Work Psychology Group
Thinking differently

Analysis of the Situational Judgement Test for Selection to the Foundation Programme 2021

Technical report

May 2021

Sam Sheridan
Amelia Powell
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. This is built upon initial pilots conducted in 2020. The SJT, in combination with the Educational Performance Measure (EPM), was used to rank applicants applying for Foundation Year One (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 2021, 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) Situational Judgement Test (SJT) was delivered for selection to FP 2021 during two testing windows which lasted from the 7th to the 19th of December 2020 and from the 18th to the 23rd January 2021. In total, 8,209 candidates sat the SJT, 4,399 completed operational Paper A and 3,810 completed operational Paper B.

1.4 In light of the context round Covid-19, 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 Foundation Programme 2021 National Recruitment Process.

Analysis

1.6 The psychometric analysis of the 2021 operational SJT is positive and shows consistency when compared to previous versions of the SJT for entry into 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 **good level of internal reliability** (.80 on both papers), which is appropriate for tests administered in high stakes selection context such as FP. The SJT was **capable of differentiating between candidates**, providing a sufficient spread of scores to support decision making as part of selection into FP.

1.8 Candidates were allowed 2 hours and 20 minutes to complete the 75-scenario test (which includes 10 pilot scenarios). The test completion analysis showed that the **test was not speeded**, with 99.9% of candidates completing the last question on Paper A and 99.8% of candidates completing the last question on Paper B.

1.9 In relation to group **differences**, the SJT results show significant differences for gender (small effect size), ethnicity (medium effect size) and country of qualification (large effect size). The EPM results also show significant differences for gender (less than small effect size), ethnicity (small effect size) and country of qualification (small effect size). In some cases, this may be exacerbated due to the uneven sizes of the subgroup categories.

1.10 **Pilot analysis.** In 2021, 86 scenarios were piloted across all three item types; Ranking, Multiple Choice Questions, and Rating. 73% (n=33) of the Ranking scenarios were added to the operational item bank. 54% (n=7) of the MCQ items were added to the operational item bank. 52% (n=86) of the rating responses 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 7 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. As part of the ongoing development of the FY1 Situational Judgement Test (SJT), the UK Foundation Programme Office (UKFPO) agreed to transition the SJT to be computer-delivered. Since July 2019, Work Psychology Group (WPG) have been working in partnership with UKFPO to develop, implement and evaluate a revised Situational Judgement Test (SJT) as part of live selection into Foundation Year One 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 new SJT, which was used operationally for the first time in December 2020, following a successful pilot in January 2020.

2.2. The objectives of this project were to:

- Develop an operational SJT for live use in 2021, to support selection of candidates into Foundation Year One (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 FP2021 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 Work Psychology Group. These draw upon the latest research as well as WPG's expertise in assessment design in high-stakes environments. Results of the 2019-2020 SJT pilot were reported to UKFPO and the SJT Oversight Group. The results indicated that the newly developed SJT items would be an appropriate measure for use as part of selection into the Foundation Year One training programme.

3.3. The Foundation Programme SJT assesses 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 Team1. 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 be part of the live selection process and to be weighted equally with the EPM to determine 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. All video scenarios were included at the end of the section.

3.4.4. **Test Length.** Two 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 summarises the development and review process undertaken for the new items that were trialled alongside the FP 2021 operational delivery.

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, Work Psychology Group (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 2020, 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 subject matter experts in August of 2020.

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

4.10. The main criterion for categorising a ranking or MC 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).

5. Operational Test Construction

5.1. The operational delivery of the FY1 SJT required the production of two 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 two 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 versions were developed to be as similar as possible in terms of content parameters.

5.3. Each operational test version consisted of 65 operational scenarios (32 ranking, 18 multiple choice 15 rating). Candidates also answered 10 pilot SJT scenarios, which did not contribute to their overall SJT score.

[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.

5.4. To allow for sufficient piloting of new content, there were 13 forms created in total, each with a different set of pilot items. A unique set of 6 text-based pilot scenarios were included across each of the 13 forms. In addition, 4 video-based pilot scenarios were included in each paper. Each candidate saw 2 live action videos and 2 animated videos. For each video, approximately half the candidates saw each version (live action/animated).

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 item, based on how close responses were to the key, 12 points for each multiple response item (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	13	14	14	12	12
Paper B	13	14	14	12	12

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, 8,209 candidates sat the SJT in 2021 during two testing windows which lasted from the 7th to the 19th of December 2020 and from the 18th to the 23rd January 2021. 4,399 completed operational Paper A and 3,810 completed operational Paper B.

6.2. The majority of candidates provided demographic data. With regards to gender, 55.2% (n=4531) of the sample indicated that they were female and 40.7% (n=3345) indicated that they were male. The ages of the sample from those who responded ranged from 21 to 56 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	Unavailable	Other
4286 (52.2%)	2422 (29.5%)	380 (4.6%)	399 (4.9%)	223 (2.7%)	252 (3.1%)

Table 4: Breakdown of Candidates' Place of Education

Educated within the UK	Educated outside of the UK	Unavailable
7546 (91.9%)	611 (7.4%)	52 (0.6%)

Test Level Results

6.3. Table 5 reports the descriptive statistics for the two operational FY1 2021 SJT papers, using raw scores.

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

	SJT Paper A	SJT Paper B
Total N	4399	3810
Mean score	896.25	903.52
Maximum possible score	1074	1080
Mean score as %	83.45%	83.66%
Standard deviation	34.91	34.79
Range	476-986	574-984
Reliability	.802	.802

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]:

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

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. Both operational papers showed an excellent level of internal reliability of (α=0.802 for both papers), which is the desired level of reliability for an operational test. While it is important to note that the format of the test has changed significantly since 2020, it is positive that the reliability remains in line with previous years (e.g. 2020 Paper One α=.74; Paper Two α=.81).

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 83.45% (i.e. mean score of 896.25 out of a total possible total raw score of 1074), and Paper B is 83.66% (mean score of 903.52 out of a possible total raw score of 1080). This indicates that the two paper versions exhibit comparable levels of difficulty. This level of difficulty is comparable to the 2020 operational FP SJT.

Timing Analysis

6.10. The standard time allowed for completion of the SJT was 140 minutes. For Paper A, 99.9% and for Paper B, 99.8% of candidates completed the last operational question. On average, candidates took 132 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 108 seconds to complete. MCQ and ranking items took, on average, approximately 110 and 102 seconds, respectively. These findings indicate that each item type is similar in terms of timing, and demonstrates that introduction of rating scenarios has not impacted the amount of time needed by candidates to complete the test.

Distribution of Scores

6.12. SJT total scores for operational Paper A and B showed a close to normal distribution, although both samples are slightly negatively skewed (see Figures 1 and 2 overleaf).

Figure 1: Distribution of SJT Scores in Paper A

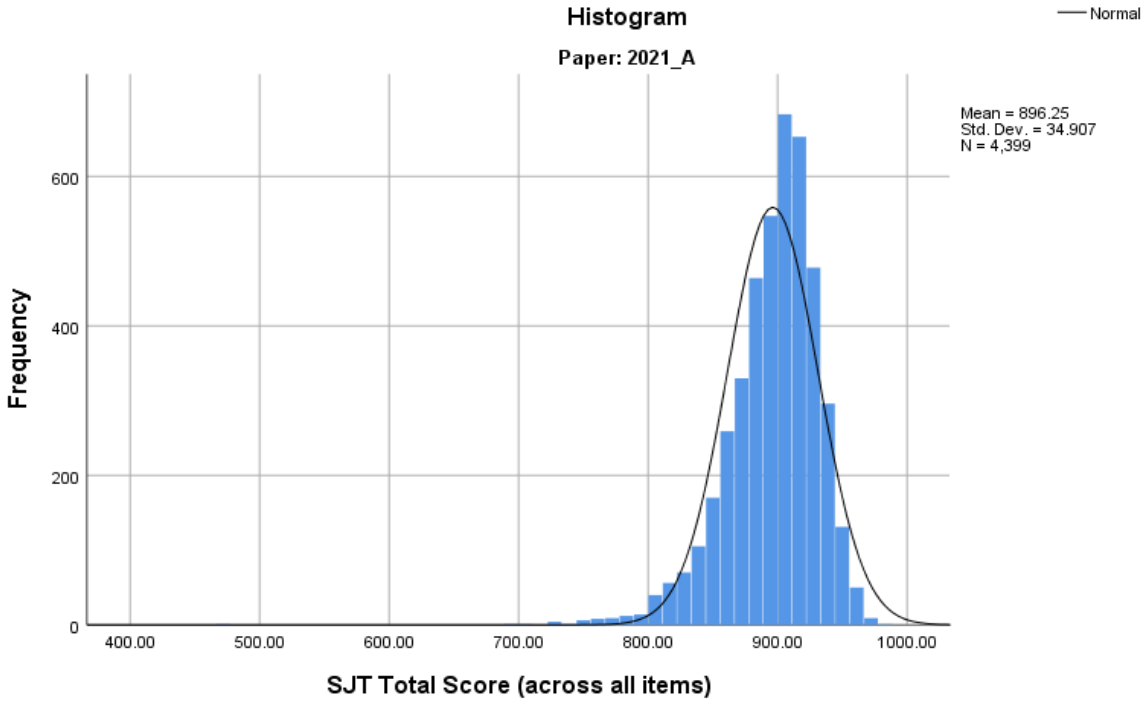
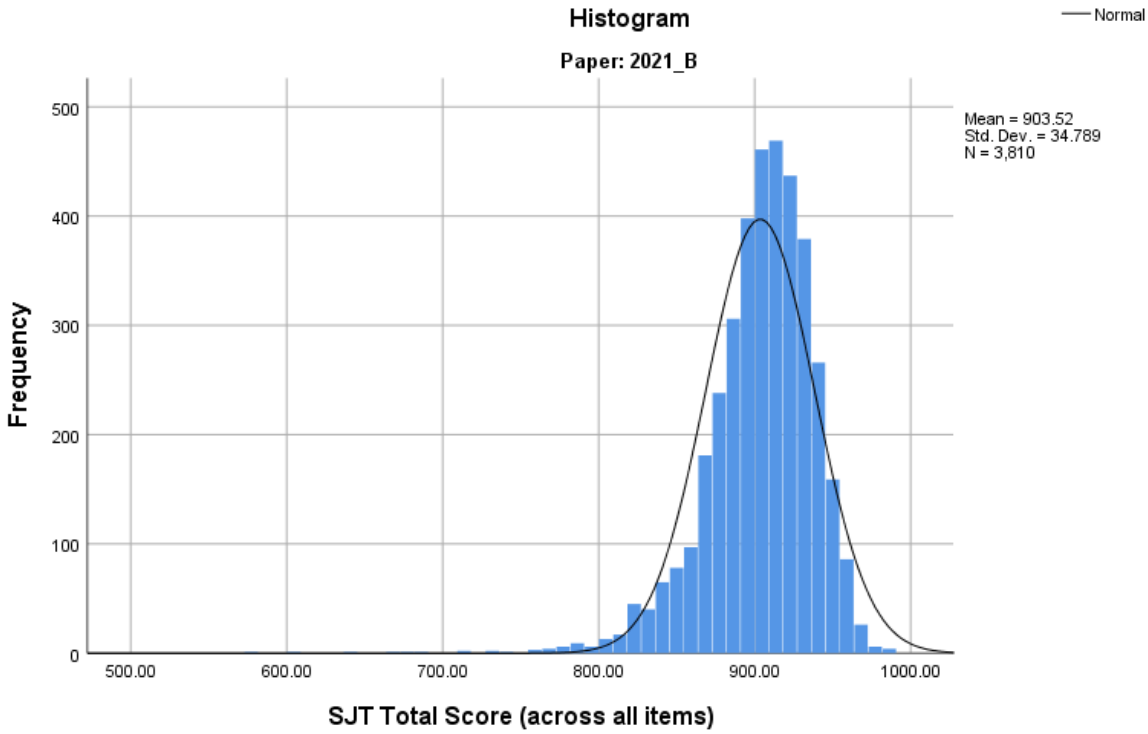


Figure 2: Distribution of SJT Scores in Paper B



Test Equating

6.13. While the two 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 two groups and these can be used to rescale the scores on the unique portion of Paper B 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 both versions, 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; the item facilities provided below are on a scale of 0-1, with 0 being the highest level of difficulty, and 1 being the lowest. Item facilities, split by paper version, are shown in Table 9.

Table 9: Item Facility by Paper Version

	Ranking			Multiple Response			Rating		
Paper	Mean Facility	Min	Max	Mean Facility	Min	Max	Mean Facility	Min	Max
A	17.42	15.70	19.46	9.73	7.72	11.58	2.79	1.41	3.99
B	17.62	15.08	19.58	9.61	8.07	11.46	2.79	1.33	4.00

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

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

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

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 10.

Table 10: Summary of Item Quality by Paper Version

	Paper A	Paper B	Paper A	Paper B	Paper A	Paper B
	Ranking Items		MCQ Items		Rating Items	
Mean	**.20**	**.23**	**.20**	**.20**	**.12**	**.12**
Good	29%	16%	22%	11%	5%	3%
Satisfactory	48%	58%	44%	56%	26%	16%
Moderate	10%	19%	17%	22%	18%	25%
Limited	13%	6%	17%	11%	51%	56%

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 2021 operational test shows a slight decrease when compared to the 2020 operational test. In 2020, the mean item partial was 0.22 across each paper. It is noteworthy that in 2020 the test consisted of Ranking and Multiple-Choice items. In 2021, when looking specifically at the item partials across item types; Multiple-Choice and Ranking items both had average partials of .20 for Paper A, and .20 and .23, respectively, for Paper B. These results are similar to 2020. Across both papers, rating items had lower average partials (.12). 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.

6.22. 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. Group Differences

7.1. In order to examine fairness issues regarding the use of the SJT and EPM, group differences in performance within the candidate sample were analysed on the basis of gender, ethnicity and place of education using total EPM scores and equated SJT scores, after outliers (candidates with very low scores and high missing data) were removed.

Group Differences in Performance on the SJT

7.2. **Gender**: Table 11 shows differences in performance on the SJT based on gender. 105 candidates (0.1%) were excluded from this analysis as data regarding their gender was unavailable. An independent t-test showed **a significant difference in performance on the SJT between female and male candidates** ($t(7859)=-10.892$, $p<0.01$), with female candidates scoring marginally higher than male candidates. The observed difference in scores represents a small effect size ($d = -0.25$). Cohen's d[5], which quantifies the magnitude of the difference between the mean SJT scores for males and females, can be classified as a small effect size. This is in line with the 2020 operational results and within other similar SJTs for selection into healthcare roles.

[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: 0.2 = small effect, 0.5 = medium effect and 0.8 = large effect.

Table 11: Gender

	Female	Male
N	4519	3342
Mean equated SJT total	900.49	892.30
Std. Deviation	32.76	33.20

7.3. **Ethnicity:** Table 12 shows differences in performance on the SJT based on ethnicity. For the purposes of the analysis to ensure a reasonable sample size in each comparison category, candidates of various Asian backgrounds (including Chinese), Black, Mixed or 'Other' backgrounds were grouped as BME. 532 candidates (0.6%) were excluded from this analysis as data regarding their ethnicity was unavailable. An independent t-test showed a **significant difference in performance on the SJT between White and BME candidates** ($t(6685.11)=30.802$, $p<0.001$), with White candidates scoring higher than BME candidates. The observed difference in scores represents a medium effect size (d =0.71). This is similar to other SJTs for selection into healthcare roles.

Table 12: Ethnicity

	White	BME
N	4277	3448
Mean equated SJT total	907.37	885.06
Std. Deviation	28.33	34.07

7.4. **Place of Education:** Table 13 shows differences in performance on the SJT based on place of education. 115 candidates (0.1%) were excluded from this analysis as data regarding their place of training was unavailable. For the purposes of the analysis to ensure a reasonable sample size in each comparison category, candidates educated outside of the UK were grouped at 'International'. An independent t-test showed a **significant difference in performance on the SJT between UK and International candidates** ($t(658.074)=29.506$ $p<0.001$), with UK educated candidates scoring higher than International. The observed difference in scores represents a large effect size (d =1.38).

Table 13: Place of Education

	United Kingdom	International
N	7541	601
Mean equated SJT total	900.37	852.39
Std. Deviation	30.03	38.96

7.5. **Ethnicity (UK only):** Table 14 shows differences in performance on the SJT based on Ethnicity when controlling for Place of education. An independent t-test showed **a significant difference**

in performance on the SJT between White and BME UK educated candidates (t(5920.32)=29.466 p<0.001), with White candidates scoring higher than BME candidates. The observed difference in scores represents a medium effect size (d =0.71).

Table 14: Ethnicity (UK-educated only)

	White	BME
N	4089	3080
Mean equated SJT total	909.53	889.32
Std. Deviation	25.70	30.85

Group Differences in Performance on the EPM

7.6. **Gender:** Table 15 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(7918)=-5.241, p<0.001), with female candidates scoring higher than male candidates. The observed difference in scores represents an effect size that does not reach the threshold to be considered small (d=-0.12).

Table 15: Gender

	Female	Male
N	4549	3371
Mean EPM total score	41.40	40.93
Std. Deviation	3.85	4.04

7.7. **Ethnicity:** Table 16 shows differences in performance on the EPM based on ethnicity. An independent t-test showed a **significant difference in performance on the EPM between White and BME candidates** (t(7339.34)=17.531, p<0.001), with White candidates scoring higher than BME candidates. The observed difference in scores represents a small effect size (*d* =0.40).

Table 16: Ethnicity

	White	BME
N	4306	3476
Mean EPM total score	41.91	40.37
Std. Deviation	3.78	3.90

7.8. **Place of Education:** Table 17 shows differences in performance on the SJT based on place of education. An independent t-test showed a **significant difference in performance on the EPM between UK and International candidates** (t(8202)=9.368 p<0.001), with UK educated candidates scoring higher than International. The observed difference in scores represents a small effect size (*d* =0.34).

Table 17: Place of Education

	United Kingdom	International
N	7570	634
Mean EPM total score	41.31	39.80
Std. Deviation	3.83	4.88

7.9. **Ethnicity (UK only):** Table 18 shows differences in performance on the EPM based on ethnicity with the UK educated cohort of candidates. An independent t-test showed a **significant difference in performance on the EPM between White and BME candidates** ($t(6366.71)=17.653$, $p<0.001$), with White candidates scoring higher than BME candidates. The observed difference in scores represents a small effect size (d =0.42).

Table 18: Ethnicity

	White	BME
N	4089	3080
Mean EPM total score	42.02	40.43
Std. Deviation	3.61	3.88

Differential Item Functioning (DIF)

7.10. 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.

7.11. DIF analysis was completed using multiple regression, and was used to examine whether demographic variables (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, R^2 change values of .01 and above were sought.

7.12. For Paper A, 1 item showed a gender difference (favouring females) and 1 showed an ethnicity difference (favouring BME candidates). For Paper B, no items showed a gender difference and 2 items showed a difference across ethnicity (one item favouring White candidates and one favouring BME candidates). Regarding place of education, Paper B had one item that favoured candidates educated in the UK. 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.

7.13. 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. 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 explained by the correlation, given the differences between the two measures.

8.3. The SJT showed a statistically significant correlation with the EPM score ($r=.35$, $p<.001$). 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.

9. Psychometric Analysis: Pilot

Piloting Overview

9.1. 86 scenarios were piloted alongside the 2021 operational tests. Each operational paper version piloted a different set of video pilot items consisting of 1 multiple choice item and 3 ranking items. There were also 13 further sets of 6 pilot questions, used across different forms of Papers A and B, consisting of 2 ranking scenarios, 1 multiple choice scenario and 3 rating

scenarios. 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 19 below. 73% (n=33) of the Ranking scenarios were added to the operational item bank. 54% (n=7) of the MCQ items were added to the operational item bank. 52% (n=86) of the rating items were added to the bank. However, decisions regarding appropriateness of adding the 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 19: Summary of Item Level Statistics for Pilot Items

	Ranking (n=45)			MC (n=13)			Rating (n=166)		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
Item Partial	.16	.00	.32	.15	.08	.23	.08	-.12	.27
Item Facility	17.41	14.49	18.87	9.34	6.73	11.62	3.03	1.00	4.00

Video based scenarios

9.3. Four video-based scenarios were among the piloted items in 2021. They were presented both as animations and live action clips. When comparing both formats it was noted that the scenarios performed similarly. There were three ranking format video-based scenarios piloted. Of these, the animated versions had mean facility of 16.42 and the live action versions had a mean facility of 16.34. There was one rating video-based scenario piloted. Across individual response items, the animated version had mean item facility of 2.64 and the live action version had a mean item facility of 2.65. Partials were also very similar regardless of live action or animated format.

9.4. Nonetheless, as with all pilot items there is some level of redundancy expected. In future, video-based items could first be piloted as text-based to reduce the risk of rejection after the investment has been made to create a video-based scenario.

10. Psychometric Analysis: Item Response Theory (IRT)

Background

10.1. The 2020 SJT Pilot Delivery aimed to model two different scoring methods; Classical Test Theory (CTT) and Item Response Theory (IRT), to determine which would be most appropriate for operational delivery. Unfortunately the sample size of the 2020 Pilot was not sufficient to allow modelling of IRT, therefore CTT was used for operational delivery. However, it was agreed that IRT would be modelled using the live data, to provide a comparison between the two approaches.

10.2. CTT technique has been used for all past operational deliveries for the FY1 SJT. The theory relies on conducting analyses on the test as a whole. Item test statistics are used to determine the performance effectiveness of different items, but they include the performance and error only within the existing sample on that test.

10.3. Item Response Theory (IRT) focuses on the theory of how items behave at different levels of test taker ability, rather than purely at the test level. It models the response of each test taker of a given ability to each item in the test. As such, the modelling is sample free (item statistics are not dependent on the specific test situation that generated from them). This allows greater flexibility in the design and scoring of test forms.

Modelling

10.4. There are a number of areas of consideration in determining whether IRT models can be used with a test. The test needs to meet the assumptions of IRT models; the model needs to appropriately fit the test; and consideration should be given to how the models compare in terms of accuracy of scoring and equating (i.e. compared to current procedures).

10.5. Two families of unidimensional models are considered in this report:

10.6. **Polytomous**: Item scores were converted to rating scales where the lowest score attained was assigned to zero, the next score 1, the next score 2 etc.

10.7. **Dichotomous**: For rating items, the top score was assigned correct and anything else was assigned zero. For ranking items, the top two categories were assigned correct and lower scores were assigned incorrect. There were a handful of individual items where this allocation resulted in all or nearly all candidates receiving the same score. For these items an individual scheme of dichotomisation was created to maximise the item variance in the dichotomised scores.

Checking IRT models assumptions

10.8. Analysis was undertaken to ascertain whether the underlying assumptions of the IRT models could be met. This includes the assumption that the test is unidimensional (i.e. that all items measure a single construct) and that items are independent of one another. A principle

components factor analysis was used to provide indications of conformity for both of these assumptions.

10.9. In relation to unidimensionality, with both polytomous and dichotomous data, it was found that the first factor accounted for less than 5% of the variance in scores and the first 10 factors accounted for less than 20% of the variance. A best-practice rule of thumb is that to demonstrate unidimensionality, the first factor is expected to account for ~20% of variance and be significantly larger than the subsequent factors.

10.10. In relation to item independence, some evidence was found of local dependence between items, relating to rating scenarios which contain multiple items. This can appear as factors loading dependent items. However, this did not account for all of the small factors identified.

10.11. In summary the FY1 SJT data does not meet the model assumptions for IRT. This is a strong indication against using IRT models. However, an attempt was made to fit IRT models to see if the empirical results were of value.

Fitting IRT models

10.12. Eight different IRT models were explored, in order to ascertain whether any were an appropriate fit for the FY1 SJT.

10.13. 180 items were included in the modelling with the full data set for maximum accuracy. A chi square statistic was used to consider individual item fit for the polytomous models and a Z test of residuals for the dichotomous models ($p<0.01$). In addition to fit to the model, the item parameters were checked to see if they were in a broadly desirable range. Items which failed this check might be identified as having extreme difficulty or very low discrimination. This is summarised in Table 20.

10.14. None of the polytomous models showed good fit across the item pool. The Samejima Graded Response model had reasonable fit for half the items, but inspection of the parameters showed that many were rated as having low discrimination or extreme difficulty. Rasch Partial Credit model had the next most items (~40%) which fit reasonably well to the model. As Rasch models are simpler and use global parameters there are no individual item parameter issues. For the dichotomous models the three parameter model showed better fit with ~70% of items passing the basic fit test to the model.

10.15. These results show that, while some models fit better than others, the fit is relatively poor. These results are not surprising, given that the data does not meet the underlying assumptions of the models.

Table 20: Summary of IRT Model Fit

Model	Description	Number of items with no significant misfit (out of 180)	Well-fitting items with parameter issue	Usable items (out of 180)
Samejima Graded Response (Polytomous)	Allows the rating scale to be defined for each individual item in terms of both discrimination and difficulty of each score point	92	57	35
Generalised Rating Scale (Polytomous)	Allows the rating scale to be defined for each individual item in terms of discrimination but uses common score point band widths for the rating scale	4	1	3
Rasch Rating Scale (Polytomous)	Sets the difficulty parameter for each item but holds the rating scale and discrimination constant across all items	15	n/a	15
Rasch Partial Credit (Polytomous)	Sets the difficulty parameter and band width for individual items but holds discrimination constant across all items	75	n/a	75
Generalised Partial Credit (Polytomous)	Similar to the Rasch Partial Credit Model but includes an individual discrimination parameter for each item	109	77	32
Three Parameter (Dichotomous)	Estimates the difficulty, discrimination and possibility of guessing for each item independently	170	57	113
Two Parameter (Dichotomous)	Estimates the difficulty and discrimination for each item independently but does not model correct responses by guessing.	138	68	70
One Parameter (Dichotomous)	Estimates the difficulty for each item independently but assumes all items have the same discrimination and does not model correct responses by guessing.	27	n/a	27

Equating and scoring

11.1. One of the benefits of IRT models is that they can be used to equate test forms to provide scores on a common scale, thereby negating the need for post hoc equating procedure currently used in CTT.

11.2. For each of the IRT models, the difference between scores on the two test forms were examined. Of the polytomous models all but the GRSM are able to reduce the differences between the forms at least as well as the post hoc equating currently in use. The Three Parameter model is the only dichotomous model which equated well, with the mean difference in the calibrated scores the same as for the post hoc equating.

11.3. When examining the spread of scores between the two forms, it is expected that the standard deviation will be equivalent for the two forms. Using CTT, there is a small difference in SD between the two forms. Most IRT models have a similar result to the CTT currently in use, though the Rasch Partial Credit eliminated differences in spread of scores. It is important to note that the difference observed is very small and could be a real difference in the samples, rather than a failure of the current equating process.

11.4. Overall the IRT models that show the best fit are also effective at equating the scores across forms.

11.5. In relation to the scoring, IRT calibration required adjustment to the item level scoring, meaning there is a difference in total scores. The relationship between the operational scores (i.e. using CTT) and each of the IRT models was examined. the polytomous models that use the same number of points as in the original scoring are closer to the original scores with all correlations above 0.95. The dichotomous scoring models each have correlations of ~0.9, suggesting a slight deviation in what is being measured.

Conclusion

11.6. The operational SJT data has shown that the SJT does not conform well to the unidimensionality requirements of IRT.

11.7. None of the polytomous IRT models fit the data well and the best fit of the dichotomous models was the three parameter model with over 60% of items showing statistical fit. However use of a dichotomous model requires substantial simplification of the response scale, which is potentially problematic.

11.8. Many of the IRT models equated well and had strong correlations with the CTT scores.

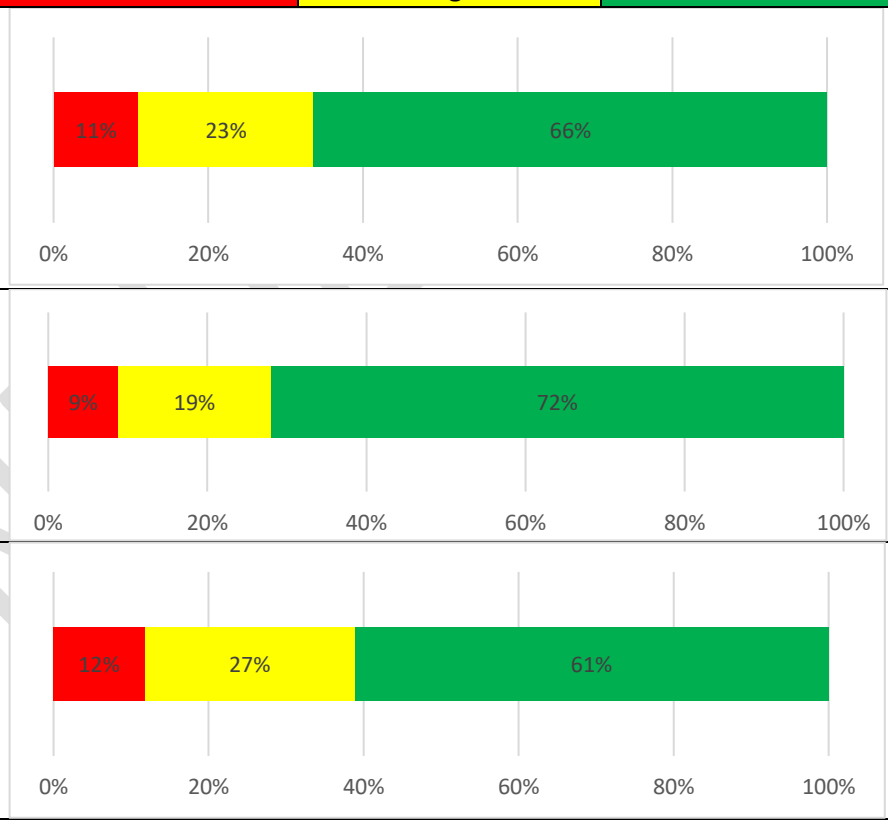
11.9. The initial evidence suggests that IRT would not be suitable, based on the lack of unidimensionality and poor statistical fit relating to the item-level data. Based on current evidence, the use of IRT is not recommended at this time.

12. Candidate Feedback

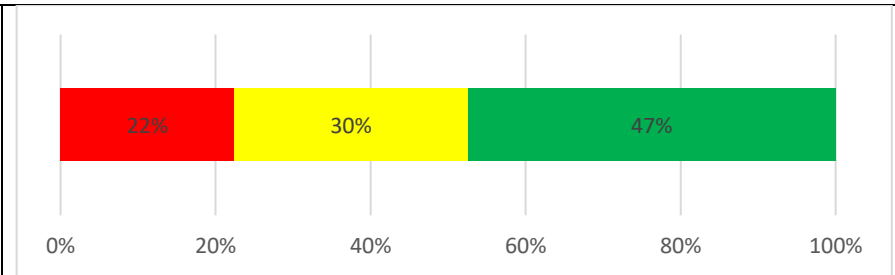
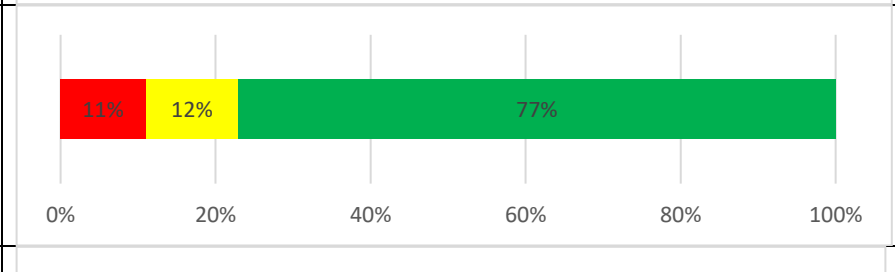
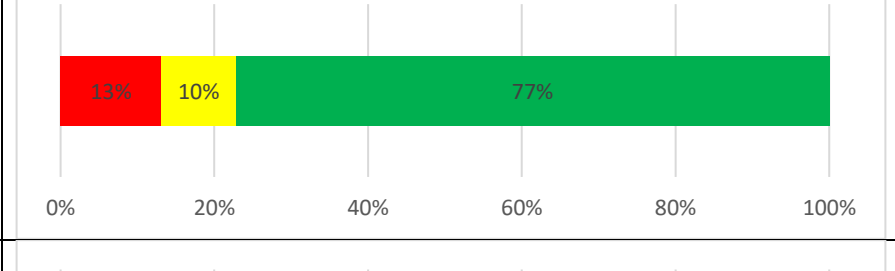
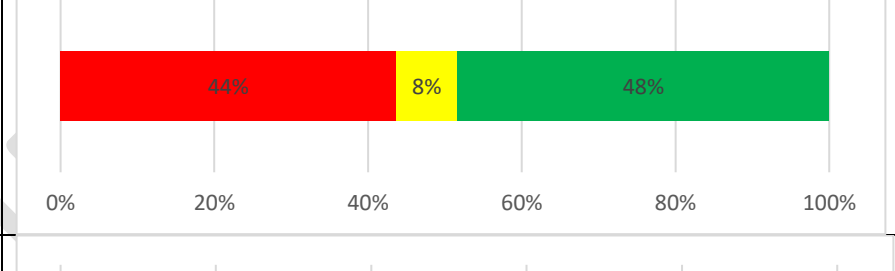
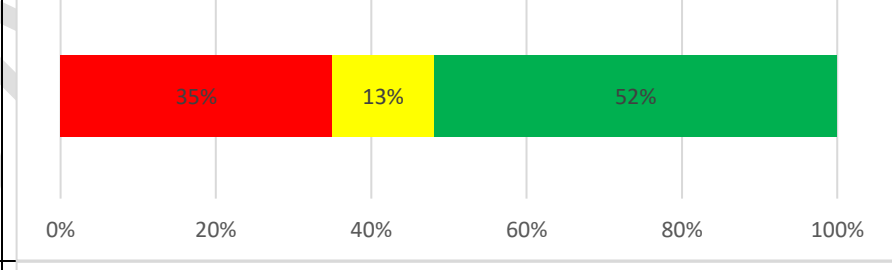
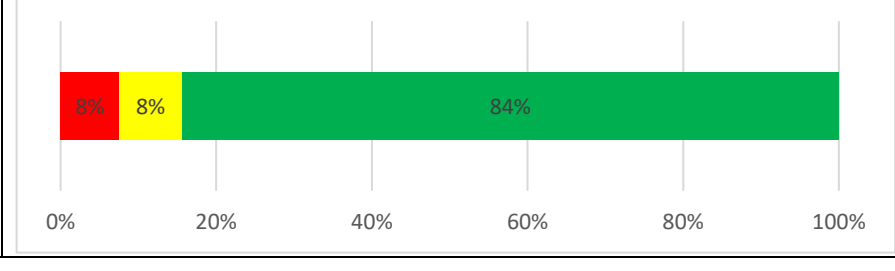
12.1. Participants who completed the operational SJT were asked to complete an evaluation questionnaire regarding their perceptions of the SJT. This feedback has been collated and reported in four key sections below. Overall, 7638 (92.94%) of participants provided feedback. The breakdown of responses to each question can be seen in Table 21. 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.

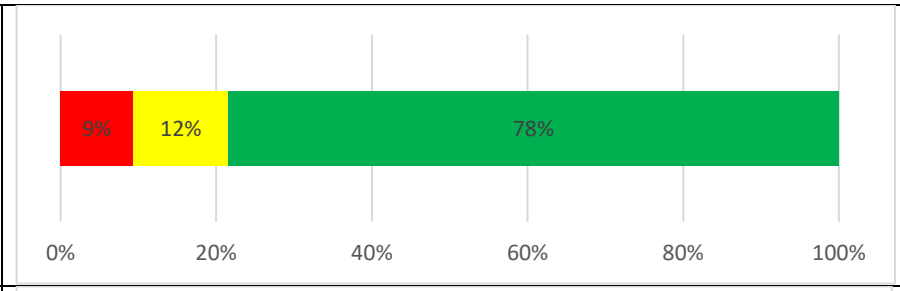
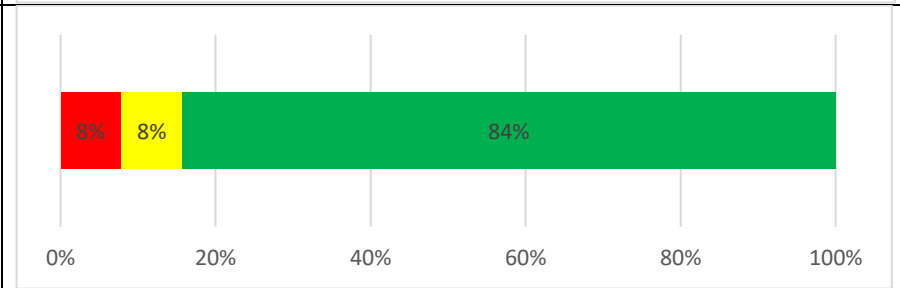
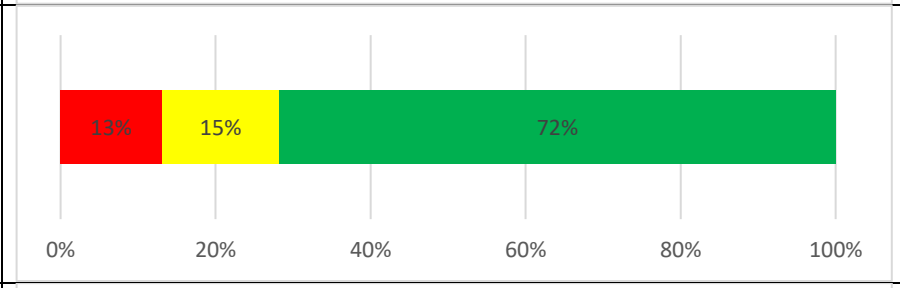
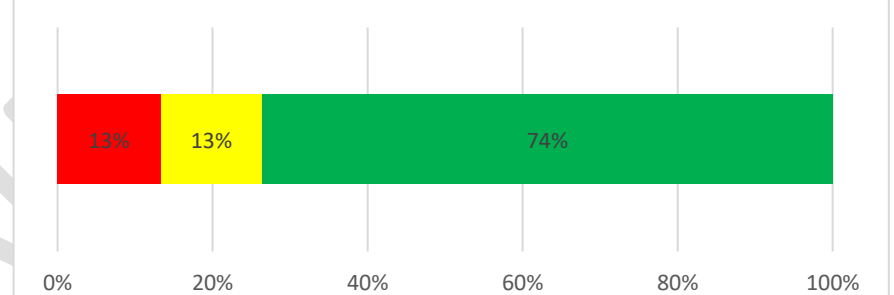
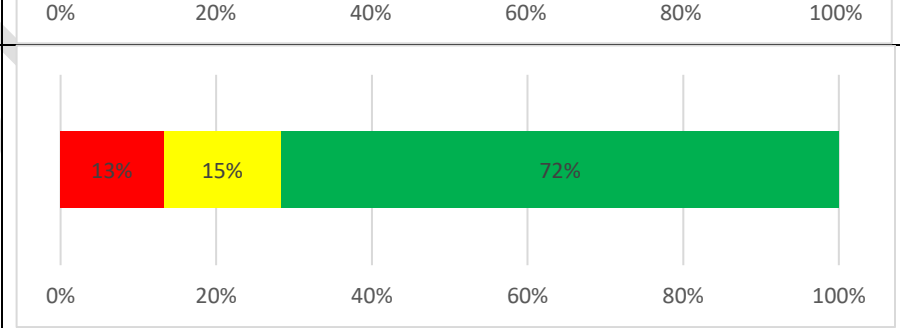
Table 21: Participant feedback on overall test content[6]

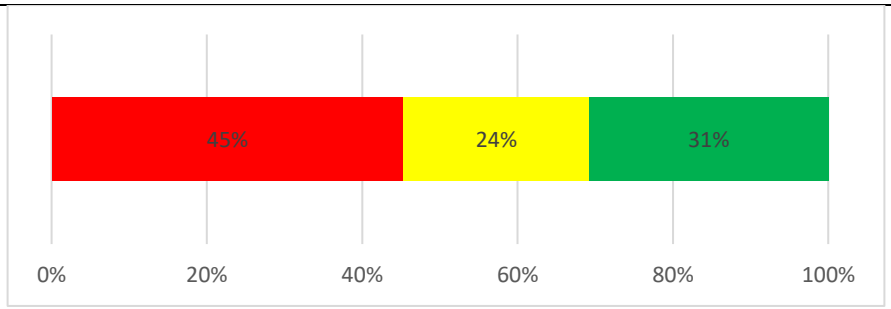
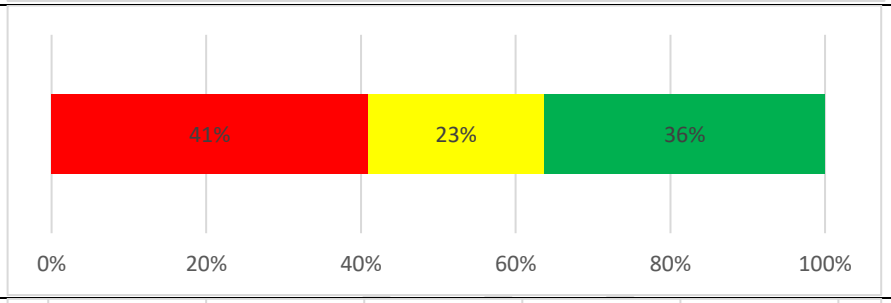
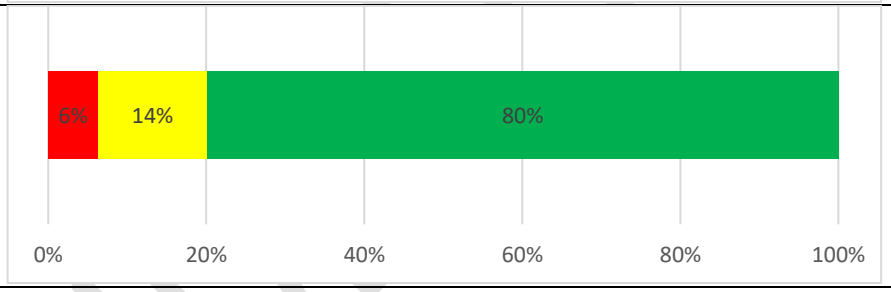
	% Disagree	% Neither Agree Nor Disagree	% Agree
The information I read in the Applicant Guide about the SJT was clear and helpful (n=7638)	11%	23%	66%
The content of the Situational Judgement Test (SJT) was relevant to the role of Foundation Year 1 doctor (n=7605)	9%	19%	72%
The content of the SJT was an appropriate level of difficulty for my training level (n=7584)	12%	27%	61%



[6] For each question, those that did not respond or selected N/A were excluded.

The content of the SJT was fair for all candidates (n=7526)	
The instructions for the SJT were clear and easy to understand (n=7572)	
There was a sufficient amount of time to complete the test (n=7558)	
Booking the test online was straight forward (n=7547)	
I was able to book an appointment that was convenient for me (n=7553)	
I found it easy to read the information/questions on screen (n=7544)	

Computer-based testing is an appropriate way to complete the SJT (n=7537)	
The venue and facilities were appropriate (N/A if you completed at home) (n=6358)	
The online proctoring system was a suitable way to sit the SJT (N/A if you completed in test centre) (n=2112)	
The format for answering the questions was straightforward (n=7521)	
I was comfortable with being asked questions from a range of different response formats (n=7502)	

I am in favour of the "video scenarios" (n=7518)	
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=7491)	
I am in favour of the "evolving" questions, in which some scenarios have multiple parts (e.g. Part A and Part B) (n=7499)	

12.2. **Instructions:** 66% of candidates agreed that the information available in the Applicant Guide about the SJT was clear and helpful. Similarly 77% agreed the instructions were clear and easy to understand. Some comments did request more clarity about the process, for example, **"more clear instructions on what is allowed for those taking the online exam e.g. not allowing watches, headphones, paper and pens. I didn't feel this was clear to me before sitting the paper".**

12.3. **Test administration:** 48% felt that booking the test online was straight forward and 52% were able to book an appointment that was convenient. Of those that completed the test in a test centre, 84% 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.

12.4. **Test content and format:** Candidates generally provided positive feedback towards the overall test content. 72% agreed that the content of the SJT was relevant to the FY1 role and 61% agreed it was appropriately difficult. However, only 47% of candidates agreed the content of the SJT was fair for all candidates. Many of the open-text comments relating to this referred to the introduction of a new format without sufficient practice materials in particular referring to the rating section of the exam; **"I do not think it was fair to introduce a new type of question without providing any guidance or practice papers as to the rationale of the answers".** 77% also felt there was a sufficient amount of time to complete the test.

12.5. Candidates were also asked about the format for answer questions and about the various new scenario types. 74% 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, 31% were in favour of video scenarios, 36% were in favour of speech responses, and 80% were in favour of evolving questions; **"I think the evolving stations are a good way of making scenarios more lifelike, particularly when done as a video format".**

12.6. **Computer-based testing and the testing platform:** 84% 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 comments did speak to the benefits of having a break from looking a the screen mid-way through the exam; **"Allow a short break from the screen to have a drink and use the bathroom if one wished. This is especially important given it is advisable that people working at a desk on a computer should take regular breaks from the screen for eye health, and posture."**

13. Summary

Summary

13.1. This report details the operational use of the SJT for selection to FP 2020, as well as the development of new items which were trialled alongside FP 2020.

13.2. The psychometric analysis of the 2021 operational SJT is positive and shows consistency when compared to previous versions of the SJT for entry into FY1 training 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.

13.3. The SJT demonstrated an overall **good level of internal reliability** (.80 on both papers), which is appropriate for tests administered in high stakes selection context such as FP. **The SJT was capable of differentiating between candidates**, providing a sufficient spread of scores to support decision making as part of selection into FY1 Training.

13.4. Candidates were allowed 2 hours and 20 minutes to complete the 75-scenario test (which includes 10 pilot scenarios). The test completion analysis showed that the **test was not speeded**, with 99.9% of candidates completing the last question on Paper A and 99.8% of candidates completing the last question on Paper B.

13.5. In relation to group differences, the SJT results show significant differences for gender (small effect size), ethnicity (medium effect size) and country of qualification (large effect size). The EPM results also show significant differences for gender (less than small effect size), ethnicity (small effect size) and country of qualification (small effect size). In some cases, this may be exacerbated due to the uneven sizes of the subgroup categories.

13.6. Significant correlations were found between SJT scores and EPM scores. Whilst these correlations are significant, indicating a degree of shared variance/commonality between the assessment methods, there is also a large amount of variance that is not explained by any commonality, indicating that the SJT appears to be assessing different constructs to that of the EPM. This is consistent with the findings of the initial predictive validity study for selection to the Foundation Programme[7].

13.7. In 2021, 86 scenarios were piloted across all three item types; Ranking, Multiple Choice Questions, and Rating. 73% (n=33) of the Ranking scenarios were added to the operational item bank. 54% (n=7) of the MCQ items were added to the operational item bank. 52% (n=86) of the rating responses were added to the bank. Decisions regarding appropriateness of adding the rating items to the operational bank were based on reviewing the individual partials of all the responses to a rating scenario.

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