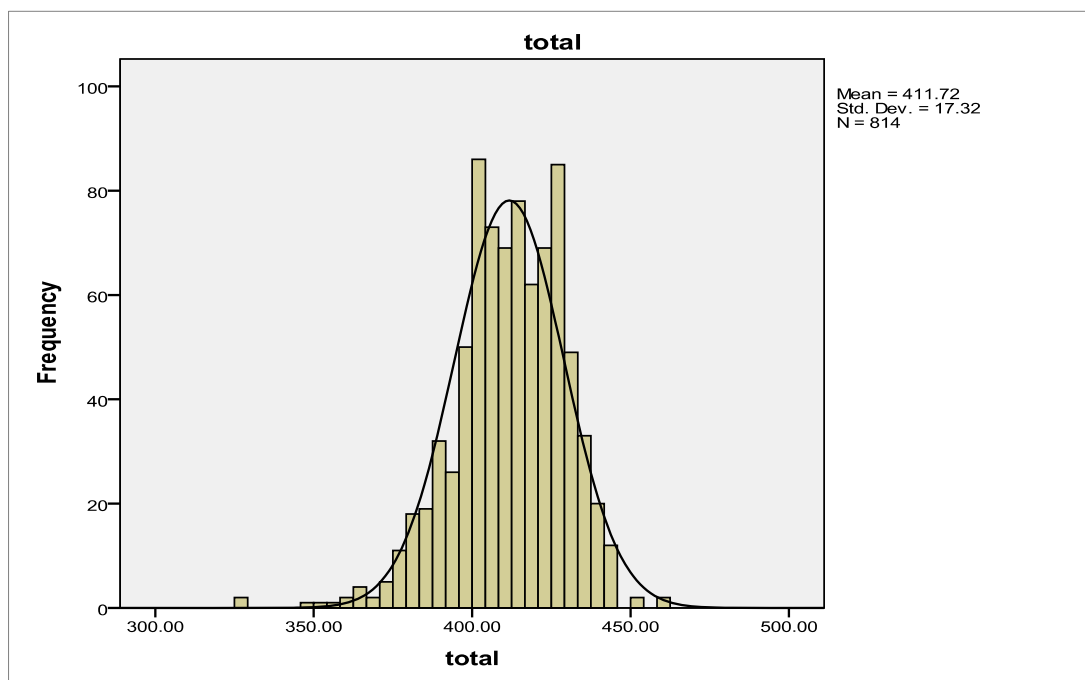
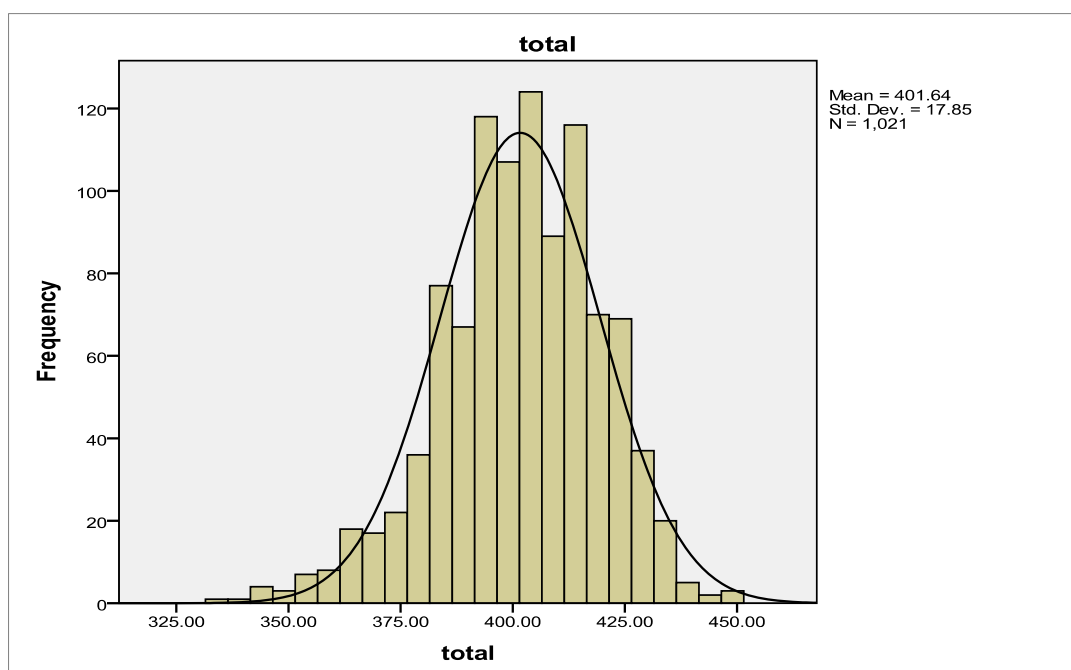


Figure 15: Distribution statistics for Paper Seven

3.9 Item Level Analysis

- 3.9.1 Item analysis was used to look at the difficulty and quality of individual SJT items. Together these can help identify how well the items differentiate between participants and the results are used to further refine the items and scoring keys.
- 3.9.2 Item facility (difficulty) is shown by the mean score for each item (out of a maximum of 20 for ranking items and 12 for multiple response items). If the facility value is very low, then the item may be too difficult and may not yield useful information. This may also indicate that the scoring key for these particular items may need to be examined again, as there may be little consensus between participants and the concordance panel. If the facility value is very high, then the item may be too easy and may not provide useful information or differentiate between participants. A range of item facilities is needed for an operational test, with few very easy (characterised by a mean score of greater than 90% of the total available score) or very difficult (characterised by a mean score of less than 30% of the total available score) items. However, this specification depends on the purpose of the test; if one was only selecting the very best performers, then more difficult items would be needed. However, in this context where participants are being ranked, it is not necessary to have many very difficult items. Within this test, 'very easy' equates to a score of 18 for ranking and 10.8 for multiple choice, and 'very difficult' equates to 11.6 for a ranking and 3.6 for a multiple choice.
- 3.9.3 The standard deviation of an item should also be considered. If an item's standard deviation is very small, it is likely to not be differentiating between participants. The standard deviation for an item should be at least 1.0. If the standard deviation 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.

3.9.4 Table 17 outlines the item level statistics for all seven papers. For the ranking items, the mean facility value was very similar across all papers. The range of facility values for 6 of the 7 papers is broadly similar. Paper three had a very small range (16.2-17.7) indicating that there were no particularly 'difficult' ranking items. The standard deviation range is broadly similar for all papers; with the exception of paper one. One ranking item within this test has a very high standard deviation (coupled with a relatively low mean and a low item partial), suggesting this is a poor item.

Table 17: Item level statistics

	N	Ranking			Multiple Choice		
		Mean	Facility Range	SD Range	Mean	Facility Range	SD Range
Paper One	1176	16.6	14.6-17.7	1.9-4.1	8.4	7.6-9.5	2.2-3.1
Paper Two	867	16.5	14.6-18.2	1.7-3.0	7.9	6.5-8.6	2.4-3.0
Paper Three	847	16.8	16.2-17.7	1.9-2.7	8.6	6.8-10.4	1.9-2.9
Paper Four	1177	16.9	14.8-18.2	1.7-3.7	8.0	6.3-8.7	2.1-2.9
Paper Five	880	16.6	14.8-18.7	1.8-3.1	9.0	6.3-10.0	2.2-2.7
Paper Six	814	16.5	13.4-18.3	1.6-2.9	9.0	8.0-10.1	2.2-2.7
Paper Seven	1021	16.4	13.7-18.7	1.7-3.1	8.2	7.2-10.3	2.2-3.2

3.9.5 For the multiple choice items, the mean facility value was similar across all items, with Paper two having the lowest mean facility value (7.9) and Paper five and six having the highest mean facility value (9.0). The range of facility values differ across the papers.

3.9.6 When constructing an operational test where the psychometric properties of all items are known, items with a range of facility values will be used.

3.9.7 Item quality was determined by the correlation of the item with overall SJT score, not including the item itself (item partial)¹¹. This analysis compares how individuals perform on a given item with how they perform on the test overall. You would expect that high scoring participants overall would select the correct answer for each item more often than low scoring participants, i.e. the item discriminates between good and poor participants. This would show a good to moderate correlation/partial. A poor correlation would indicate that performance on the individual item does not reflect performance on the test as a whole. Table 18 below outlines how items performed for each of the seven papers and overall.

3.9.8 Although the item partial provides vital information in terms of how well as item is performing, this needs to be taken into consideration with a number of other statistics

¹¹ 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.

(item facility, SD) and information (distracter analysis using available data). It is also recommended that item partials are balanced with other considerations, e.g. need to provide coverage of all target domains.

Table 18: Item level statistics

	Overall	Paper One	Paper Two	Paper Three	Paper Four	Paper Five	Paper Six	Paper Seven
Range of Item Partial	-.03-.36	.11 - .27	.03 - .30	-.01 - .32	-.01 - .31	.04-.30	-.03-.34	.05 - .36
Mean Item Partial	0.17	0.18	0.16	0.18	0.17	0.18	0.17	0.17
Good (>0.17)	111 (53%)	12 (40%)	15 (50%)	18 (60%)	18 (60%)	17 (57%)	17 (57%)	14 (47%)
Moderate (0.13-0.17)	42 (25%)	7 (23%)	5 (17%)	8 (27%)	5 (17%)	9 (30%)	3 (10%)	5 (17%)
Item requires further review (<0.13)	57 (27%)	11 (37%)	10 (33%)	4 (13%)	7 (23%)	4 (13%)	10 (33%)	11 (37%)

3.9.9 Papers one, two and seven all have the fewest items with partials above .17. Paper six also has a relatively large proportion of items below .13.

3.9.10 111 of the 210 (53%) items are deemed as having good psychometric properties with regards to item quality and it is likely that the majority of these will enter the item bank, after final review and following consideration of the other available data (e.g. item facility, SD). 42 (25%) of the items are deemed as moderate. These items will be reviewed further, and where deemed appropriate, taking into consideration all the information about the item, some of these items may enter the item bank in their current form. . Those 'moderate' items that are not deemed suitable to enter the item bank in their current form, and all items requiring further review will undergo further review (please see section 6 for further detail).

3.9.11 57 of the 210 (27%) of the items require further review and 42 (25%) are deemed as having moderate psychometric properties. However this may be in part due to the origin of some of the items. Out of the 210 piloted items, 56 of the items (27%) that were piloted were already known to have moderate or poor psychometric properties prior to any reviews/changes to the items:

- 27 of the items in the pilot are items that previously had item partials of .13 or below. These items were reviewed and went to a concordance stage.
- 13 of the items in the pilot are items that previously had item partials of .17 or below. The items were reviewed and the majority went to a concordance stage.
- 16 items that were not deemed sufficient to go to the spring pilot following concordance were also reviewed and used in the PRE.

3.9.12 Although these were all reviewed and felt to be sufficient to be piloted following alterations, it was likely that a proportion of these items would still not have sufficient psychometric properties.

3.9.13 Items deemed as 'moderate' or requiring 'further review' were further broken down in terms of origin. This indicates that there is no real difference between 'new' items and items previously known to have poor or moderate psychometric properties.

- Out of the 99 items that were deemed as 'moderate' or requiring 'further review'; 27 of the items (27%) were known to previously have poor or moderate psychometric properties and 72 (73%) of the items were new items,
- Thus; 27 of the 56 items (48%) that previously had moderate or poor psychometric properties were classified as needing further review or as moderate.
- 72 of the 154 (47%) 'new' items were classified as needing further review or as moderate.

3.10 Group Differences

3.10.1 In order to examine fairness issues regarding the use of a SJT for selection into the FP, group differences in performance within the participant sample were analysed on the basis of sex, ethnicity and age for each of the seven papers. Group differences for all seven papers overall is not provided as the three tests had not been equated and therefore the data would not be meaningful.

3.10.2 Table 19 shows group differences in performance on the SJT based on sex. For all seven papers female participants scored slightly higher than male participants. This effect was greatest for Paper One and weakest for Paper Five and Paper Six. However, based on t-test results in combination with analysis of effect size using Cohen's D (>30) it is determined that the differences in the mean SJT scores for males and females were not significant indicating that performance on the SJT does not appear to be influenced by sex differences.

Table 19: Group differences by sex

	Sex	N	Mean	SD	Sig Difference
Paper One	Male	451	394.55	24.13	Ns
	Female	668	400.84	22.09	
Paper Two	Male	344	395.51	23.09	Ns
	Female	479	400.27	19.92	
Paper Three	Male	330	413.99	17.53	Ns
	Female	425	417.06	16.34	
Paper Four	Male	452	407.03	18.97	Ns
	Female	650	411.21	21.47	

Paper Five	Male	354	411.61	19.04	Ns
	Female	487	413.95	20.87	
Paper Six	Male	318	408.63	19.57	Ns
	Female	458	411.74	24.84	
Paper Seven	Male	405	399.66	18.50	Ns
	Female	550	402.85	18.31	

3.10.3 Table 20 shows group differences in performance on the SJT based on ethnicity by White and Black and Minority Ethnic (BME) groups. White participants scored higher than BME participants on all seven tests and t-tests showed that this difference was statistically significant ($p < 0.01$) for all seven papers. The effect was greatest for Paper Five, followed by Paper One, Paper Three and Paper Six and weakest for Paper Seven. A richer understanding of the implications of the observed groups differences in practice (for sex and ethnicity) is needed as an impetus for future research. Without detailed systematic research in this area, causal factors cannot be reliably identified.

Table 20: Group differences by ethnicity

	Ethnicity	N	Mean	SD	T-test Sig.	Cohen's d
Paper One	White	626	403.80	19.3	$p < 0.01$	0.53
	BME	461	392.07	25.18		
Paper Two	White	491	401.73	18.69	$p < 0.01$	0.40
	BME	311	393.37	23.58		
Paper Three	White	521	418.17	15.45	$p < 0.01$	0.50
	BME	217	410.08	19.04		
Paper Four	White	814	411.84	16.88	$p < 0.01$	0.40
	BME	262	402.98	28.29		
Paper Five	White	600	416.99	17.67	$p < 0.01$	0.71
	BME	228	402.67	22.38		
Paper Six	White	385	415.78	15.19	$p < 0.01$	0.51
	BME	366	404.81	27.99		
Paper Seven	White	718	403.09	17.43	$p < 0.01$	0.36
	BME	221	396.23	20.78		

- 3.10.4 In terms of age, there was a positive correlation ($r=-.08$ - Spearman's rho, $p<0.05$) between age and scores on the SJT for Paper Three; that is older participants scored significantly better on the SJT than younger participants. There were no significant correlations for the remaining six papers.
- 3.10.5 Differential Item Functioning (DIF) was examined at an item level. The DIF analysis is a procedure used to determine if test items are fair and appropriate for assessing the ability of various demographic groups. It is based on the assumption that test takers who have similar ability (based on total test scores) should perform in similar ways on individual test items regardless of their sex or ethnicity. DIF is a necessary but not sufficient condition for bias: bias only exists if the difference is illegitimate, i.e., if both groups should be performing equally well on the item. An item may show DIF but not be biased if the difference is due to actual differences in the groups' ability to answer the item, e.g. if one group is high proficiency and the other low proficiency, the low proficiency group would necessarily score much lower.
- 3.10.6 DIF, undertaken using a multiple regression analysis, was used to examine whether the demographic variable (e.g. sex) significantly predicts performance on each item once overall test performance has been controlled for (i.e. is there a difference in item performance beyond that which expected due to differences between groups on the test overall?).
- 3.10.7 61 (29%) items were flagged for sex differences (Males performed better on 30 items and females on 31 items). 8 of the items were in Paper 1, 13 of the items were in Paper 2, 6 of the items were in Paper 3, 8 of the items were in Paper 4, 8 of the items were in Paper 5, 7 of the items were in Paper 6 and 11 of the items were in Paper 7.
- 3.10.8 57 (27%) items were flagged for ethnicity differences (White applicants performed better on 31 items and Black and Minority Ethnic applicants on 26). 12 of the items were in Paper 1, 6 of the items were in Paper 2, 6 of the items were in Paper 3, 8 of the items were in Paper 4, 8 of the items were in Paper 5, 7 of the items were in Paper 6 and 10 of the items were in Paper 7. As items on which males and females performed better, and items on which white and BME applicants performed better are present in equal proportions, this strongly suggests that the test is not biased.
- 3.10.9 These items will be reviewed in light of these results following the pilot to identify whether there appears to be any bias in the item content. Once reviewed, if the items do appear to demonstrate bias (as outlined above, DIF is a necessary but not sufficient condition for bias), items will either be adjusted and re-piloted or will be removed from the item bank.

3.11 Correlations with Quartiles and Application Form

- 3.11.1 Correlations were run between SJT total scores and current FP selection methods. The selection method used for FP 2012 includes quartiles and a competency based application form. Quartiles are calculated using examination scores from preceding years at medical school. Students are awarded one of four quartile scores (34, 36, 38, 40). The application form consists of 6 questions. Question 1 covers career history and qualifications. Questions 2 to 6 are 'white space' competency based questions, scored by clinicians against an agreed national scoring criteria, based on the Foundation doctor person specification.

3.11.2 At the $p < 0.01$ level significant correlations were found between SJT scores and quartile scores (Spearman rho) for all seven papers and between SJT scores and the application form (Pearson r) for five of the seven papers. Although these correlations are significant, indicating some shared variance/commonality between the assessment methods, there is also a large amount of variance not explained, therefore the SJT appears to be assessing somewhat different constructs to the other methods.

Table 21: Correlations between SJT total scores and current selection methods

	Current selection methods	SJT total scores
Paper One	Quartiles ¹²	.23**
	Application form ¹³	.063*
Paper Two	Quartiles	.24**
	Application form	Ns
Paper Three	Quartiles	.18**
	Application form	.11**
Paper Four	Quartiles	.16**
	Application form	Ns
Paper Five	Quartiles	.25**
	Application form	.15**
Paper Six	Quartiles	.31**
	Application form	.17**
Paper Seven	Quartiles	.18**
	Application form	.12**

* Significant at the $p < 0.05$ level

** Significant at the $p < 0.01$ level

3.12 Participant Evaluation

3.12.1 All participants who participated in the PRE were asked to complete an evaluation questionnaire regarding their experience and perceptions of the SJT. A total of 6788 (99.2%) participants completed the questionnaire.

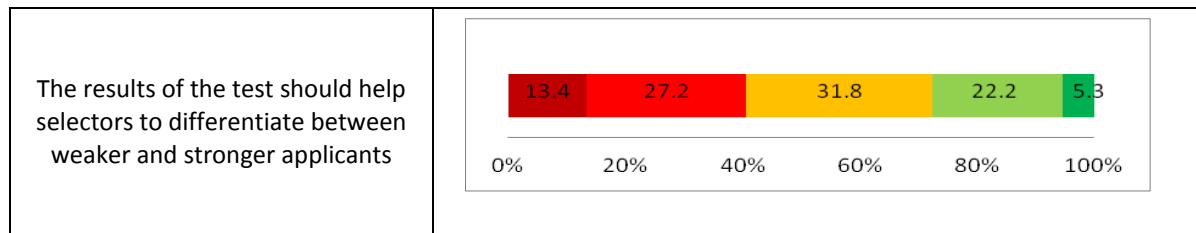
3.12.2 Participants were asked to indicate their level of agreement with several statements regarding the content of the SJT paper, the results of which are shown in the Table 22.

¹² All correlations between SJT scores and quartiles use Spearman Rho's correlation coefficient

¹³ All correlations between SJT scores and application form use Pearson r's correlation coefficient

Table 22: Participant evaluations

% participants (N=6788)														
Strongly Disagree %	Disagree %	Neither %	Agree %	Strongly Agree %										
The information I received in advance of the SJT was clear and helpful		<table border="1"> <tr> <td>Strongly Disagree</td> <td>Disagree</td> <td>Neither</td> <td>Agree</td> <td>Strongly Agree</td> </tr> <tr> <td>3.1</td> <td>18.6</td> <td>23.4</td> <td>42.6</td> <td>22.0</td> </tr> </table>			Strongly Disagree	Disagree	Neither	Agree	Strongly Agree	3.1	18.6	23.4	42.6	22.0
Strongly Disagree	Disagree	Neither	Agree	Strongly Agree										
3.1	18.6	23.4	42.6	22.0										
The instructions for the test were clear and easy to understand		<table border="1"> <tr> <td>Strongly Disagree</td> <td>Disagree</td> <td>Neither</td> <td>Agree</td> <td>Strongly Agree</td> </tr> <tr> <td>1.3</td> <td>3.7</td> <td>10.7</td> <td>48.9</td> <td>34.8</td> </tr> </table>			Strongly Disagree	Disagree	Neither	Agree	Strongly Agree	1.3	3.7	10.7	48.9	34.8
Strongly Disagree	Disagree	Neither	Agree	Strongly Agree										
1.3	3.7	10.7	48.9	34.8										
The test seemed well-run and well-invigilated		<table border="1"> <tr> <td>Strongly Disagree</td> <td>Disagree</td> <td>Neither</td> <td>Agree</td> <td>Strongly Agree</td> </tr> <tr> <td>2.3</td> <td>3.1</td> <td>19.4</td> <td>46.7</td> <td>38.4</td> </tr> </table>			Strongly Disagree	Disagree	Neither	Agree	Strongly Agree	2.3	3.1	19.4	46.7	38.4
Strongly Disagree	Disagree	Neither	Agree	Strongly Agree										
2.3	3.1	19.4	46.7	38.4										
The content of the test seemed relevant to the Foundation Programme		<table border="1"> <tr> <td>Strongly Disagree</td> <td>Disagree</td> <td>Neither</td> <td>Agree</td> <td>Strongly Agree</td> </tr> <tr> <td>2.2</td> <td>2.4</td> <td>14.1</td> <td>50.4</td> <td>28.8</td> </tr> </table>			Strongly Disagree	Disagree	Neither	Agree	Strongly Agree	2.2	2.4	14.1	50.4	28.8
Strongly Disagree	Disagree	Neither	Agree	Strongly Agree										
2.2	2.4	14.1	50.4	28.8										
The scenario content seemed appropriate for my training level		<table border="1"> <tr> <td>Strongly Disagree</td> <td>Disagree</td> <td>Neither</td> <td>Agree</td> <td>Strongly Agree</td> </tr> <tr> <td>1.8</td> <td>3.0</td> <td>15.9</td> <td>53.6</td> <td>23.7</td> </tr> </table>			Strongly Disagree	Disagree	Neither	Agree	Strongly Agree	1.8	3.0	15.9	53.6	23.7
Strongly Disagree	Disagree	Neither	Agree	Strongly Agree										
1.8	3.0	15.9	53.6	23.7										
The level of difficulty of the test was appropriate		<table border="1"> <tr> <td>Strongly Disagree</td> <td>Disagree</td> <td>Neither</td> <td>Agree</td> <td>Strongly Agree</td> </tr> <tr> <td>1.9</td> <td>7.5</td> <td>24.1</td> <td>51.3</td> <td>15.1</td> </tr> </table>			Strongly Disagree	Disagree	Neither	Agree	Strongly Agree	1.9	7.5	24.1	51.3	15.1
Strongly Disagree	Disagree	Neither	Agree	Strongly Agree										
1.9	7.5	24.1	51.3	15.1										
The content of the test appeared to be fair		<table border="1"> <tr> <td>Strongly Disagree</td> <td>Disagree</td> <td>Neither</td> <td>Agree</td> <td>Strongly Agree</td> </tr> <tr> <td>8.4</td> <td>18.5</td> <td>31.4</td> <td>32.8</td> <td>8.8</td> </tr> </table>			Strongly Disagree	Disagree	Neither	Agree	Strongly Agree	8.4	18.5	31.4	32.8	8.8
Strongly Disagree	Disagree	Neither	Agree	Strongly Agree										
8.4	18.5	31.4	32.8	8.8										



3.12.3 84% of the participants felt that the test instructions were clear and easy to understand, while 65% thought that the information given about the pilot was clear and helpful. 79% of participants also agreed or strongly agreed that the content of the SJT seemed relevant to the FP. 77% felt that the scenario content was appropriate for their level of training and 66% considered that the difficulty level was appropriate. If they felt that the level of difficulty was not appropriate, participants were asked to indicate whether they felt that the test was too hard or too easy. 695 (19.4%) participants responded; 135 participants felt that the test was too easy and 560 felt that it was too difficult. Overall, 41.6% of participants agreed or strongly agreed that the content of the test was fair, with 31.4% neither agreeing nor disagreeing with this statement. When considering whether the results of the test would help differentiate between the strong and weak participants, 27.3% agreed or strongly agreed, whilst 42% neither agreed nor disagreed with this statement.

Part Three: Summary & Recommendations

4 Summary

- 4.1 The Parallel Recruitment Exercise was undertaken for a number of reasons, but for the purpose of this report, principally to maximise the item bank to ensure that there was a sufficient number of items to use in the live selection round in 2013. As such the purpose was not to evaluate the use of the SJT for Selection to the Foundation Programme, although full evaluation of the tests was carried out.
- 4.2 The psychometric analysis presented in this report is evidence that the SJT is a reliable measurement methodology. Test level analysis was consistent with findings from previous reports and was as expected based on a 30 item test; half the length of the operational test. The range for a 30 item test was as expected, and based on the number of items, was able to differentiate sufficiently between applicants.
- 4.3 Item-level analysis showed that a large proportion of the SJT items worked well. A total of 111 (53.0%) of the items were deemed to have sufficient item partials. A more in depth review of these items, including analysis of facility values and DIF analysis will take place and it is expected that a large proportion of these will enter the item bank, ready to be used in live selection.
- 4.4 The relationships between the SJT score and the application form and quartile score were examined. This analysis showed that generally those that scored higher on the SJT also achieved higher marks in the application form and the quartile score. However, the correlation coefficient indicates that they are testing different constructs.

5 Item Writing Methodology

- 5.1 Two methods of item writing methodology were trialled during the development phase of the PRE; Item development Interviews and Item Writing Workshops. An outline of the two methodologies and the number of items developed using each of these methodologies is outlined in Section 3.
- 5.2 In summary, it is the authors' recommendation that the item development interviews allow for a sustainable and efficient process. It is also proposed to evaluate the item writing workshops approach further which will have other benefits including stakeholder buy-in and developing expertise in item writing amongst the academic audience.
- 5.3 Considerations for item writing methodologies include:
- Scope for scenario diversity
 - Motivation of item writer
 - Scope for involvement with or without ongoing commitment
 - Engagement with clinicians
 - Direct and indirect clinician costs
 - Scope to increase item quality over time
 - Security risk
 - Generation and sharing of knowledge of best practice in SJT item writing principles
 - Efficiency and cost efficiency in terms of number of items generated per 'day'
- 5.4 We recommend continued use of review workshops to allow for the detailed group review of items resulting in enhanced item quality. The review workshops also allow input from FY2; thus negating the need for separate focus groups.
- 5.5 In any development process, considerations around representation of those involved and how this may impact on potential discrimination need to be addressed. Recommendations include:
- Monitor ethnic representation of all involved in the development process (interviewees, reviewers, concordance panel)
 - Undertake targeted approach for each of the development phases with regard to under-represented groups
 - Continue to monitor group differences (sex, ethnicity, age)
 - Continue to undertake Differential Item Functioning (DIF) analysis to identify potential bias in items

6 Clinician review

- 6.1 Following the large scale pilot, all items were reviewed by an expert in SJT development and clinicians, also experienced in SJT development; this is termed the clinician review. This review consists of two aspects;
- A review of the items to ensure that the item looks relevant and unambiguous. As all items have gone through a review and concordance stage, it is not expected at this stage that any major concerns will be raised. However it is important to remember that the pilot is still part of the development process items and as such some updates to items may be made at this stage
 - The second, and the predominant stage of the clinician review, is using the statistics derived from the pilot to help interpret and improve poorly performing items. This in-depth review involves the experts examining all item statistics, including graphs) in an attempt to identify where changes can be made to help the item differentiate between applicants.
- 6.2 Those items from the spring 2011 pilot with insufficient psychometric properties were the main focus of this review, although a number of items that were deemed as having 'sufficient' psychometric properties were also reviewed and some updates made. As a result of this review, the content of some of items was altered and/or the scoring key reviewed and as such these items were re-piloted as part of the PRE. 89 items piloted in the PRE were items from this pool.
- 6.3 In some cases the changes made at the clinician review resulted in the item not performing as well psychometrically in the PRE as it had done previously in earlier pilots. In these cases, a review will be made of the item to see why this may be the case, why alterations were made (e.g. because it was ambiguous, or because changes were made in an attempt to improve the item psychometrically) and whether the item should be returned to its original form and entered into the item bank, or rejected.

7 Item bank

- 7.1 Following the PRE, there will be more than a sufficient number of items in the item bank for live selection in 2013. The equating strategy is yet to be finalised but it is anticipated that roughly 100 items will be required in each year, on the current assumptions of two national dates (plus a third date for those with extenuating circumstances) and 60 live items in a test.
- 7.2 Currently, not including items from the PRE, the item bank consists of 79 items (27 from the initial pilot and 52 from the main pilot). Although 104 items from the main pilot were deemed as having sufficient item partials, this included repeat anchor and feedback items. In 31 cases, items from the main pilot that had sufficient item partials were reviewed and small changes made based on the item statistics. As such these were either re-concorded and piloted, or just re-piloted.
- 7.3 From the PRE, a total of 111 items were deemed as having 'good' item partials. A more in depth review of these items, including analysis of facility values and DIF analysis will take place and it is expected that a large proportion of these will enter the item bank, ready to be used in live selection.