# The fairness, effectiveness and acceptability of selection for specialty training in the UK

***This article examines the fairness, effectiveness (validity and reliability) and acceptability of the 2009 selection processes for 10 hospital specialties, using data from a nationwide evaluation of 33 selection processes across the UK.***

Selection for specialty training in the UK is 'high stakes' as it decides whether junior doctors enter their specialty of choice and who will be the consultants and GP principals of the future. Selection is also competitive: in 2009, 11 417 applicants made 17 465 applications for 6580 entry-level training posts (Centre for Workforce Intelligence, personal communication, 2010). Selection processes therefore need to be fit for purpose. The General Medical Council (2011) states that 'processes for recruitment, selection and appointment must be open, fair and effective', and it is also essential that selection processes are acceptable to junior doctors and those responsible for their selection (Tooke, 2008).

With respect to fairness, existing UK evidence is limited to analyses comparing GP selection centre performance by gender, and whether initial medical training was in the UK or overseas (Brown et al, 2001, 2003; Fraser et al, 2009). Effectiveness incorporates both reliability and validity. Reliability can be assessed in several ways and most UK studies report coefficients for internal consistency and inter-rater reliability above 0.8 (Goodyear et al, 2007; House of Commons Health Committee, 2007; Rao, 2007; Gallagher et al, 2008; Johnston and French, 2008; Onyon et al, 2009; Patterson et al, 2009a, b; Gale et al, 2010).

There is no evidence on pass mark or cut score reliability, the percentage of applicants scoring within one standard error of measurement (SEM) of the cut score (Holsgrove, 2010), and about whom there is uncertainty as to whether or not they should be offered a post. With respect to predictive validity, the published UK evidence suggests that while shortlist scores are generally good predictors of interview or selection centre scores, evidence as to whether shortlisting or interview or selection centre scores predict performance in post is mixed (Reeve et al, 1993; Patterson et al, 2005, 2009a, b; Davison et al, 2006; Bindal et al, 2007; House of Commons Health Committee, 2007; Pashayan et al, 2007; Johnston and French, 2008; Gale et al, 2010). On acceptability, and despite the importance attached to re-building confidence in selection processes (Tooke, 2008), the studies of this criterion are largely confined to single specialty surveys (Randall et al, 2006a, b; Humphrey et al, 2008; Johnston and French, 2008; Gale et al, 2010).

This article reports key findings from the National Evaluation of Specialty Selection that assessed the first recruitment round of the 2009 selection processes for 10 hospital specialties.

## Methods

The National Evaluation of Specialty Selection was a concurrent mixed methods cross-sectional study focused on UK recruitment at specialty trainee or core trainee year 1 (ST/CT1) in core medical training, core surgical training, trauma and orthopaedics, paediatrics, psychiatry, obstetrics and gynaecology, and histopathology. Cardiothoracic surgery (ST3), emergency medicine (ST4) and a pilot selection process for general surgery (ST3) were also included. For the four smaller specialties (histopathology, cardiothoracic surgery, emergency medicine and general surgery), national data were obtained; for the five large specialties, data collection was undertaken in five deaneries: the West Midlands and London were selected because they are two of the largest deaneries, Oxford as the most competitive, and North Western and Yorkshire and The Humber as they hosted national selection for several smaller specialties. Data for trauma and orthopaedics were obtained from the four deaneries involved with a pilot national process.

The specialty and deanery combinations meant that a total of 33 selection processes were included, using selection methods that varied across specialties and, in core surgical training, across deaneries. However, all included longlisting on eligibility criteria, shortlisting based on information provided in the application form and a selection centre process, in which applicants rotated around 3–8 objective structured clinical examination-type stations of 5–30 minutes, with 1–3 assessors at each station. Selection processes were designed implicitly or explicitly using a blueprint in order to assess the competencies identified on the person specification for each specialty (available at www.mmc.nhs.uk). Some specialties made offers to the top N applicants in order to fill N vacancies (and would subsequently go further down their ranked list if an offer was declined). Other specialties set a minimum competency score for appointment and would therefore make offers to the top N applicants provided that they achieved this minimum score.

The data sources for the analyses reported here are questionnaires completed by applicants and assessors at selection centres between January and March 2009, docu-

**Professor Hywel Thomas** *is Professor of the Economics of Education and* **Dr Ian Davison** *is Lecturer in the School of Education, The University of Birmingham, Birmingham,* **Professor Harry Gee** *was Head of the Postgraduate School of Obstetrics and Gynaecology, West Midlands Workforce Deanery,* **Professor Janet Grant** *is Professor of Education in Medicine, CenMEDIC, London and* **Dr Celia Taylor** *is Senior Lecturer in Medical Education in the School of Clinical and Experimental Medicine, The University of Birmingham, Birmingham B15 2TT*

*Correspondence to: Dr C Taylor (c.a.taylor@bham.ac.uk)*

ments on selection processes and selection score data. All applicants and assessors attending a selection centre in one of the study sites were eligible to participate in the study. Applicants were asked to provide their General Medical Council numbers on their questionnaires to indicate consent to link their views and demographics to their selection scores.

To assess fairness, applicants' demographics were linked to their selection scores to assess whether any particular group of applicants had lower selection scores than others. To compare results across specialties, selection scores were standardized within each selection process to have a mean of 0 and a standard deviation of 1. Multiple linear regression for each specialty was then undertaken with standardized selection score as the dependent variable. The independent variables were sex, ethnicity (grouped as white, Asian or other), place of initial medical training (UK, other EU or non-EU), current post (foundation year 2 or other), age and years of postgraduate experience.

Effectiveness was assessed in terms of reliability and validity. Four measures of reliability were used. First, Cronbach's alpha evaluates the internal consistency of selection systems, i.e. the extent to which different stations assess the same underlying competency. Second, inter-rater reliability assesses the level of agreement between different assessors on the same station using consistency intra-class correlation coefficients with 'average measures' (McGraw and Wong, 1996; Howell, 2002). This article reports two measures of pass mark reliability, which is the reliability of the decision whether or not to offer an applicant a post. Cut score reliability identifies the proportion of applicants in each of four categories around the cut score (the lowest score for which an applicant was actually offered a post): clear pass (score ≥1 SEM above the cut score), borderline pass (<1 SEM above the cut score), borderline fail (<1 SEM below the cut score) and clear fail (≥1 SEM below the cut score) (Holsgrove, 2010). Where specialties also set a minimum competency score for appointment, the percentage of applicants offered a post with a score less than 1 SEM above this minimum score – minimum score reliability – was also calculated.

The predictive validity of shortlist scores is reported in relation to the scores obtained at the subsequent selection centre using Pearson's correlation coefficients corrected for unreliability of the criterion and restriction of range (Bobko, 1983).

For acceptability, this article reports specialty-level responses by applicants and assessors to the statement 'this selection process was fair'; and, for applicants only, 'this selection process enabled me to show how my skills and abilities make me suitable for specialty training', on a 6-point Likert scale from 'strongly disagree' (1) to 'strongly agree' (6). While the data are summarized using means, statistical significance was assessed non-parametrically using Kruskal–Wallis one-way analysis of variance (ANOVA) with correction for ties; post-hoc Mann–Whitney U-tests with the Bonferroni adjustment were used to compare each specialty against all other specialties combined.

Statistical analyses used SPSS v17 and STATA v11. In general, $P<0.05$ was considered statistically significant, although $P<0.01$ was used where there were some duplicate questionnaires from applicants who completed a questionnaire in more than one deanery. $P<0.01$ was also used in the analysis of fairness since five separate regression analyses were undertaken.

The National Research Ethics Service advised that NHS research ethics was not required. Ethical approval was therefore obtained through the University of Birmingham. The results for effectiveness have been anonymised as required by the Department of Health.

## Results

Questionnaire data were collected from 4041 applicants (estimated 90% response rate) and 519 assessors (estimated 65% response rate). Response rates are estimated from planned attendance figures, as final applicant and assessor numbers were not always provided (some applicants, for example, did not attend their interview). While specialties and deaneries were generally forthcoming with requests for score data, complete data were not obtained for every selection process. *Table 1* summarizes the data used for each evaluation criterion, while *Supplementary table 1* (available from www.bjhm.co.uk) summarizes the characteristics of the applicants and assessors completing questionnaires.

**Table 1. Data summary**

| Criterion | Data type | No. of specialities | No. of selection processes | Sample size |
|---|---|---|---|---|
| Fairness | Applicant questionnaires and selection centre scores linked by General Medical Council numbers in large enough specialties to enable robust analysis (mean score used for applicants with a selection score from more than one deanery) | 5 | 23 | 1553 |
| Effectiveness: internal consistency | Selection centre scores at station level | 10 | 26 | 3701 |
| Effectiveness: inter-rater reliability | Selection centre scores with assessor-level data | 4 | 4 | 388 |
| Effectiveness: cut score reliability | Selection centre scores at station level and pass/fail decision | 5 | 7 | 918 |
| Effectiveness: minimum score reliability | Selection centre scores at station level, pass/fail decision and minimum competency score | 4 | 5 | 817 |
| Effectiveness: predictive validity | Shortlist and selection centre scores | 8 | 13 | 2411 |
| Acceptability: applicant | Applicant questionnaires (includes 'duplicates', i.e. those attending more than one selection event) | 10 | 33 | 4041 |
| Acceptability: assessor | Assessor questionnaires | 10 | 30 | 519 |

## Fairness

In all, 46% of applicants responding to the questionnaire in the five larger specialties provided consent to link their demographics to their selection scores. The largest effect on selection scores was place of initial medical training (*Supplementary table 2*). The unweighted mean coefficient across specialties was 0.63 standard deviations higher for UK trained applicants compared with EU trained applicants and 0.78 standard deviations higher compared with those trained elsewhere.

White applicants generally achieved higher scores than Asian applicants (mean +0.31 standard deviations) and those from other ethnic backgrounds (mean +0.35 standard deviations). Overall there was a small negative effect for males (mean -0.11 standard deviations), but the coefficient was only statistically significant in one specialty. Increasing age had a small negative effect on scores (mean -0.04 standard deviations per year) which was statistically significant in four specialties, but the effects of postgraduate experience and current post (foundation year 2 or other) were not statistically significant in any of the five specialties. The percentage of variance in scores explained by these demographics varied from 8% to 45% across specialties.

## Effectiveness

### Internal consistency

Cronbach's alpha coefficients ranged from 0.35 to 0.83; 2/26 (8%) selection processes had coefficients above 0.8 as recommended for high stakes medical examinations (Holsgrove, 2010) (*Supplementary table 3*).

### Inter-rater reliability

Of the 17 station-level absolute intra-class correlations, 16 (94%) were 0.8 or above (*Supplementary table 3*).

### Pass mark reliability

Based on the actual cut score for appointment, the percentage of borderline applicants ranged from 12% to 55% (*Figure 1* and *Supplementary table 3*).

### Minimum score reliability

The percentage of appointed applicants scoring <1 SEM above the minimum competency score ranged from 0% to 20% (*Supplementary table 3*).

### Predictive validity of shortlisting scores

Following corrections for restriction in range and unreliability of the criterion, Pearson correlation coefficients between shortlist and selection centre scores ranged from 0.55 to 0.91, all above 0.5, which would be considered 'excellent' (Cooper et al, 2003) if job performance, rather than selection centre scores, was used as the criterion (*Supplementary table 3*).

## Acceptability

Specialty-level mean responses by applicants and assessors to the statement 'this selection process was fair' ranged from 4.2 to 5.1 and 4.8 to 5.4 respectively (where 4=slightly agree and 5=agree). Kruskal–Wallis one-way ANOVAs were $\chi^2(9)=53$, $P<0.001$ for applicants and $\chi^2(9)=29$, $P<0.001$ for assessors, suggesting statistically significant differences between specialties (*Supplementary table 4*). Specialty-level mean applicant responses to the question 'this selection process enabled me to show how my skills and abilities make me suitable for specialty training' ranged from 4.3 to 5.0. The Kruskal–Wallis one-way ANOVA gave $\chi^2(9)=65$, $P<0.001$, again suggesting statistically significant differences between specialties (*Supplementary table 4*). Overall, there was no correlation between applicants' ratings of fairness and their selection scores (data not shown).

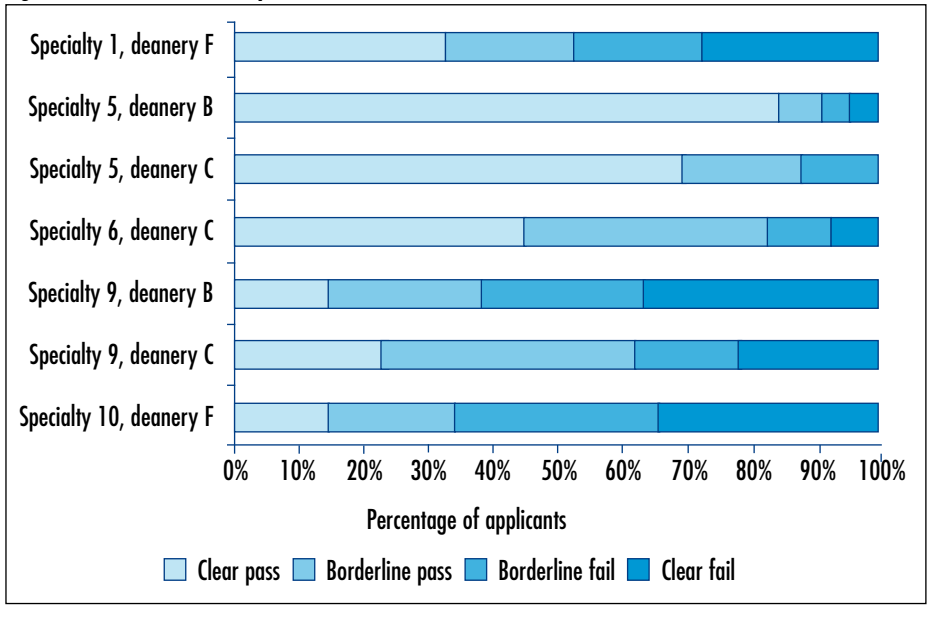
# Discussion

The National Evaluation of Specialty Selection is the largest evaluation of specialty selection undertaken in the world. The range of specialties and large sample size add substantially to an existing database of predominantly small-scale specialty-specific studies, none of which include all three of the criteria applied in this study: fairness, effectiveness and acceptability.

The results are a snapshot of the 2009 selection processes in 10 hospital specialties, focussing on evidence collected at the point of selection. While data were not obtained for all 33 selection processes, it is the authors' view that specialties and deaneries shared accessible data and did not withhold unfavourable results. For the five large specialties (core medical training, core surgical training, obstetrics and gynaecology, paediatrics and psychiatry) and trauma and orthopaedics the analyses of fairness and acceptability combined data across deaneries. With the exception of core surgical training, the national coordination of selection in these specialties justifies this approach. The selection process for core surgical training was determined at deanery level, so that aggregation may hide differences between deaneries.

On fairness, applicants who qualified overseas and, to a lesser extent, older, non-white, male applicants achieved lower selection centre scores. That the largest effect is for place of initial training concurs with the existing literature (Brown et al, 2001, 2003). The included demographics accounted for 45% of score variance in one specialty, which is worryingly high.

***Figure 1. Pass mark reliability.***



However, these results do not necessarily indicate bias. Skills such as communication and empathy have been included in the curriculum of UK medical schools to meet the needs of the NHS, so one possible explanation is that these skills have less prominence in training in other countries. It is also possible that these results are affected by response bias, since not all applicants consented for their demographics to be linked to their selection scores. Long-term follow up is required to determine if the selection processes are biased by examining whether the differences in selection scores reported here are mirrored by differences in performance in post.

The data on the effectiveness criterion show inter-rater reliability is generally good. Caution is required since these coefficients may be inflated through implicit or explicit collusion between raters within a station. The results on internal consistency contrast those of other studies (Gallagher et al, 2008; Johnston and French, 2008; Patterson et al, 2009a; Gale et al, 2010) with 24 of the 26 selection centre processes having internal consistency lower than the recommended 0.8 minimum: worrying results for a process regarded by many candidates as 'the most high stakes assessment of their careers' (Postgraduate Medical Education and Training Board, 2010).

The selection processes with internal consistency above 0.8 included six or seven stations, although there was not a clear relationship between the number of stations and internal consistency. The low internal consistency resulted in low cut score reliability with about half the applicants being borderline in four out of seven selection processes, suggesting assessors have difficulty distinguishing between these applicants. A smaller proportion of applicants were appointed whose true scores might be below the minimum competency score, although only one selection process used a formal standard setting process to identify the minimum score required for appointment.

The correlation coefficients for the predictive validity of shortlist to interview scores are at least on a par with those reported in the literature (Davison et al, 2006; House of Commons Health Committee, 2007; Pashayan et al, 2007; Johnston and French, 2008; Patterson et al, 2009a, b). These positive correlations suggest that few applicants who are rejected at shortlisting would have scored high enough in the selection centre to be offered a post. The ultimate test of selection processes is prediction of performance in post. This requires a cohort study which was outside the time horizon of this study and is its principal weakness.

Applicants gave positive responses to statements regarding fairness and opportunity to demonstrate skills and abilities. Assessors also responded positively to the statement on fairness, findings that concur with the existing evidence (Randall et al, 2006a, b; Humphrey et al, 2008; Johnston and French, 2008; Gale et al, 2010). This provides evidence that selection processes are now broadly acceptable to applicants and assessors, countering concerns cited by what was the Postgraduate Medical Education and Training Board (2010). However, this study identified differences in applicants' views across specialties, suggesting that specialties could benefit from sharing best practice.

Selection processes in most specialties underwent radical change after the Medical Training Application Service was abandoned in 2008. Changes since then have been evolutionary not revolutionary, such that the findings reported here remain pertinent and provide a benchmark against which further changes can be evaluated. A major change has been an increase in the level of national coordination of selection and this has two major benefits. First, specialties are now able to invite almost all longlisted applicants to a selection centre rather than having to rely on shortlisting. Second, a significant reduction in duplication and opportunities to exploit economies of scale should result in cost savings without reducing fairness or effectiveness.

The key area for improvement identified from these results is the need to increase internal consistency and hence pass mark reliability. However, care is needed in how to achieve this as reliability is not an end in itself. A series of almost identical stations might be highly reliable, but would not assess the range of competencies required for specialty training, i.e. validity would be compromised. An almost cost-neutral approach to improving reliability would be to have more stations but with only one assessor in each, as is often the norm for objective structured clinical examination assessments (Newble, 2004). Increasing the number of stations is far more important for improving reliability than increasing the number of assessors per station (Newble, 2004). Other ways of improving reliability include providing additional assessor training (e.g. calibration exercises), reviewing the scoring systems used and implementation of formal standard setting processes in order to ensure that an appropriate minimum competency score is applied. BJHM

## KEY POINTS

- The evidence described in this article indicates a commitment in the profession to continuing improvement. The evidence from the National Evaluation of Specialty Selection study points to five ways in which selection can be further improved.
- The wealth of data from each year's selection round could provide information on the effectiveness of specific selection processes, which would have most benefit if shared between specialties; relying on published papers is not enough as the bias toward reporting effective outcomes can hide less effective practices.
- Methods of assessor training need piloting; the variability of current practice provides a rich source for studies designed to identify the most effective methods.
- Specialties should consider using more stations but with only one assessor in each in order to improve reliability.
- The General Medical Council should consider whether the assessments involved in selection require specialties to set and monitor relevant standards.
- A cross-specialty longitudinal cohort study following successful and unsuccessful applicants would enable assessment of long-term predictive validity.

*This is independent research commissioned and funded by the Policy Research Programme in the Department of Health (Award number 016 0114). The views expressed are not necessarily those of the Department. The Policy Research Programme team suggested a number of specialties to be included in this project but participation of all*

*specialties was voluntary. The Department of Health required that the results for validity and reliability be anonymised. The NESS project team also included: Stephen Field, Andrew Malins, Laura Pendleton and Elizabeth Spencer. The authors would like to thank the specialties, deaneries, applicants and assessors who participated in this study. Volker Patent and Ros Searle from The Open University contributed to questionnaire development and data collection. A number of colleagues from The University of Birmingham also assisted us with data collection: Julie Bedward, Sarah Burke, Sandra Cooke, Vickie Firmstone, Julie Foster, Amirta Johal, Natasha MacNab, Liz Potts and Penny Smith. The project administrators Magdalena Skrybant and Amy Snooks coordinated data collection and brought the report together. Finally, the advisory group have provided valuable comments on the project and their support has helped us work collaboratively with the medical profession.*
*Conflict of interest: none.*

Bindal T, Wall D, Goodyear HM (2007) Performance of paediatric Senior House Officers following changes in recruitment. *Med Teach* **29**: 498–500

Bobko P (1983) An analysis of correlations corrected for attenuation and range restriction. *J Appl Psychol* **68**: 584–9

Brown CA, Wakefield SE, Bullock AD (2001) The selection of GP trainees in the West Midlands: Audit of assessment centre scores by ethnicity and country of qualification. *Med Teach* **23**: 605–9

Brown CA, Wakefield SE, Bullock AD (2003) The selection of GP trainees in the West Midlands: second audit of assessment centre scores by ethnicity and country of qualification. *Med Teach* **25**: 649–53

Cooper D, Robertson IT, Tinline G (2003) *Recruitment and Selection: A Framework for Success.* Thomson Learning, London

Davison I, Burke S, Bedward J, Kelly S (2006) Do selection scores for general practice registrars correlate with end of training assessments? *Educ Prim Care* **17**: 473–8

Fraser A, Calvert M, Wilkinson M, Freemantle N (2009) Standardised patient assessments on consecutive days during high-stakes GP training interviews: is there any evidence of candidates sharing information? *Educ Prim Care* **20**: 285–90

Gale T, Roberts M, Sice P et al (2010) Predictive validity of a selection centre testing non-technical skills for recruitment to training in anaesthesia. *Br J Anaesth* **3**: 3

Gallagher AG, Neary P, Gillen P, Lane B, Whelan A, Tanner WA, Traynor O (2008) Novel method for assessment and selection of trainees for higher surgical training in general surgery. *ANZ J Surg* **78**: 282–90

General Medical Council (2011) *The Trainee Doctor.* General Medical Council, London

Goodyear HM, Jyothish D, Diwakar V, Wall D (2007) Reliability of a regional junior doctor recruitment process. *Med Teach* **29**: 501–3

Holsgrove G (2010) Reliability issues in the assessment of small cohorts. GMC supplementary guidance. General Medical Council, London

House Of Commons Health Committee (2007) *Modernising Medical Careers: Volume II Written Evidence*. The Stationery Office Limited, London

Howell D (2002) *Statistical Methods for Psychology.* Duxbury, Pacific Grove

Humphrey S, Dowson S, Wall D, Diwakar V, Goodyear HM (2008) Multiple mini-interviews: opinions of candidates and interviewers. *Med Educ* **42**: 207–13

Johnston PW, French FH (2008) Analysis of selection tools for appointment of specialty trainees in histopathology in Scotland. *Br J Hosp Med* **69**: 101–5

McGraw KO, Wong SP (1996) Forming Inferences About Some Intraclass Correlation Coefficients. *Psychol Methods* **1**: 30–46

Newble D (2004) Techniques for measuring clinical competence: objective structured clinical examinations. *Med Educ* **38**: 199–203

Onyon C, Wall D, Goodyear HM (2009) Reliability of multi-station interviews in selection of junior doctors for specialty training. *Med Teach* **31**: 665–7

Pashayan N, Duff C, Mason BW (2007) Selection into specialty training in public health: performance of the Medical Training Application Service shortlisting. *J Public Health* **29**: 331–7

Patterson F, Ferguson E, Norfolk T, Lane P (2005) A new selection system to recruit general practice registrars: preliminary findings from a validation study. *BMJ* **330**: 711–14

Patterson F, Baron H, Carr V, Plint S, Lane P (2009a) Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. *Med Educ* **43**: 50–7

Patterson F, Carr V, Zibarras L et al (2009b) New machine-marked tests for selection into core medical training: evidence from two validation studies. *Clin Med* **9**: 417–20

Postgraduate Medical Education and Training Board (2010) *Final report of the selection into Specialty Training. Postgraduate Medical Education and Training Board*, London

Randall R, Davies H, Patterson F, Farrell K (2006a) Selecting doctors for postgraduate training in paediatrics using a competency based assessment centre. *Arch Dis Child* **91**: 444

Randall R, Stewart P, Farrell K, Patterson F (2006b) Using an assessment centre to select doctors for postgraduate training in obstetrics and gynaecology. *The Obstetrician and Gynaecologist* **8**: 257–62

Rao R (2007) The structured clinically relevant interview for psychiatrists in training (SCRIPT): A new standardized assessment tool for recruitment in the UK. *Acad Psychiatry* **31**: 443–6

Reeve PE, Vickers MD, Horton JN (1993) Selecting anaesthetists: the use of psychological tests and structured interviews. *J R Soc Med* **86**: 400–3

Tooke J (2008) *Aspiring to Excellence: Findings and national recommendations of the independent inquiry into Modernising Medical Careers.* Universities UK, London

## Supplementary table 1. Participant characteristics (based on questionnaire responses)

| Characteristic | | Applicants (*n*=4041) | Assessors (*n*=519) |
|---|---|---|---|
| Gender *n* (%) | Male | 1842 (45.6) | 350 (67.4) |
| | Female | 2167 (53.6) | 151 (29.1) |
| | Not stated | 32 (0.8) | 18 (3.5) |
| Ethnicity *n* (%) | White | 1715 (42.4) | N/A |
| | Asian | 1420 (35.1) | |
| | Other | 662 (16.4) | |
| | Not stated | 244 (6.0) | |
| Initial medical training *n* (%) | UK | 2751 (68.1) | N/A |
| | Other EU | 243 (6.0) | |
| | Non-EU | 1013 (25.1) | |
| | Not stated | 34 (0.8) | |
| Years of postgraduate experience median (IQR) (*n*=3112) | | 2.0 (2.0–2.5) | N/A |
| Number of selection days as an assessor median (IQR) (*n*=455) | | N/A | 2 (0–5) |
| Age median (IQR) (*n*=3112) | | 37.0 (25.0–30.0) | N/A |
| Current post *n* (%) | Foundation year 2 | 2928 (72.5) | N/A |
| | Other | 984 (24.4) | |
| | Missing | 129 (3.2) | |

IQR = interquartile range

## Supplementary table 2. Fairness results (multiple regression analyses by specialty)

| | Coefficient (*P* value) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Specialty | Male (comparator = female) | Asian (comparator = white) | Other ethnic group (comparator = white) | EU trained (comparator = UK trained | Non-EU trained (comparator = UK trained | FY2 post (comparator =other post) | Age (years) | Postgraduate experience (years) | Constant | Sample size/no. of selection scores | Adj. $R^2$ |
| Core medical training | -0.09 (0.11) | **-0.43 (<0.01)** | **-0.46 (<0.01)** | **-0.98 (<0.01)** | **-0.85 (<0.01)** | 0.17 (0.04) | **-0.03 (<0.01)** | 0.02 (0.45) | **1.35 (<0.01)** | **808/1620** | **0.40** |
| Core surgical training | **-0.38 (<0.01)** | -0.19 (0.21) | -0.31 (0.05) | -0.24 (0.31) | -0.72 (0.05) | -0.20 (0.36) | 0.02 (0.65) | -0.38 (0.02) | 1.01 (0.25) | 246/709 | 0.08 |
| Psychiatry | -0.21 (0.03) | -0.07 (0.59) | -0.21 (0.23) | **-1.09 (<0.01)** | **-1.10 (<0.01)** | 0.11 (0.38) | **-0.02 (<0.01)** | 0.03 (0.24) | **1.32 (<0.01)** | **224/484** | **0.45** |
| Obstetrics and gynaecology | 0.04 (0.76) | **-0.42 (0.01)** | -0.24 (0.13) | **-0.75 (<0.01)** | **-0.72 (<0.01)** | 0.21 (0.25) | **-0.06 (<0.01)** | 0.01 (0.89) | **2.00 (<0.01)** | **181/421** | **0.30** |
| Paediatrics | 0.07 (0.77) | -0.42 (0.11) | -0.36 (0.13) | -0.11 (0.78) | -0.53 (0.13) | 0.20 (0.49) | **-0.09 (<0.01)** | 0.00 (0.99) | **2.83 (<0.01)** | **94/154** | **0.24** |
| Unweighted mean coefficient | -0.11 | -0.31 | -0.35 | -0.63 | -0.78 | 0.10 | -0.04 | -0.06 | | | |

Data from the five key National Evaluation of Specialty Selection deaneries (London, North Western, Oxford, Yorkshire & The Humber and West Midlands) are included in all regressions except paediatrics, which only includes data from North Western, Oxford and West Midlands deaneries and psychiatry, which does not include London. Statistically significant results are shown in bold. FY2 = foundation year 2.

**Supplementary table 3. Effectiveness results by selection process**

| Selection process: specialty and deanery (no. of applicants) | No. of stations or assessments | Internal consistency (Cronbach's alpha) | Inter-rater reliability (consistency) | Pass mark reliability - fairness (% 'borderline') | Minimum score reliability - competency (%<1 SEM above min score) | Predictive validity (uncorrected/corrected correlation coefficient; 95% CI) |
|---|---|---|---|---|---|---|
| Specialty 1, deanery F (*n*=144) | 7 | 0.834 | | 30% | 20% | 0.61/0.78 (0.71–0.84) |
| Specialty 2, deanery A (*n*=125) | 3 | 0.679 | | | | 0.44/0.58 (0.43–0.72) |
| Specialty 2, deanery C (*n*=43) | 3 | 0.459 | | | | |
| Specialty 3, deanery F (*n*=69) | 6 | 0.826 | | | | 0.50/0.55 (0.39–0.71) |
| Specialty 4, deanery F (*n*=47) | 5 | 0.664 | Portfolio: 0.904<br>Audit presentation: 0.729<br>Academic/research: 0.894<br>Clinical interview: 0.889<br>Leadership/team work: 0.863<br>Telephone: 0.889<br>Technical skills: 0.825 | | | |
| Specialty 5, deanery A (*n*=77) | 3 | 0.620 | | | | |
| Specialty 5, deanery B (*n*=190) | 3 | 0.632 | | 12% | 7% | 0.39/0.66 (0.53–0.80) |
| Specialty 5, deanery C (*n*=33) | 3 | 0.519 | | 30% | 3% | |
| Specialty 5, deanery D (*n*=62) | 3 | 0.729 | | | | |
| Specialty 5, deanery E (*n*=51) | 3 | 0.691 | | | | 0.65/0.91 (0.86–0.97) |
| Specialty 6, deanery A (*n*=60) | 3 | 0.776 | | | | 0.48/0.67 (0.49–0.84) |
| Specialty 6, deanery C (*n*=40) | 3 | 0.349 | | 48% | 0% | |
| Specialty 6, deanery D (*n*=62) | 3 | 0.778 | Portfolio (pilot station): 0.954 | | | |
| Specialty 6, deanery E (*n*=55) | 3 | 0.560 | | | | |
| Specialty 7, deanery A (*n*=228) | 3 | 0.697 | Portfolio review: 0.883<br>Suitability for specialty: 0.882<br>Clinical scenario: 0.911<br>Communication: 0.859<br>Ethical scenario: 0.918<br>Professionalism: 0.886 | | | 0.48/NA |
| Specialty 7, deanery B (*n*=784) | 3 | 0.700 | | | | 0.66/0.81; 0.79–0.84 |
| Specialty 7, deanery C (*n*=151) | 3 | 0.707 | | | | |
| Specialty 7, deanery D (*n*=259) | 3 | 0.734 | | | | |
| Specialty 7, deanery E (*n*=392: internal consistency; *n*=174: predictive validity) | 3 | 0.729 | | | | 0.56/NA |
| Specialty 8, deanery D (*n*=51) | 7 | 0.763 | Portfolio: 0.951<br>Scenario and probity: 0.864<br>Commitment to specialty: 0.901 | | | |
| Specialty 8, deanery E (*n*=27) | 7 | 0.602 | | | | |
| Specialty 9, deanery B (*n*=410) | 3 | 0.529 | | 49% | 8% | 0.45/0.78 (0.71–0.85) |
| Specialty 9, deanery E (*n*=40) | 3 | 0.575 | | | | 0.37/0.63 (0.31–0.94) |
| Specialty 9, deanery C (*n*=55) | 6 | 0.608 | | 55% | X | |
| Specialty 9, deanery A (*n*=90) | 3 | | | | | 0.17/NA |
| Specialty 9, deanery D (*n*=200) | 4 | 0.471 | | | | |
| Specialty 10, deanery F (*n*=46) | 6 | 0.473 | | 51% | X | 0.44/0.81 (0.60–1.00) |

X = no minimum score for appointability set; NA = Not available – data required for correcting the coefficient not available. The scores of all applicants attending the selection centres for which data are available (*Table 1*) are included in the analyses. CI = confidence interval; SEM = standard error of measurement.

## Supplementary table 4. Acceptability results by specialty

| | 'This selection process was fair' | | | 'This selection process was fair' | | | 'This selection process enabled me to show how my skills and abilities make me suitable for specialty training' | | |
|---|---|---|---|---|---|---|---|---|---|
| Specialty | Applicant mean score (range) | Mann–Whitney z (*P*) | No. of responses/no. of questionnaires | Assessor mean score (range) | Mann–Whitney z (*P*) | No. of responses/no. of questionnaires | Applicant mean score (range) | Mann–Whitney z (*P*) | No. of responses/no. of questionnaires |
| All specialties combined | 4.68 (1–6) | N/A | 3996/4041 | 4.97 (1–6) | N/A | 386/519 | 4.54 (1–6) | N/A | 4006/4041 |
| Core medical training | 4.65 (1–6) | -2.27 (0.023) | 1688/1696 | 4.84 (1–6) | -1.50 (0.134) | 124/148 | 4.56 (1–6) | 0.08 (0.940) | 1689/1696 |
| Core surgical training | 4.61 (1–6) | -2.39 (0.017) | 656/671 | 4.93 (3–6) | -0.47 (0.634) | 41/54 | **4.41 (1–6)** | **-3.41 (<0.001)** | **659/671** |
| Psychiatry | 4.75 (2–6) | 2.02 (0.043) | 478/482 | 4.75 (2–6) | -2.35 (0.019) | 56/66 | 4.57 (1–6) | 0.85 (0.395) | 480/482 |
| Trauma and orthopaedics | 4.84 (2–6) | 2.70 (0.007) | 130/135 | 5.27 (3–6) | 2.32 (0.021) | 26/58 | 4.84 (1–6) | 1.84 (0.066) | 132/135 |
| Histopathology | 4.53 (1–6) | -1.52 (0.129) | 121/124 | 5.30 (4–6) | 1.25 (0.210) | 10/11 | 4.29 (1–6) | -2.24 (0.025) | 122/124 |
| Obstetrics and gynaecology | 4.68 (1–6) | 0.45 (0.655) | 338/347 | 4.91 (3–6) | -0.90 (0.369) | 33/45 | 4.54 (1–6) | 2.54 (0.011) | 339/347 |
| Paediatrics | 4.82 (2–6) | 3.08 (0.002) | 433/433 | 4.93 (3–6) | -0.91 (0.364) | 41/61 | 4.67 (1–6) | 3.08 (0.002) | 433/433 |
| Cardiothoracic surgery | 4.21 (2–6) | -2.70 (0.007) | 43/43 | 5.36 (4–6) | 2.32 (0.021) | 22/22 | 4.26 (1–6) | -1.84 (0.066) | 43/43 |
| Emergency medicine | **5.06 (4–6)** | **3.39 (<0.001)** | **62/63** | 5.44 (5–6) | 2.30 (0.022) | 17/21 | **5.03 (3–6)** | **4.29 (<0.001)** | **62/63** |
| General surgery pilot | 5.04 (3–6) | 2.96 (0.003) | 47/47 | 5.41 (5–6) | 2.39 (0.017) | 16/33 | 4.70 (3–6) | 0.89 (0.376) | 47/47 |

Scores range from (1) strongly disagree to (6) strongly agree; Mann–Whitney U tests compare each specialty with all other specialties combined. *P* values before correcting for multiple comparisons using the Bonferroni adjustment are shown. Results statistically significant at $P<0.01$ (applicants) or $P<0.05$ (assessors) after correcting for multiple comparisons are shown in bold.