Work Psychology Group
Thinking differently

Analysis of the Situational Judgement Test for Selection to the Foundation Programme 2022

Technical report

May 2022

Melissa Washbrook
Natalie Christodoulidou
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. This is built upon the initial pilot conducted in 2020 and operational delivery in 2021, 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 applicants 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, 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.

1.3 The Foundation Programme (FP) SJT was delivered for selection to the FP 2022 during two testing windows which lasted from the 6th to the 18th of December 2021 and from the 17th to the 22nd of January 2022. In total, N=9,109 candidates sat the SJT; n=3,085 completed operational Paper A, n=2,993 completed operational Paper B, and n=3,031 completed operational Paper C.

1.4 As a result of the ongoing impact of the COVID-19 pandemic, 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 applicants 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 2022 National Recruitment Process.

Analysis

1.6 The psychometric analysis of the 2022 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 and can be used **to guide the continued development of the operational SJT** for use as part of the National Recruitment of FY1 doctors.

1.7 The SJT demonstrated an overall **excellent level of internal reliability** (α=.842 Paper A; α=.846 Paper B; α=.837 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.

1.8 Candidates were allowed 140 minutes to complete the 75-scenario test (which includes 10 pilot scenarios). The test completion analysis showed that the **test was not speeded**, with 99.8% of candidates completing the last question on Paper A, 99.7% of candidates completing the last question on Paper B, and 99.9% 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** (moderate 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). The EPM results also show significant differences for **gender** (negligible effect size), **ethnicity** (small effect size) and **place of education** (moderate 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.

1.10 **Candidate feedback and practice materials.** 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 inclusion of **video-based scenarios** within the test. Open text comments focused on the perceived lack of online resources available to help candidates prepare for the SJT, including answer keys and rationale statements for practice materials.

1.11 **Pilot analysis.** In 2022, 120 scenarios were piloted across all three item types; Ranking, Multiple Choice, and Rating. 69% (n=25) of the Ranking scenarios were added to the operational item bank, 79% (n=19) of the Multiple Choice scenarios were added to the operational item bank, and 39% (n=165) of the Rating items were added to the bank.

Table of Contents

2. Introduction

Overview & Objectives

2.1. An SJT has been used for selection into Foundation Year One (FY1) Training for the past 8 years. The SJT, in combination with the Educational Performance Measure (EPM), is used to rank applicants 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 revised computer-based Situational Judgement Test (SJT) as part of live selection into FY1 Training. This provided an opportunity to enhance engagement by introducing new SJT item types and multimedia elements, ensuring the SJT continues to remain innovative whilst retaining its good quality psychometric properties. This report aims to evaluate the performance of the SJT, following the second operational used in December 2021 – January 2022. This follows the successful operational implementation in December 2020 – January 2021, and successful pilot in January 2020.

2.2. The objectives of this project were to:

- Develop an operational SJT for live use in 2022, 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.

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

- Confirmation of the Test Specification
- Item Development
- Operational Test Construction
- Scoring and Psychometric Analysis (including both operational and pilot 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 National Recruitment Process.

3. Test Development

Confirmation of the Test Specification

3.1. 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.

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 applicant 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 piloted in January 2020 to determine the suitability of question and response types identified by WPG and was used operationally in 2021. These draw upon the latest research as well as WPG's expertise in assessment design in high-stakes environments. The results indicated that the newly developed SJT items would be an appropriate measure for use as part of selection into the FY1 training programme.

3.3. The Foundation Programme SJT is designed to assess five of the nine attributes from the Foundation Programme 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.*

[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/).

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.*
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 Foundation Programme, 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; ranking, multiple choice, and rating. 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:

- **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).
- **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).
- **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).

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.

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 applicants when a scenario is going to have multiple parts.

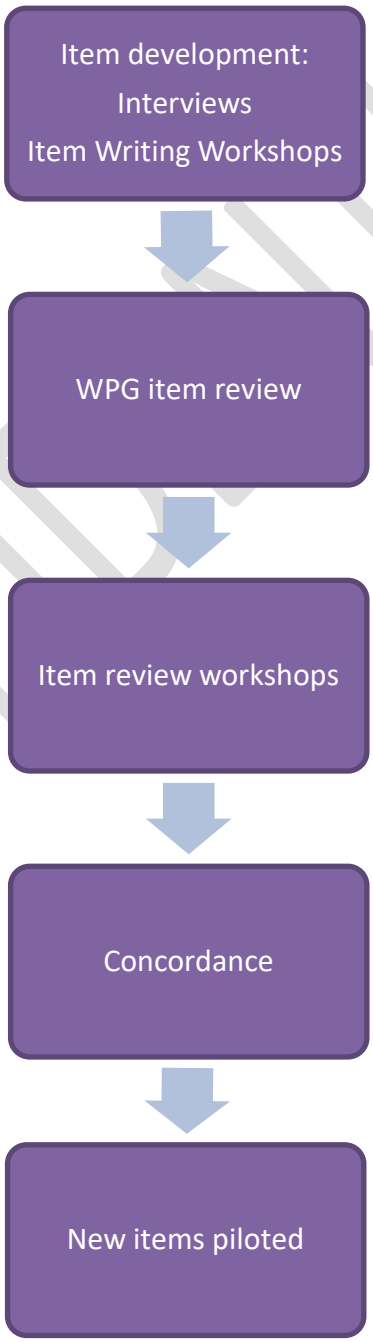
Finally, while the majority of scenarios were presented as text, the computer-based SJT introduced a small number of video-based scenarios. The scenarios presented within the videos were very similar in nature to the text-based scenarios, but candidates had the added benefit of being able to see and hear the characters' actions. Video scenarios were included at the start or end of the sections.

3.4.4. **Test Length.** Three papers, each consisting of 75 scenarios (65 operational items; 10 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 the new items that were trialled alongside the FP 2022 operational delivery.

Figure 1: The development and review process for trial items



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 Development Interviews (IDIs), using the Critical Incident Technique (CIT), were conducted to develop SJT items. CIT interviews aim to elicit, from SMEs, scenarios or incidents involving FY1 doctors who demonstrate particularly effective or ineffective behaviour and that reflect the SJT target attributes. Using CIT interviews has numerous benefits, including the 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.

4.4. In addition to telephone interviews, item writing workshops were also held, 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.

4.5. Following the interviews and item development, 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.

4.6. In addition to developing items for operational use, a practice paper was also developed for applicants use. The practice paper was designed to familiarise applicants with the structure of the SJT, as it is a full-length test including a range of question types, much like the operational papers. The practice paper is hosted online by PearsonVUE, therefore offering a very similar experience to the operational test. It is important to note that the practice paper was not a revision tool as each SJT scenario presents a unique dilemma and therefore applicants are not expected to revise with regards to how they should answer, but rather use their judgement, based on the unique context provided within the scenarios themselves.

Item Review and Concordance

4.7. Item review workshops were held in May 2021, to ensure that all SJT items developed as part of item development 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.

4.8. It is important that the response keys (answers) for the SJT items are finalised based on expert consensus. In addition to agreeing an initial key during the item development process, a concordance study was also conducted to examine the degree of consensus on the item keys between SMEs in August of 2021.

4.9. The concordance test paper was delivered using an online survey platform, Key Survey. In order to implement this, WPG facilitated 2 online concordance sessions for SMEs, which included a short presentation summarising the purpose and process of concordance. The SMEs who attended these sessions were then emailed a link to complete the concordance test paper on Key Survey, in their own time.

4.10. The main criterion for categorising a Ranking or Multiple Choice item as having satisfactory levels of concordance, was the use of a significant Kendall's W[2]. For Rating items, the concordance level was determined based on the means, where 50% or above was deemed satisfactory. If the level of concordance was satisfactory, then the concordance key was compared against the existing key.

4.11. As expected, for some items, the key favoured by the concordance panel differed from the item writer key; this was considered as part of the concordance analysis. Final pilot keys were determined by psychometric experts from WPG, based on detailed qualitative and quantitative analysis of the concordance key and item writer / review workshop key. Alternative keys were then used, for some items, if the psychometric analysis supported their use (i.e., item partial, facility).

[2] 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, 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.

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. Three operational papers were developed in 2022, compared to the two operational papers used in the 2021 SJT.

5.3. Each operational test version consisted of 65 operational scenarios (35 Ranking, 17 Multiple Choice, and 13 Rating scenarios). 7 animated videos (2 Rating, 2 Multiple Choice, and 3 Ranking scenarios) were included within each operational test.

5.4. Candidates also answered 10 pilot SJT scenarios (3 Ranking, 2 Multiple Choice, and 5 Rating scenarios), which did not contribute to their overall SJT score. To allow for sufficient piloting of new content, there were 12 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. There was a maximum of 20 points available for each Ranking scenario, based on how close responses were to the key, 12 points for each Multiple Choice scenario (points awarded for each correct option identified) and a maximum of 3 or 4 points for each rating item (dependant on 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 items 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	12	12	13
Paper B	14	14	13	12	12
Paper C	14	14	12	12	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=9,109 candidates sat the 2022 SJT during two testing windows which lasted from the 6th to the 18th of December 2021 and from the 17th to the 22nd January 2022. n=3,085 completed operational Paper A, n=2,993 completed operational Paper B and n=3,031 completed operational Paper C.

6.2. The majority of candidates provided demographic data. With regards to gender, 56.4% (n=5,139) of the sample indicated that they were female, 39.5% (n=3,597) indicated that they were male, and 4.1% (n=373) did not specify or their data was unavailable. The ages of the sample from those who responded ranged from 21 to 58 years. 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
4354 (47.8%)	2964 (32.5%)	482 (5.3%)	424 (4.7%)	350 (3.8%)	535 (5.9%)

Table 4: Breakdown of Candidates' Place of Education

Educated within the UK	Educated outside of the UK	Unavailable
7979 (87.6%)	1062 (11.7%)	68 (0.7%)

Test Level Results

6.3. Table 5 reports the descriptive statistics for the three operational FY1 2022 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	3085	2993	3031
Mean score	928.06	942.74	926.61
Maximum possible score	1106	1115	1102
Mean score as %	83.91%	84.55%	84.08%
Standard deviation	39.40	40.40	39.88
Range	430 - 1008	495 - 1035	476 - 1021
Reliability	.842	.846	.837

Reliability

6.4. 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.

6.5. A commonly accepted rule of thumb for describing internal reliability or internal consistency, using Cronbach's alpha, is as follows[3]:

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

6.6. Following best-practice procedure, a small number of items were removed prior to scoring based on their psychometric performance and detraction from the overall reliability of each paper.

6.7. All three operational papers showed excellent levels of internal reliability (Paper A α=0.842; Paper B α=0.846; and Paper C α=0.837), which is the desired level of reliability for an operational test. It is positive that the reliability remained in line with previous years (α=0.802 for both operational papers used in 2021).

6.8. WPG will continue to monitor the performance of each item year-on-year and select the items that perform best psychometrically for future use. This will become more feasible as the bank expands through continued item development.

Test Difficulty

6.9. The difficulty level for Operational Paper A is 83.91% (i.e., mean score of 928.06 out of a total possible total raw score of 1106), Paper B is 84.55% (mean score of 942.74 out of a possible total raw score of 1115) and Paper C is 84.08% (mean score of 926.61 out of a possible total raw score of 1102). This indicates that the three paper versions exhibit comparable levels of difficulty, which is consistent with the difficulty level observed in the 2021 operational FP SJT.

Timing Analysis

6.10. The standard time allowed for completion of the SJT was 140 minutes. For Paper A, 99.8%, Paper B 99.7% and Paper C 99.9% of candidates completed the last operational question. On

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

average, candidates took 130 minutes to complete the test. These findings indicate that the time allowed to complete the test is sufficient.

6.11. On average, Rating scenarios took 112 seconds to complete. Multiple Choice and Ranking items took, on average, approximately 100 and 102 seconds, respectively. These findings indicate that each item type is similar in terms of timing and demonstrates that candidates have sufficient time to complete both operational and pilot items included within the test.

Distribution of Scores

6.12. SJT total scores for operational Paper A, B and C showed a close to normal distribution, although all three samples are slightly negatively skewed (see Figures 2, 3 and 4 below).

Figure 2: Distribution of SJT Scores in Paper A

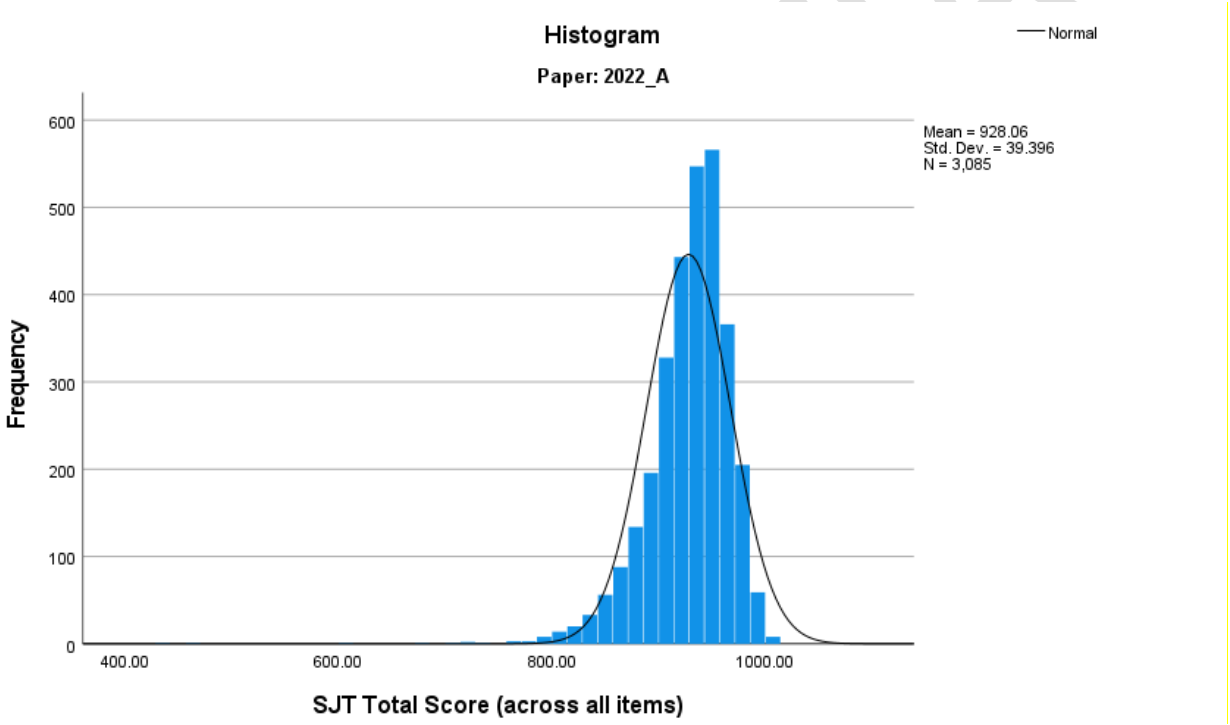


Figure 3: Distribution of SJT Scores in Paper B

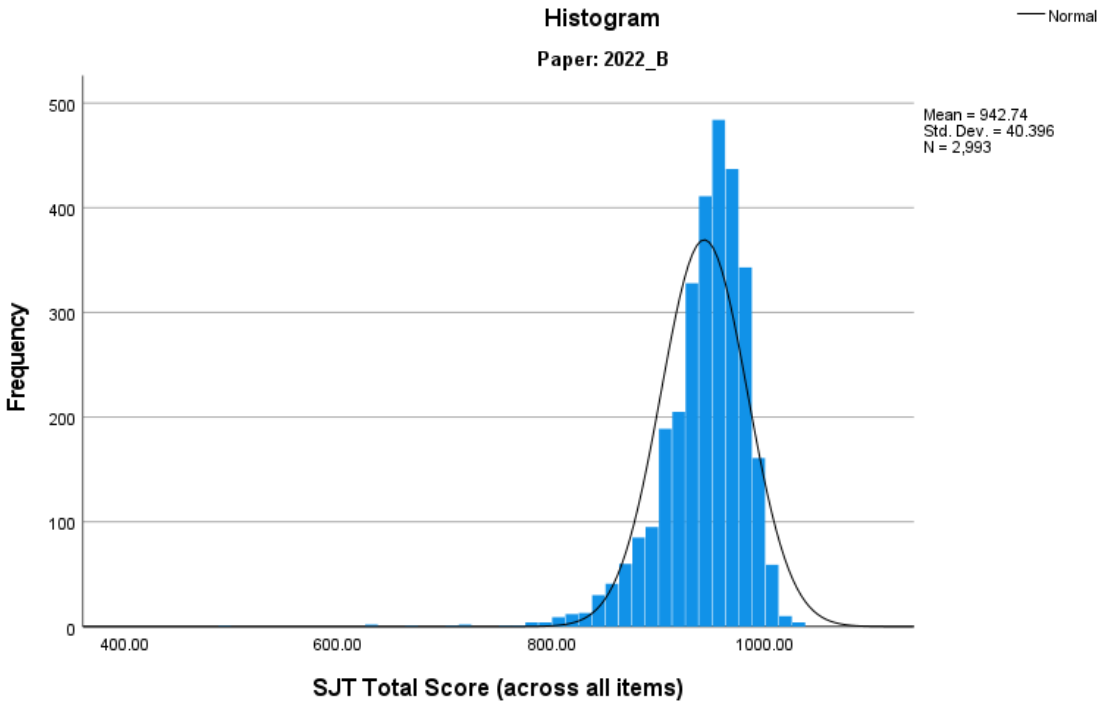
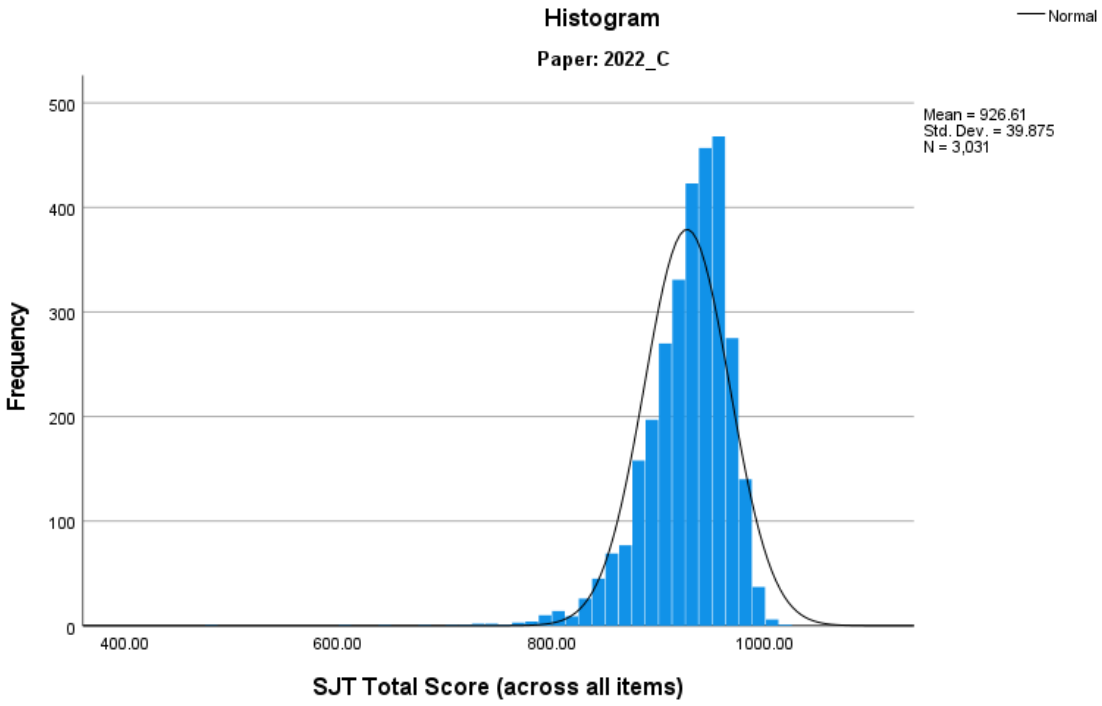


Figure 4: Distribution of SJT Scores in Paper C



Test Equating

6.13. 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.

6.14. There are a number of 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[4]

6.15. 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

6.16. 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

	Ranking			Multiple Choice			Rating		
Paper	**Mean Facility**	**Min**	**Max**	**Mean Facility**	**Min**	**Max**	**Mean Facility**	**Min**	**Max**
A	17.41	14.72	19.44	9.71	7.98	11.37	2.72	1.48	3.99
B	17.54	14.91	19.31	9.72	8.10	11.57	2.79	1.34	4.00
C	17.53	14.89	19.43	9.49	7.92	11.42	2.78	1.30	4.00

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

[4] The data of a small number of candidates who were extreme outliers are not included within this section of the report.

Item Quality

6.18. 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**

Moderate = Correlation of **.13 to .16**

Limited = Correlation of **.12 or below**

6.19. 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
	Ranking Items			Multiple Choice Items			Rating Items		
Mean	**.20**	**.21**	**.22**	**.22**	**.23**	**.22**	**.15**	**.15**	**.14**
Good	34%	29%	37%	47%	41%	29%	11%	12%	13%
Satisfactory	34%	46%	37%	41%	53%	47%	33%	31%	25%
Moderate	26%	11%	20%	6%	0%	12%	11%	15%	15%
Limited	6%	14%	6%	6%	6%	12%	46%	42%	47%

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

6.21. The overall item quality for the 2022 operational test shows a slight increase for Multiple Choice Items and Rating items, when compared to the 2021 operational test. In 2021, the mean item partial was .20 for Multiple Choice items and .12 for Rating items for both Paper A and B. The mean partials across all three 2022 papers showed an improvement in 2022 for both Multiple Choice (.22, .23, and .22) and Rating items (.15, .15, and .14). The item quality for Ranking Scenarios in 2022 (.20, .21, and .22) is consistent with that observed in 2021 (.20 and .23 for Paper A and B respectively).

Across all papers, rating items had lower average partials (Paper A and B, .15, Paper C, .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.

Video-based scenarios

6.22. Seven animated videos (2 Rating, 2 Multiple Choice, and 3 Ranking scenarios) were included within the operational test. Six videos were consistent across Paper A, B and C ('anchor items') and one unique Ranking video-based item was included in each paper. The mean facility and partials for each item type were similar across all three papers, which is highlighted in Table 8 below.

Table 8: Summary of Item Partial and Facility for Video-Based Scenarios by Paper Version

	Ranking			Multiple Choice			Rating		
Paper	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Item Partial	0.15	0.18	0.15	0.12	0.13	0.12	0.13	0.13	0.12
Item Facility	16.49	16.46	17.20	10.16	10.09	10.17	2.69	2.67	2.68

6.23. WPG will continue to review the current operational item bank and the existing process for item and test development to maintain and enhance the overall quality of the test. Items classified as being of limited quality will require further review and may be repiloted or excluded from future operational versions of the SJT. 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.

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

7.1. 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

7.2. **Gender**: Table 9 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** ($t(7526.09)=-10.80$, $p<.001$), with female candidates scoring significantly higher than male candidates. Cohen's d[5], which quantifies the magnitude of the difference between the mean SJT scores for males and females, shows a small effect size ($d=.24$). This is in line with the 2021 operational results and within other similar SJTs for selection into healthcare roles.

Table 9: Gender

	Female	Male
N	5104	3572
Mean equated SJT total	931.77	923.42
Std. Deviation	34.73	35.88

7.3. **Ethnicity:** Table 10 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** ($F(4,8511)=300.95, p<.001$). Eta-squared[6], which is a measure of effect size, shows a moderate effect ($\eta^2=.12$).

[5] 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.

[6] 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.

Post-hoc testing (Tukey HSD) revealed significant differences between those candidates describing themselves as 'White' and those candidates describing themselves either as 'Asian' (p<.001), 'Black' (p<.001), 'Mixed' (p<.001), or 'Other' (p<.001). Additionally, significant differences were found between candidates describing themselves as 'Asian' and those describing themselves as 'Black' (p<.01), 'Mixed' (p<.001), or 'Other' (p<.05). Significant differences were also found between candidates describing themselves as 'Black' and those describing themselves as 'Mixed' (p<.001). Furthermore, significant differences were also observed between candidates describing themselves as 'Mixed' and those describing themselves as 'Other' (p<.001). 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 10: Ethnicity

	White	Asian	Black	Mixed	Other
N	4339	2942	471	419	345
Mean equated SJT total	940.34	916.44	911.11	932.81	911.13
Std. Deviation	29.00	35.94	37.60	32.50	40.33

7.4. **Place of Education:** Table 11 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(1185.75)=45.18, p<.001), with UK educated candidates scoring significantly higher than International candidates. The observed difference in scores represents a large effect size (d=1.82). These results are consistent with those observed in 2021.

Table 11: Place of Education

	United Kingdom	International
N	7954	1023
Mean equated SJT total	934.57	878.84
Std. Deviation	29.51	38.00

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 12 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** ($F(4,7592)= 212.32$, $p<.001$). A moderate effect size ($\eta^2=.10$) was observed.

Post-hoc testing (Tukey HSD) revealed significant differences between UK educated candidates describing themselves as 'White' and those candidates describing themselves either as 'Asian' ($p<.001$), 'Black' ($p<.001$), 'Mixed' ($p<.05$), and 'Other' ($p<.001$). Additionally, significant differences were found between UK educated candidates describing themselves as 'Asian' and those describing themselves as 'Mixed' ($p<.001$). Significant differences were also found between UK educated candidates describing themselves as 'Black' and those describing themselves as 'Mixed' ($p<.001$). Furthermore, significant differences were observed between UK educated candidates describing themselves as 'Mixed' and those describing themselves as 'Other' ($p<.001$). 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 12: Ethnicity (UK only)

	White	Asian	Black	Mixed	Other
N	4123	2454	374	379	267
Mean equated SJT total	942.97	924.46	921.05	938.38	921.88
Std. Deviation	25.91	29.99	30.17	26.43	33.66

Differences in Performance on the EPM

7.6. **Gender:** Table 13 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** ($t(8734)=-4.17$, $p<.001$), with female candidates scoring marginally higher than male candidates. The observed difference in scores represents an effect size that does not reach the threshold to be considered small and is therefore considered negligible ($d=.09$). These results are consistent with those observed in 2021.

7.7. **Table 13: Gender**

	Female	Male
N	5139	3597
Mean EPM total score	41.33	40.99
Std. Deviation	3.77	3.88

7.8. **Ethnicity:** Table 14 shows differences in performance on the EPM 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 scores on the

EPM compared to other ethnic groups. A one-way ANOVA **found a significant overall effect of ethnicity on EPM scores** ($F(4,8569)=115.71$, $p<.001$), with a small effect size ($\eta^2=.05$) observed.

Post-hoc testing (Tukey HSD) revealed significant differences between those candidates describing themselves as 'White' and those candidates describing themselves either as 'Asian' ($p<.001$), 'Black' ($p<.001$), 'Mixed' ($p<.001$), and 'Other' ($p<.001$). Additionally, significant differences were found between candidates describing themselves as 'Asian' and those describing themselves as 'Mixed' ($p<.001$). Differences were also found between candidates describing themselves as 'Black' and those candidates describing themselves as 'Mixed' ($p<.05$). Significant differences were also found between candidates describing themselves as 'Mixed' and those describing themselves as 'Other' ($p<.01$). 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 14: Ethnicity

	White	Asian	Black	Mixed	Other
N	4354	2964	482	424	350
Mean EPM total score	42.03	40.23	40.40	41.19	40.29
Std. Deviation	3.61	3.83	3.62	3.91	3.89

7.9. **Place of Education:** Table 15 shows differences in performance on the EPM based on place of education. An independent t-test showed a **significant difference in performance on the EPM between UK and International candidates** ($t(1417.37)=16.59$, $p<.001$), with UK educated candidates scoring higher than International candidates. The observed difference in scores represents a moderate effect size ($d=.51$). A small effect size ($d=.34$) was observed in 2021, indicating the performance gap between UK and International candidates has increased in 2022.

Table 15: Place of Education

	United Kingdom	International
N	7979	1062
Mean EPM total score	41.41	39.50
Std. Deviation	3.80	3.49

7.10. **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 16 shows differences in performance on the EPM based on ethnicity for those candidates educated in the UK. 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 scores on the EPM compared to other ethnic groups. A one-way ANOVA found a **significant overall effect of ethnicity on EPM scores for UK educated candidates** ($F(4,7614)=79.81$, $p<.001$), with a small effect size ($\eta^2=.04$) observed.

Post-hoc testing (Tukey HSD) revealed significant differences between UK educated candidates describing themselves as 'White' and those candidates describing themselves either as 'Asian' ($p<.001$), 'Black' ($p<.001$), 'Mixed' ($p=.001$) and 'Other' ($p<.001$). Additionally, significant differences were found between UK educated candidates describing themselves as 'Asian' and those describing themselves as 'Mixed' ($p=.001$). Differences were also found between UK educated candidates describing themselves as 'Black' and those candidates describing themselves as 'Mixed' ($p<.05$). 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 16: Ethnicity (UK only)

	White	Asian	Black	Mixed	Other
N	4128	2462	377	382	270
Mean EPM total score	42.12	40.52	40.51	41.33	40.78
Std. Deviation	3.61	3.85	3.69	3.82	3.97

Differential Item Functioning (DIF)

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

7.12. DIF analysis was completed using multiple regression and was used to examine whether demographic variables (including gender, ethnicity, and place of education) significantly predict performance on each item individually, controlling for overall test performance (i.e., 'is there a difference in item performance beyond that which expected due to differences between groups on the test overall?'). To determine significant effects sizes, R2 change values of .01 and above were sought.

7.13. For Paper A, 1 item showed a gender difference (favouring females), 4 items showed an ethnicity difference (3 favouring BME; 1 favouring White candidates), and 2 items showed a place of education difference (favouring UK educated candidates). For Paper B, 1 item showed

a gender difference (favouring males), 2 items showed an ethnicity difference (favouring BME candidates), and 2 items showed a place of education difference (favouring UK educated candidates). For Paper C, 2 items showed a gender difference (favouring females), 2 items showed an ethnicity difference (favouring BME candidates), and 2 items showed a place of education difference (favouring UK educated candidates).

7.14. Given the number of statistical tests involved, there is a risk that random differences may reach statistical significance (type 1 error). For this reason, positive results are treated as 'flags' for further investigation, rather than confirmation of difference or bias. A further internal review of these items will be carried out by the WPG team. Once reviewed, if the items do appear to demonstrate bias (as outlined above, DIF is a necessary but not sufficient condition for bias), items will be removed from the item bank if deemed appropriate.

7.15. Overall, the small proportion of items identified as exhibiting DIF suggest that there is not risk of bias at the item level.

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*=.10 to *r*=.29 is considered weak, a correlation between *r*=.30 to *r*=.49 is considered moderate, and a correlation between *r*=.50 to *r*=1.00 is considered strong**[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=.39, p<.001). The results show that the SJT is related to the EPM component of the selection

[7] Pearson, K. (2008). Encyclopedia of public health. Pearson's correlation coefficient. Dordrecht: Springer, 1090-1.

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 2021 (r=.35, p<.001).

9. Psychometric Analysis: Pilot

Piloting Overview

9.1. 120 scenarios were piloted alongside the 2022 operational tests. There were 12 sets of 10 pilot questions, used across different forms of Papers A, B and C. Each pilot set consisted of 3 Ranking scenarios, 2 Multiple Choice scenarios, and 5 Rating Scenarios (including between 4 and 8 items per scenario). Within each pilot set, scenarios were loosely allocated to achieve balance across the target criteria.

Pilot Items Analysis

9.2. Analysis was conducted at the item level to evaluate the quality of the pilot items. A summary of the item level statistics is shown in Table 17 below. 69% (n=25) of the Ranking scenarios were added to the operational item bank. 79% (n=19) of the Multiple Choice items were added to the operational item bank. 39% (n=165) of the rating items were added to the bank. However, decisions regarding the appropriateness of adding rating items to the operational bank were based on reviewing the individual partials of the responses to each scenario as a whole and whether the removal of one or two responses would still result in a suitable rating scenario. Many of the items show acceptable SD and facility values which indicate that they are capable of differentiating candidates. In some cases, items deemed inappropriate for addition to the operational item bank will be refined and repiloted as part of the next cycle of development.

Table 17: Summary of Item Level Statistics for Pilot Items

	Ranking (n=36)			Multiple Choice (n=24)			Rating (n=420)		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
Item Partial	.17	-.05	.37	.18	.01	.36	.07	-.26	.33
Item Facility	17.42	15.45	19.09	9.62	6.40	11.59	2.94	.55	4.00

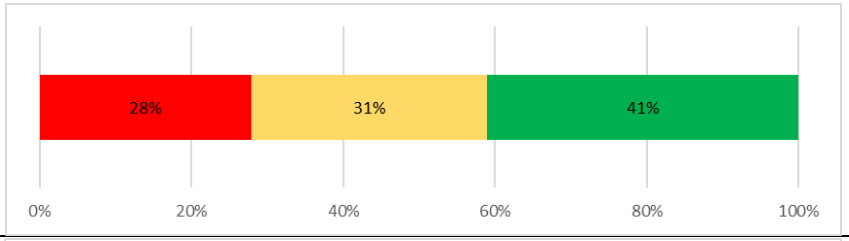
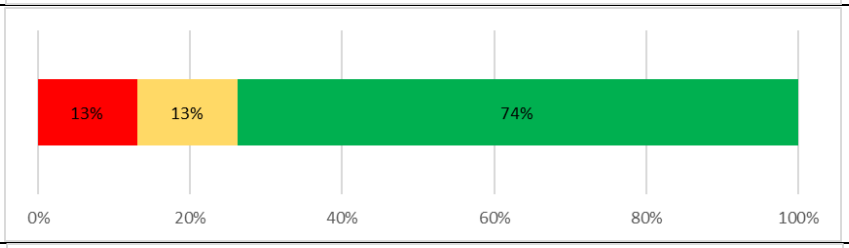
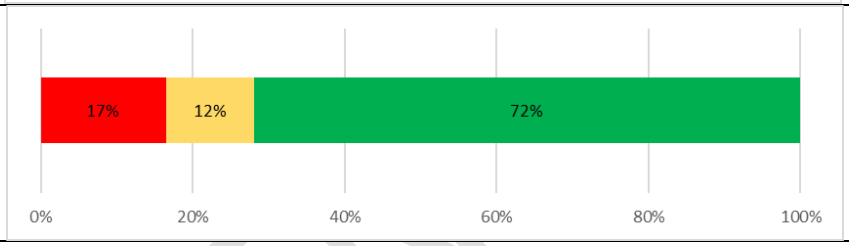
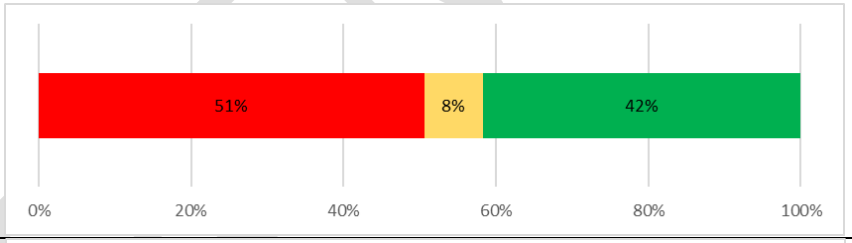
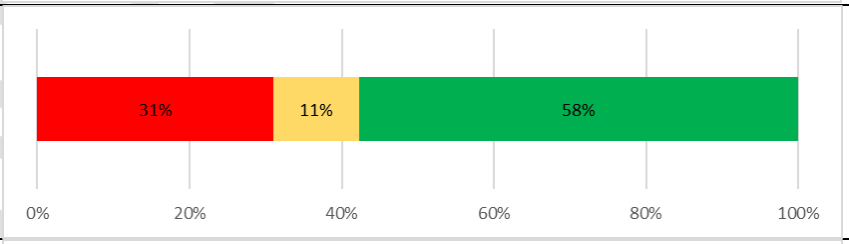
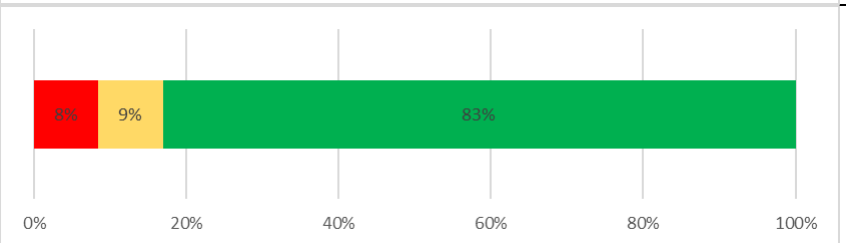
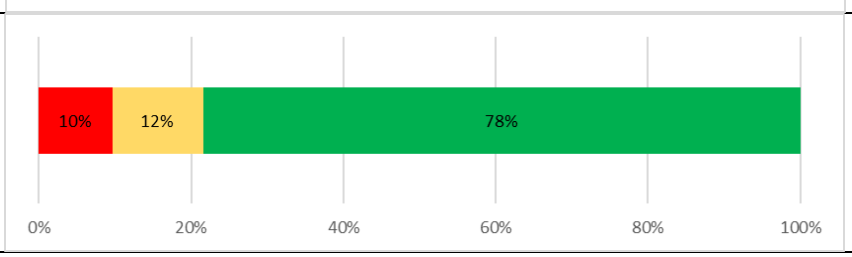
10. Candidate Feedback

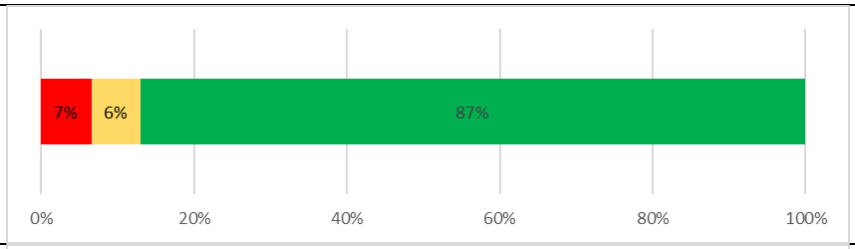
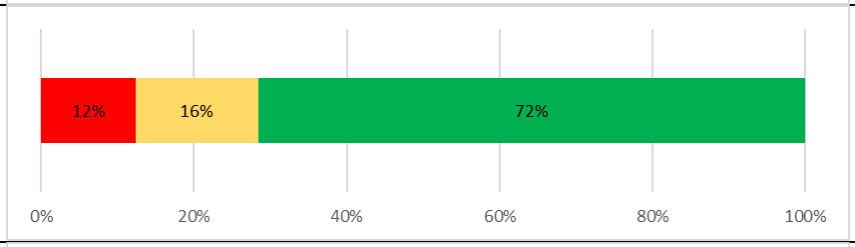
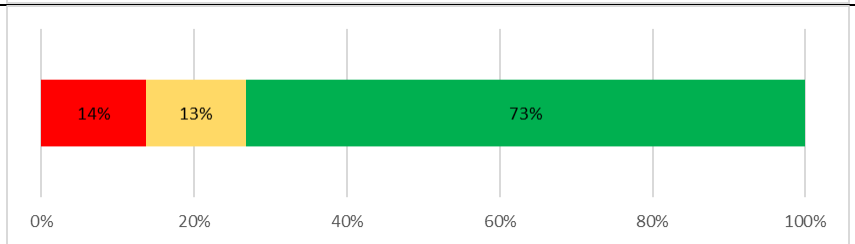
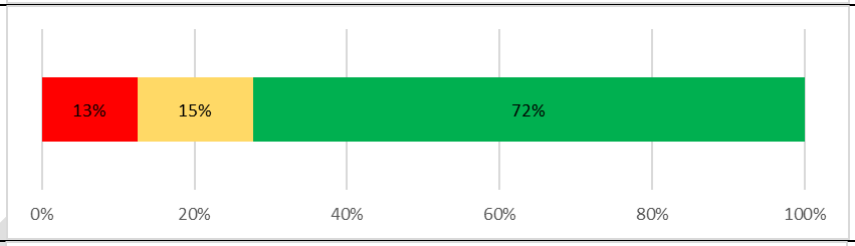
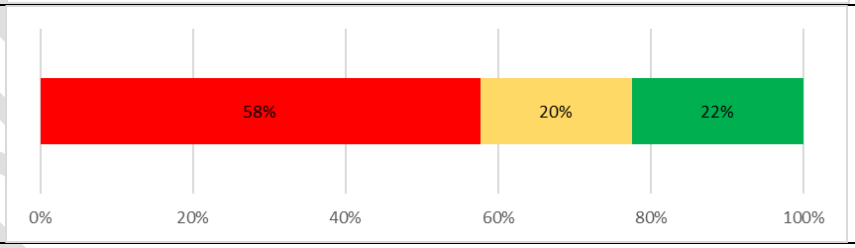
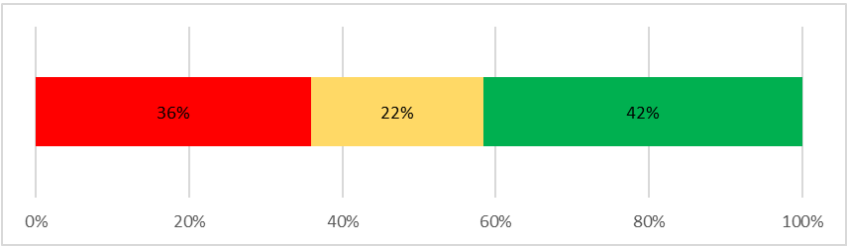
10.1. Participants who completed the operational SJT were asked to complete an evaluation questionnaire regarding their perceptions of the SJT. This feedback has been collated and reported in four key sections below. Overall, n=8120 (89.14%) of participants provided feedback. The breakdown of responses to each question can be seen in Table 18 below. Qualitative feedback was also gathered from candidates to provide further insight and context about their perceptions of the SJT. Some of these comments have been provided in the commentary below.

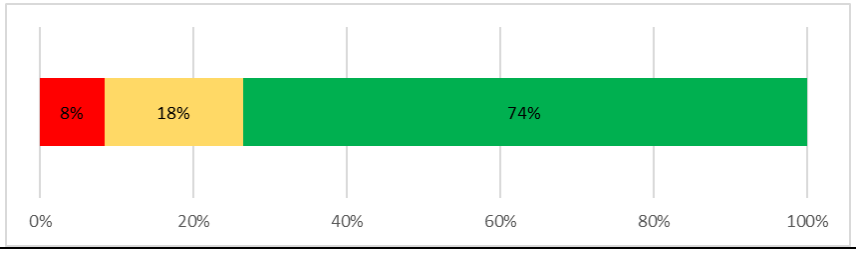
Table 18: Participant feedback on overall test content[8]

	% Disagree	% Neither Agree Nor Disagree	% Agree
The information I read in the Applicant Guide about the SJT was clear and helpful (n=8,120)	12%	23%	65%
The content of the Situational Judgement Test (SJT) was relevant to the role of Foundation Year 1 doctor (n=8,055)	12%	21%	67%
The content of the SJT was an appropriate level of difficulty for my training level (n=8,037)	15%	30%	55%

[8] For each question, those that did not respond, or selected 'Not Applicable' were excluded.

The content of the SJT was fair for all candidates (n=7,968)	
The instructions for the SJT were clear and easy to understand (n=8,029)	
There was a sufficient amount of time to complete the test (n=8,006)	
Booking the test online was straightforward (n=7,970)	
I was able to book an appointment that was convenient for me (n=7,994)	
I found it easy to read the information/questions on screen (n=7,979)	
Computer-based testing is an appropriate way to complete the SJT (n=7,976)	

The venue and facilities were appropriate (N/A if you completed at home) (n=7,049)	
The online proctoring system was a suitable way to sit the SJT (N/A if you completed in test centre) (n=2,038)	
The format for answering the questions was straightforward (n=7,970)	
I was comfortable with being asked questions from a range of different response formats (n=7,953)	
I am in favour of the "video scenarios" (n=7,962)	
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=7,907)	

I am in favour of the "evolving" questions, in which some scenarios have multiple parts (e.g., Part A and Part B) (n=7,943)	

10.2. **Instructions:** 65% of candidates agreed that the information available in the Applicant Guide about the SJT was clear and helpful. Similarly, 74% agreed the instructions were clear and easy to understand. Some comments did request more clarity about the process, for example; **"It would have been nice to know the test was in two halves prior to commencing the exam, and it was unclear where the 10-minute gap would be. Test centre was also unaware of the 10-minute break in the middle."**

10.3. **Test administration:** 42% felt that booking the test online was straight forward and 58% were able to book an appointment that was convenient. To further improve the booking process, some candidates suggested that **"It would be helpful to release a time at which the booking for the SJT slots will open so that everyone has a fair chance of booking a slot that is convenient for them"**. Of those that completed the test in a test centre, 87% felt the venue and facilities were appropriate. Of those that sat it remotely, 72% felt the online proctoring system was a suitable way to sit the SJT.

10.4. **Test content and format:** Candidates generally provided positive feedback towards the overall test content. 67% agreed that the content of the SJT was relevant to the FY1 role and 55% agreed it was appropriately difficult. However, only 41% of candidates agreed the content of the SJT was fair for all candidates. Many of the open-text comments related to the availability of practice materials for all question types, and rationales for the correct responses in advance of sitting the test; **"there should be more practice questions available with rationales: for example, the only online mock had no answers which does not allow us to prepare for the rating questions at all".** 72% of candidates felt there was a sufficient amount of time to complete the test.

10.5. Candidates were also asked about the format for answer questions and about the various new scenario types. 73% agreed that the format for answering the questions was straightforward and 72% reported that they felt comfortable being asked questions from a range of different response formats. In terms of specific scenario formats, 74% of candidates were in favour of evolving questions and 42% of candidates were in favour of speech responses. Only 22% of candidates were in favour of video scenarios, with comments from candidates relating to the time taken to complete the video scenarios; "**the questions with video scenarios often took longer to answer as watching the video took more time than reading the question. It would be ideal if we were able to either speed up the video or fast forward."**

10.6. **Computer-based testing and the testing platform:** 83% of candidates found it easy to read the information/questions on screen and 78% felt computer-based testing is an appropriate way to complete the SJT. Some candidates did provide qualitative comments regarding the test centre conditions, which included; noisy test centres, which was disruptive during the test, poor facilities such as toilets, and slow computers. "**Test centres need to be up to the standard for online exams. My test centre had a few issues: slow computer, no headphones available - however was requested and given promptly."**

11. Summary

Summary

11.1. This report details the operational use of the SJT for selection to the Foundation Programme 2022, as well as the development of new items which were trialled alongside the 2022 operational SJT.

11.2. The psychometric analysis of the 2022 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 and can be used **to guide the continued development of the operational SJT** for use as part of the National Recruitment of FY1 doctors.

11.3. The SJT demonstrated an overall **excellent level of internal reliability** (α=.842 Paper A; α=.846 Paper B; α=.837 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.

11.4. Candidates were allowed 140 minutes to complete the 75-scenario test (which includes 10 pilot scenarios). The test completion analysis showed that the **test was not speeded**, with 99.8% of candidates completing the last question on Paper A, 99.7% of candidates completing the last question on Paper B, and 99.9% of candidates completing the last questions on Paper C.

11.5. In relation to our **Equality, Diversity and Inclusion (ED&I) analysis,** the SJT results show significant differences for **gender** (small effect size), **ethnicity** (moderate 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). The EPM results also show significant differences for **gender** (negligible effect size), **ethnicity** (small effect size) and **place of education** (moderate 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.

11.6. **Candidate feedback and practice materials.** 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 inclusion of **video-based scenarios** within the test. Open text comments focused on the perceived lack of online resources available to help candidates prepare for the SJT, including answer keys and rationale statements for practice materials.

11.7. **Pilot analysis.** In 2022, 120 scenarios were piloted across all three item types; Ranking, Multiple Choice, and Rating. 69% (n=25) of the Ranking scenarios were added to the operational item bank, 79% (n=19) of the Multiple Choice scenarios were added to the operational item bank, and 39% (n=165) of the Rating items were added to the bank.