SJT REVIEW

AN INDEPENDENT REVIEW OF EVENTS RELATING TO THE USE OF SITUATIONAL JUDGEMENT TESTS IN THE *SELECTION FOR FOUNDATION* PROCESS 2013

COMMISSIONED BY HEALTH EDUCATION ENGLAND

LED BY

PROFESSOR JOHN C. MCLACHLAN

WITH

PROFESSOR JAN ILLING

CENTRE FOR MEDICAL EDUCATION RESEARCH
DURHAM UNIVERSITY

FINAL REPORT

Contents

List of Acronyms

DH	Department of Health
GMC	General Medical Council
HEE	Health Education England
ISFP	Improving Selection for Foundation Programme
MSC	Medical Schools Council (Also BMA 'Medical Student Committee)
NES	NHS Education Scotland
SJT	Situational Judgement Test
UKFPO	United Kingdom Foundation Programme Office
WPG	Work Psychology Group

Executive Summary

This Review into the errors affecting the use of Situational Judgement Tests during the Selection for Foundation Process 2013 was commissioned by Health Education England (HEE), at the request of the Medical Schools Council (MSC) and the United Kingdom Foundation Programme Office (UKFPO). It was carried out by Professor John McLachlan, assisted by Professor Jan Illing, both of the Centre for Medical Education Research, Durham University. The Review strategy included telephone interviews with stakeholders including HEE, MSC, UKFPO, General Medical Council (GMC), Improving Selection for Foundation Programme (ISFP), the Work Psychology Group (WPG), Stephen Austin and Sons, and Trax UK, along with representatives from individual medical schools. Copies of relevant e-mails were requested and assembled into a chronological log. Copies of relevant documents were requested (over 80 in total) were also assembled into a chronological and indexed record. Stakeholders were offered the opportunity of submitting a 'statement of perspective', identifying key issues and remediation steps from their point of view. All respondents were frank and open about events, voluntarily identifying areas where errors were made, and suggesting improvements for the future.

The focus throughout was on understanding what went wrong, and identifying what can be done better. While students were the most significantly affected stakeholders, the well-being of future patients must always be the paramount consideration.

The recommendations arising from this Review are listed below, numbered according to Review Sections in which they are found.

Recommendation 1.1: that the use of SJTs in selection for Foundation be continued while evidence is gathered as indicated below.

Recommendation 1.2: that evidence for the validity of SJTs in selection for Foundation context be gathered as a matter of high priority.

Recommendation 1.3: that UKFPO and MSC explore with HEE, DH, GMC and other stakeholders the possibilities of longitudinal tracking of students and doctors through their subsequent careers.

Recommendation 1.4: that an independent psychometrician with relevant health care experience be appointed to the Rules group.

Recommendation 3.1: that the RULES group specifies unequivocally the nature of the process intended. If it is a 'selection for employment' process, then the criteria by which candidates are deemed unappointable should be made explicit beforehand.

Recommendation 3.2: that student members of the Rules Group be invited to contribute to all strategic discussions relating to the use of SJTs.

Recommendation 3.3: that further research is carried out into the nature of very low scores on the SJT.

Recommendation 3.4: that subsequent to the research described in Recommendation 3.3, the Rules Group consider if a remediation programme might be developed for failing candidates.

Recommendation 3.5: that candidates who receive very low scores but are still deemed appointable, should be able to avail themselves of any remediation programmes available for failing candidates, with a view to addressing issues on commencing employment.

Recommendation 4.1: that UKFPO should review sympathetically the cases of candidates affected in this first year of operation of the process, and that these candidates should be able to apply for vacancies that become available through the reserve process.

Recommendation 6.1: that irrespective of whichever company is contracted to carry out printing and scanning in the future, the Medical Schools Council should brief them beforehand on the nature and purposes of the SJT and should remain in close liaison throughout the entire process, not just the end stages of reporting.

Recommendation 7.1: multi and missing mark procedures should be established as a matter of urgency by MSC with the provider companies for the next round of selection.

Recommendation 7.2: realistic timescales and deadlines should be agreed with the commercial providers, taking into account the experiences gathered this year.

Recommendation 7.3: that irrespective of whichever company is contracted to carry out scanning in the future, the Medical Schools Council should brief them beforehand on the nature and purposes of the SJT and should remain in close liaison throughout the entire process.

Recommendation 7.4: that the scanning company be sent both 'attendance' and 'absence' lists to ease the task of checking candidate forms.

Recommendation 7.5: MSC should generate a separate .CSV ('comma-separated values' file storing tabular data as plain text) file for each individual exam date showing only the candidates taking the exam on that date, again to ease checking.

Recommendation 7.6: that adequate time is allowed, not just for scanning, but also for checking the results prior to allocation and informing the candidates of the outcomes.

Recommendation 8.1: Written Risk Analysis and Mitigation policies should be developed to cover the handling of results, including discussions with experienced OMR systems operators.

Recommendation 11.1: the validity review should consider all the components of selection for Foundation, including the decile ranking, and additional points, separately and in combination

Recommendation 11.2: no transcription of responses should be permitted after the expiry of the set time. Consistency of practice across sites is essential.

Introduction

Review Strategy

The errors around use of SJTs in 2013 were distressing and upsetting for the student cohort, at a time when Finals were looming in the minds of many. Those who had allocations changed or even withdrawn were particularly affected. In this review, medical students are therefore regarded as crucial stakeholders, and the significant negative consequences for them must be fully recognised. Events were also difficult and challenging for many of the staff involved in the process. However, in the end, it is the safety and wellbeing of future patients that must be the paramount consideration.

The Terms of Reference arose through negotiation with the Lead Reviewer (JMcL), and were drawn widely rather than narrowly, in order to review the use of SJTs in selection for Foundation as a whole, rather than just the errors occurring this year. The full Terms of Reference are provided in Appendix A. A major aim of the review has been to identify improvements for the future, rather than focusing on the retrospective attribution of blame.

Telephone interviews were held with a wide variety of individuals, including representatives of the MSC, UKFPO, ISFP, GMC, BMA, WPG, Stephen Austin and Sons, Ltd (the printing company responsible for the papers), Trax UK, Ltd (the scanning company), and representatives from individual Medical Schools. Relevant e-mail logs were requested from key participants, and a master chronological record was created of e-mail communication. A document log of all documents submitted, arranged chronologically, was also constructed, and currently contains eighty six individual documents, running from June 2010 till the present. The MSC provided a copy of a Chronological Incident Report, which proved invaluable. No significant challenges to this chronology arose during the Review, and it has therefore been adopted as a reliable account, and will not be re-iterated within the review.

All discussions were conducted in a non-confrontational way. Perhaps as a consequence, all respondents were frank and open about events, voluntarily identifying areas where errors were made, and suggesting improvements for the future.

A constraint was the time scale available, with the HEE indicating a deadline of 19th April 2013 for the receipt of the Final Report. The Lead Reviewer therefore re-arranged his work and vacation schedules to be able to devote the maximum time possible to the Review. By these expedients, the total working time available was commensurate with a longer, more reflective approach. The time available influenced the Review strategy, with telephone interviews being employed, rather than face to face meetings. These interviews were followed by submission of invited written documentation. In particular, a

number of stakeholders were offered the opportunity to submit a 'statement of perspective', outlining their perceptions of events in general. These were used in order to speed up the identification of areas requiring further investigation, rather than following an iterative process in which the Lead Reviewer drafted a summary of conversations, and then sent them to participants to review.

The Lead Reviewer conducted all of the telephone interviews, reviewed all the documentation, and drafted the Review. Professor Jan Illing drafted the summary of events, read and commented on the draft manuscript, and provided an invaluable sounding board for discussion during the process. A number of her valuable suggestions have been incorporated as Recommendations. The first person singular has been adopted throughout, for consistency. The Lead Reviewer accepts responsibility for the final Review, and any errors therein.

The time available precluded draft copies being submitted to stakeholders for their comments. This version must be viewed as provisional in advance of this process, and the right is reserved to issue a modified version in the light of any substantive errors of fact or interpretation which emerge. Some information, particularly about the psychometrics of this year's test were not available before the deadline, and may be issued as a supplement to this Review.

This Report is structured around the Terms of Reference (Appendix A). One feature of the process, the method of scaling and aggregation, was not included in the Terms of Reference but was raised by the BMA Students in their evidence to the Review, and this has therefore received consideration. Several other issues not within the Terms of Reference also proved germane, and receive comment, but not formal recommendations.

A Glossary of Technical Terms is provided in Appendix B.

Summary Timeline of Events

This is derived from the Chronological Incident Report produced by the Medical Schools Council[i] which is accepted as a true record of events.

19.2.13
One error was identified for a single applicant from Imperial College, London. The error was considered to be an isolated incident; therefore the contingency plans to re-scan all 8,162 answer sheets and postpone the matching algorithm did not go ahead. However, checks were performed on all the low scoring applicants, as well as all the Imperial candidates. Manual checks were made to identify any creases on all papers, the cause or the initial error. The supplier confirmed a second scan reported identical results.

20.2.13
The decision to run the matching algorithm was made.

22.2.13
The names of applicants to receive FP places and those to be placed on reserve list were confirmed via FPAS to foundation school managers. Following this, Imperial requested remarking of 9 applicants.

25.2.13
The results of the matching algorithm were sent out via FPAS to all applicants. Imperial were sent the data requested to check the nine applicants (only nine rows of data per applicant were sent, based on previous agreement with MSC and UKFPO). Following this Imperial reported ten errors in the data. The errors identified were that the answers provided were not recorded by the scanner and rubbed out answers were incorrectly marked.

The MSC and the UKFPO decided to determine the extent of the marking errors. Examination of the whole dataset identified three types of errors, affecting 1250 applicants.

26. 2.13
Various explanations were put forward to explain the errors:

- Fast turnaround by the supplier
- Lack of direct communication between Trax UK (who had been subcontracted by Stephen Austin & Sons) and the MSC
- The scanner had initially identified a need for a high number of manual checks on running the first batch of answer sheets. As a single person was operating the machine this was not done. Instead the sensitivity of the scanning machine was changed from 4 to 8 which would result in picking up faint graphite marks including rubbings out.

A teleconference with MSC, UKFPO, Stephen Austin & Sons and Trax UK was held and the following actions were agreed:

- The scanning error to be shared with the BMA and the four CMOs
- The error was communicated to applicants and stakeholders and placed on the UKFPO website
- Plans were made to re-run the algorithm
- The answer sheets were recalled. In particular the 1250 score sheets that were found to have at least one error were checked.
- Staff were recruited from the Universities and MSC and signed a form stating their involvement and access to the confidential data. One person reported having a conflict of interest.

The data was divided by University and pasted into separate workbooks to be examined by one member of staff and checked by a second for accuracy. The original answer sheets were obtained and matched with the electronic copies and checked individually by hand. Some answer sheets were missing. These were obtained by contacting the

medical school and gaining a scanned copy. Errors were recorded and coded. A log was made of the type of error.

A further error (3rd type) was noted during the checking process. This error involved marking two or more responses; however a single response was recorded. Following the early identification of more than 50 errors, applicants were notified and they were alerted to a potential re-running of the algorithm.

Corrections were made to a master copy of scanned data.

27.2.13
MSC issued an apology to applicants (approved by UKFPO and DH). The scanning firm Trax UK reported the turnaround was impossibly fast and that they had difficulty with the different question formats. They also changed the scanning process both before and after Christmas without informing MSC.

Errors were identified by the scanning company before Christmas but in an attempt to correct them; further errors had inadvertently been made following the change in sensitivity to the graphite reading.

A decision was made not to re-scan due to time but also due to awareness that manual verification would still need to be carried out to be confident all errors were identified and corrected.

It was agreed that all medical schools would check the answer sheets with the excel file.

28.2.13
Between 28 February and 4 March each answer sheet was checked with the database and corrected where necessary. In addition to the errors already identified a further error identified the scanner had duplicated answers from one candidate to another. Nine applicants were affected by this error. In total 353 applicants had been affected by scanning errors[ii].

Further Considerations

Challenging and deeply regrettable as the circumstances were, they could easily have been much worse. If the error had not emerged for several more weeks, until the appeal process had commenced, students would have been much further advanced in the process of allocation, and would have made many more irreversible decisions. The crease error, in association with the policy of excluding some candidates with low scores and certain criteria, had the effect of drawing attention to the possibility of errors, and the checking process thus initiated led to the emergence of the further errors.

Review

As indicated in the Introduction, the main body of the Review is organised around the Terms of References, as itemised below.

1. The reasons for selecting Situational Judgment Tests (SJTs) in the first place for ranking students into the 2013 Foundation programme.

The previous use of 'white space' statements for Foundation was viewed as unsatisfactory for a number of reasons[iii]. No validity evidence was available for the use of the white space statements, and in general, evidence does not support the use of personal statements in high stakes selection in health care contexts[iv]. The scoring process was subjective, and therefore likely to be of low reliability[v]. There were also concerns about authenticity, in that candidates could seek assistance in writing their statements, or indeed plagiarise them, and veracity, in that some categories of assertion required further checking. Assessing white space statements was also expensive, requiring considerable expenditure of clinician time in training, scoring and quality assurance. There was also the possibility of administrative errors, such as transcription errors, although I understand there were extensive checking processes in place.

As a consequence, the Medical Schools Council was commissioned by the Department of Health (DH) to lead a Steering Group, comprising the major stakeholders, to carry out an option appraisal for the best approach for selection of applicants into the UK Foundation Programme and allocation to Foundation Schools[vi]. This concluded, *inter alia*, that "*Work should begin to develop and pilot Situational Judgement Tests (SJT) which would assess, under invigilated conditions, the professional behaviour, judgement and fitness for purpose of applicants based on a detailed job analysis*".

This was in association with a new Educational Performance Measurement, which divided candidates into deciles, and included additional points for prior educational achievement such as earlier qualifications, publications and conference participation. These recommendations were supported by DH.

By contrast with the use of 'white space' statements, situational judgement tests have demonstrated validity in a variety of other settings (particularly in the selection of GP trainees). They also have demonstrated reliability in this and other settings, including the SJT pilots carried out during the ISFP process. An extensive review produced by the Work Psychology Group[vii] is an appropriate summary of evidence in this regard.

A comprehensive and professional development process was undertaken by ISFP, in identifying the person specification[viii]. This was used in preparing relevant test materials for pilots, as examples for candidates and for use in the live test.

An appropriate programme of checking assessor concordance on the answers was pursued[ix]. The acceptability of SJTs was explored extensively with stakeholders including students, and in general received very positive evaluations. An extensive series of trials was carried out, which demonstrated that an appropriate degree of reliability could be achieved. The decision to use SJTs in selection for Foundation was therefore fair and appropriate, and the items themselves were as good as they could reasonably have been expected to be.

Recommendation 1.1: that the use of SJTs in selection for Foundation be continued while evidence is gathered as indicated below.

However, no *direct* evidence of validity for SJTs in selection for Foundation was available at the point of their introduction. This is an inevitable consequence of the introduction of a new and large scale assessment process. While SJTs have been demonstrated to be reliable in pilots, their validity is inferential in nature only. This is not inappropriate for a new test, but carries with it the obligation to carry out an organised schedule of continuing validity studies from the time of introduction. There are significant risks in not pursuing a validation strategy.

Recommendation 1.2: that evidence for the validity of SJTs in selection for Foundation context be gathered as a matter of high priority.

I am aware that approaches to this are already in hand. I will not specify the exact nature of such validity tests but will offer some examples. It would, for instance, be possible to compare SJT scores (particularly for low scoring candidates) with exam finals results, with effect from June 2013. A case-control approach could be adopted. By September 2014, it will be possible to compare SJT scores for individual students with subsequent performance in Foundation Year One. Since the construct under consideration is that SJTs are an appropriate measure of performance in Foundation Year 1, it would be plausible to assume that doctors who had gained experience in that setting would be more capable of responding appropriately to the items than final year students. Construct validity could be explored by having Foundation Year 2 doctors undertake an SJT drawn from the item bank, and comparing their performance with that of the students. This could also provide information on the distribution of scores obtained. The assumption that doctors of increasing seniority should perform correspondingly better on the SJT could be tested by enrolling registrars and other senior doctors with relevant experience in the same trial. As indicated later in this report, it will also be essential to collect evidence for the validity of the other components of selection for Foundation (the decile ranking, academic achievements, and previous degrees) separately and in combination.

In the long term, cohort studies would be invaluable in determining the validity of this and other selection tools. Such processes would be greatly aided by a consistent recording process by which medical students were assigned a medical student number by the GMC at an early stage. Associated with this, it may be necessary to collect consent data from the beginning of such a process in order to address issues arising

from ethical considerations, and to explore Data Protection and other legal considerations with regard to information transfer.

Recommendation 1.3: that UKFPO and MSC explore with HEE, DH, GMC and other stakeholders the possibilities of longitudinal tracking of students and doctors through their subsequent careers.

I understand that discussions are already underway in the form of a Medical Selection Outcomes Research Database between NES, HEE and the MSC. The recommendation here is for the centralisation of all data regarding medical careers. This poses formidable challenges, but in the view of the reviewers is essential to long term evaluation of medical training and performance, and is therefore essential to long term patient safety.

Some of the data which will arise from such validation processes will be at once technical and complex in nature, and highly important in strategic and practical terms. I believe that the Rules would benefit from further independent psychometric advice, independent of the Work Psychology Group, who are contracted to provide a service.

Recommendation 1.4: that an independent psychometrician with relevant health care experience be appointed to the Rules group.

2. The design and psychometric properties of this particular SJT test as seen in the pilots.

Much work remains to be done in regard to selection for medical training[x]. An extensive process of piloting and testing was carried out[xi], from 'micropilots' to establish the parameters, through item design, concordance testing, delivery and analysis. These pilots led on to a Parallel Recruitment Exercise[xii], in which the initial pilot observations were confirmed and extended. Without going into technical detail (contained in these last two references), these established that SJTs were reliable in this context, and permitted the development of a large number of technical recommendations concerning delivery. It is difficult to see how this developmental process could have been improved, and it accords with best practice in the field. Piloting of this nature can offer good evidence of reliability, and of the general nature of the outcomes with regard to distributions of responses. It is harder to use pilot tests of this kind to explore validity, however, since only a genuine test has full authenticity for candidates.

3. The decision to use cut offs determined from the mean and standard error of measurement and the process by which this decision was arrived at.

I understand that this possibility was canvassed from early in the process. This was by way of a continuation of the option present in the previous 'white space' selection method. In this, candidates viewed as highly unsatisfactory on the basis of worrying white space statements could be withdrawn from the process. However, the decision to

consider a cut off in the SJTs has been controversial. In a technical report submitted to UKFPO and MSC in April 2012[xiii], the Work Psychology Group indicated that:

> *'Because any cut inherent in these methods will be somewhat arbitrary in nature, it is not recommended that the SJT is used unilaterally to reject candidates. Rather it is recommended that those identified [i.e. flagged as low scorers] should be subject to more detailed assessment to determine their suitability, perhaps through interview or using a clinical skills OSCE approach.'*

This issue, however, requires detailed consideration. Crucially, a distinction should be drawn between an *assessment* process and a *selection for employment* process (see also *Purposes – Competency and Discrimination* in the Glossary).

In assessment processes, criterion referencing and formal standard setting methods are usually employed. Candidates may be deemed to have *failed* or *passed*. Consideration of false-positives and false-negatives are important around the cut score, and some form of re-assessment process is generally employed. Stability of standards from year to year is highly desirable.

Selection for employment processes are usually norm referenced, as a pool of candidates are matched to a smaller number of available posts, and candidates are generally ranked as a consequence. There are three categories of outcome, rather than just *pass* or *fail*.

Candidates may be deemed *unappointable*, *appointable but not appointed*, or *appointed*. The second category will arise when there are fewer vacancies than there are appointable applicants, and will depend on the position of candidates in the ranking. Such *appointable but unappointed* candidates have not *failed* in the way that candidates deemed *unappointable* have.

Re-assessment is not normally used in selection for employment. For instance, candidates unsuccessful at interview are not normally re-interviewed, despite the low reliability of interview processes. Standards for appointment may also vary from occasion to occasion, since it depends on the pool of applicants. A selection for employment process requires a clear person specification, particularly with regard to candidates identified as unappointable.

Is selection for Foundation an assessment process or a selection for employment process? In the Rules Group Minutes for 23rd February 2012, the distinction between assessment and selection for employment is made, and it is concluded, and indeed, apparently insisted on by the student representatives, that the SJT is an application for employment, not a medical school assessment.[xiv] But in this current round, it was possible that all candidates could be allocated a place eventually, albeit through a waiting list. Therefore, the decision this year could only be one of *appointable* or *not appointable*. In future years, a more complex situation might arise, if the number of

candidates exceeds the number of places[1]. But in these circumstances, the SJT should not be used as a work force planning tool.

While it would be commonplace in selection for employment to employ a ranking process without an opportunity for reassessment, it is essential that candidates are aware in advance of the nature of the process. Where criteria are available these should be clearly signalled to the candidates in advance.

It was not absolutely clear from the information to candidates provided in the Handbook which of these two processes is underway, as discussed in the next section.

Candidates deemed unappointable in this round had the opportunity of re-taking the SJT in the subsequent year, so this decision was not quite as final as is normally the case in employment issues.

Could the SJT bear the weight of selection for employment?

It would generally be viewed as inappropriate to use a norm-referenced approach in selection for employment to determine which candidates are unappointable, without some other kind of further review process[2]. No matter what the absolute quality of the cohort, some candidates will always fall below any defined cut off.

However, it is clear that the Rules Group envisaged the use of person specification criteria in making the decision about withdrawing candidates from this round. These criteria were discussed at the Rules Group meeting of 22nd November 2012, and are summarized in a document marked "Management Confidential" of January 24th 2013, produced in advance of the Review of low scorers.

These criteria include, but are not limited to:

- an inability to comprehend written English at a speed appropriate to the requirements of an F1 role and at a level to allow effective communication;
- inability to deal effectively with pressure and/or challenge;
- inability to prioritise tasks and information and take appropriate decisions;
- lack of familiarity with or failure to demonstrate an understanding of the major principles of the GMC's *Good Medical Practice (2009)*;
- failure to demonstrate sufficient attributes of an F1 doctor.

[1] A further complication, outside the scope of this review, lies in the nature of Foundation Year 1 posts, which represent both employment and the 6th year of medical training. On this basis, it seems inappropriate that candidates who graduate from medical school in good standing, and perform acceptably in the SJT as to cause concern, should face the risk of not gaining a Foundation Year 1 post. In terms of the above discussion, a candidate who was viewed as 'appointable but not appointed' would also have been deprived of the chance to complete training in a timely manner. Since this well outside my Terms of Reference, I can make no formal Recommendation. However, I believe that HEE should address this issue, in association with other stakeholders, including students.

[2] Note however, that the typical distribution of SJT outcomes is unusual. It is highly leptokurtic and negatively skewed, with a long tail of very low scoring candidates, and may deviate from the requirements for normality.

These were operationalized as the following outcomes:

a. ***30% or more of the questions have not been answered***.

b. ***Answers have been given in a systematic random fashion*** (e.g. A,B,C,D,E in response to each question).

c. ***The worst option has been selected first for 50% or more of the questions In Part 1 of the paper*** (sort five options into the most appropriate sequence).
For example, if there are 40 questions in Part 1, the applicant would need to have answered 20+ questions in this way to fall into this category.

d. ***The score for Part 2 of the paper is a third or less of the total score available for Part 2*** (select the three most appropriate options out of five).
For example, if there are 20 questions in Part 2 and each has a maximum score of 12, there is a total score available of 240. Any applicant scoring 80 or less would fall into this category.

As I understand it, what happened in practice was that a norm referenced cut off (2.5 Standard Deviations, SD, below the mean) was mooted, but used in practice as a screen. Candidates falling below this value were reviewed individually against these criteria, derived from the person specification. Eventually, on the basis of this individual review, it was determined that those candidates who fell 4 SD below the mean *also* failed on the criteria. In retrospect, it may not have been necessary to invoke a norm referenced cut score at all for these candidates, and this may merely have added to the confusion. Equally, given that patient safety is the paramount consideration, it could be considered appropriate to have the ability to determine whether or not a particular candidate should proceed to employment on the basis of worrying responses in the Situational Judgement Test.

Whether or not these particular criteria and their operational equivalents are appropriate to this task is a matter of expert judgement. In my view, against proposed standards for defensibility (see *Defensible* in the Glossary); the Rules Group were credible judges; there is sufficient research evidence to support the use of SJTs in high stakes medical selection; the methods were practicable; due diligence was present in delivery; and the outcomes were not implausible. With regard to this last point, it is by no means implausible to expect that out of over 8,000 applicants, most (but not all) of whom will graduate from medical school, there might nonetheless remain 12 potential 'false positives' (See Glossary) detectable by the SJT under the criteria described above.

It has been represented to me that the SJT should not bear a greater weight than the preceding five years of medical training and assessment. While this view has credibility, it is nonetheless possible for a candidate to demonstrate evidence of behavior giving rise to concern with regard to patient safety in a number of ways, of which the SJT is one. Inevitably, there will be 'false positives' graduating from medical school, where all of the assessment outcome decisions are made as expert judgements against a social or educational construct. The SJT is an additional test to graduating from medical

school, not a replacement test. In the end, patient safety should be the over-riding consideration. This is fully acknowledged by all participants.

For future years it might be appropriate to confirm explicitly that a particular cut off determined on a norm referenced basis will be used only as a starting point for screening candidates to confirm whether or not they conform to the person specification. However a norm referenced cut off should not be used as the sole determinant of whether or not a candidate should continue in the process. It is the nature of norm referencing that some candidates will always fall below any given cut score determined by use of standard deviations, but this does not necessarily say anything about their absolute level of performance.

Recommendation 3.1: that the RULES group specifies unequivocally the nature of the process intended. If it is a 'selection for employment' process, then the criteria by which candidates are deemed unappointable should be made explicit beforehand.

Perhaps unfortunately, at various Rules Group meetings to discuss the use of cut offs in this way, BMA student representatives were not present, on the grounds that confidential information might also be discussed. Evidently, student representatives should not be present when individual SJT Items, or individual students, are discussed. But since the use of cut offs is a strategy decision rather than a tactical issue on question content, it would have been better if student representatives had been present during this discussion.

Recommendation 3.2: that student members of the Rules Group be invited to contribute to all strategic discussions relating to the use of SJTs.

It is nonetheless of considerable interest as to why some candidates recorded such low scores on the SJT. It would be invaluable to explore this as a research project, possibly including the failing candidates themselves.

Recommendation 3.3: that further research is carried out into the nature of very low scores on the SJT.

It is also relevant to consider if these reasons for low scores are remediable, in order to assist failing candidates understand how they might improve.

Recommendation 3.4: that subsequent to the research described in Recommendation 3.3, the Rules Group consider if a remediation programme might be developed for failing candidates.

Candidates who score very low, but are still deemed appointable, may well have issues that could usefully be addressed before they commence work. If there is a remediation programme in place for failing candidates, then perhaps this could be extended to those low scoring candidates in advance of commencing work. Perhaps future employers

would also be able to assist in development of structured support in the work place for these low scoring candidates.

Recommendation 3.5: that candidates who receive very low scores but are still deemed appointable, should be able to avail themselves of any remediation programmes available for failing candidates, with a view to addressing issues on commencing employment.

4. Information circulated to candidates in advance of this decision.

Notwithstanding the appropriateness of the decision to withdraw some candidates from the process on the basis of their performance in the SJT, there is a separate issue about the transparency of the process.

The Applicants' Handbook provided the following information to students:

> *"If you achieve an exceptionally low score compared to the rest of your cohort (i.e. very extreme outliers, from 0-1%) you will have your paper reviewed, but we anticipate this will only be a very small number of applicants. We are not actively looking to exclude applicants in this way, but if the situation arises we will review your answer sheet, and you may be asked to undertake additional assessment so that we can be sure that you do not pose a risk to patients and their safety. If this additional assessment is unsatisfactory, you may be withdrawn from the process, but will be able to reapply the following year".*

This paragraph indicates that there may be a reassessment opportunity but does not confirm that this will definitely be the case. It does indicate that patient safety is the overriding concern, but does not offer guidance as to how this might be assessed. I understand that UKFPO's view was that a degree of ambiguity was intentional in order to preserve flexibility within a new process. However, the consequence was a lack of clarity for candidates, which is undesirable in a high stakes process. In my view it would be appropriate for UKFPO to recognize this, as a gesture of goodwill. I am aware that an Appeals process is under way. Notwithstanding the outcome of these appeals, I believe it would be appropriate for UKFPO to allow these candidates to apply for vacancies that become available through the reserve process.

Recommendation 4.1: that UKFPO should review sympathetically the cases of candidates affected in this first year of operation of the process, and that these candidates should be able to apply for vacancies that become available through the reserve process.

5. The psychometric properties of the SJT test as delivered (including for example Theta Curves showing sensitivity at different score values).

The full psychometric analysis of the SJT could not be made available to me on the time scale of this review. The Lead Reviewer has requested this information to be made

available to him as soon as possible, and will consider releasing a Supplement to the Review should it prove appropriate.

6. The selection of the organisations to operationalize the handling and delivery of the SJT marking.

Full documentation was made available to me on this process, including some commercially sensitive material.

The field available for award of these specialised contracts was very limited, especially since the project is comparatively small compared to other national and international contracts such as that for GCSEs. The printing company selected (Stephen Austin) had a successful track record in this area. The decision to use this company was therefore appropriate at the time. Similarly, the scanning company sub-contracted by Stephen Austin, Trax UK, also had a successful record in this area, and was also a justifiable choice at the time. As far as I can ascertain, the tendering and contract processes were carried out appropriately, given the very limited options available.

The decision to allocate subsequent contracts for the SJT process will of course be a commercial decision. However, I do not recommend that the current companies involved be excluded from this process. In the way that a reflective doctor who discovers s/he has made a mistake may well be less likely to repeat that error in the future, an organization which has openly recognised errors, and addressed them, may well be an appropriate choice to continue with the process on an on-going basis.

It is clear from my discussions with representatives of Stephen Austin and Trax UK that Trax UK in particular did not fully appreciate the nature of the assessment, its high stakes nature, and the ranking process which resulted. They indicate that they would value direct liaison with the client through all stages in the process.

Recommendation 6.1: that irrespective of whichever company is contracted to carry out printing and scanning in the future, the Medical Schools Council should brief them beforehand on the nature and purposes of the SJT and should remain in close liaison throughout the entire process, not just the end stages of reporting.

7. The operational execution of the SJT marking and an analysis of the errors that occurred.

A detailed account of the errors which arose is given in the MSC SJT Scanning Error Incident Report prepared for the Rules Group meeting of 21st March 2013. No contrary evidence arose in the course of my investigations, and I therefore accept this as a true record, and will not re-iterate its detailed findings. In summary, there were a variety of issues that arose.

1) A clock mark printing error on two response sheets, on one of which candidate's answers were transposed by one cell, leading to a very low score.
2) Duplication of two sets of responses.
3) Errors involving the grey scale scanning setting which was changed during the process.
4) Operator errors in making manual interventions.
5) A degree of confusion over the checking processes.

The majority of the errors arose with respect to (3) and (4). This is a relatively common challenge in optical mark reading systems, and in retrospect, clear agreed policies should have been in place to deal with events such as the presence of multimarks. The MSC SJT Scanning Error Incident Report indicates a number of detailed options for change to address this and other errors in the future. I will not attempt to constrain this option appraisal, since it depends on factors such as accuracy subsequent to any changes, and cost, which cannot be assessed at the moment. I will merely make the general recommendation that these be explored fully and as a matter of urgency.

While the events at the point of scanning with regard to sensitivity settings and operator entry error may seem the most egregious, in reality there were a number of latent errors in the system, particularly involving timelines, and communication, which require to be addressed. See Section 8.

Stephen Austin and Sons indicated that it would be possible to send the response sheets direct to the scanning company, removing one source of error and possible delay.

Trax UK have also indicated that provision of lists both of candidates attending and candidates absent would be helpful in speeding up the process.

The net effect of reviewing the whole process is to confirm that it conforms to a James Reason style 'Swiss Cheese' error[xv], in which the active errors took place during the scanning process, but there were a number of latent errors present throughout the system.

In retrospect there were a number of key vulnerabilities present in the process.

First, there was lack of clear policy on handling exceptions such as missing marks and multimarks.

Recommendation 7.1: multi and missing mark procedures should be established as a matter of urgency by MSC with the provider companies for the next round of selection.

Second, and with hindsight, the timescales available for the process seem to have been determined by considerations of reporting deadlines rather than practical deadlines for the delivery of the work. Stephen Austin and Trax had both accepted these deadlines in advance, but these proved impractical, particularly when slippage began to occur.

Recommendation 7.2: realistic timescales and deadlines should be agreed with the commercial providers, taking into account the experiences gathered this year.

Third, communication between MSC, Stephen Austin and Trax could clearly have been improved particularly at the early stages of the process, with regard to the nature of the test. The scanning company should clearly understand the uses to which the data will be put, the high stakes nature of the process, and the implications of a ranking system. In assessment processes such as GCSEs, candidates are divided into grade categories. An error will affect an individual student, and will only have a significant consequence if they lie at a grade boundary. Since schools are liable to challenge any unexpected results, there is a candidate-led scrutiny process for unexpected decisions. With Selection for Foundation, however, an absolute ranking is created; and an error for an individual candidate will as a consequence displace all candidates lower in the ranking. These candidates may lie at a high stakes decision boundary, such as that between being allocated first and second choices or indeed being placed on the reserve list. This makes the accuracy required much higher than in a grade based system, and this should be entirely clear to the companies involved beforehand.

Recommendation 7.3: that irrespective of whichever company is contracted to carry out scanning in the future, the Medical Schools Council should brief them beforehand on the nature and purposes of the SJT and should remain in close liaison throughout the entire process.

Other detailed suggestions have emerged during the course of this Review, not covered by the MSC Report, and those with particular merit are indicated below.

Recommendation 7.4: that the scanning company be sent both 'attendance' and 'absence' lists to ease the task of checking candidate forms.

Recommendation 7.5: MSC should generate a separate .CSV ('comma-separated values' file storing tabular data as plain text) file for each individual exam date showing only the candidates taking the exam on that date, again to ease checking.

General accuracy

A query was raised in the MSC SJT Scanning Error Incident Report about the generic accuracy to be expected in OMR scanning processes. The company had quoted an accuracy of 99.99 %, which sounds very impressive. If however, this is the transaction error rate, where a transaction is each entry made by a candidate, and with over 8000 candidates making 260 responses, then MSC calculate that over 200 errors will be made, each of which might affect a different candidate.

However, another way to look at it relates to the *category* of error made. Here, issues (1) and (2) affected 2 candidates each, while issues (3) and (4) affected 416 errors in total (these are grouped together because they are linked). If issues (3) and (4) had been avoided, then the accuracy would have been very high.

It should be humanly possible to put policies in place to ensure that these particular errors are avoided in future. Undoubtedly, other category errors may emerge, since only experience tends to reveal these, although capturing the experience of others may help. And no system operated by humans will be free of individual error. It would seriously harm the credibility of the process, however, if a significant number of errors arose next year. It is therefore essential that adequate time is allowed, not just for scanning, but also for checking the results prior to allocation and informing the candidates of the outcomes. I will not constrain the choices of further checking mechanisms (as identified in the MSC Scanning Error Incident Report), since again there are unpredictable issues of consequential accuracy and cost.

Recommendation 7.6: that adequate time is allowed, not just for scanning, but also for checking the results prior to allocation and informing the candidates of the outcomes.

8. The existence and content of any Risk Analyses and Risk Mitigation policies in place in advance of the execution of the programme.

An extensive set of Risk Analyses had been prepared throughout the process, under the guidance of the Project Director. These are a good model as far as they go. As far as I can determine no formal risk analysis or risk mitigation policies were in place with respect to handling the results of the SJT. My understanding is that UKFPO had taken the view that the extensive trialing process had represented an exploration of possible risks in the implementation of the process. Whilst this is understandable, in that it is very difficult to foresee what might go wrong in a novel process, it would nonetheless be desirable to implement a further Risk Analysis with respect to the handling of the SJT results, with corresponding Risk Mitigation actions identified in advance. This Risk Analysis could usefully be informed by discussion with other large organisations using Optical Mark Reading for high stake purposes in health care education, who may already have encountered the kinds of errors that can arise. Examples might include large UK medical schools or Royal Colleges. The contacts here should not be confined to senior staff: often it is the operators of the equipment themselves who have relevant practical experience.

Recommendation 8.1: Written Risk Analysis and Mitigation policies should be developed to cover the handling of results, including discussions with experienced OMR systems operators.

9. The responses to the discovery of errors.

On discovery of the unfolding catalogue of problems, a number of possible strategies presented at each stage. In particular, one crucial choice was on whether to re-scan the entire set of responses, or to re-check manually. Despite the challenges it presented, I believe the decision to re-check the scores manually was undoubtedly the correct one, and that any other decision would have posed serious problems further down the line.

I was particularly struck by the evident hard and devoted work of the senior administrative staff in UKFPO and MSC, particularly Janet Brown, Sharon Witts and Siobhan Fitzpatrick, and the concern they showed for students during the process.

10. The communication strategy subsequent to discovering errors

Communication was frank and open, and this is to be commended. As sequential errors came to light, and each one was signaled in turn to Medical Schools and candidates, with regular changes, sometimes contradictory of previous communications, in the information provided. An impression may have been conveyed of confusion within the MSC and UKFPO. One comment from the BMA Medical Student Committee evidence was critical of the timing of a particular e-mail, in the evening. However, this was plainly a judgement call on the part of UKFPO to begin communication as soon as possible.

The rapid communication policy represented an appropriate attempt to keep participants informed in a rapidly developing situation. In the spirit of full disclosure which is most appropriate to helping avoid such situations arising in the future, this is the proper policy. I also note that an apology was issued to candidates by the MSC, and this full acknowledgement of errors is also appropriate and commendable.

11. Other issues arising

Scaling and aggregation

This issue was not part of the original Terms of Reference, but since it featured in the student evidence to this review, it will be considered here.

This subject is technical in nature, but the essence is that if two scores are aggregated, the one with the highest variance will contribute most to the weighting, even if this is not intended. A valuable report on this issue was produced by the Work Psychology Group, comparing the impact of various different scaling algorithms, but not unequivocally

recommending one. However, this report focused on the SJTs and the decile rankings for candidates: in reality, significant components of the final EPM are the additional points for previous degrees and academic achievements. The distribution of these aspects does not seem to have been considered separately, and is essential to this discussion. Further, although the 'design principle' of the ISFP was that the EPM and SJTs were to be weighted equally, as far as I know, this decision was arbitrary. If, for instance, the total EPM were to prove to be of lower validity *and* reliability compared to the SJTs, then this might appropriately be considered in the final weighting. Such issues cannot be resolved on the evidence available: in the absence of validity information, all choices are essentially arbitrary. But the validity studies recommended should consider all the components separately and in combination, not just the SJTs.

Recommendation 11.1: the validity review should consider all the components of selection for Foundation, including the decile ranking, and additional points, separately and in combination

Transcribing responses outside the time period of the SJT.

In a very small number of cases, some candidates had recorded their responses on the answer sheet, and sought additional time to transcribe them to the response sheet. Unfortunate though this is, allowing such a practice for individuals creates inequity between candidates, who in general were working to the time provided for all aspects of the test. No further transcription should take place at the expiry of the normal assessment time, or the extended time for candidates with special needs, and this should be made particularly clear to candidates on the commencement of the exam.

Recommendation 11.2: no transcription of responses should be permitted after the expiry of the set time. Consistency of practice across sites is essential.

12. Recommendations for future policies and procedures.

(See Executive Summary for all of the Recommendations gathered together)

Appendix A. Terms of Reference

1. The reasons for selecting Situational Judgment Tests (SJTs) in the first place for ranking students into the 2013 Foundation programme.

2. The design and psychometric properties of this particular SJT test as seen in the pilots.

3. The decision to use cut offs determined from the mean and standard error of measurement and the process by which this decision was arrived at.

4. Information circulated to candidates in advance of this decision.

5. The psychometric properties of the SJT test as delivered (including for example Theta Curves showing sensitivity at different score values).

6. The selection of the organisations to operationalize the handling and delivery of the SJT marking.

7. The operational execution of the SJT marking and an analysis of the errors that occurred.

8. The existence and content of any Risk Analyses and Risk Mitigation policies in place in advance of the execution of the programme.

9. The response processes implemented on discovery of errors.

10. The communication strategy subsequent to discovering errors.

11. Summary recommendations for future policies and procedures.

Aspects not included were the use of the Educational Performance Measure, and the award of additional points reflecting other aspects of educational performance such as previous degrees, publications, etc.

Appendix B. Glossary.

Arbitrary

The word 'arbitrary' has several meanings, one of which certainly is 'capricious', but another accords with 'judgement', as in 'arbitration'. See **Standard**.

Criterion Referenced

Based (in principle) on some absolute standard of knowledge or performance. See also **Norm Referenced, Context Referenced**.

The usual steps are:

- First, *define* a group of experts
    - (knowledge of subject, knowledge of context, knowledge of assessment, knowledge of students)
- Then establish the minimum required size of the expert group
    - Frequently taken to be about 8, some evidence that 10 are needed if there is no feedback on item or candidate performance, 6 if there is.
- The experts may make judgements on *test items (rational)* or *test takers* (*pragmatic*)
- In principle, judgements on test items are *prospective,* judgements on test takers are *retrospective*.

Defensible

If assessment is always arbitrary, what are the characteristics of a "Defensible" standard? Norcini and Shea (1997) suggest the following exploratory questions:

- Are the judges credible?
- Is the method used supported by a body of research evidence and data?
- Is the method practicable (too complex a method can lead to errors)
- Can 'Due Diligence' be demonstrated? (e.g. exam security, lack of bias)
- What are the outcomes? ("If you have an outcome which violates common sense then there is something wrong with the standard")

Experts, Expert Panel

The idea of the 'expert' involved in standard setting can be defined in different ways ((http://www.edmeasurement.net/5221/Angoff%20and%20Ebel%20SS%20%20-%20TDA.pdf)). However, I propose a simpler definition. The individual expert must be an expert in the domain under assessment, must have at least a basic understanding of assessment processes (including the particular assessment under consideration), and most crucially of all, be thoroughly familiar with the level at which candidates should be expected to operate. This requires familiarity with the normal capabilities of those working at the level of the candidates. Criterion referenced methods fail when there are unrealistic positive or negative expectations of the appropriate level of performance by the candidates.

False Negative

This term refers to candidates whose 'true score' would meet or exceed the required threshold, but whose actual score (the 'true score' plus the 'error score) on a particular occasion does not reach the threshold. The implication is that those candidates would be appropriate to go into practice, but do not have the opportunity.

False Positive

This term refers to candidates whose 'true score' would not meet or exceed the required threshold, but whose actual score (the 'true score' plus the 'error score) on a particular occasion does reach the threshold. The implication is that those candidates would not be appropriate to go into practice.

Grade

A Grade represents a qualitative description of performance on an assessment (see **Score**). For instance, 'acceptable' and 'unacceptable' might be awarded or more complex outcomes such as 'unsatisfactory', 'borderline', 'satisfactory' and 'merit'. There is no fixed relationship between a score and a grade (so the pass mark is not always 50%). The term 'mark' conflates the concepts of score and grade, and is avoided in this report. 'Cut score' is frequently used as defining the boundary between one grade and another, in preference to 'pass mark'.

High Stakes

When important consequences arise from an assessment, it is generally described as 'high stakes'. Summative assessments in medicine are almost by definition high stakes. A high stakes exam should be clearly defined as to purpose. It should be 'blue printed' i.e. matched against a body of knowledge which must itself be defined in advance. The development of assessment items requires assessors to be trained, benchmarked and audited. Assessment items should be field tested, and there should be a feedback loop which allows for performance (see below) to be evaluated. The size of the assessment

must be suitable to the task. Appropriate standard setting methods must be employed, involving expert staff. Storage and delivery of the assessment items must be secure.

Item Performance

Assessment items can be more or less easy. This property is called *Facility*. If the question is easy, then most candidates can answer it correctly (high facility). Conversely, if a question is difficult, few students can answer it (low facility).

The *Discrimination* of a question shows the range of responses it receives. It might be helpful to think of discrimination as being like the standard deviation of the distribution of the answers, while facility is in some ways like the mean.

Finally, a question may be answered correctly by strong candidates and incorrectly by weak candidates. This can be thought of as a correlation (and for MCQs, is calculated as the *Point Biserial*). The situation of interest occurs when strong candidates tend to get an individual item wrong, suggesting that there is something wrong with the item.

A sophisticated way of looking at the performance of each individual assessment item is *Item Response Theory*. This approach is used by professional testing organisations, such as the Australian Counsel for Educational Research (ACER) and the National Board of Medical Examiners (NBME) in the USA.

Once the performance of individual items has been determined, these can be combined in various ways according to the purpose of the assessment. For instance, a competency assessment can be designed to be most sensitive in the pass-fail zone, while a discriminator assessment might combine items with a much wider range of facilities and strong discrimination properties.

Low Stakes

A test which does not in itself lead to serious consequences. It is frequently considered that lower assessment standards may be required of a low stakes test. A number of low stakes assessments may be aggregated to give a 'high stakes' outcome. In such cases an approach such as Generalisability Theory must be used to confirm that a sufficient number of tests are employed to give valid and reliable outcomes.

Norm referenced; normative

These terms refer to standard setting based on how an examinee performs against a reference population (e.g. those who took the test). See **Criterion Referenced; Context Referenced**

Criterion referencing is more commonly employed in medical assessment. However, norm referencing is still entirely appropriate where a set number of places have to be filled (as a ranking). It is more reliable than criterion referencing, especially with high performing students, and is more robust than criterion referencing (hawks and doves

often agree on the relative ranking, but disagree on the absolute grading). It may be used serially as in a Progress Test. An interesting question is whether a minimum number of candidates are required for its employment, and there is no clear context-independent answer to this.

Purposes – Competency and Discrimination

Assessments can be intended either to assess competence ('do all candidates meet a minimum standard?') or to discriminate between candidates ('where do candidates fall with respect to each other on a particular scale?'). Each assessment should be designed for its purpose. For instance, a competence assessment should be most sensitive at the borderline between pass and fail. Discriminator assessments, by contrast, may be designed to be most sensitive in the middle of the range, where most candidates are found. And, naturally, the scoring and reporting scales are different for each kind of assessment. For competence assessments, only two scale points are required – pass/fail, competent/not competent, both for individual assessment items and for the assessment items as a whole. For discriminator assessments, many more points are necessary, and the fineness of the scale required relates to the number of candidates and the intended purposes of the discrimination.

Competency Assessments benefit from Criterion Referencing approaches, while Discriminator Assessments benefit from Norm Referencing.

Purposes – Formative and Summative

Similarly, the distinction between formative and summative purposes is well known – formative assessments offer feedback to candidates and summative assessments determine progression. A widely agreed assessment principle is that formative and summative tests should be kept separate. For instance, Stern (2006) says "*Evaluators must decide the purpose of evaluation prior to developing an evaluation system…Educators planning both formative and summative assessments should use separate and independent systems*". However, all summative assessments can have formative consequences.

Reliability

Reliability is the degree to which an assessment measures with consistency.
There are several different ways of approaching this.

In Classical Test Theory (also known as Classical Measurement Theory, 'True Score' Theory), it is assumed that any given Score consists of a True Score plus an Error. The error is treated as being of one kind, and it is assumed that the Error can be estimated. Typical tools for exploring this kind of error are Test-Retest estimates, Cronbach's Alpha and tests of inter-rater reliability such as Kappa.

In Generalisability Theory, errors are treated as arising from a number of sources, each of which can be explored and measured separately. More technically, it considers all sources of error (factors) and their interactions, e.g. candidate, marker, item, student-with-item, marker-with item, marker-with-student, and marker-with-student-with-item.

In Item Response Theory, the underlying construct is that there is a relationship between the probability of a candidate answering the question correctly, and the ability of the student. This is expressed as the Item Characteristic Curve. This sophisticated, powerful but complex interpretation is widely but probably exclusively used in national and large commercial testing organisations.

Score

A Score is the raw performance on an assessment (see **Grade**). There is no fixed relationship between a score and a grade. The term 'mark' conflates the concepts of score and grade, and is avoided in this report. 'Cut score' is frequently used as defining the boundary between one grade and another.

Standard

A standard is a statement about whether an examination performance is good enough for a particular purpose. It is based on expert judgement against a social or educational construct, and in that sense, as Case and Swanson (1996) state: "Standard setting is always arbitrary but should never be capricious". See 'Arbitrary' in this regard.

Utility

The Utility of an assessment was helpfully summarised by van der Vleuten (1996) as
Utility = V x R x E x A x C
Where

V = Validity
R = Reliability
E = Educational Impact
A = Acceptability
C = Cost

However, this might better be described as a general relationship than an equation, and the construct of Defensibility (capable of withstanding professional or legal challenge) should be added. Hence, a better formulation might be:

Utility is a function of Validity, Reliability, Educational Impact, Acceptability and Cost and Defensibility.

Validity

Overall, Validity is the degree to which a test measures what it is intended to measure. It relates to Reliability in somewhat complex ways - a measure with low Reliability is sometimes described as being excluded from having high Validity - but Reliability and Validity cannot be traded off against one another in a simple way as is sometimes assumed.

There are a variety of sub-types of validity. Their meanings may sometimes be controversial, but the following operational definitions are used here.

Face Validity: Whether an item makes sense to a panel of experts. One can usefully ask this of one item or question.

Content Validity: Whether the items in an assessment accurately represent the domain being tested e.g. fair sampling. One can usefully ask this of one test or group of items.

Criterion Validity: Drawing inferences between scale scores and some other measure of the same construct. One can usefully ask this of one or more tests.

There are two sub-varieties of criterion validity:

> **Concurrent Validity** is when correlation of one measurement is observed against another measure of known or supposed validity at the same time.
> **Predictive Validity** is when correlation of one measurement is observed against another measure of known or supposed validity at a future time.

Construct Validity: A test of the underlying construct. One can usefully ask this of one or more tests. This is the hardest to understand, but an example of a construct is that in a test, higher scores will be progressively obtained by those with increasing levels of expertise. So a test of construct validity would be to give a medical performance test to 1st year students, 5th Year students, Foundation Year 2 doctors, registrars and consultants.

Convergent Construct Validity should be positive where tests are assumed to measure the same construct and Divergent Construct Validity should be negative where tests are assumed to measure different constructs.

References for Glossary

Case SM, Swanson DB. (1996) Constructing written test questions for the basic and clinical sciences. *National Board of Medical Examiners, Philadelphia*.

Norcini JJ. (2003) Setting standards on educational tests. *Medical Education,* 37: 464-469.

Stern DT, Friedman Ben-David M, Norcini J, Wojtczak A, Schwarz MR. (2006) Setting school-level outcome standards. *Medical Education*, 40: 166-172.

van der Vleuten C. (1996) The assessment of professional competence: developments, research and practical implications. *Advances in Health Science Education*, 1: 41–67.

[i] MSC Scanning Incident Error Report, papers produced for the UKFPO Rules Group meeting March 2013

[ii] Ibid., p 13

[iii] NHS high quality review.

[iv] GMC selection review

[v] A study based on 210 questions second-marked by the QA panel in the Northern Deanery Foundation School with regard to August 2008 applications showed agreement of less than 50% between the panel and the original scorers. This does not include instances where the QA panel could not agree.

[vi] http://www.isfp.org.uk/AboutISFP/Documents/ISFP_Option_Appraisal_Final_Report_Full_Appendices.pdf

[vii] http://www.isfp.org.uk/ISFPDocuments/Documents/Situational%20Judgement%20Tests%20Monograph%20FINAL%20Oct%202012.pdf

[viii] http://www.isfp.org.uk/AboutISFP/Documents/Appendix_D_-_FY1_Job_Analysis.pdf

[ix] http://www.isfp.org.uk/SJT/WhatIsTheSJT/Pages/WhatIsTheSJT.aspx

[x] Moore CG (2012) Can we predict which doctors will fail? Cardiff University MBA Dissertation

[xi] Patterson F, Archer V, Kerrin M, Carr V, Faulkes L, Stoker H, & Good D. (2010) "Improving Selection to the Foundation Programme" (Initial Pilot Report). Work Psychology Group and the University of Cambridge.

[xii] http://www.isfp.org.uk/AboutISFP/Documents/Final_report_of_PRE_Full_Appendices.pdf

[xiii] Patterson F, Baron H, Ashworth V, Knight A. Work Psychology Group. FY1 SJT Development: Scaling Issues, Combining EPM/SJT Scores and Cut Scores. April 2012 (page 13, point 3.2).

[xiv] Item 15, AOB: Individual SJT Scores.

[xv] Reason J. Human error: models and management. BMJ. 2000;320:768–70.