Work Psychology Group
Thinking differently

Analysis of the Situational Judgement Test for Selection to the Foundation Programme 2017

Technical Report

Prof Fiona Patterson
Amy Aitkenhead
Rachael Shaw
Vicki Ashworth
Anna Rosselli

April 2017

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 selection to FP 2017 in December 2016 and January 2017, over three administration sessions. The SJT, in combination with the Educational Performance Measure (EPM)[1], was used to rank applicants applying for Foundation Year One (F1) training and allocate them to foundation schools. This is the fourth year during which the SJT has been used operationally.

1.1.2 The SJT must be developed and validated in accordance with accepted best practice, so that it provides an effective, rigorous and legally defensible method of selection. This technical report therefore provides an overview of the results from the operational delivery of the FP 2017 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.

1.2 Background

1.2.1 The Foundation Programme is a two-year generic training programme, which forms the bridge between medical school and specialist/general practice training. An SJT was introduced to the Foundation Programme selection process for entry to the Foundation Programme in 2013. The Foundation Programme SJT assesses five domains from the Foundation Programme person specification: Commitment to Professionalism, Coping with Pressure, Patient Focus, Effective Communication, and Working Effectively as Part of a Team[2].

1.2.2 Following each recruitment cycle, an evaluation of the SJT is undertaken to enable ongoing monitoring of the test's suitability to be used in this context and to identify any potential future recommendations. The evaluation results are outlined in a technical report which is produced each year[3].

[1] 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 publications.

[2] See F1 Job Analysis report 2011 for full details of how domains were derived and what comprises each domain (http://www.isfp.org.uk/ISFPDocuments/Pages/FinalreportofPilots.aspx).

[3] See Patterson, F., Ashworth, V., Murray, H., Empey, L., & Aitkenhead, A. (2013). *Analysis of the Situational Judgement Test for Selection to the Foundation Programme 2013: Technical Report*.

See Patterson, F. Murray, H., Baron, H., Aitkenhead, A., & Flaxman, C. (2014*). Analysis of the Situational Judgement Test for Selection to the Foundation Programme 2014: Technical Report*.

See Patterson, F., Aitkenhead, A., Edwards, H., Flaxman, C., Shaw, R., & Rosselli, A. (2015). *Analysis of the Situational Judgement Test for Selection to the Foundation Programme 2015: Technical Report.*

Part One: Item Development

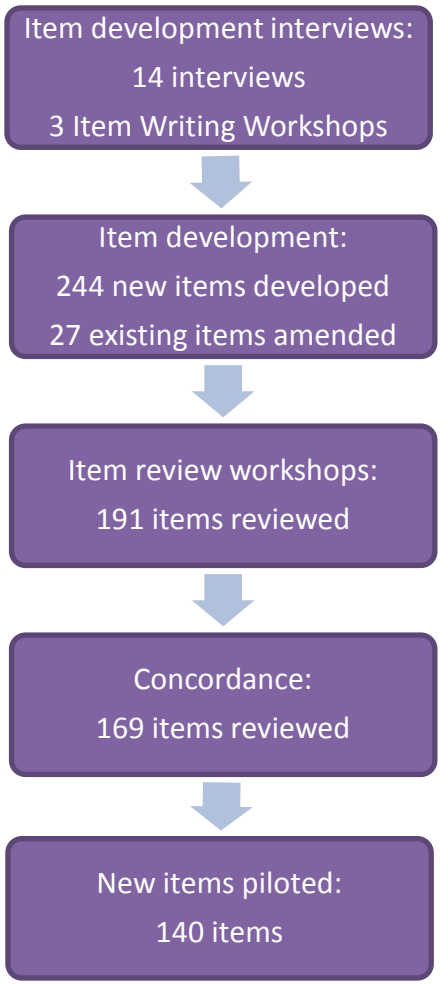
2 Development of Trial Items

2.1 Process Overview

2.1.1 To ensure that there is a sufficient number of items within the 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 FP 2017 operational delivery.

Figure 1: Item development and review process



2.2 Item Development

2.2.1 Prior to item development, a review of the current operational bank (n=322), containing items that were established through the previous trials (between 2010 and 2016), 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.

2.2.2 SJTs are a measurement methodology, and as such there is a variety of different lead-in instruction formats that can be used. There are no specifications with regard to the number of items written within each lead-in format, with the item content leading the decisions relating to lead-in format. The four different ranking lead-in formats (all with five responses to be ranked) that have been used are:

- **Rank Appropriateness of Actions:** '*Rank in order the appropriateness of the following actions in response to this situation (1= Most appropriate; 5= Least appropriate)'*
- **Rank Agreement with Statements:** *'Rank in order the extent to which you agree with the following statements in this situation (1= Most agree with; 5= Least agree with)'*
- **Rank Importance of Considerations:** *'Rank in order the importance of the following considerations in the management of this situation (1= Most important; 5= Least important)'*
- **Rank Order of Actions:** *'Rank the order in which the following tasks should be undertaken (1= Do first; 5= Do last)'*

2.2.3 **Telephone Interviews**

2.2.4 Item Development Interviews (IDIs), using Critical Incident Technique (CIT), were held to develop SJT items. CIT interviews aim to elicit, from Subject Matter Experts (SMEs), scenarios or incidents involving F1 doctors that demonstrate particularly effective or ineffective behaviour and that reflect the SJT target domains. Using CIT interviews has a number of benefits, including involvement of a broad range of individuals from across the country in the design process, without the need for a significant commitment in terms of time and effort. SMEs who had been nominated by Foundation School Leads, in addition to others who had previously been involved, or had expressed an interest in being involved in the process, were invited to take part in the interviews.

2.2.5 In total, 14 interviews were conducted by trained interviewers. Foundation Year Two (F2) doctors were deliberately targeted, as these individuals are closest to the role and therefore are well placed to provide relevant and realistic scenarios.

2.2.6 The telephone interviews lasted between 30 and 45 minutes. During the call, a trained interviewer asked the interviewee to describe a number of scenarios, providing as much information as possible. This included the precursor 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.

2.2.7 **Item Writing Workshops**

2.2.8 Item writing workshops were also held, with an aim for clinicians to develop SJT item content. Prior to the workshop, SMEs were asked to spend some time in preparation for the workshop, thinking of example situations that could be used as a basis for scenario content. During the workshop, SMEs were

introduced to SJT item writing principles and, independently or in pairs, each SME wrote a number of scenarios and responses.

2.2.9 Using item writing workshops has a number of benefits, including: the generation of a large number of items per SME; the opportunity for SMEs to work together and gain ideas from each other for new item content; the ability to tailor the content of items, helping to avoid scenarios that have not worked well in the past or that there are already a large number of within the item bank; and the development of expertise within the SME item writer pool. In FP 2017, the inclusion of item writing workshops broadened the range of SMEs involved in the item development process, and provided greater opportunity for WPG facilitators to support the development of wide-ranging scenario content.

2.2.10 Three item writing workshops took place; one workshop was held in Cambridge, one in Glasgow and one in Birmingham. A total of 18 individuals took part in the workshops, including three F2 doctors. All participants who volunteered to take part were sent briefing material that outlined the purpose of the item writing workshop and their role on the day. All participants also signed a confidentiality and code of conduct agreement.

2.2.11 Table 1 shows the range of job roles held by the SMEs involved in the item development process (telephone interviews and item writing workshops).

Table 1: SMEs' job roles

Job Role	Number
Clinical/Educational Supervisor	18
F2 doctor	6
Director of Medical Education	1
Clinical Teaching Fellow	1
Foundation Training Programme Director	4
Fellow in Medical Education	1

2.2.12 Table 2 shows the range of the primary specialties of SMEs involved in the item development process. SMEs from a breadth of specialties were involved in developing items. The breadth of specialties helps to ensure that the bank of items represents a wide spread of topic areas and medical settings.

Table 2: SMEs' primary clinical specialties (note: some individuals indicated that they had more than one primary clinical specialty)

Specialty	Number
Acute Medicine and Infectious Diseases	1
Care of the Elderly	1
Diabetes and Endocrinology	1
Emergency Medicine	2
General and Colorectal Surgery	1
General Medicine	1
General Practice	1
Haematology	1
Nephrology	2
Obstetrics and Gynaecology	3
Paediatrics	3
Rheumatology	1
Trauma and Orthopaedics	2
Urology	1
Not specialty specific / F2 doctor / Core Medical Trainee	13

2.3 Previous Item Review

2.3.1 Items that had been trialled previously, but had not been deemed suitable to enter the operational item bank at that point, were also reviewed and re-written by trained SJT item writers. Following initial review and re-write, these items then entered the item development process at the SME review workshop stage. Items were selected for potential re-trialling based on the information gathered from previous pilots, i.e. how the items performed. To identify these items, and inform the re-write process, a 'distractor analysis' was carried out. This provides detailed information about how different applicants responded to the individual items. Based on this, and in combination with reviewing the item content, item writers suggested potential refinements to the items. These suggested changes were reviewed by a clinician before the item was deemed suitable to enter the item development process at the review workshop stage alongside the newly developed items. A total of 27 items were identified and subsequently re-written. Amendments included clarifying any possible ambiguity in the scenario or trying to differentiate two responses by making one option 'worse' or 'better'.

2.4 Item Review

2.4.1 All newly developed items were reviewed by the core team of item reviewers from Work Psychology Group (WPG). Each scenario was reviewed in accordance with SJT item writing principles and the design

specification of the F1 SJT. In particular, scenarios were reviewed to ensure that they addressed one of the target domains and were set at an appropriate level for an F1.

2.4.2 Following the introduction of item writing workshops for FP 2017, an increased level of item redundancy was observed at this reviewing stage. This was due to the varying quality of items written at these workshops, with many being written in a workable format, but a larger number not meeting the standard required of an SJT item.

2.5 Review Workshops

2.5.1 The aims of the review workshops were for SMEs to review SJT items for relevance, fairness and face validity and to gain agreement on the scoring key. A small number of F2 doctors attended the workshops (one to two per group) to provide additional input in terms of the items' relevance and realism to the F1 role.

2.5.2 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 attended the review workshops were awarded six continuing professional development (CPD) points.

2.5.3 Six review workshops took place across three days in May and June 2016. Two workshops were held in Manchester, two in London and two in Cardiff. A total of 33 individuals took part in the workshops, including eight F2 doctors.

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

2.5.5 Each of the six groups, with the aid of a facilitator, reviewed 25-35 items. Attendees were asked to consider the scenario content and the responses, without sight of the answer key that was initially proposed following the IDI. They were also asked to discuss a possible answer key, which was compared with the original answer key. Their comments and suggestions were recorded by the facilitator and updates were made to items. In a small number of cases, items that generated a large amount of debate were reviewed by a second group before a decision was made. Items were only progressed to the next stage if the group agreed with a key, following discussions regarding the correct answer.

2.5.6 A total of 191 items were reviewed during the review workshops. During the course of the review workshops, it was agreed that 20 items should not be taken forward due to issues with either relevance or fairness.

2.5.7 For FP 2017, a final review workshop was held with a stakeholder group, to review all trial items before they entered the subsequent stage. The group was split into pairs and the 171 items were split amongst the pairs of reviewers. All items were discussed within pairs, with items highlighted as requiring review being flagged for discussion with the wider group. During these group discussions, it was agreed that a further two items would be removed from the item development process.

2.6 Concordance Panels

2.6.1 Concordance panels were held following the review workshop stage. Concordance panels involve SMEs, in this case senior clinicians working closely with F1s, completing a paper consisting of trial SJT 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. Generally, those items that exhibit high levels of consensus go forward to be trialled. Items exhibiting low levels of consensus are subject to further review, with changes made if necessary. Concordance panels also provide the opportunity for additional feedback to be gathered about the items, regarding fairness and relevance.

2.6.2 The criteria for SME involvement in the concordance panel were that the clinicians worked closely with F1 doctors and were very familiar with the responsibilities and tasks, as well as the necessary skills and attributes, required for the role. Unlike the earlier item development stages, F2s were not invited to take part in the concordance panels.

2.6.3 Three concordance panels were held, each in two sessions, with one paper reviewed at each panel. Paper One consisted of 62 items, Paper Two consisted of 62 items and Paper Three consisted of 45 items; therefore, a total of 169 items were answered by the clinicians who attended the concordance sessions. At this stage, the papers were not constructed as final tests, i.e. less consideration was given to the spread of item topics or domains in comparison with operational paper construction, as the aim of the concordance panels was to analyse individual items. This was made clear to those attending the panels.

2.6.4 A total of 44 individuals attended the concordance stage. Panel One consisted of 12 individuals and Panel Two consisted of 17 individuals. Fifteen individuals attended Panel Three, however not all individuals managed to complete the paper. A minimum of 10 individuals is required for robust concordance analysis, with ideally 12 or more undertaking each paper; this level was mostly achieved, with 15 out of the 169 items being sat by less than 10 individuals, all of which were within Paper Three.

2.6.5 After completing a confidentiality form and a short calibration session to ensure that individuals had understood the instructions, the panel was asked to complete the SJT items without discussing them. A separate feedback sheet was provided for comments about individual items. There was no set time limit for completing the papers.

2.6.6 The concordance panel was also invited to leave written feedback on each of the items and a short individual discussion was facilitated at the end of the testing session to allow the attendees to provide comments about the test more generally.

2.6.7 Following the concordance panels, concordance analysis was undertaken to investigate the experts' level of agreement over the keyed response for each trial item. This process involves a number of stages and takes into account both quantitative and qualitative data.

2.6.8 The main criterion for selecting an item for use was a significant Kendall's *W*[4] of above .50, therefore, following best practice processes, any item that produced a low and non-significant Kendall's *W* was removed from the test (n=2) due to unsatisfactory levels of consensus.

2.6.9 One-hundred-and-one items with a significant Kendall's value of above .50 were eligible to be piloted based on this criterion. A qualitative review (including reviewing SME feedback) of items with a Kendall's value of above 0.50 deemed 18 items to be unsuitable based on feedback from SMEs that indicated that

[4] Kendall's *W* (also known as Kendall's coefficient of concordance) is a non-parametric statistic. If the test statistic *W* is 1, then all the survey respondents have been unanimous, and each respondent has assigned the same order to the list of concerns. If *W* is 0, then there is no overall trend of agreement among the respondents, and their responses may be regarded as essentially random. Intermediate values of *W* indicate a greater or lesser degree of unanimity among the various responses. In this context (and with 12 respondents), 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.

the items needed substantial changes, e.g. due to issues of relevance, difficulty, ambiguity or fairness. One additional item was deemed inappropriate for the trialling process due to the item being regarded as too easy, based on a high Kendall's value and feedback from SMEs, thus this item was also removed from the item development process.

2.6.10 Given that there is a 'tolerance' around the inclusion criterion figure (as the criterion of +.50 and the associated significance level is dependent on a number of factors, including the number of participants), it is also important to look at those items that have a significant Kendall's *W* but one that is below .50, as, whilst below .50, these may exhibit satisfactory levels of concordance given that the coefficient is significant. Following a qualitative review of these 47 items (including feedback provided by SMEs) and detailed review of the statistics, a further eight items were removed from the item development process at this stage. Thirty-nine remained in the FP 2017 item development process, some of which were amended slightly in accordance with feedback that was obtained from the SMEs who attended the concordance panels.

2.6.11 Based on the process outlined above, 140 items (82.8% of all items) were deemed to be successful following concordance review and analysis and 29 items (17.2% of all items) were not deemed as suitable to continue in the FP 2017 item development process due to low consensus amongst experts and/or based on feedback from SMEs. These items will be further reviewed and amended to ascertain the appropriateness of them entering the FP 2018 item development process.

2.6.12 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 one established after the concordance stage, then the key will be reviewed with the assistance of SMEs.

2.6.13 The breakdown of items at this final stage of the development process, relevant to each of the target domains, was as follows:

- Commitment to Professionalism – 33
- Coping with Pressure – 20
- Effective Communication – 20
- Patient Focus – 31
- Working Effectively as Part of a Team – 36

2.6.14 The different lead-in formats are reviewed each year and their performance is monitored. The number of items developed for trialling in FP 2017 using each of these lead-ins was as follows:

- Rank Actions – 88
- Rank Agreement – 0
- Rank Considerations – 7
- Rank Order – 3
- Multiple Choice - 42

Part Two: Scoring, Analysis and Evaluation

3 Operational Test Structure and Construction

3.1 Test Construction Overview

3.1.1 All SJT items used operationally have been part of an item bank that has been developed between 2010 and 2016. Every item within the operational item bank has been deemed to have sufficient psychometric properties to be used operationally and is reviewed annually to ensure it has current clinical relevance.

3.1.2 Three administrations of the SJT were undertaken, requiring the production of three versions of the test paper, which were subsequently equated. Paper three was included as a 'mop up' paper, and comprised items from both Paper one and Paper two.

3.1.3 In line with previous years, for each paper version, 70 items were administered. Of these, 60 were operational items and 10 were trial items. There were 40 ranking operational items and 20 multiple choice operational items for each paper, plus trial items. The paper versions were designed with specific overlaps ('anchor' questions), which can be used to compare populations and equate the different papers.

3.2 Spread of Items across Papers

3.2.1 The three papers were developed to be as similar as possible in terms of content, psychometric properties and difficulty. The process for selecting operational items for the tests and splitting these between papers to maximise equivalence is illustrated in Figure 2.

3.2.2 Minor differences between the papers are unavoidable and therefore a statistical technique known as test equating is undertaken as part of the scoring process. Test equating is used to align scores on multiple versions of the test whilst controlling for differences in ability across groups (see section 4.3 for more information about test equating).

3.2.3 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 *within* each paper. The split of target domains is outlined in Table 3, which shows an identical domain split for all three papers, and an almost equal split of domains overall. Of the 99 items that were used across the three papers the spread of target domains was as follows: 20 are categorised under the 'Commitment to Professionalism' domain, 23 'Coping with Pressure', 12 'Effective Communication', 23 'Patient Focus' and 21 'Working Effectively as Part of a Team'. It must be noted that items may relate to more than one domain (e.g. to work effectively as part of a team you often have to demonstrate effective communication). However, the item is categorised in accordance with the domain most appropriate for the main issue at the centre of the dilemma.

Figure 2: Operational test construction process

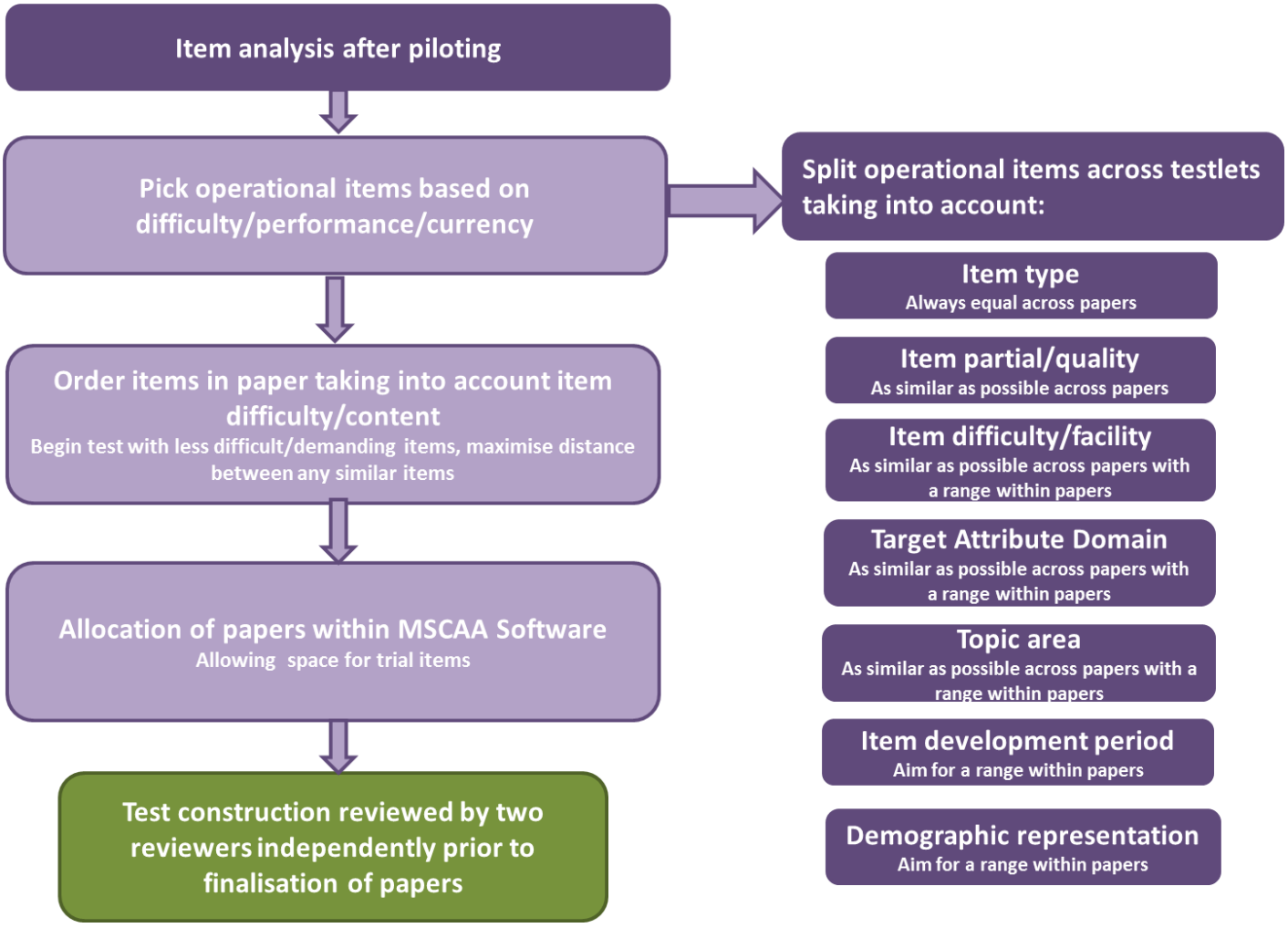


Table 3: Spread of target domains within each paper

Paper	Commitment to Professionalism	Coping with Pressure	Effective Communication	Patient Focus	Working Effectively as Part of a Team
1	14	14	8	12	12
2	12	13	9	14	12
3	14	12	8	13	13

3.2.4 As previously mentioned, as well as selecting items based on their domains and topics, attention was paid to ensuring that the range of 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	15.31	19.15	17.43	7.16	11.31	9.20	.160	.270	.203
2	15.34	19.15	17.38	7.98	11.05	9.49	.140	.270	.202
3	15.31	19.15	17.43	7.98	11.05	9.49	.140	.270	.202

3.2.5 In addition to maximising equivalence across the different versions of the test, test construction involves the process of determining the position within the test where each item should be placed. The content of the first few items in each operational paper was reviewed to ensure that they would not be viewed as particularly distressing. This is to avoid placing undue anxiety on the applicants as they commence the test. As far as possible, the distance between items with any perceived similarity in content was also maximised (e.g. two items with different topic codes and target attributes, yet both involving the treatment of a young child, would be positioned so as to not be close to each other).

3.2.6 Following initial test construction, the papers were reviewed for overlap and spread. This process was undertaken by two individuals who were not part of the initial test construction process. Where any overlap was identified (e.g. there appeared to be two items in a paper that address a similar topic) or there were potential issues with the spread of domains (e.g. there appeared to be a 'grouping' of items that could be viewed as similar), these were flagged and adjustments were made to the placing of items where possible, whilst still taking into account the other required parameters.

3.2.7 Trial items were assigned to operational papers across 14 'testlets', each containing 10 trial items (seven ranking, three multiple choice). At this stage, another review was undertaken to identify any overlap (e.g. two items in a paper that address a similar topic) across trial and operational items. These were flagged and adjustments were made to the placing of the trial items where possible.

3.2.8 A review of the item order within the papers was undertaken internally at WPG. Additionally, in response to a request from the UKFPO, a full review of the operational bank took place in addition to a final review of test paper construction from UKFPO appointed clinicians. The review of the bank recommended five items were removed from the operational item bank, due to no longer being relevant. This was implemented for FP 2017 test construction. Following the review of test construction, minor updates were made based on feedback received.

3.3 Equality and Diversity within Test Construction

3.3.1 Following a recommendation from the independent equality and diversity review that took place during FP 2014 item development, a process is undertaken in which all items are tagged according to the broad demographic characteristics of the individuals portrayed in the scenario. Gender and ethnicity characteristics are assigned based on the given names and allocated to relevant 'tags'. In addition,

whether the item portrays the individual(s) concerned in a positive/neutral or a negative way is also recorded. This is to help ensure that the tests do not unintentionally present certain groups favourably or unfavourably and to monitor the way in which gender is linked to job roles portrayed in the items.

3.3.2 Specifically, the tagging process allows for the spread of demographic characteristics to be reviewed for each operational paper, and for this to be taken into consideration during test construction. Whilst the psychometric properties of the item and target domain were prioritised in terms of selecting items for each paper, a check was also undertaken to ensure that, where possible, each paper provided an appropriate spread of demographic characteristics. Table 5 shows the representation of ethnicity across the three papers along with the number of items in which the ethnic group was presented in a negative way, or a neutral or positive way. In each paper, between 22 and 26 items were considered to be neutral with regards to ethnicity. This means that from the information included in the item, it was not possible to infer a particular ethnicity in these items (for example, if no names were referred to).

Table 5: Ethnicity representation within each paper

Ethnicity/ characteristic portrayed		Paper		
		1	**2**	**3**
Neutral		26 (40%)	22 (36%)	24 (38%)
White	Neutral/positive	21 (32%)	22 (36%)	19 (30%)
	Negative	9 (14%)	10 (16%)	12 (19%)
BME	Neutral/positive	8 (12%)	6 (10%)	6 (10%)
	Negative	1 (2%)	2 (3%)	2 (3%)

3.3.3 Table 6 shows the representation of gender across different character types for each of the three papers. This also shows the number of items in which each gender was presented in a negative way, or a neutral or positive way. In each paper, between 6 and 15 items were considered to be neutral with regards to gender. This means that the gender of the individual(s) within these items was not referred to. Paper 2 contained a higher number of items with female Foundation Programme colleagues who are displaying negative traits in comparison to males, although Paper 3 contained a larger number of male Foundation Programme colleagues displaying negative behaviours in comparison to females. Similarly, for senior colleagues, Papers 1 and 2 contained a larger number of items with males displaying negative characteristics compared to females, however Paper 3 contained a larger number of items with females displaying negative characteristics compared to males. For patients, Papers 2 and 3 contained a higher number of items with males who are displaying negative characteristics compared to females. For nurses, Papers 2 and 3 contained a higher number of males displaying negative traits compared to females.

3.3.4 For friends/relatives there was a greater number of items with female friends/relatives who were displaying positive or neutral behaviours in comparison to males, however this difference has reduced since the previous year.

3.3.5 To counteract the overrepresentation of any biases within the item bank, each year the items within the bank are reviewed with regards to their representation of gender and ethnicity and guidance is provided for item writers to follow to try to address the balance. However, it is important to note that, while frequency of representativeness is important, it is the negative portrayal of individuals that is the primary focus of this monitoring and there are a relatively similar number of items displaying negative traits for both genders across Papers 1, 2 and 3.

Table 6: Gender representation within each paper

			Paper		
			1	**2**	**3**
Neutral			15 (19%)	6 (7%)	13 (17%)
FP Colleague	Female	Neutral/positive	4 (5%)	5 (6%)	5 (6%)
		Negative	1 (1%)	4 (5%)	2 (3%)
	Male	Neutral/positive	4 (5%)	3 (4%)	1 (1%)
		Negative	1 (1%)	1 (1%)	3 (4%)
Friend/relative	Female	Neutral/positive	3 (4%)	2 (4%)	2 (3%)
		Negative	0 (0%)	0 (0%)	0 (0%)
	Male	Neutral/positive	0 (0%)	0 (0%)	0 (0%)
		Negative	0 (0%)	1 (1%)	1 (1%)
Nurse	Female	Neutral/positive	6 (8%)	4 (5%)	5 (6%)
		Negative	1 (1%)	1 (1%)	1 (1%)
	Male	Neutral/positive	2 (3%)	2 (2%)	2 (3%)
		Negative	1 (1%)	2 (2%)	2 (3%)
Patient	Female	Neutral/positive	4 (5%)	13 (16%)	9 (11%)
		Negative	0 (0%)	1 (1%)	0 (0%)
	Male	Neutral/positive	8 (10%)	16 (19%)	14 (18%)
		Negative	0 (0%)	3 (4%)	4 (5%)
Senior Colleague	Female	Neutral/positive	8 (10%)	4 (5%)	7 (9%)
		Negative	5 (6%)	3 (4%)	4 (5%)
	Male	Neutral/positive	7 (9%)	8 (10%)	5 (6%)
		Negative	6 (8%)	5 (6%)	0 (0%)

3.4 Status of Items within Operational Item Bank for FP 2017

3.4.1 Following FP 2017 test construction, it is possible to ascertain how each of the items within the existing operational bank has been utilised across the five administrations of the SJT. Preceding the delivery of the FP 2017 SJT, of the 317 items in the operational bank:

- 5 have been used in all five operational administrations (i.e. FP 2013, FP 2014, FP 2015, FP 2016 and FP 2017)
- 24 have been used in four operational administrations
- 21 have been used in three operational administrations
- 49 have been used in two operational administrations
- 103 have only been used in one operational administration
- 115 remain unused operationally

3.4.2 Test construction on a yearly basis is a complex process that takes into account multiple factors, such as spread of item topics and domains and ensuring psychometric equivalence across the different test papers. To seek to continually improve the test, especially in terms of reliability and differentiation, prioritising items that have the highest item partials (taking into account all other factors) can result in items that have been deemed appropriate to enter the item bank not being used in that particular year. In addition, due to the smaller sample sizes with trial items, statistics are less stable and thus best

practice test construction principles advise limiting the proportion of 'new trial' items within a test. WPG will seek to work with MSC Assessment to ensure that utilisation of items is maximised to promote both robust test construction and item efficiency.

4 Scoring and Test Equating

4.1 Following the scanning of all responses and a series of quality checks[5] undertaken by MSC Assessment, the raw responses were received by WPG for scoring.

4.2 The scoring quality assurance (QA) procedure follows the process summarised below:

- **Scoring syntax QA:** this includes a check for typographical/SPSS errors, item type, key, number of options and tied scores. In advance of receiving the operational data, dummy data are also run, to test that 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 have 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 that they are in the expected range and that there are no anomalies.

4.3 Whilst 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. This equating process also ensures equivalence across the papers. 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 is used to ensure that the scores are fully equivalent.

4.4 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.

4.5 The raw equated SJT scores were transformed onto a scale that was similar to the EPM score scale, whilst preserving the original distribution. The scale was set to be from 0.00 to 50.00, with a mean and standard deviation (SD) that were as similar as possible to the EPM mean and SD, and with scores rounded to two decimal places. 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 55, similar to FP 2016 (67) and much reduced from FP 2015 (149), FP 2014 (144) and FP 2013 (147). This reduction was achieved through a change in the equating process since FP 2016 (equating Paper One to Paper Two, rather than the reverse, as had been done previously), which leaves the Paper Two applicants' scores the same, and changes the Paper One scores. As equating is undertaken using three different sections of the papers, individuals with the same raw score on one paper can obtain a different equated score, due to differences in sub-scores on the three equated sections. As the largest proportion of applicants sat Paper One, this paper has previously been the largest source of ties. As a result of using this method of equating, it is possible to separate applicants who have an identical score on Paper One, and the small

[5] See http://www.foundationprogramme.nhs.uk/download.asp?file=FP_2014_SJT_scanning_FINAL.pdf for further details on the quality checks undertaken by MSC Assessment following scanning.

differences in scores on the different sections are reflected by the decrease in tied scores this year. Keeping two decimal places only affects scores on Paper One and makes it much less likely that a score from Paper One will fall on a number that could be a Paper Two score.

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 test's ability 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 for use in future years. As a new methodology was introduced for item development for FP 2017, it is important to monitor whether this has had any impact on the psychometric properties of the items written.

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
- Relationships between the EPM and the SJT
- Evaluation of applicant reactions

5.3 Sample

5.3.1 There were a total of 7,713 applicants who took the FP 2017 SJT. They were final year medical students, including students who had been pre-allocated to a Defence Deanery FP, 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 7. 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 7: Number of applicants taking each paper

	No. of applicants	Percentage of Overall Sample
Paper One	5,594	72.5%
Paper Two	2,114	27.4%
Paper Three	5	0.1%

5.3.4 Applicant demographic data were collected from the Oriel application system, however demographic data were not available for all applicants as the fields are not mandatory to complete.

5.3.5 Table 8 outlines the breakdown of applicants by gender. Overall, more females completed the test (52.8%) than males (43.0%), reflecting the male/female split of applicants to the Foundation Programme.

Table 8: Applicant gender by paper

		Male	Female	Not declared
Overall	No. of applicants	3,316	4,072	325
	% of applicants	43.0%	52.8%	4.2%
Paper One	No. of applicants	2,434	2,923	237
	% of applicants	43.5%	52.3%	4.2%
Paper Two	No. of applicants	881	1,146	87
	% of applicants	41.7%	54.2%	4.1%
Paper Three	No. of applicants	1	3	1
	% of applicants	20.0%	60.0%	20.0%

5.3.6 Table 9 outlines the breakdown of applicants by ethnicity. Overall, the majority of applicants reported their ethnicity as 'White' (61.4%), with the smallest proportion of applicants (3.4%) reporting themselves as being from 'Black' backgrounds. Four hundred and eighty-five (6.3%) applicants did not declare their ethnicity. The proportion of individuals in each ethnic group was roughly equivalent in Paper One and Paper Two. Paper Three had a very small sample size and reflected just two ethnic backgrounds; 'White' (n=2) and 'Black' (n=1).

Table 9: Applicant ethnicity by paper

		White	Asian	Black	Mixed	Other	Not declared
Overall	No. of applicants	4,734	1,368	259	297	570	485
	% of applicants	61.4%	17.7%	3.4%	3.9%	7.4%	6.3%
Paper One	No. of applicants	3,362	1,037	190	219	435	351
	% of applicants	60.1%	18.5%	3.4%	3.9%	7.8%	6.3%
Paper Two	No. of applicants	1,370	331	68	78	135	132
	% of applicants	64.8%	15.7%	3.2%	3.7%	6.4%	6.2%
Paper Three	No. of applicants	2	0	1	0	0	2
	% of applicants	40.0%	0.0%	20.0%	0.0%	0.0%	40.0%

5.3.7 Table 10 outlines the breakdown of applicants' ethnicity when classified either into the 'White' or 'Black and Minority Ethnic (BME)' group. Four thousand, seven hundred and thirty-four (61.4%) applicants were

classified as White and 2,494 (32.3%) applicants were classified as being from BME groups. Paper One had 33.6% BME applicants, Paper Two had 28.9% BME applicants and Paper Three had 20.0% BME applicants.

Table 10: Applicant ethnicity by paper

		White	BME	Not declared
Overall	No. of applicants	4,734	2,494	485
	% of applicants	61.4%	32.4%	6.3%
Paper One	No. of applicants	3,362	1,881	351
	% of applicants	60.1%	33.6%	6.3%
Paper Two	No. of applicants	1,370	612	132
	% of applicants	64.8%	28.9%	6.2%
Paper Three	No. of applicants	2	1	2
	% of applicants	40.0%	20.0%	40.0%

5.3.8 Table 11 outlines the breakdown of applicants by their country of medical education (UK and non-UK medical schools). 94.3% of applicants were from UK medical schools and 5.0% of applicants were from non-UK medical schools. Paper One had 3.8% non-UK applicants, Paper Two had 8.0% non-UK applicants and Paper Three had 0.0% non-UK applicants.

Table 11: Applicant country of medical education by paper

		UK	Non-UK	Not declared
Overall	No. of applicants	7,273	384	56
	% of applicants	94.3%	5.0%	0.7%
Paper One	No. of applicants	5,336	215	43
	% of applicants	95.4%	3.8%	0.8%
Paper Two	No. of applicants	1,933	169	12
	% of applicants	91.4%	8.0%	0.6%
Paper Three	No. of applicants	4	0	1
	% of applicants	80.0%	0.0%	20.0%

5.3.9 The mean age of the sample was 24.7 years (SD: 3.09) and the median was 24.0 years, with a range of 21 – 56 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 12 provides an overview of test completion across all of the test papers. Across all test papers, 99.6% of applicants attempted the last operational item on the test. 99.2% answered all items and less than 0.2% failed to answer four or more items.

5.4.2 Test completion was also examined by paper, through identifying the proportion of applicants who did not attempt the last operational item. 0.1% of applicants in Paper One and 0.2% of applicants in Paper Two did not answer the last operational question on the first section of the paper. 0.3% of applicants in

Paper 1 and 0.6% of applicants in Paper 2 did not finish the last question in the second section of the paper. 99.3% of applicants in Paper One answered all items and 99.0% of applicants in Paper Two answered all items. 0.2% of applicants in Paper One and 0.4% of applicants in Paper Two failed to answer four or more items. Therefore, it seems that there is a very similar completion rate for Paper One and for Paper Two. 100% of applicants in Paper Three completed all items in the test, with no items missed.

5.4.3 These results are comparable with previous years (99.1% completion rate (answered all items) for FP 2016, 99.7% for FP 2015, 98.8% for FP 2014 and 97.9% for FP 2013) and indicate that the SJT is a power test, rather than a speed test. This indicates that 140 minutes continues to be an appropriate length of time to complete 70 items.

Table 12: Analysis of Test Completion

		Attempted last item	Answered all items	Failure to answer four or more items
Overall	No. of applicants	7,686	7,652	17
	% of applicants	99.6%	99.2%	0.2%
Paper One	No. of applicants	5,580	5,554	9
	% of applicants	99.7%	99.3%	0.2%
Paper Two	No. of applicants	2,101	2,093	8
	% of applicants	99.4%	99.0%	0.4%
Paper Three	No. of applicants	5	5	0
	% of applicants	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 13 illustrates the test level descriptive statistics presented alongside the results from FP 2016, FP 2015, FP 2014 and FP 2013. Due to the extremely low number of applicants who completed Paper Three, the reliability, Standard Error of Measurement (SEM), skew and kurtosis have not been reported for this paper, as the sample size means that this analysis would not be robust.

5.5.2 **Test level Facility (Difficulty)**: Mean scores are broadly similar between Paper One and Paper Two, with Paper One exhibiting a slightly higher raw mean score. The *percentage correct* (the mean raw score, expressed as a percentage of the total maximum score available) is highly comparable across the three papers, with Paper One again exhibiting a slightly higher percentage correct (Paper One 86.3%, Paper Two 85.7%). This indicates that the papers have comparable levels of difficulty, even prior to the equating process. These findings are broadly comparable to previous years; the mean percentage correct for Paper One was 87.3% for FP 2016, 85.7% for FP 2015, 84.0% for FP 2014 and 83.1% for FP 2013, whilst for Paper Two it was 85.3% for FP 2016, 84.8% for FP 2015 and 83.4% in both FP 2014 and 2013. The equating strategy that follows scoring takes into account any differences between the papers in terms of difficulty and the maximum available score for each paper.

5.5.3 For FP 2017, the pattern of increasing scores following each operational use of the SJT that has previously been observed each year was not evident. This could be due to the item development process since 2015

having sought to address the observed increase through the production of items with additional complexity.

5.5.4 Results from FP 2017 outline that the raw mean score for Paper One was 0.6% higher than Paper Two. Any differences based on the difficulty of the test content are addressed during the equating process, however these differences continue to exist following equating. This is therefore likely to be due to genuine differences in the populations undertaking the two papers. This is a pattern that is consistent across administration years.

Table 13: Operational level descriptive statistics by paper (raw or unequated)

	N	Reliability (α)	SEM	Mean	Mean % correct	Skew	Kurtosis	SD	Min	Max	Max Score	N items
Paper One	5,594	.73	15.19	893.8	86.3	-1.04	3.30	29.23	675	968	1,036	60
Paper Two	2,114	.77	15.78	887.4	85.7	-1.65	7.03	32.90	586	954	1,036	60
Paper Three	5	n/a	n/a	877.4	84.7	n/a	n/a	33.07	823	904	1,036	60
Overall FP 2017	*7,713*	*.75**	*15.49**	*890.6**	*86.0**	*-1.35**	*5.17**	*31.07**	*586*	*968*	*1,036*	*60*
Overall FP 2016	*7,807*	*.74**	*15.23**	*896.1**	*86.3**	*-1.34**	*6.16**	*30.07**	*568*	*978*	*1,038/ 1,039*	*60*
Overall FP 2015	8,088	.71*	16.23*	884.8*	85.2*	-1.31*	4.99*	29.92*	611	965	1,036/ 1,039	60
Overall FP 2014	7,957	.69*	15.80*	842.0*	83.7*	-2.09*	14.64*	28.17*	498	914	996/ 1,016	58/59[6]
Overall FP 2013	8,162	.72*	16.71[7]*	859.1*	82.8*	-1.69*	8.51*	31.56*	546	935	1,037/ 1,038	60

** Average across Papers One and Two*

5.5.5 **Spread of Scores**: The range of scores is largest for Paper Two, and smallest for Paper Three. However, the SD is a much better indicator of the spread of scores than the range, as the range can be strongly affected by a single outlier.

5.5.6 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: 29.23) is lower than that for Paper Two (SD: 32.90). This suggests a slightly greater variation in scores for 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, which may account for this difference in variance. The SD for Paper Three (SD: 33.07) is higher, but any measure of distribution will be unstable in such a small sample. Overall, the values of the SDs are as expected and, given that the SD is affected by the number of items, this can be considered comparable with FP 2016 (SD: 30.07), FP 2015 (SD: 29.92), FP 2014 (SD: 28.17) and FP 2013 (SD: 31.56).

[6] The overall number of items for FP 2014 was lower, as two operational items were removed from Paper One and one operational item was removed from Paper Two as a result of them having negative item partials.

[7] SEM calculated using the mean of the SEM for Paper One and Paper Two. In FP 2013, this was calculated using the mean of the standard deviation and reliability across Paper One and Paper Two.

5.5.7 **Reliability**: The mean reliability for FP 2017 is α=.75, which is sufficient for the use of an operational SJT (Paper One α=.73, Paper Two α=.77). Paper Two had higher reliability than Paper One, consistent with previous years. 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 (as with Paper Two), reliability coefficients tend to be higher. In this case, since Paper Two applicants exhibit a slightly greater spread of scores (indicated by the higher SD), the reliability coefficient is also slightly higher. Inspection of the SEM[8] indicates that the underlying accuracy of scores on the two papers is comparable but that Paper One has a slightly lower SEM than Paper Two (15.19 and 15.78, respectively), indicating slightly lower measurement error in Paper One.

5.5.8 Overall, the reliability has increased slightly in comparison to FP 2016 (Paper One α=.71, Paper Two α=.77), FP 2015 (Paper One α=.69, Paper Two α=.72), FP 2014 (Paper One α=.67, Paper Two α=.70) and FP 2013 (Paper One α=.67, Paper Two α=.76). The slight decrease in the reliability for FP 2014 may have reflected the removal of two operational items from Paper One and one operational item from Paper Two.

5.5.9 **Distribution of Scores**: Figures 3 and 4 illustrate the distribution of scores for Papers One and Two, which are slightly negatively skewed. This is also reflected in the skew value presented in Table 13 above. A negative skew indicates that the tail on the left side is longer than the right side. The extent of the skew for FP 2017 is larger for Paper Two (i.e. the tail of lower scorers is more pronounced, with more extreme low scorers). The overall extent of the skew for FP 2017 is comparable to FP 2016 and FP 2015.

5.5.10 In looking at the distribution of scores, we can also examine the kurtosis[9] figure presented in Table 13. This indicates that the distribution has a slightly higher peak with scores more clustered around the mean than would be expected in a normal distribution. Again, for Paper Two the kurtosis value is higher than in Paper One, suggesting that the Paper One scores are more in line with what we would expect of a normal distribution. The overall kurtosis is comparable to FP 2016 and FP 2015.

[8] The 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.

[9] Kurtosis is a measure of the peak of a distribution, and indicates how high the distribution is around the mean. Positive values indicate that the distribution has a higher peak than would be expected in a normal distribution; negative values indicate that the distribution has a lower peak than would be expected in a normal distribution.

Figure 3: Distribution Statistics for Paper One

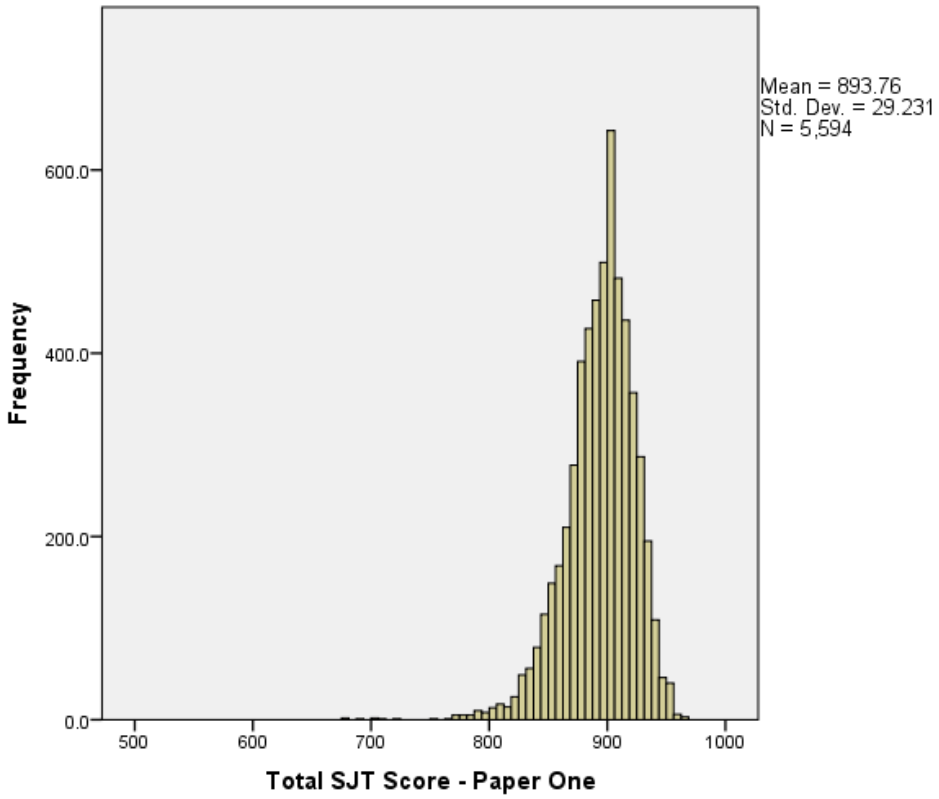
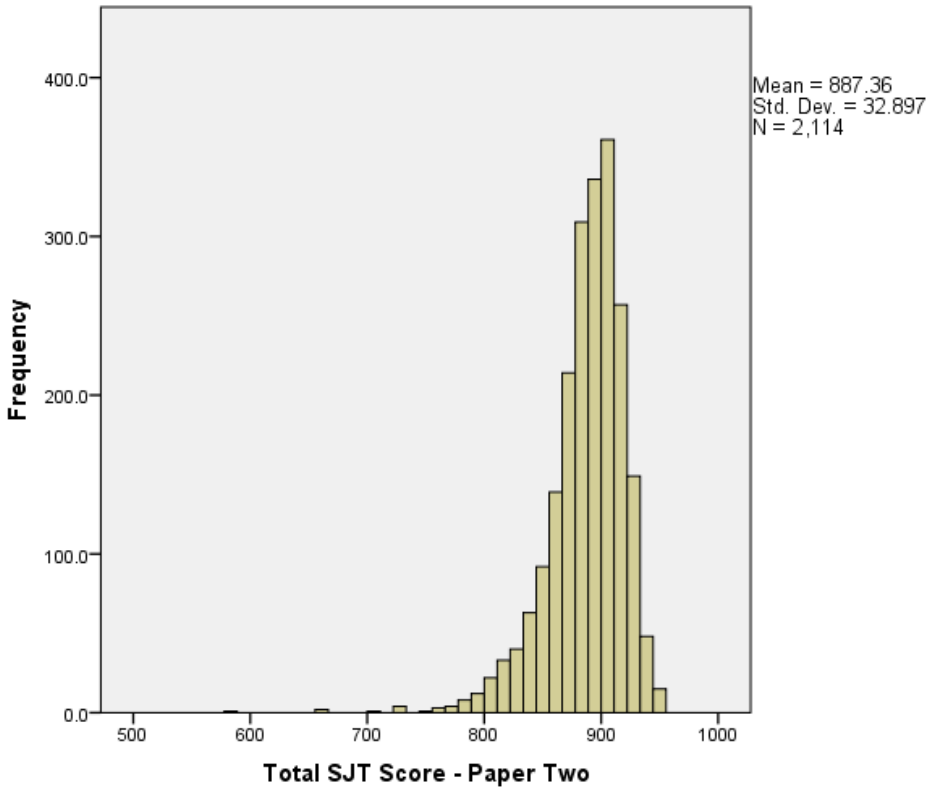


Figure 4: Distribution Statistics for Paper Two



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 sample size for completed 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 and Spread of Scores:** 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 choice items). Although test construction strives to include items that are challenging, 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, again, may not provide useful information or differentiate between applicants. A range of item facilities is sought for an operational test, with very few items categorised as very easy (a mean score of greater than 90% of the total available score) and very few items categorised as very difficult (a mean score of less than 10% of the total available score).

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. Prior to operational delivery, all operational items fell within these parameters, based on their psychometric properties from the piloting stages.

5.6.4 Table 14 outlines the item level statistics for Papers One and Two, once outliers have been excluded[10]. As a comparison, the overall item level statistics for FP 2016, FP 2015, FP 2014 and FP 2013 are also provided. Paper Three has not been included, as the small sample size can skew the overall results.

5.6.5 The mean item facility for ranking items is 17.4 and the mean item facility for multiple choice items is 9.7. The facility ranges and SDs for both ranking and multiple choice items are in line with expectations. The facility values are fairly consistent with FP 2016, when the mean facility values were 17.5 for ranking and 9.9 for multiple choice.

5.6.6 Items that can be categorised as 'easy' (more than 90% of the total available score) for both ranking and multiple choice are reviewed to ensure that they are sufficiently differentiating between applicants (through examination of the item partial) and are therefore providing useful information. If this is not the case, then they are removed from the operational bank. Additionally, items with low SDs (below 1 SD) are also reviewed and removed if they are deemed to be no longer appropriate.

[10] For the purposes of item level analysis and in line with best practice, seven outliers were excluded from Paper One and two outliers were excluded from Paper Two.

Table 14: Item level statistics: Facility values

	Ranking			Multiple Choice		
	Mean Facility	Facility Range	SD Range	Mean Facility	Facility Range	SD Range
Paper One	17.5	15.3-19.4	1.32-2.85	9.6	7.1-11.3	1.64-2.56
Paper Two	17.4	15.2-19.4	1.36-3.01	9.6	8.1-11.2	1.69-2.77
Overall FP 2017	17.4	15.2-19.4	1.34-2.91	9.7	7.1-11.3	1.64-2.77
Overall FP 2016	17.5	14.8-19.3	1.26-2.97	9.9	7.4-11.6	1.30-2.71
Overall FP 2015	17.3	15.3-19.0	1.37-3.26	9.7	7.5-11.3	1.66-2.72
Overall FP 2014	17.1	14.4-19.1	1.41-2.73	9.2	5.2-11.2	1.69-2.67
Overall FP 2013	16.9	14.5-19.0	1.48-2.78	9.1	5.8-11.3	1.67-2.67

5.6.7 **Item Quality**: Item quality is determined by the correlation of the item with the overall operational SJT score, not including the item itself (item partial)[11]. This analysis compares how the cohort performs on a given item with how they perform on the test overall, and is a good indication of 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 therefore 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 15 outlines how items performed for each of the two papers and overall. As a comparison, the overall item performance for FP 2016, FP 2015, FP 2014 and FP 2013 is also included.

Table 15: Operational Item level statistics: Item partials

	Range of Item Partials	Mean Item Partial	Good (>.17)	Moderate (.13-.17)	Unsatisfactory for operational bank (<.13)
Paper One	.06-.23	.17	34 (56.7%)	20 (33.3%)	6 (10%)
Paper Two	.06-.30	.19	47 (78.3%)	5 (8.3%)	8 (13.3%)
Overall FP 2017	.06-.30	.18	67 (67.7%)	21 (21.2%)	11 (11.1%)
Overall FP 2016	.00-.27	.17	61 (61.6%)	21 (21.2%)	17 (17.2%)
Overall FP 2015	.05-.26	.16	41 (41.4%)	34 (34.34%)	24 (24.2%)
Overall FP 2014	.02-.41	.16	40 (41.2%)	32 (33.0%)	25 (25.8%)
Overall FP 2013	.04-.33	.17	45 (45.5%)	36 (36.0%)	18 (18.2%)

5.6.8 Sixty-seven of the 99 (67.7%) operational items are deemed to have good psychometric properties with regards to item quality, whilst 21 (21.2%) are deemed as moderate. Eleven (11.1%) of the 99 items were

[11] With regards to acceptable levels of correlations for item partials, guidelines suggest, in general, .2 or .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 levels of correlation to be contributing to the reliability of the test.

deemed to have unsatisfactory item partials and required further review. However, as these items were not detracting from the overall reliability of the test, they were not removed from operational scoring. This is an improvement on previous years.

5.6.9 **Review of Items**: The recommendation to remove items from the operational item bank is based on a combination of psychometric information, including the item partial, item facility and SD; however, the three statistics are typically linked. In general, the following criteria are used in combination to assess whether an item should be removed:

- Item partial below .13
- Item facility above 90% and below 10% of the total available mark
- SDs of below 1 and above 3

5.6.10 A level of item redundancy is to be expected each year and is in line with SJTs used in other contexts. This has been accounted for within the test development process, with regards to the building of the item bank. Following review of the psychometric properties, it is recommended that 11 items are removed from the operational item bank. This is a decrease in comparison to FP 2016 and FP 2015, where 17 and 24 items were removed following operational delivery.

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 age, gender, ethnicity and country of medical education.

5.7.2 **Age**: In terms of age, there is a negative correlation between age and the SJT scores (r = -.16, p < .001), with younger applicants scoring significantly higher on the SJT than older applicants. However, this correlation represents a weak relationship between age and SJT score (Davis, 1971[12]). This finding is in line with previous findings from FP 2016 (r = -.13, p < .001), FP 2015 (r = -.06, p < .001) FP 2014 (r = -.11, p < .001) and FP 2013 (r = -.075, p < .001), and therefore the effects of age on SJT performance should continue to be monitored.

5.7.3 **Gender**: Table 16 shows group differences in performance on the SJT based on gender. Overall, female applicants scored significantly higher than male applicants by 0.19 SDs. A t-test[13] revealed that the difference is statistically significant (p < .001, t = 8.29, d = 0.19). Cohen's d[14], which quantifies the magnitude of the difference between the mean SJT scores for males and females, can be classified as a small effect size. This difference is consistent with the difference observed for other selection and assessment methods used at various stages of the medical career pathway[15]. The difference is also

[12] Davis, J. A. (1971). *Elementary survey analysis*. Englewood Cliffs, NJ: Prentice–Hall.

[13] 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 p value indicates the probability of finding a difference of the given magnitude or greater in a sample where there is no actual difference between the groups. By convention, p values below .05 are said to indicate statistical significance – i.e. low likelihood of a similar finding happening by chance.

[14] Cohen's d is an effect size statistic used to estimate the magnitude of the difference between the two groups. In large samples even negligible differences between groups can be statistically significant. Cohen's d quantifies the difference in SD units. The guidelines (proposed by Cohen, 1988) for interpreting the d value are: 0.2 = small effect, 0.5 = medium effect and 0.8 = large effect.

[15] Patterson, F., Zibarras, L., & Ashworth, V. (2016). Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. *Medical teacher*, *38*(1), 3-17.

comparable with that found during FP 2016 ($p < .001$, $d = 0.20$), FP 2015 ($p < .001$, $d = 0.26$) and FP 2014 ($p < .001$, $d = 0.22$). In FP 2013, the observed difference between males and females was non-significant. DIF analysis (see 5.8) provides further insight into group differences and indicates that the gender differences are minimal at the item level.

Table 16: SJT group differences by gender

	Gender	N	Mean	SD	T-test Sig.	Cohen's *d*
Equated SJT score	Male	3,308	885.52	29.67	$p < .001$	0.19
	Female	4,066	891.23	29.19		

5.7.4 **Ethnicity**: Table 17 shows group differences in performance on the SJT based on ethnicity, when applicants are grouped into two categories: White and BME. White applicants scored significantly higher than BME applicants by 0.74 SDs. A t-test revealed that the difference is statistically significant ($p < .001$, $t = 29.24$, $d = 0.90$). Cohen's *d,* which quantifies the magnitude of the difference in the mean SJT scores between White and BME applicants, can be classified as a large effect size. This effect size has increased compared with that found during FP 2016 ($p < .001$, $d = 0.77$), FP 2015 ($p < .001$. $d = 0.61$), FP 2014 ($p < .001$, $d = 0.50$) and FP 2013 ($p < .001$, $d = 0.55$). Again, this difference is consistent with the difference observed for other selection and assessment methods used at various stages of the medical career pathway and a recent review of the research evidence suggests that SJTs used in medical selection can reduce group differences observed[16]. Whilst large differences are found for ethnicity, country of medical qualification confounds these differences, and therefore ethnicity differences are also examined split by country of medical qualification in section 5.7.7 to 5.7.15. The DIF analysis (see 5.8) provides further insight into group differences and indicates that there are minimal differences at the item level based on ethnicity.

Table 17: SJT group differences by ethnicity (two groups)

	Ethnicity	N	Mean	SD	T-test Sig.	Cohen's *d*
Equated SJT score	White	4,731	896.3	25.52	$p < .001$	0.90
	BME	2,483	875.0	31.18		

5.7.5 To provide a comparison, Table 18 shows group differences in performance on the EPM (both decile score and total EPM score) based on ethnicity, when applicants are grouped into the same categories; White and BME. Similar to the SJT, White applicants score higher than BME applicants by 0.38 SDs on the EPM decile scores and by 0.39 SDs on the total EPM score. T-tests reveal that these differences are statistically significant (Decile: $p < .001$, $t = 16.53$, $d = 0.47$; Total EPM: $p < .001$, $t = 16.14$, $d = 0.38$). Cohen's d can be classified as a small effect size for both the decile score and the total EPM score.

Table 18: EPM group differences by ethnicity (two groups)[17]

	Ethnicity	N	Mean	SD	T-test Sig.	Cohen's *d*
EPM Decile	White	4,726	39.0	2.81	$p < .001$	0.47
	BME	2,483	37.8	2.90		

[16] Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education*, 50(1), 36-60.

[17] In line with best practice, five outliers were removed from the analysis.

Total EPM score	White	4,726	41.4	3.75	$p < .001$	0.38
	BME	2,483	39.9	3.77		

5.7.6 **Country of Medical Education**[18]: Table 19 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 perform significantly better than those from non-UK medical schools by 1.50 SDs. A t-test reveals that the difference is statistically significant ($p < .001$, $t = 22.92$, $d = 2.30$). This is a large effect size, and larger than the differences in performance between ethnic groups. There is a level of verbal comprehension that is inherent in the SJT, therefore differences could be due, in part, to English language ability differences; although effort is made to ensure SJT items use simple language and several main principles of the SJT Style Guide support this requirement. In addition, the SJT is developed based on the cultural values of the NHS. As International Medical Graduates attain medical training centred on a working culture of a different healthcare system, this may also be contributing to the observed difference in SJT scores.

Table 19: SJT group differences by country of medical education

	Country	N	Mean	SD	T-test Sig.	Cohen's *d*
Equated SJT score	UK	7,266	890.6	27.54	$p < .001$	2.30
	Non-UK	376	847.3	36.07		

5.7.7 **Ethnicity by Country of Medical Education**: As outlined in Table 20, a greater proportion of UK applicants are categorised as White, and a greater proportion of non-UK applicants are categorised as BME. As such, this might explain some of the differences seen between groups. In other words, ethnicity is likely to be confounded by country of medical education. It is important to note, however, that the sample sizes for UK and non-UK applicants are very uneven, with 19 times more UK than non-UK applicants. Therefore, analysis of differences between these groups should be interpreted with caution.

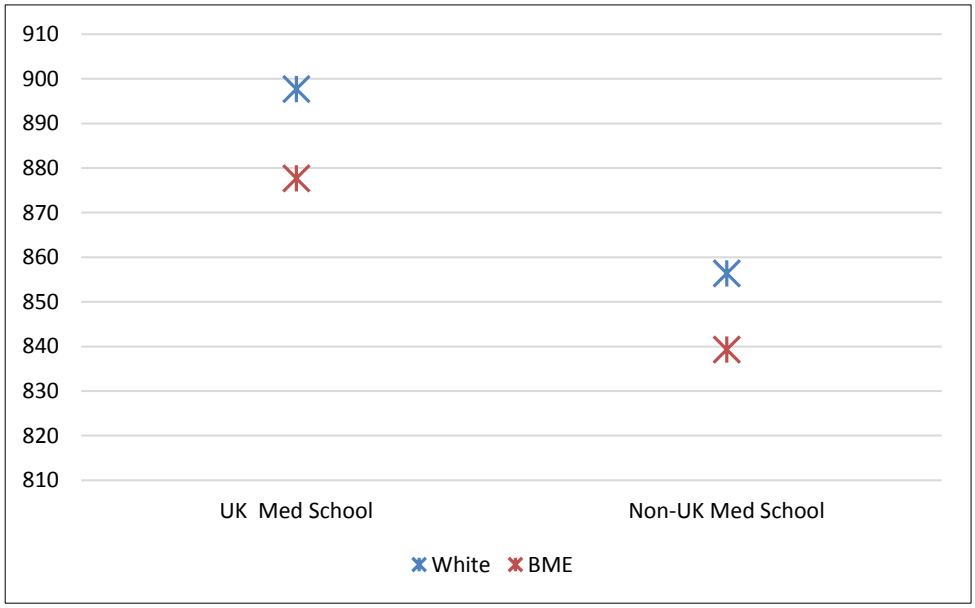
5.7.8 Table 20 shows the mean scores split by applicants' country of medical education and ethnicity; this is also illustrated graphically in Figure 5. In general, applicants who trained in the UK outperformed applicants who trained elsewhere, regardless of ethnicity, thus indicating that country of medical education plays a role in the ethnicity effects that have been found. Again, the results of this analysis should be interpreted with caution due to the small sample size of the non-UK group of applicants.

Table 20: SJT mean scores by ethnicity and country of medical education

		UK			Non-UK		
		N	Mean	SD	N	Mean	SD
Ethnicity	White	4,569	897.7	24.12	162	856.4	31.11
	BME	2,312	877.7	29.08	171	839.3	36.29

[18] Country of medical education was derived using medical school. All statistical analyses involving country of medical education (i.e. those reported in 5.7.6, 5.7.7, 5.7.8, and 5.7.10) should be **treated with caution**. This is because the variances for UK and non-UK applicants are very different; this violation of the assumptions of the analysis, together with the very uneven sample sizes for the groups (with over 19 times more UK than non-UK applicants), means that the results of these analyses are not robust and should be treated with caution.

Figure 5: Mean scores by ethnicity and country of medical education



5.7.9 Regression analyses were conducted to explore the contribution of country of medical education (UK, non-UK) and ethnicity (White, BME) to SJT performance in greater detail. A linear regression was conducted first, to analyse the amount of variance in SJT scores that each of the variables predicted independently. Place of medical education accounted for 10.0% of the variance. A separate linear regression demonstrated that ethnicity accounted for 11.8% of the variance in SJT score. Therefore, when analysed separately, medical education and ethnicity explained comparable proportions of the variance in SJT score.

5.7.10 Following on from this, a hierarchical regression was conducted[19]. Country of medical education was entered into the regression equation first in Model One, followed by ethnicity in Model Two. After the 9.6% of SJT score variance that country of medical education accounted for ($F(1,7212) = 764.10$, $p < .001$), ethnicity (White, BME) accounted for a further 10.3% of score variance when entered into the model ($F(2,7211) = 893.44$, $p < .001$). These results indicate that ethnicity still accounts for a significant proportion of the variance in SJT scores after accounting for place of medical education. This is also illustrated in Figure 7, which shows a clear difference in scores by ethnicity for both UK and non-UK groups. However, the proportion of variance explained by ethnicity, once place of medical education has been controlled for, is slightly lower than when looking at ethnicity alone, indicating that some of

[19] When conducting a hierarchical regression, the variables of interest (in this case, country of medical education and ethnicity) are entered into the analysis in two separate steps to determine the amount of variance in scores that they each explain. Only applicants with data for both variables will be included throughout all the steps. Therefore, slight variations in the regression coefficient for country of medical education can be seen compared to the linear regression above, because fewer applicants will have been included in the analysis overall (i.e. those with complete data for country of medical education, but missing data for ethnicity are excluded from the hierarchical regression).

the variance in ethnicity is explained by place of medical education. However, as ethnicity and medical education are highly correlated and the UK and non-UK groups are unbalanced in terms of sample size, these results should be interpreted with caution.

5.7.11 The analysis outlined above has contrasted White and BME groups, given that this is both typically how ethnic group differences are reported and that some analyses (e.g. t-test, regressions) require dichotomous grouping. However, grouping the ethnic groups other than White together in this way can lose data relating to the differences in performance between these different BME groups, who may have different backgrounds and educational experiences. Therefore, to further explore these differences in performance, applicants' ethnicity was broken down into five categories: White, Asian, Black, Mixed, and Other.

5.7.12 Table 21 displays the differences in equated SJT scores when ethnicity is grouped into five categories. The table also splits applicants by country of medical education, to facilitate comparison with the previous analyses. Applicants describing themselves as White or Mixed have the highest level of performance overall and across the UK and non-UK categories. Overall, those describing themselves as Black had the lowest mean score.

Table 21: SJT group differences by ethnicity (five groups) and country of medical education

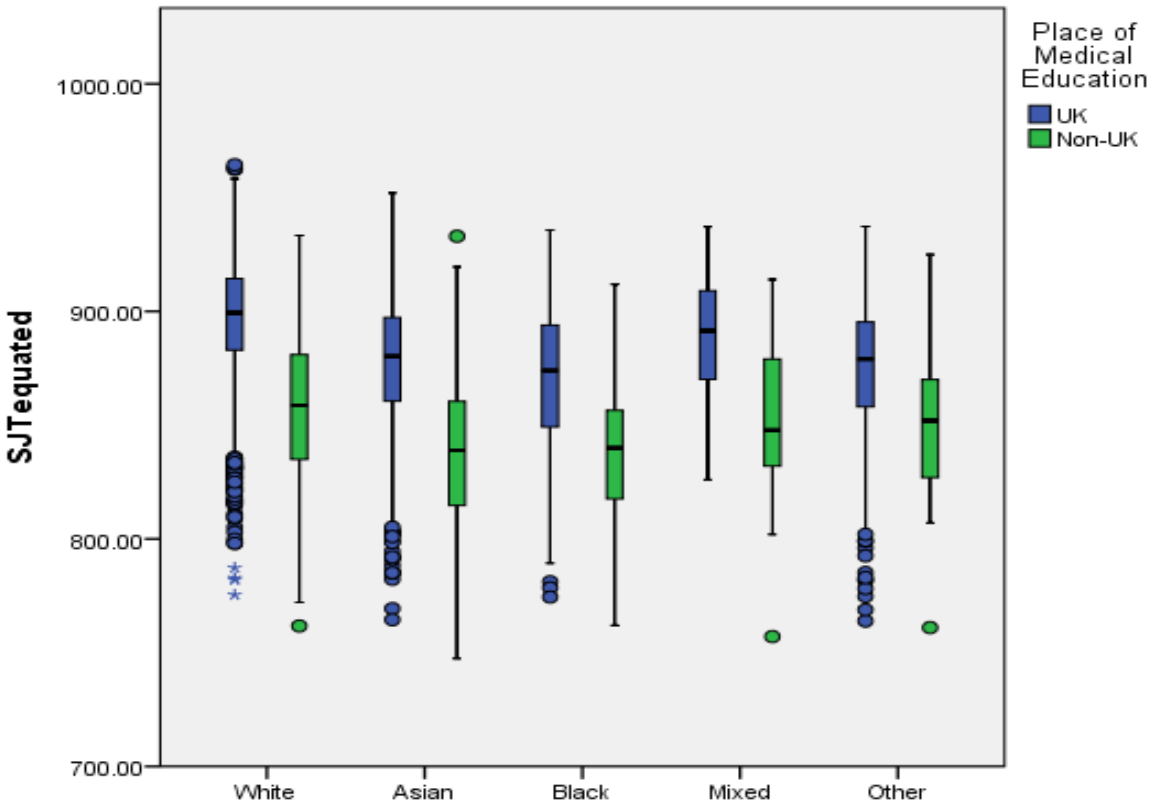
	Country of Education						**Overall**		
	UK			**Non-UK**					
Ethnicity	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	N	**Mean**	**SD**
White	4,569	897.7	24.12	162	856.4	31.11	4,731	896.3	25.52
Asian	1,270	877.5	28.33	92	836.2	35.68	1,362	874.7	30.68
Black	212	869.8	32.84	44	838.2	35.07	256	864.4	35.24
Mixed	283	888.9	26.28	14	850.5	41.87	297	887.1	28.32
Other	547	875.2	29.08	21	847.8	37.53	568	874.2	29.85

5.7.13 A one-way analysis of variance for ethnicity found a significant difference in scores for ethnic group (White, Asian, Black, Mixed, Other). Those who described themselves as 'White' scored significantly higher than all groups except for applicants who described themselves as 'Mixed' and those who described themselves as 'Mixed' scored significantly higher than all groups except for applicants who described themselves as 'White' ($F(4,7204) = 45.86$, $p < .001$), with a small but significant effect size (partial eta-squared of 0.03 and an eta-squared of 0.02). Due to the small sample sizes in the majority of the non-UK groups broken down by ethnicity, a two-way analysis of variance was not conducted to investigate an interaction effect between the two.

5.7.14 The effect size for ethnic group is small when broken down into five categories. This is in comparison to when ethnicity is analysed in a t-test, which shows a medium effect size. This indicates that the effect of ethnicity on SJT scores is complex, and that grouping all BME groups together can potentially reduce the amount of information available. However, it is important to note, when making comparisons between these analyses, that the numbers within some of the groups are relatively small and therefore the findings from the t-tests can be interpreted with a greater degree of confidence due to the increased comparability of the sample sizes for this statistical analysis.

5.7.15 Table 21 and Figure 6 show that for each ethnic group, there is a larger spread of scores for those trained outside the UK compared to UK applicants and that non-UK applicants are scoring lower than the UK applicants across all ethnic groups, which is consistent with the results from FP 2016, FP 2015 and FP 2014.

Figure 6: SJT score variance by ethnicity (five groups) and country of medical education[20]



5.8 Differential Item Functioning

5.8.1 One explanation for test level group differences is that SJT item content discriminates against particular groups. Items are designed to avoid content that might discriminate (e.g. avoiding the use of colloquial words/phrases, which might disadvantage particular groups) and item development follows the recommendation of the FP 2014 independent equality and diversity review, with the use of ethnicity and gender in items monitored at item and test development stages (see 3.3). Another explanation for group differences in performance is that real differences exist between groups of applicants, which can be due to differences in experience, attitudes or differential self-selection.

5.8.2 DIF analysis was performed to identify whether individual items are differentially difficult for members of different groups (i.e. based on gender and ethnicity). DIF analysis considers whether the prediction of an item's score is improved by including the background grouping variable in a regression equation after

[20] For each group, the box shows the score range from the 25th to the 75th percentile, with the line within the bar representing the median score. The whiskers show the range to the 5th and 95th percentiles, with scores outside this range shown as separate points (i.e. outliers).

total score has been entered. A positive result suggests that people with similar overall scores from different groups have different success rates on the item. However, because of the number of statistical tests involved, there is a danger that random differences may reach statistical significance (type 1 error). For this reason, positive results are treated as 'flags' for further investigation of items, rather than confirmation of difference or bias. Items exhibiting R-squared values with a negligible effect size, even where these differences are significant, are unlikely to indicate a meaningful difference in the performance between the groups. As such, for FP 2017 only items exhibiting at least a small effect size are reported, as determined by an R-squared value of 0.02 or above (Cohen, 1988[21]). No items were flagged for ethnicity or gender differences, which suggests that group differences at a test level are not likely the result of the test being more difficult for some groups. Therefore, it is recommended that other explanations of group difference are considered.

5.9 Correlations with the EPM

5.9.1 The relationship between SJT equated total scores and the EPM, the second tool for selection to FP 2017, was assessed using correlations[22]. Due to the low number of applicants who completed Paper Three, the correlations have not been reported for this paper, as the sample size means that this analysis would not be robust. A summary of the results can be found in Table 22, below.

Table 22: Correlations between SJT total scores and the EPM

	Current selection methods (EPM)	**SJT total scores**
Overall	Total Score	r = .31*
	Decile	r_s = .32*
Paper One	Total Score	r = .33*
	Decile	r_s = .32*
Paper Two	Total Score	r = .28*
	Decile	r_s = .31*

* Significant at the $p < .001$ level

5.9.2 At the $p < .001$ level, significant correlations were found between SJT scores and EPM decile scores for Paper One and Paper Two, and between SJT scores and total EPM score for Paper One and Paper 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. These overall correlations are comparable with FP 2016 (r = .33; r_s = .34), FP 2015 (r = .34; r_s = .35), FP 2014 (r = .30; r_s = .32) and FP 2013 (r = .30; r_s =.30).

[21] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates

[22] 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), while 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 EPM total score), the Pearson product-moment correlation coefficient is used (r). For non-normally distributed data (i.e. the EPM decile), the Spearman's rank correlation coefficient is used (r_s).

5.10 Item Level Analysis – Trial Items

5.10.1 Fourteen sets of items were trialled in the FP 2017 SJT, with each set consisting of seven ranking and three multiple choice items. Ten sets were trialled alongside operational Paper One, and ten alongside Paper Two (of which two sets were also trialled alongside Paper One). Two sets from Paper One and two sets from Paper Two were also trialled alongside Paper Three. The number of applicants completing each trial set ranged from 488 to 726, providing an appropriate sample size to enable robust analysis to take place. Trial sets were allocated across different schools and therefore the samples are not randomly allocated. Given that the operational data indicate that there are differences in performance between the two populations, even after equating test difficulty, this does mean that trial item performance may also be influenced by this.

5.10.2 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. A distractor analysis was also undertaken, to identify the answer key selected by the best performing applicants, which is reviewed alongside the content of each item before determining the final key, to ensure that the correct answer key is considering the appropriate level for the applicants.

5.10.3 **Item Facility**: Table 23 outlines the item level statistics for all 14 sets of trial items. Overall, for ranking and multiple choice questions, the mean scores are slightly higher than those for FP 2016. For the ranking items, the mean facility value was broadly similar across papers. The lowest was Paper 14 at 15.47, and the highest was Paper 10 at 17.50. Neither of these deviates greatly from the overall mean (16.9). For the multiple choice items, the mean facility was also broadly similar across papers: the lowest was Paper 14 at 7.22, and the highest was Paper 10 at 10.61.

5.10.4 Items that are at the 'easier' end of the scale (above 90% of the total available score) are 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 23: Trial Item level statistics: Facility values

		Ranking			Multiple Choice		
	N	Mean	Facility Range	SD Range	Mean	Facility Range	SD Range
Paper 1	523	16.98	14.77-17.93	1.64-2.44	9.24	7.76-11.38	1.72-2.38
Paper 2	507	16.91	15.02-18.41	1.43-2.65	9.77	9.26-10.19	2.17-2.36
Paper 3	488	17.40	16.70-18.16	1.61-2.26	8.35	6.97-10.01	1.40-2.22
Paper 4	523	17.05	15.43-18.46	1.66-2.41	8.35	7.80-9.19	1.92-2.77
Paper 5	502	16.82	14.59-18.35	1.58-2.81	9.14	8.76-9.40	2.03-2.33
Paper 6	539	16.76	14.75-19.04	1.32-3.00	9.71	9.02-10.84	1.88-2.54
Paper 7	503	16.85	15.98-18.08	1.62-2.68	9.12	8.50-9.53	2.47-3.07
Paper 8	498	17.38	16.19-18.92	1.59-2.48	9.35	8.92-9.59	2.19-2.33
Paper 9	490	16.65	15.94-17.74	1.27-2.34	9.72	8.91-11.13	1.88-2.62
Paper 10	516	17.50	15.08-18.36	1.42-2.14	10.61	10.19-11.03	1.92-2.19
Paper 11	497	17.20	15.58-18.32	1.49-2.46	10.52	9.80-11.15	1.75-2.22
Paper 12	726	16.46	14.66-18.15	1.63-2.77	9.08	7.28-10.09	2.25-2.68
Paper 13	717	17.05	15.56-18.82	1.52-2.65	10.42	9.89-10.74	1.94-2.42
Paper 14	663	15.47	11.02-17.66	1.67-2.62	7.22	4.51-10.81	2.09-2.63
Overall (FP 2017)	7,692	16.9	11.02-19.04	1.27-3.00	9.3	4.51-11.38	1.40-3.07
Overall (FP 2016)	7,761	16.9	12.54-18.81	1.13-2.93	8.7	4.66-11.18	1.62-2.99
Overall (FP 2015)	8,045	16.9	9.5-19.2	1.40-3.28	9.1	5.4-11.6	1.18-3.05
Overall (FP 2014)	7,925	17.2	13.1-19.3	1.21-3.21	9.3	4.0-11.0	1.78-2.91
Overall (FP 2013)	8,156	16.7	13.56-19.1	1.36-3.30	8.7	6.61-10.67	2.11-3.05

5.10.5 **Item Quality**: 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 (i.e. item facility and SD), as well as how the best performing applicants performed (i.e. if the best performing applicants have a very different key from that of the SMEs then it suggests that there may be problems with the item).

5.10.6 Table 24 outlines how items performed for each of the 14 papers and overall. Trial Papers 2, 3 and 13 have the fewest items with partials above .17, and Trial Paper 3 has the largest proportion of items below .13.

5.10.7 Forty-four of the 140 (31.4%) trial items were deemed as having good psychometric properties, with regards to item quality. Thirty-eight of the 140 (27.1%) items were deemed as having moderate psychometric properties. Fifty-eight of the 140 (41.4%) items were deemed as requiring further review. Comparing this to the analysis of FP 2016 trial items (when 30.0% of items were deemed as having good

psychometric properties, 21.4% moderate and 48.6% poor) indicates a slight increase in the number of successful trial items. The proportion of items requiring further review is broadly in line with other SJTs and is an acknowledged aspect to an SJT development process. Despite all trial items undergoing thorough review, with high levels of content and face validity, item level analysis provides important information about how well the item is differentiating psychometrically. In this context in particular, item level differentiation can be reduced when the sample population is derived from a homogenous group (thereby reducing the amount of variability between applicants' performance). Given that all those who sit the SJT are seeking a place on the Foundation Programme, it could be argued that the variability in applicant performance is reduced compared to SJTs used, for example, into selection for specialty training, where applicants may be applying for multiple specialties. As such, a redundancy rate of between 40-50% is not unexpected.

Table 24: Trial item level statistics: Item partials

	Range of Item Partials	Mean Item Partial	Good (>.17)	Moderate (.13-.17)	Item requires further review (<.13)
Paper 1	.08 - .29	.16	5 (50.0%)	1 (10.0%)	4 (40.0%)
Paper 2	.03 - .16	.10	0 (0.0%)	5 (50.0%)	5 (50.0%)
Paper 3	.00 - .20	.11	1 (10.0%)	3 (30.0%)	6 (60.0%)
Paper 4	.09 - .20	.14	2 (20.0%)	3 (30.0%)	5 (50.0%)
Paper 5	.07 - .28	.14	3 (30.0%)	2 (20.0%)	5 (50.0%)
Paper 6	.00 - .28	.14	2 (20.0%)	5 (50.0%)	3 (30.0%)
Paper 7	.10 - .29	.21	8 (80.0%)	1 (10.0%)	1 (10.0%)
Paper 8	.03 - .23	.13	3 (30.0%)	3 (30.0%)	4 (40.0%)
Paper 9	.02 - .19	.13	4 (40.0%)	2 (20.0%)	4 (40.0%)
Paper 10	.03 - .27	.15	4 (40.0%)	2 (20.0%)	4 (40.0%)
Paper 11	-.03 - .21	.15	5 (50.0%)	2 (20.0%)	3 (30.0%)
Paper 12	.08 - .20	.13	2 (20.0%)	4 (40.0%)	4 (40.0%)
Paper 13	.03 - .23	.13	1 (10.0%)	4 (40.0%)	5 (50.0%)
Paper 14	-.08 - .23	.10	4 (40.0%)	1 (10.0%)	5 (50.0%)
Overall (FP 2017)	-.08 - .29	.14	44 (31.4%)	38 (27.1%)	58 (41.4%)
Overall (FP 2016)	-.01 - .26	.13	42 (30.0%)	30 (21.4%)	68 (48.6%)
Overall (FP 2015)	-.04 - .31	.14	43 (30.7%)	36 (25.7%)	61 (43.6%)
Overall (FP 2014)	-.08 - .38	.12	38 (27.1%)	26 (18.6%)	76 (54.3%)
Overall (FP 2013)	-.15 - .47	.17	61 (43.6%)	39 (27.9%)	40 (28.6%)

5.10.8 **Analysis of item level statistics across each year:** Table 25 summarises these differences between trial item statistics from FP 2013-17, split by ranking and multiple choice items. Overall, the observed item level statistics are generally stable across each year of administration.

Table 25: Summary of differences in trial item statistics: FP 2013, FP 2014, FP 2015 and FP 2016

Item type and development period	N	Facility	Facility Range	Mean SD	SD Range	Mean Partial	Partial Range	Good (>.17)	Moderate (.13-.17)	Further review (<.13)
All ranking items FP 2017	98	16.9	11.02-19.04	2.01	1.27-3.00	.13	-.08-.29	27 (27.6%)	26 (26.5%)	45 (45.9%)
All ranking items FP 2016	98	16.9	12.54-18.81	1.98	1.13-2.93	.13	.00-.26	31 (31.6%)	20 (20.4%)	47 (48.0%)
All ranking items FP 2015	98	16.9	9.5-19.2	1.99	1.40-3.28	.14	-.04-.03	30 (30.6%)	23 (23.5%)	45 (45.9%)
All ranking items FP 2014	98	17.2	13.1-19.3	1.97	1.21-3.21	.12	-.08-.27	22 (22.5%)	21 (21.4%)	55 (56.1%)
All ranking items FP 2013	98	16.7	13.6-19.1	2.23	1.36-3.30	.17	-.15-.47	45 (45.9%)	23 (23.5%)	30 (30.6%)
All MCQ items FP 2017	42	9.3	4.51-11.38	2.21	1.40-3.07	.15	-.05-.29	17 (40.5%)	12 (28.6%)	13 (31.0%)
All MCQ items FP 2016	42	8.7	4.66-11.18	2.37	1.62-2.99	.13	-.01-.25	11 (25.2%)	10 (23.8%)	21 (50.0%)
All MCQ items FP 2015	42	9.1	5.4-11.6	2.30	1.18-3.05	.14	.03-.30	13 (31.0%)	13 (31.0%)	16 (38.1%)
All MCQ items FP 2014	42	9.3	4.0-11.0	2.35	1.78-2.91	.14	-.03-.38	16 (38.1%)	5 (11.9%)	21 (50.0%)
All MCQ items FP 2013	42	8.7	6.6-10.7	2.58	2.11-3.05	.17	-.07-.42	18 (42.9%)	14 (33.3%)	10 (23.8%)

5.10.9 **Review of Trial Items:** Following further review by the WPG team of the available data (e.g. item partial of above .13, item facility below 90% and above 10%, SD between 1 and 3, in addition to the answer key of the best performing applicants), 80 (57.1%) of the items are deemed to be appropriate to enter the operational item bank. Twenty items (14.3%) were not deemed to have suitable performance statistics but have been reviewed and amended in collaboration with a clinician and have been deemed suitable to be entered into the FP 2018 item development process. Forty items (28.6%) will not re-enter the item development process. This is a slight increase in the number of successful trial items in comparison to FP 2016, where 72 (51.4%) items were deemed to be appropriate to enter the operational item bank.

5.10.10 Table 26 provides a summary of various item characteristics at the point of item trialling (n=140) and at the point of entry into the operational item bank (n=80) for FP 2017. This provides an insight into which types of items were most and least likely to be successful throughout the development, review and trialling process.

5.10.11 Table 26 illustrates that, in general, multiple choice items are performing slightly better than ranking items during the trial item analysis. With respect to the target domains, there is a similar level of success, with those written under the 'Commitment to Professionalism' and 'Patient Focus' domains demonstrating a slightly lesser tendency than the other domains to be successful after trialling.

5.10.12 Items that have previously been piloted are less likely to be successful during the review process and trial item analysis. A significant amount of time is invested into the development of each item, therefore it is important to use information gleaned from the pilot analysis to try to refine and improve suitable items that were not successful. WPG will therefore continue to review items but to remove 'Previous Pilot'

items that require further review (i.e. those that have been piloted twice; N=5 for FP 2017), as they have already been through at least one previous review and refinement process and, after two unsuccessful pilot attempts, efforts would be more appropriately placed into the development of new pilot items.

5.10.13 No items were developed asking applicants to rank the extent to which they agreed with difference statements ('Agreement' items), with no items of this type being trialled. Therefore, it is still not possible to draw any particular conclusions about this item type. Consequently, this item type will continue to be developed until more data are available to establish its overall performance.

5.10.14 For FP 2018, WPG will continue to develop items with a similar split of all item types, however it is acknowledged that the split each year is partly dependent upon the topic content of items developed.

Table 26: Summary of proportions of items that were successful at different stages of the item development and trialling process

	Total	Item Type		Target Domain					Ranking lead-in				Development Process	
		Ranking	Multiple Choice	Commitment to Professionalism	Coping with Pressure	Effective Communication	Patient Focus	Working Effectively as Part of a Team	Actions	Agreement	Considerations	Order	Previous pilot	Newly developed
Trialled	140	98	42	33	20	20	31	36	88	0	7	3	21	119
To enter operational item bank *(% of those trialled*)	80 (57.1%)	51 (52.0%)	29 (69.0)	15 (45.5%)	13 (65.0%)	12 (60.0%)	16 (51.6%)	24 (66.7%)	46 (52.3%)	0 (0.0%)	2 (28.6%)	3 (100%)	16 (76.2%)	64 (53.8%)

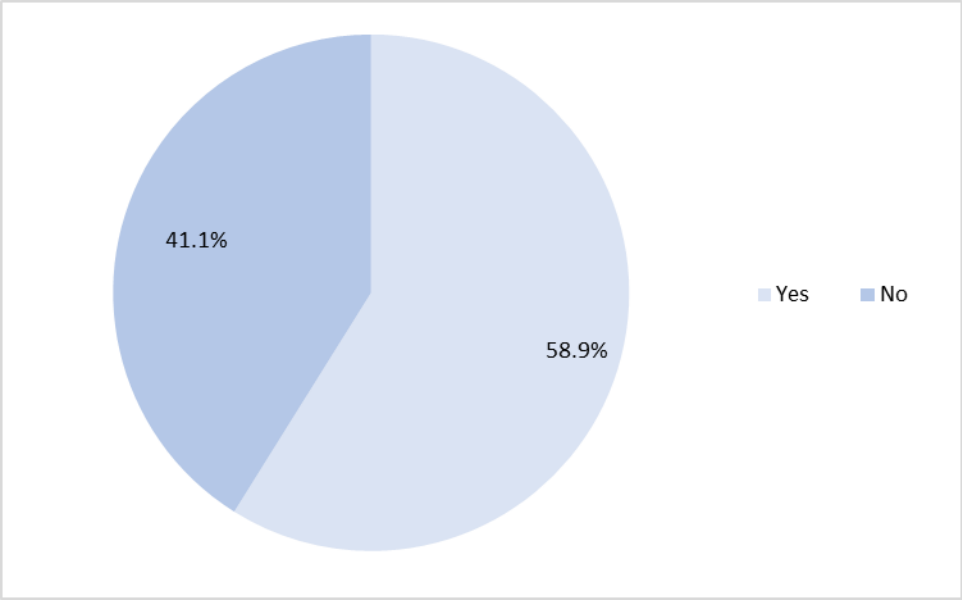
*Includes those which were re-written/refined following previous pilots

5.11 Applicant Reactions

5.11.1 All applicants who participated in the SJT were asked to complete an evaluation questionnaire regarding their perceptions of the SJT. There was a reduction in the number of questions asked of applicants, in comparison to previous years and a change in response format. A total of 7,303 applicants (97.4% of applicants) completed the questionnaire, however due to a technical issue with the new system, only 4,698 responses were collected (60.9%).

5.11.2 Applicants were asked to indicate whether they thought that the content of the SJT appeared to be fair for selection to the Foundation Programme, the results of which are shown in Figure 7. Of those who responded, 2,768 agreed (58.9%) that the content was fair and 1,930 disagreed (41.1%). In FP 2016, 37.7% of applicants either agreed or strongly agreed with this statement, 31.3% neither agreed nor disagreed and 31.5% disagreed or strongly disagreed with this statement. Applicant reaction data are typically collected on a scale to allow differentiation or strength of opinion to be captured. The use of a dichotomous scale can potentially mask candidates' true perceptions and should be interpreted with caution (e.g. a relatively large number of candidates in previous years' administrations neither agreed nor disagreed with this statement).

Figure 7: Applicant Evaluation Responses



Part Three: Summary & Recommendations

6 Summary

6.1 This report details the operational use of the SJT for selection to FP 2017, as well as the development of new items which were trialled alongside FP 2017.

6.2 Test completion analysis revealed that the majority of applicants (99.2%) answered all items within the test, indicating that the time limit of 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 differentiate sufficiently between applicants. Test level analysis demonstrates that the three different versions of the test were broadly similar, however test equating techniques are undertaken to ensure equivalence in scores across the different versions.

6.4 The mean reliability for FP 2017 slightly increased to 0.75 (FP 2016 = 0.74; FP 2015 = 0.71; FP 2014 = 0.69; FP 2013 = 0.72) and operational item level analysis revealed an increase in the number of items of a high quality, with 88.9% classed as good or moderate in terms of their psychometric properties. These results are indicative that the Foundation Programme SJT is a well-established and robust test.

6.5 Review of the operational analysis of previous administrations of the FP SJT had indicated that the way in which applicants are responding to items each year was resulting in increased item facilities at the item level and an increased mean percentage correct at the test level. However, in FP 2017, the mean score decreased slightly, which may have been a result of item content having been developed with an increased level of complexity for the last two years, with the level of difficulty being a focus of review within SME item review workshops to ensure appropriateness. The spread of scores suggests that the test is still differentiating between applicants and the SD (i.e. spread of scores) is comparable to that observed in the previous three administrations of the SJT.

6.6 Group differences analysis reveals significant differences between performance in the test based on ethnicity, country of medical education, age and gender. Female applicants outperformed male applicants; White applicants outperformed BME applicants; applicants from UK-based medical schools outperformed applicants from non-UK-based medical schools; and younger applicants outperformed older applicants. For gender and age, these effects were small. The effects for White versus BME applicants were large, demonstrating an observed increase in the size of this effect for FP 2017. Similar differences in applicant performance according to ethnicity have been observed for both undergraduate and postgraduate assessments in medical education[23]. Test content is unlikely to be the only explanation for this difference, which could be due to a number

[23] Menzies L, Minson S, Brightwell A, Davies-Muir A, Long A, Fertleman C. (2015). An evaluation of demographic factors affecting performance in a paediatric membership multiple-choice examination. *Postgraduate Medical Journal, 91,* 72-76.

Wakeford R, Denney ML, Ludka-Stempien K, Dacre J, McManus C. (2015). Cross-comparison of MRCGP & MRCP(UK) in a database linkage study of 2,284 candidates taking both examinations: assessment of validity and differential performance by ethnicity. *BMC Medical Education,* 15, 1.

of complex social factors. For example, there may be bias in terms of which groups are getting access to support or obtaining funding to access coaching courses for the SJT; a second freely available practice paper will be developed ahead of FP 2018, to increase the preparation material available to all applicants. In addition, experiences during undergraduate training, both on the wards and in medical school (e.g. negative stereotyping from colleagues and teachers), can contribute to the differential attainment often observed[24]. The observed effect was also large for UK versus non-UK applicants. The performance of applicants who have received their medical education outside of the UK may be affected by a lower fluency in the English language or differences in the working cultures of the healthcare systems in a different country of medication education.

6.7 Significant correlations were found between SJT scores and EPM decile scores, and between SJT scores and total EPM scores. Whilst these correlations are significant, indicating a degree of shared variance/commonality between the assessment methods, there is also a large amount of the variance that is not explained, indicating that the SJT appears to be assessing different constructs from the EPM. This is consistent with the findings of the initial predictive validity study for selection to the Foundation Programme[25].

6.8 One hundred and forty items were trialled alongside the operational items during FP 2017; 57.1% of these items are deemed to be appropriate to enter the operational item bank and 14.3% have been amended and re-entered into the trial item development process for FP 2018. These items will be entered at the review workshop stage, ensuring a group of SMEs review the item before it is re-piloted.

6.9 Only one question was asked of applicants in FP 2017, which explored whether they thought that the content of the SJT was fair. An alternative response scale was used this year. Thirty-eight percent of applicants either agreed or strongly agreed with this question in FP 2016, however 59% answered 'yes' to this question in FP 2017.

[24] Woolf, K., Cave, J., Greenhalgh, T, Dacre, J. (2008). Ethnic stereotypes and the underachievement of UK medical students from ethnic minorities: qualitative study British Medical Journal; 337.

[25] Cousans, F., Patterson, F., Edwards, H., McLachlan, J.C. & Good, D. Evaluating the Complementary Roles of an SJT and Academic Assessment for Entry into Clinical Practice. *Advances in Health Sciences Education* https://doi.org/10.17863/CAM.4578

7 Recommendations

7.1 Re-validating SJT Competencies

7.1.1 In 2011, a systematic, multi-method job analysis was conducted to define the professional attributes expected in the role of an F1 doctor. It is best practice to revisit this regularly, to ensure that it accurately reflects the competencies required for success in a role. Following recent political tensions in relation to the role, in addition to a focus on recruiting for values in the National Health Service, it may be timely to re-validate the competencies that that were previously deemed to be important for success in the role to ensure that they are still relevant and important. Consideration should also be given to whether there are any additional attributes that it might be beneficial for the SJT to assess.

7.2 Item Writing Methodology

7.2.1 A mixed methodology for item writing was used for FP 2017. It is recommended that this mixed methodology, comprising both item development telephone interviews and item writing workshops, continues to be used for FP 2018. During the workshops, a trained item developer introduces the clinicians to item writing principles, and outlines the parameters in terms of the type of items that are sought to be generated to ensure an appropriate spread of domain/topics. In previous item development cycles, certain topics were over-represented during the telephone interviews. Holding workshops allowed a trained facilitator to encourage the development of a broader range of scenarios, tapping into each of the domains and a broad range of item topics. A core group of item writers should be sought to consistently participate each year, with the aim of developing the SJT item writing expertise of these individuals, thus seeking to improve the quality of items developed annually.

7.3 Reducing the Impact of Group Differences

7.3.1 Following the observed group differences in FP 2017, care should continue to be taken to minimise any unconscious bias throughout the item development process, e.g. through training item writers and reviewers not to use colloquial language. The demographic representation of SMEs involved in all stages of item development should also be monitored, from idea generation through to the concordance panels. Consideration could be given around how to target and incentivise SMEs from minority groups to encourage them to participate in the item development process.

7.4 Decreasing Homogeneity of Pilot Paper Samples

7.4.1 Currently, pilot papers are not randomly allocated to applicants, with each pilot paper being sat by a small number of universities (2-3). This may be having some impact on the difference observed in the psychometric properties of pilot items across the different papers. The pilot analysis may be of more value if the pilot samples were more reflective of the overall population, therefore consideration could be given to logistical arrangements that would enable the random allocation of pilot papers across medical schools. If the SJT were to be administered online, this would be a good opportunity to review the current administration process that determines the allocation of pilot papers.