Work Psychology Group
Thinking differently

Analysis of the Situational Judgement Test for Selection to the Foundation Programme 2023

Technical Report

October 2023

Melissa Washbrook
Lana Delic
Emma Dougan
Professor Fiona Patterson

1.   Executive Summary

Overview

1.1.   The aim of this project was to develop, implement and evaluate a Situational Judgement Test (SJT) as part of live selection into The Foundation Programme (FP) for 2022-23. This is built upon the initial pilot conducted in 2021 and operational delivery in 2022, following the introduction of a computer-based SJT using a revised test specification, for selection into Foundation Year One (FY1) training. The SJT, in combination with the Educational Performance Measure (EPM), was used to rank candidates applying for FY1 training and allocate them to foundation schools.

1.2.   The objectives of this project were to:

- Develop an operational SJT for live use in 2022-23, to support selection of candidates into the Foundation Programme.

- Continue to test a bank of SJT items based on the agreed test specification.

- Evaluate the SJT in terms of test and item performance, including reliability, validity and fairness.

- Develop additional practice material to support candidates in preparing for the SJT.

1.3.   The Foundation Programme (FP) SJT was delivered for selection to the FP 2022-23 during two testing windows which lasted from the 7th to the 20th of December 2022 and from the 18th to the 23rd of January 2023, with contingency dates held from 25-30th January 2023. In total, N=9,692 candidates sat the SJT; n=3,174 completed operational Paper A, n=3,216 completed operational Paper B, and n=3,302 completed operational Paper C.

1.4.   The exam was delivered both at PearsonVUE (PV) testing centres and using PV's OnVUE online testing solution. This allowed the SJT to be delivered directly to candidates in a home setting supported by PV online proctoring, negating the need to travel to a test centre, for those who were unable to attend.

1.5.   The main sections of this report outline the test development process and details evaluation results of the operational SJT used during the FP 2023 National Recruitment Process.

Analysis

1.6.   The psychometric analysis of the 2022-23 operational SJT is positive and shows consistency when compared to previous versions of the SJT for entry into the FP. The results show good evidence that the test specification is suitable for this context.

1.7.   The SJT demonstrated an overall excellent level of internal reliability (α=.860 Paper A; α=.838 Paper B; α=.830 Paper C), which is appropriate for tests administered in high stakes selection context such as the Foundation Program. The SJT was capable of differentiating between

candidates, providing a sufficient spread of scores to support decision making as part of selection into the Foundation Program.

1.8. Candidates were allowed 140 minutes to complete the 75-scenario test (which includes 9 pilot scenarios). The test completion analysis showed that the test was not speeded, with 99.7% of candidates completing the last question on Paper A, 99.9% of candidates completing the last question on Paper B, and 99.8% of candidates completing the last questions on Paper C.

1.9. In relation to our Equality, Diversity and Inclusion (ED&I) analysis, the SJT results show significant differences for gender (small effect size), ethnicity (large effect size), and place of education (UK or International) (large effect size). Differences based on ethnicity were still observed, though the differences were smaller, when place of education was controlled for (moderate effect size).

1.10. The EPM results also show significant differences for gender (small effect size), and ethnicity (small effect size). Similar to the SJT results, differences based on ethnicity were still observed for the EPM results, though the differences were smaller, when place of education was controlled for (small effect size). In some cases, the differences seen may be exacerbated due to the uneven sample size within subgroup categories. No differences were observed in EPM scores based on place of education (UK or International).

1.11. Candidate feedback was generally positive with regards the contents and relevance to the FY1 role, though there was less agreement in terms of perceptions of fairness and the of difficulty of the test. Most candidates found the question format appropriate although there was less consensus on the suitability of speech responses.

Table of Contents

2.   Introduction

Overview & Objectives

2.1.   A Situational Judgement Test (SJT) has been used for selection into Foundation Year One (FY1) Training for the past 9 years. The SJT, in combination with the Educational Performance Measure (EPM), has been used to rank candidates applying for FY1 training and allocate them to foundation schools. Since July 2019, Work Psychology Group (WPG) have been working in partnership with the UK Foundation Programme Office (UKFPO) to develop, implement and evaluate a computer-based Situational Judgement Test (SJT) as part of live selection into FY1 Training. This report aims to evaluate the performance of the SJT, following its operational used in December 2022 – January 2023.

2.2.   The objectives of this project were to:

- Develop an operational SJT for live use in 2022-23, to support selection of candidates into FY1.

- Continue to test a bank of SJT items based on the agreed test specification.

- Evaluate the SJT in terms of test and item performance, including reliability, validity and fairness.

- Develop additional practice material to support candidates in preparing for the SJT.

2.3.   The main phases of this project have consisted of:

- Confirmation of the Test Specification

- Item Development

- Operational Test Construction

- Scoring and Psychometric Operational Analysis

- Reporting

2.4.   The main sections of the current report outline the test development process and provide the evaluation results of the operational SJT used during the FP 2022-23 National Recruitment Process.

3. Test Development

Confirmation of the Test Specification

3.1. The FP 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 FP selection process for entry to the FP in 2013.

3.2. As part of the ongoing development of the FY1 SJT, an investment was made in 2019 to develop a new computer-based SJT. This provided an opportunity to enhance candidate engagement by introducing new SJT item types and multimedia elements, ensuring the SJT continues to remain innovative whilst still retaining its good quality psychometric properties. This process involved a number of different development stages, and input from a range of stakeholders and Subject Matter Experts (SMEs). The SJT was originally piloted in January 2020 to determine the suitability of question and response types identified by WPG and has been used operationally since 2021. These draw upon the latest research as well as WPG's expertise in assessment design in high-stakes environments.

3.3. The FP SJT is designed to assess five of the nine attributes from the FP person specification: Commitment to Professionalism, Coping with Pressure, Patient Focus, Effective Communication and Working Effectively as Part of a Team[1]. These attributes are detailed in Table 1.

Table 1: Target Attributes

Commitment to Professionalism. *Takes responsibility for own actions. Displays honesty, integrity, awareness of confidentiality and ethical issues. Demonstrates motivation and desire for continued learning.*
Coping with Pressure. *Capability to work under pressure and remain resilient. Demonstrates ability to adapt to changing circumstances and manage uncertainty. Remains calm when faced with confrontation. Develops and employs appropriate coping strategies and demonstrates judgement under pressure. Demonstrates awareness of the boundaries of their own competence and willing to seek help when required, recognising that this is not a weakness. Exhibits appropriate level of confidence and accepts challenges to own knowledge.*
Patient Focus. *Ensures patient is the focus of care. Demonstrates understanding and appreciation of the needs of all patients, showing respect at all times. Takes time to build relationships with patients, demonstrating courtesy, empathy and compassion. Works in partnership with patients about their care.*

---

[1] See FY1 Job Analysis report 2011 for full details of how attributes were derived and what comprises each attribute (https://isfp.org.uk/final-report-of-pilots-2011/).

Effective Communication. *Actively and clearly engages patients and colleagues in equal/open dialogue. Demonstrates active listening. Communicates verbal and written information concisely and with clarity. Adapts style of communication according to individual needs and context. Able to negotiate with colleagues and patients effectively.*

Working Effectively as Part of a Team. *Capability and willingness to work effectively in partnership with others and in multi-disciplinary teams. Demonstrates a facilitative, collaborative approach, respecting others' views. Offers support and advice, sharing tasks appropriately. Demonstrates an understanding of own and others' roles within the team and consults with others where appropriate.*

3.4. Key elements of the test specification framework include:

3.4.1 **Test Purpose.** To design and implement an SJT to be used as part of the live selection process and to be weighted equally with the EPM to determine candidate rankings.

3.4.2 **Test Content.** The scenarios are set within the context of the FP, but do not require prior experience FY1 training. The scenarios do not aim to assess clinical knowledge or facts but are pitched at a level that candidates will feel some degree of challenge.

3.4.3 **Item Types and Response Formats.** Three item types are used; rating, multiple choice, and ranking. Candidates are asked what they should do in response to the situation presented. The papers were split into three sections, based on the three different response formats:

- **Rating:** Candidates were asked to independently rate each of the 4-8 response options, in order of their appropriateness or importance, in responding to the situation (e.g., Rate the importance of the following considerations in the management of this this situation).
- **Multiple choice:** Candidates were asked to select the three most appropriate response options, from the 8 presented, which together will best resolve the situation presented (e.g., Choose the THREE most appropriate actions to take in this situation).
- **Ranking**: Candidates were asked to rank the 5 response options presented in order of their appropriateness or importance in response to the situation on a scale from one to five (e.g., 1= Most appropriate; 5= Least appropriate).

3.4.4 Within each section, there were a range of different response types. The three response types are summarised below:

- **Actions:** Candidates were asked to judge the appropriateness of a range of actions in response to the given situation.
- **Considerations:** Candidates presented with a list of considerations and asked to judge how important each consideration is in the management of the given situation.

- **Speech:** Candidates were presented with a series of speech responses (i.e., quotes) and asked to judge the appropriateness of these in the given conversation.

3.4.5   Throughout the test, there were some 'evolving' scenarios, comprised of up to 3 scenarios, which are linked by a common context. Candidates respond to each scenario independently, as new information is presented, but each of the scenarios is related to one another (e.g., may relate to the same patient or same colleague). These scenarios are therefore considered to be more representative of real workplace dilemmas, which tend to be multi-faceted. Clear instructions are provided to ensure it is clear to candidates when a scenario is going to have multiple parts.

3.4.6   **Test Length.** Three papers, each consisting of 75 scenarios (66 operational items; 9 pilot items) to be completed in 140 minutes.

4.   Item Development

4.1.   Trialling of new items takes place alongside the operational SJT each year, 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. Figure 1 below summarises the development and review process undertaken for new items that were trialled alongside operational delivery in 2023.

Figure 1: The development and review process for trial items

Item development:
Item Writing Workshops

↓

WPG item review

↓

Item review workshops

↓

New items piloted
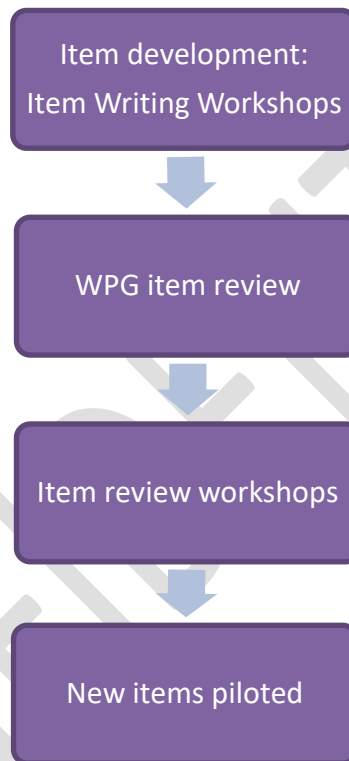
4.2.   The process allowed for the development of enough items that at each stage, if an item was not performing, it could be made redundant.

4.3.   Scenarios were developed in collaboration with Subject Matter Experts (SMEs) from a range of specialties to ensure that the SJT is relevant for all candidates entering FY1 Training. Item writing workshops were held with SMEs, with an aim for clinicians to develop SJT item content. Prior to each workshop, SMEs were asked to spend some time in preparation 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, wrote a number of scenarios and responses. Using item writing workshops has a number of benefits, including: efficient generation of a large number of items; the opportunity for SMEs to work together and gain ideas from each other to form 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. 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.

4.4.   Following the item development workshops, WPG conducted internal reviews of each SJT item, to ensure they were of high quality based on the best-practice principles of SJT item writing, and to ensure that they were suitable based on the test specification.

Item Review

4.5.   Item review workshops were held in June 2022, to ensure that all SJT items developed as part of item development process were thoroughly reviewed by SMEs with the appropriate expertise, prior to piloting. More items than were needed were brought to the review workshops so that some could be dropped while still ensuring there were enough remaining items to be taken through to the concordance stage.

4.6.   Typically, concordance analysis is conducted to ensure the response keys (answers) for the SJT items are finalised based on expert consensus. Due to limited SME availability in participating in the concordance stage, it was not possible to conduct this stage prior to piloting the newly developed content. Concordance was planned to take place following the operational test to gain SME input, however due to the change in the selection process in 2024 (and the SJT no longer being used), this stage was not conducted.

Practice Material

4.7.   In additional to developing items for operational use, a full-length practice paper, which followed the same structure as the operational test with a range of question types, was also developed with input from SMEs. The practice paper provided candidates with the correct answer key and rationale statements to explain the answer to each question, to support candidates in preparing for the SJT. This was available to candidates on the UKFPO website.

4.8.   A practice paper was also hosted online by PearsonVue to help familiarise candidates with the structure of the SJT, which offered a very similar experience to the operational test.

5. Operational Test Construction

5.1.   The operational delivery of the FY1 SJT required the production of three sufficiently equivalent test versions, which allowed the equating of scores to ensure that each test version was of comparable difficulty.

5.2.   The strategy for creating three versions maximised the use of the operational item bank and diversity of items across versions, whilst retaining sufficient overlap ('anchor items') to enable equating. The three versions were developed to be as similar as possible in terms of content parameters.

5.3.   Each operational test version consisted of 66 operational scenarios (14 Rating, 20 Multiple Choice, and 32 Ranking scenarios).

5.4.   Candidates also answered 9 pilot SJT scenarios (4 Rating, 2 Multiple Choice, and 3 Ranking scenarios), which did not contribute to their overall SJT score. To allow for sufficient piloting of new content, there were 9 forms created in total, each with a different set of pilot items.

5.5.   Item keys were pre-determined based on the item writer key, concordance key and piloting or previous operational data available. There was a maximum of 3 or 4 points for each Rating item (dependant on the key), 12 points for each Multiple Choice scenario (points awarded for each correct option identified), and a maximum of 20 points available for each Ranking scenario, based on how close responses were to the key.

5.6.   Papers were developed to be as similar as possible based on content, difficulty, psychometric properties, and balanced across the target attributes. Table 2 provides a breakdown of the number of operational scenarios within each target criteria in each version.

Table 2: Number of scenarios within each target attribute

	Commitment to Professionalism	Coping with Pressure	Patient Focus	Effective Communication	Working Effectively as Part of a Team
Paper A	14	14	13	12	13
Paper B	14	14	13	11	14
Paper C	15	14	13	11	13

5.7.   Supporting documents for the SJT administration were also produced (e.g., instructions for candidates). These were integrated into the computer-based system provided by Pearson VUE. Pearson VUE also provided candidates with the option to complete a tutorial, before the test began, demonstrating how to answer questions using the 'drag and drop' format.

6. Psychometric Analysis: Operational

Candidate Sample

6.1.   In total, N=9,692 candidates sat the 2023 SJT during two testing windows which lasted from the 7th to the 20th of December 2022 and from the 18th to the 23rd January 2023. For those candidates who had extenuating circumstances and were required to complete the test outside the testing window, contingency dates were held from 25-30th January 2023.

6.2.   Across the papers, n=3,174 completed operational Paper A, n=3,216 completed operational Paper B and n=3,302 completed operational Paper C.

6.3.   The majority of candidates provided demographic data. With regards to gender, 57.2% (n=5,540) of the sample indicated that they were female, 38.4% (n= 3,723) indicated that they were male, and 4.4% (n=429) did not specify or their data was unavailable. Breakdowns of the candidates' ethnicity and place of education are provided in Tables 3 and 4, respectively.

Table 3: Breakdown of Candidates' Ethnicity

White	Asian	Black	Mixed	Other	Unavailable
4520 (46.6%)	3168 (32.7%)	531 (5.5%)	449 (4.6%)	362 (3.7%)	662 (6.8%)

Table 4: Breakdown of Candidates' Place of Education

Educated within the UK	Educated outside of the UK	Unavailable
8385 (86.5%)	1262 (13.0%)	45 (0.5%)

Test Level Results

6.4.   Table 5 reports the descriptive statistics for the three operational FY1 2023 SJT papers, using raw scores.

Table 5: Descriptive Statistics of Raw Data for Papers A, B and C

	SJT Paper A	SJT Paper B	SJT Paper C
Total N	3174	3216	3302
Mean score	888.53	890.04	878.01
Maximum possible score	1058	1065	1055
Mean score as %	83.98%	83.57%	83.22%
Standard deviation	41.86	39.61	37.95
Range	286-974	287-970	378-962
Reliability	.860	.838	.830

Reliability

6.5. Reliability refers to the extent to which assessments are consistent – for example, the internal reliability of a test assesses the consistency of results across items within a test. The values for reliability coefficients range from 0 to 1.0. A coefficient of 0 means no reliability and 1.0 means perfect reliability. Since all tests have some error, reliability coefficients never reach 1.0.

6.6. A commonly accepted rule of thumb for describing internal reliability or internal consistency, using Cronbach's alpha, is as follows[2]:

Cronbach's alpha	Internal consistency
$\alpha \geq 0.8$	Excellent
$0.7 \leq \alpha < 0.8$	Good
$0.6 \leq \alpha < 0.7$	Acceptable
$0.5 \leq \alpha < 0.6$	Weak
$\alpha < 0.5$	Unacceptable

6.7. Following best-practice procedure, two items were removed from Paper A and C prior to scoring based on their psychometric performance and detraction from the overall reliability of each paper.

6.8. All three operational papers showed excellent levels of internal reliability (Paper A α=0.860; Paper B α=0.838; and Paper C α=0.830), which is the desired level of reliability for an operational test. This is consistent with the reliability observed in 2022 (Paper A α=0.842; Paper B α=0.846; and Paper C α=0.837).

---

[2] Kline, P. (2000). The handbook of psychological testing (2nd ed.). London: Routledge.

Test Difficulty

6.9.   The difficulty level for Operational Paper A is 83.98% (i.e., mean score of 888.53 out of a total possible total raw score of 1058), Paper B is 83.57% (mean score of 890.04 out of a possible total raw score of 1065) and Paper C is 83.22% (mean score of 878.01 out of a possible total raw score of 1055). This indicates that the three paper versions exhibit comparable levels of difficulty, which is consistent with the difficulty level observed in the 2022 operational FP SJT.

Timing Analysis

6.10.   The standard time allowed for completion of the SJT was 140 minutes. For Paper A, 99.7%, Paper B 99.9% and Paper C 99.8% of candidates completed the last operational question. These findings indicate that the time allowed to complete the test is sufficient.

Distribution of Scores

6.11.   SJT total scores for operational Paper A, B and C showed a close to normal distribution, although all three samples are slightly negatively skewed, suggesting candidates tend to score towards the higher end of the scale (see Figures 2, 3 and 4 below).

Figure 2: Distribution of SJT Scores in Paper A

Histogram
Paper: 2023_A

Mean = 888.53
Std. Dev. = 41.856
N = 3,174
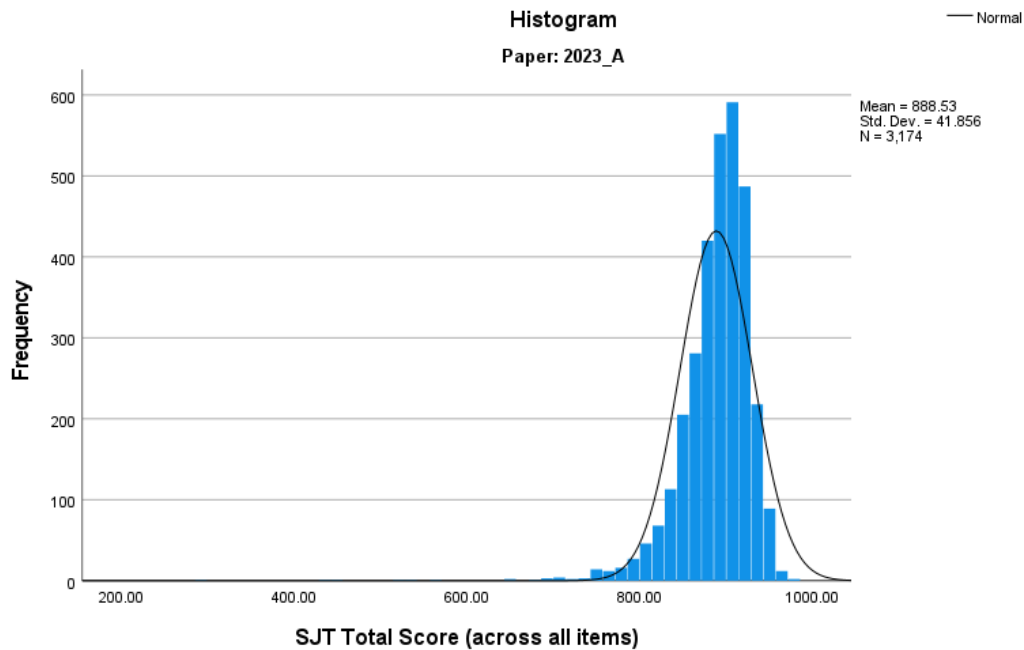
Figure 3: Distribution of SJT Scores in Paper B

Histogram

Paper: 2023_B

Mean = 890.04
Std. Dev. = 39.61
N = 3,216
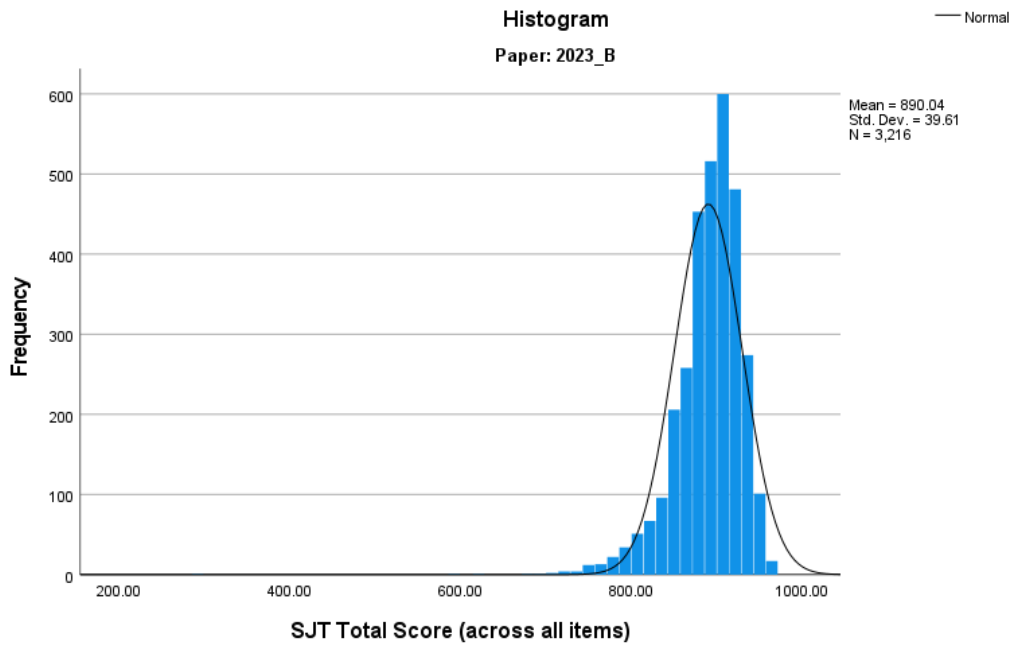
Figure 4: Distribution of SJT Scores in Paper C

Histogram

Paper: 2023_C

Mean = 878.01
Std. Dev. = 37.947
N = 3,302
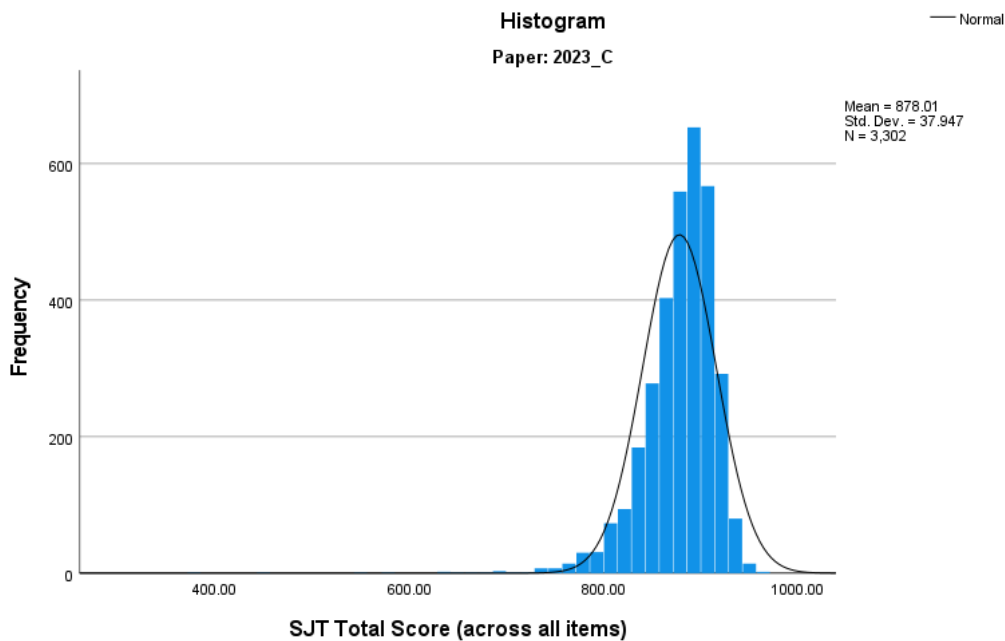
Test Equating

6.12.   While the three test versions used were developed to be as similar as possible in terms of content, statistical equating procedures are required to balance variation across papers caused by measurement error. Without this, it is not possible to determine whether small differences in scores between versions relate to random differences in populations assigned to a version or

differences in difficulty. In practice, observed differences will be a function of both sample and test differences.

6.13. There are several approaches to equating. In this instance, a chained linear equating process was used. The test papers were designed with specific overlaps ('anchor items') which could be used to compare populations and link the different test versions. The performance on the identical items enables estimation of the difference in ability between the three groups and these can be used to rescale the scores on the unique portion of Paper B and C to the scale of Paper A.

Item Level Results[3]

6.14. Item analysis was used to examine the facility (difficulty) and quality (effectiveness) of individual SJT items. For all three papers, the majority of items performed effectively and contributed to test performance.

Item Facility

6.15. Item facility is determined by the mean score for each item. Item facilities, split by paper version, are shown in Table 6.

Table 6: Item Facility by Paper Version

Paper	Rating			Multiple Choice			Ranking		
	Mean Facility	Min	Max	Mean Facility	Min	Max	Mean Facility	Min	Max
A	2.95	1.20	4.00	9.39	7.29	11.55	17.52	15.14	19.53
B	2.91	1.17	3.99	9.58	8.20	11.54	17.30	14.73	19.25
C	2.88	1.10	4.00	9.02	7.08	11.12	17.50	15.00	19.55

6.16. Overall, these results show that item facilities for items included in each version of the test were similar.

Item Quality

6.17. Item quality or effectiveness is determined by examining the item partial coefficient, which is the degree of correlation between the item and the overall mean SJT score (the mean SJT score excludes the item itself). The quality of SJT items is established according to the following four categories:

- **Good** = Correlation of **.25 or higher** between performance on the item and overall test score
- **Satisfactory** = Correlation of **.17 to .24**

---

[3] The data of a small number of candidates (n=24) who were extreme outliers are not included within this section of the report.

- **Moderate** = Correlation of **.13 to .16**
- **Limited** = Correlation of **.12 or below**

6.18. Item quality, split by paper version, is provided in Table 7.

Table 7: Summary of Item Quality by Paper Version

	Paper A	Paper B	Paper C	Paper A	Paper B	Paper C	Paper A	Paper B	Paper C
	Rating Items			Multiple Choice Items			Ranking Items		
Mean	**.15**	**.14**	**.15**	**.22**	**.22**	**.20**	**.23**	**.23**	**.20**
Good	19%	12%	10%	35%	35%	25%	34%	44%	22%
Satisfactory	21%	20%	35%	45%	40%	45%	47%	44%	44%
Moderate	23%	22%	13%	15%	15%	10%	16%	3%	25%
Limited	38%	46%	42%	5%	10%	20%	3%	9%	9%

6.19. Those items that were classified as limited did not detract from the psychometric quality of the test, and so remained in the test.

6.20. The overall item quality for the 2023 operational test is consistent with that observed in 2022. The mean partials for Rating scenarios in 2023 (.15, .14, .15) did not change significantly from 2022 (.15, .15, and .14). The item quality of Multiple Choice scenarios in 2023 (.22, .22, .20) is similar with that observed in 2022 (.22, .23, and .22). Similarly, the item quality for Ranking scenarios in 2023 (.23, .23, .20) are consistent with that observed in 2022 (.20, .21, .22).

6.21. Across all papers, rating items had lower average partials (Paper A and C, .15, Paper B, .14). There are several things to consider when interpreting this finding:

- Rating items may be a slightly different assessment of the target attributes than the other item types. The rating section represents a smaller proportion of the total marks available, therefore it is not surprising that they are less predictive of overall performance.
- Rating items may also have less variance than other formats. While there are fewer marks available, they take less time to complete per item (as each scenario includes between 4 and 8 items). Moreover, particularly poor items can be removed from the scenario, to improve the overall quality of the scenario. As such, considering the benefits and shorter timeframe needed, the rating section is still a valuable part of this test. The quality of items will continue to be monitored in future.
- By design, the SJT now has more variety in terms of item types and response types (e.g., speech-based responses) than previous iterations. Despite this, the reliability has remained high.

7.   Equality, Diversity, and Inclusion (ED&I) Analysis

Equality, Diversity and Inclusion (ED&I) analysis was conducted to investigate group difference in performance on the SJT and EPM within the candidate sample on the basis of gender, ethnicity, and place of education. In order to examine fairness issues regarding the SJT and the EPM, analysis was conducted on the equated SJT scores and total EPM scores, after outliers (candidates with very low scores and high missing data) and those whose demographic data was unavailable, were removed.

Differences in Performance on the SJT

7.1.   **Gender**: Table 8 shows differences in performance on the SJT based on gender. An independent t-test showed **a significant difference in performance on the SJT between female and male candidates** (t(7645.40)=-11.54, $p<.001$), with female candidates scoring significantly higher than male candidates. Cohen's d[4], which quantifies the magnitude of the difference between the mean SJT scores for males and females, shows a small effect size (d=.25). This is consistent with the 2022 operational results where a small effect size was also found (d=.24) and in line within other similar SJTs used for selection into healthcare roles.

Table 8: Gender

	Female	Male
N	5518	3709
Mean equated SJT total	893.12	884.39
Std. Deviation	34.44	36.43

7.2.   **Ethnicity:** Table 9 shows differences in performance on the SJT based on ethnicity. On initial glance, there are observable differences in mean scores between specified ethnic groups, with candidates who described themselves as 'White 'or 'Mixed' having higher mean SJT scores compared to other ethnic groups. **A one-way ANOVA found a significant overall effect of ethnicity on SJT scores** (F(4,8992)=355.64,$p<.001$). Eta-squared[5], which is a measure of effect size, shows a large effect (η^2=.14). The size of the effect has increased slightly compared to 2022, when a moderate effect size was observed (η^2=.12).

---

[4] 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: 0–0.19= negligible, 0.20–0.49= small effect, 0.50–0.79= moderate effect and 0.80+ = large effect.

[5] Eta-squared is a measure of effect size that is commonly used in ANOVA models. It measures the proportion of variance associated with each main effect and interaction effect in an ANOVA model. The guidelines (proposed by Cohen, 1988) for interpreting the eta-squared are: 0.01 indicates a small effect, 0.06 indicates a moderate effect, and 0.14 indicates a large effect.

7.3. Post-hoc testing (Tukey HSD) revealed that:

- Candidates who described themselves as 'White' performed significantly better on the SJT compared to those candidates who described themselves either as 'Asian' ($p<.001$), 'Black' ($p<.001$), 'Mixed' ($p<.001$), or 'Other' ($p<.001$).

- Additionally, candidates who described themselves as 'Asian' had significantly higher SJT scores than those describing themselves as 'Other' ($p=.004$).

- Significant differences were also found between candidates describing themselves as 'Mixed' and those describing themselves as either 'Asian' ($p<.001$), 'Black' ($p<.001$), or 'Other' ($p<.001$), with those describing themselves as 'Mixed' having significantly higher SJT scores.

- It is important to note the differing sample sizes between each group (which in some cases are small samples), meaning apparent differences between groups should be interpreted with caution.

Table 9: Ethnicity

	White	Asian	Black	Mixed	Other
N	4510	3154	525	448	360
Mean equated SJT total	902.65	877.48	873.32	892.08	871.07
Std. Deviation	28.62	35.90	34.28	32.91	40.16

7.4. **Place of Education:** Table 10 shows differences in performance on the SJT based on place of education. To ensure a reasonable sample size in each comparison category, candidates educated outside of the UK were grouped as 'International'. An independent t-test showed a **significant difference in performance on the SJT between UK and International candidates** (t(1416.08)=48.96, $p<.001$), with UK educated candidates scoring significantly higher than International candidates. The observed difference in scores represents a large effect size (d=1.95). These results are consistent with those observed in 2022 when a large effect size was also observed (d=1.82).

Table 10: Place of Education

	United Kingdom	International
N	8373	1237
Mean equated SJT total	897.01	839.34
Std. Deviation	27.71	40.04

7.5. **Ethnicity (UK only):** Ethnicity is confounded by place of education, and therefore differences in SJT scores based on ethnicity are examined for UK educated candidates only. Table 11 shows differences in performance on the SJT based on ethnicity, when controlling for place of education (UK educated only). On initial glance, there are observable differences in mean scores between specified ethnic groups, with UK educated candidates who described themselves as 'White' or 'Mixed' having higher mean SJT scores compared to other ethnic groups. A one-way ANOVA found a **significant overall effect of ethnicity on SJT scores for those candidates educated in the UK** ($F(4,7898)$= 260.51, p<.001). A moderate effect size (η^2=.12) was observed, which is consistent with the findings observed in 2022 (η^2=.10).

7.6. Post-hoc testing (Tukey HSD) revealed that:

- UK educated candidates who describe themselves as 'White' performed significantly better on the SJT compared to those candidates who described themselves either as 'Asian' (p<.001), 'Black' (p<.001), 'Mixed' (p<.001), or 'Other' (p<.001).
- Additionally, UK educated candidates describing themselves as 'Asian' had significantly higher SJT scores compared to those describing themselves as 'Black' (p=.002).
- Significant differences were also found between UK educated candidates describing themselves as 'Mixed' and those describing themselves as either 'Asian' (p<.001), 'Black' (p<.001), or 'Other' (p<.001), with those who described themselves as 'Mixed' performing significantly better on the SJT.
- It is important to note the differing sample sizes between each group (which in some cases are very small samples), meaning apparent differences between groups should be interpreted with caution.

Table 11: Ethnicity (UK only)

	White	Asian	Black	Mixed	Other
N	4261	2562	416	411	253
Mean equated SJT total	905.71	886.87	881.80	897.10	887.76
Std. Deviation	24.42	27.87	26.83	26.98	28.29

Differences in Performance on the EPM[6]

7.7. **Gender:** Table 12 shows differences in performance on the EPM based on gender. An independent t-test showed a **significant difference in performance on the EPM between female and male candidates** (t(7749.44)=-9.38, p<.001), with female candidates scoring marginally higher than male candidates. The observed difference in scores represents an effect size that is small (d=-.20), which has increased when compared to the 2022 scores where a negligible difference was found (d=-.09).

---

[6] EPM scores were unavailable for n=45 candidates.

Table 12: Gender

	Female	Male
N	5540	3723
Mean EPM total score	45.80	45.23
Std. Deviation	2.82	2.94

7.8. **Ethnicity:** Table 13 shows differences in performance on the EPM based on ethnicity. A one-way ANOVA **found a significant overall effect of ethnicity on EPM scores** ($F(4,9025)=97.63$, $p<.001$). A small effect size ($\eta^2=.04$) was observed, which is consistent with the findings from 2022 ($\eta^2=.05$)

7.9. Post-hoc testing (Tukey HSD) revealed that:

- Candidates who described themselves as 'White' had significantly higher EPM scores and those candidates describing themselves either as 'Asian' ($p=.001$), 'Black' ($p=.001$), 'Mixed' ($p<.001$), or 'Other' ($p<.001$).

- Additionally, significant differences were also found between candidates describing themselves as 'Mixed' and those candidates describing themselves as 'Black' ($p=.011$), with those describing themselves as 'Mixed' having significantly higher EPM scores.

- Furthermore, those candidates who described themselves as 'Other' had significantly higher EPM scores than those who described themselves as 'Black' ($p=.031$).

- It is important to note the differing sample sizes between each group (which in some cases are small samples), meaning apparent differences between groups should be interpreted with caution.

Table 13: Ethnicity

	White	Asian	Black	Mixed	Other
N	4520	3168	531	449	362
Mean EPM total score	46.16	44.97	44.76	45.35	45.32
Std. Deviation	2.78	2.84	2.92	2.78	2.91

7.10. **Place of Education:** Table 14 shows differences in performance on the EPM based on place of education. An independent t-test showed **no significant difference in performance on the EPM between UK and International candidates** ($t(1636.44)=1.87$, $p>.05$). A significant difference with a moderate effect size ($d=.51$) was observed in 2022, indicating the performance gap between UK and International candidates has decreased in 2023.

Table 14: Place of Education

	United Kingdom	International
N	8385	1262
Mean EPM total score	45.58	45.41
Std. Deviation	2.87	2.97

7.11. **Ethnicity (UK only):** Ethnicity is confounded by place of education, and therefore differences in EPM scores based on ethnicity are examined for **UK educated candidates only.** Table 15 shows differences in performance on the EPM based on ethnicity for those candidates educated in the UK. A one-way ANOVA found a **significant overall effect of ethnicity on EPM scores for UK educated candidates** (F(4,7909)=103.16, p<.001), with a small effect size (η^2=.05) observed. This is consistent with the findings observed in 2022 (η^2=.04).

Post-hoc testing (Tukey HSD) revealed that:

- UK educated candidates who described themselves as 'White' had significantly higher EPM scores compared to those candidates who described themselves either as 'Asian' (p<.001), 'Black' (p<.001), 'Mixed' (p=.001) or 'Other' (p<.001).

- Additionally, significant differences were found between UK educated candidates describing themselves as 'Asian' and those describing themselves as 'Black' (p<.001), with those describing themselves as 'Asian' having higher EPM scores.

- Furthermore, UK educated candidates describing themselves as 'Mixed' had higher EPM scores than those describing themselves as either 'Asian' (p=.02) or 'Black' (p<.001).

- Differences were also found between UK educated candidates describing themselves as 'Black' and those candidates describing themselves as and 'Other' (p<.001), with those describing themselves as 'Other' having significantly higher EPM scores.

- It is important to note the differing sample sizes between each group, meaning apparent differences between groups should be interpreted with caution.

Table 15: Ethnicity (UK only)

	White	Asian	Black	Mixed	Other
N	4266	2565	419	411	253
Mean EPM total score	46.16	44.93	44.33	45.37	45.38
Std. Deviation	2.78	2.81	2.67	2.79	3.00

8. Criterion Related Validity

8.1. The essential function of personnel selection and assessment procedures (e.g., psychometric tests) is to provide a means of estimating the likely future job performance of candidates. This is known as **criterion-related validity**. This can be completed in two ways, (1) examining the relationships between performance on selection processes and in-role performance data, called **predictive validity**, and (2) examining the relationships between performance in new selection methods and the existing selection processes, called **concurrent validity**. Predictive validity is a longer-term goal for analysis, and therefore, this section focuses on concurrent validity.

8.2. The most commonly used measure of validity is a *correlation coefficient*. The larger the correlation between selection and criterion variables, the more commonality there is in the constructs they are assessing. **Generally, within a selection context, a validity correlation between r=.20 and r=.35 is considered adequate, and r=.35 to r=.50 is considered good**[7] and demonstrates that there is a positive association between performance on both criteria.

8.3. The SJT and EPM are deigned to exhibit some overlap, as medical school performance is somewhat dependent on successfully demonstrating some of the professional attributes measured in the SJT. However, by design, it is expected that a large portion of variance will not be explained by the correlation, given the differences between the two measures.

8.4. The SJT showed a statistically significant, 'moderate' correlation with the EPM score (r=.30, p<.001)[8]. The results show that the SJT is related to the EPM component of the selection process, but that each component is measuring different attributes and capture a unique variance in performance, thereby making both useful elements of the overall selection process. This is consistent with the relationship observed between the SJT and EPM scores in 2022 (r=.39, p<.001).

---

[7] Evers, A., Hagemeister, C., & Hostmaelingen, A. (2013). *EFPA Review Model for the description and evaluation of psychological and educational tests*. Tech. Rep. Version 4.2. 6). Brussels: European Federation of Psychology Associations.

[8] Candidates who were considered an outlier due to very low SJT scores and high missing data (n=24) and those who did not have EPM scores available (n=45) were excluded from this analysis.

9. Candidate Feedback

9.1. Candidates who completed the operational SJT were asked to complete an evaluation questionnaire regarding their perceptions of the SJT. This feedback has been collated and reported in four key sections below. Overall, n=8919 (91.63%) of participants provided feedback. The breakdown of responses to each question can be seen in Table 16 below.
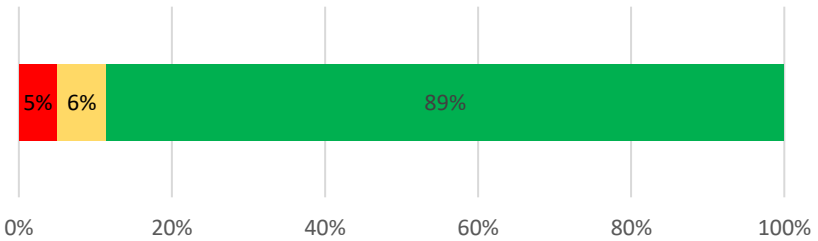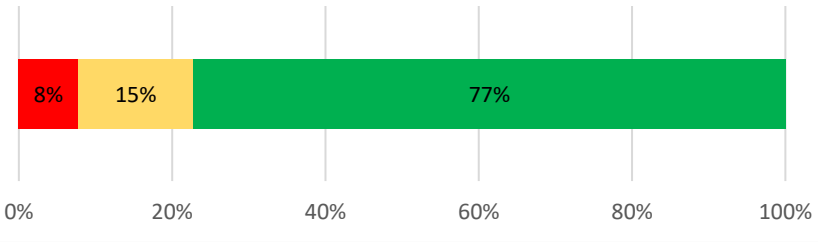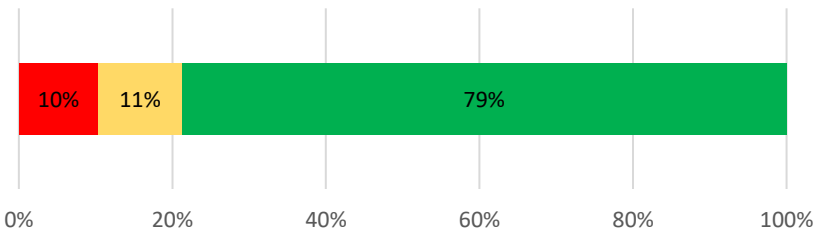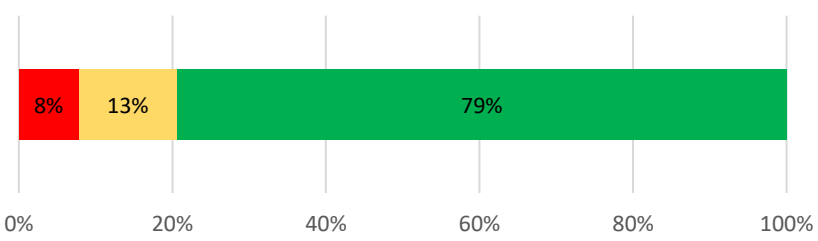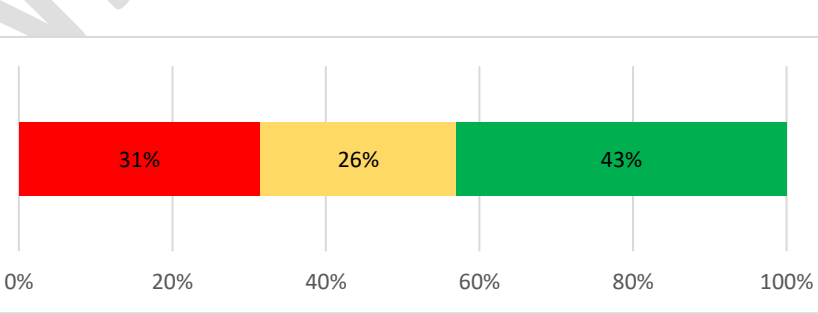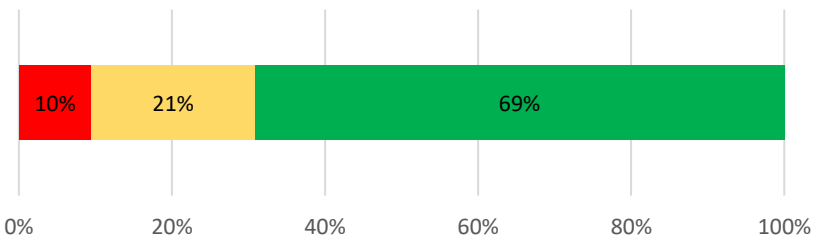
Table 16: Participant feedback on overall test content[9]

	% Disagree	% Neither Agree Nor Disagree	% Agree
The information I read in the Applicant Guide about the SJT was clear and helpful (n= 8919)	7%	19%	74%
The content of the Situational Judgement Test (SJT) was relevant to the role of Foundation Year 1 doctor (n=8883)	13%	21%	66%
The content of the SJT was an appropriate level of difficulty for my training level (n=8880)	13%	29%	57%
The content of the SJT was fair for all candidates (n=8806)	26%	31%	43%

---

[9] For each question, those that did not respond, or selected 'Not Applicable' were excluded from the analysis.

Statement	Red	Amber	Green
The instructions for the SJT were clear and easy to understand (n=8882)	10%	11%	79%
There was a sufficient amount of time to complete the test (n=8864)	16%	11%	73%
Booking the test online was straight forward (n=8821)	37%	11%	53%
I was able to book an appointment that was convenient for me (n=8843)	27%	11%	62%
I found it easy to read the information/questions on screen (n=8830)	8%	8%	84%
Computer-based testing is an appropriate way to complete the SJT (n=8825)	7%	10%	83%

The venue and facilities were appropriate (N/A if you completed at home) (n=8182)	5% / 6% / 89%
The online proctoring system was a suitable way to sit the SJT (N/A if you completed in test centre) (n=2304)	8% / 15% / 77%
The format for answering the questions was straightforward (n=8811)	10% / 11% / 79%
I was comfortable with being asked questions from a range of different response formats (n=8776)	8% / 13% / 79%
I am in favour of the "speech" questions, in which I was asked to consider the appropriateness of speech responses (provided as direct quotes) (n=8585)	31% / 26% / 43%
I am in favour of the "evolving" questions, in which some scenarios have multiple parts (e.g., Part A and Part B) (n=8747)	10% / 21% / 69%

9.2. **Instructions:** 74% of candidates agreed that the information available in the Applicant Guide about the SJT was clear and helpful. Similarly, 79% agreed the instructions were clear and easy to understand.

9.3. **Test administration:** 53% felt that booking the test online was straight forward and 62% were able to book an appointment that was convenient. Of those that completed the test in a test centre, 89% felt the venue and facilities were appropriate. Of those that sat it remotely, 77% felt the online proctoring system was a suitable way to sit the SJT.

9.4. **Test content and format:** 66% agreed that the content of the SJT was relevant to the FY1 role and 57% agreed it was appropriately difficult, whilst 29% neither agreed nor disagreed. However, only 43% of candidates agreed the content of the SJT was fair for all candidates. 73% of candidates felt there was a sufficient amount of time to complete the test.

9.5. **Item types:** Candidates were also asked about the format for answer questions and about the various new scenario types. 79% agreed that the format for answering the questions was straightforward and 79% reported that they felt comfortable being asked questions from a range of different response formats. In terms of specific scenario formats, 69% of candidates were in favour of evolving questions and 43% of candidates were in favour of speech responses, however 31% disagreed that they were in favour of the "speech" questions.

9.6. **Computer-based testing and the testing platform:** 84% of candidates found it easy to read the information/questions on screen and 83% felt computer-based testing is an appropriate way to complete the SJT.

10.   Summary and Conclusions

10.1.   This report details the operational use of the SJT for selection to the Foundation Programme in 2023.

10.2.   The psychometric analysis of the 2023 operational SJT is positive and shows consistency when compared to previous versions of the SJT for entry into the FP.

10.3.   The SJT demonstrated an overall **excellent level of internal reliability** (α=.860 Paper A; α=.838 Paper B; α=.830 Paper C), which is appropriate for tests administered in high stakes selection context such as the FP. The SJT was **capable of differentiating between candidates**, providing a sufficient spread of scores to support decision making as part of selection into the FP.

10.4.   Candidates were allowed 140 minutes to complete the 75-scenario test (which includes 9 pilot scenarios). The test completion analysis showed that the **test was not speeded**, with 99.7% of candidates completing the last question on Paper A, 99.9% of candidates completing the last question on Paper B, and 99.8% of candidates completing the last questions on Paper C.

10.5.   In relation to our **Equality, Diversity and Inclusion (ED&I) analysis,** the SJT results show significant differences for **gender** (small effect size), **ethnicity** (large effect size), and **place of education** (UK or International) (large effect size). Differences based on ethnicity were still observed, though the differences were smaller, when place of education was controlled for (moderate effect size).

10.6.   The EPM results also show significant differences for **gender** (small effect size), and **ethnicity** (small effect size). Similar to the SJT results, differences based on ethnicity were still observed for the EPM results, though the differences were smaller, when place of education was controlled for (small effect size). In some cases, the differences seen may be exacerbated due to the uneven sample size within subgroup categories. No differences were observed in EPM scores based on place of education (UK or International).

10.7.   **Candidate feedback** was generally positive with regards the **contents and relevance of the SJT to the FY1 role**, though there was less agreement in terms of perceptions of **fairness** and the of **difficulty** of the test. Most candidates found the question format appropriate although there was less consensus on the suitability of **speech responses.**