Racial and Gender Influences on Pass Rates for the UK and Ireland Specialty Board Examinations

Article *in* Journal of Surgical Education · August 2015

DOI: 10.1016/j.jsurg.2015.08.003 · Source: PubMed

6 authors, including:

David Richens
Nottingham University Hospitals NHS Trust
66 PUBLICATIONS **941** CITATIONS
SEE PROFILE

Hettie Till
University of Dundee
20 PUBLICATIONS **244** CITATIONS
SEE PROFILE

ORIGINAL REPORTS

Racial and Gender Influences on Pass Rates for the UK and Ireland Specialty Board Examinations

David Richens, MB, BS, FRCS,* Timothy R. Graham, MB, BS, FRCS,* John James,† Hettie Till, PhD,‡ Philip G. Turner, MB, BS, FRCS,* and Cara Featherstone, PhD*

*Joint Committee on Intercollegiate Examinations, Edinburgh, UK; †The Sickle Cell Society, London, UK; and ‡Assessment Support and Psychometric Services, Dundee, UK

INTRODUCTION: We explored effects of gender, ethnic origin, first language, and training status on scores in the Intercollegiate Specialty Board examinations in the UK and Ireland across the computer-marked written section and in the face-to-face oral and clinical section.

METHODS: Demographic characteristics and examination results from 9987 attempts across 177 sittings from 2009 to 2013 were analyzed in an analysis of variance by training status, gender, ethnic origin, first language, and section (computer-marked multiple-choice examination vs face-to-face oral and clinical examination).

RESULTS: We found increasing alignment between examiner and candidate characteristics during this period, with a 50% increase in examiners of Asian ethnic origin and a 60% increase in examiners whose first language is not English. The strongest factor in the analysis of variance was training status ($F[2, 9818] = 27.67$, $p < 0.001$), with candidates in training significantly outperforming others. Within "core candidates" (first attempt, in training), we found significant main effects for ethnic origin ($F[5, 4809] = 2.36$, $p = 0.04$), and first language ($F[2, 4809] = 5.29$, $p = 0.003$), but no interaction effects between these factors and section (both $F < 1$, $p > 0.05$).

CONCLUSIONS: Training status was the most important factor in candidates' results. Although the analysis showed significant effects of ethnic origin and first language within "core candidates," these differences were statistically indistinguishable between the 2 sections of the examination, suggesting that the differential attainment by these factors cannot be attributed to examiner bias in a face-to-face examination. (J Surg Ed ▮:▮▮▮-▮▮▮. © 2015 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: differential attainment, equality and diversity, surgical examinations

COMPETENCIES: Medical Knowledge, Patient Care, System-Based Practice, Interpersonal Skills and Communication

INTRODUCTION

This study investigates differential attainment in candidates sitting the Intercollegiate Specialty Board exams in the UK and Ireland. These examinations are high stakes competency based assessments taken at the completion of surgical training. They are conducted by the Joint Committee on Intercollegiate Examinations (JCIE) and regulated by the General Medical Council (GMC). Success in the appropriate Intercollegiate Specialty Board examination is a mandatory step in becoming eligible to apply for a consultant appointment in the relevant specialty. The Examination Regulations were relaxed in November 2006 to allow candidates without national training numbers and also those not in recognized training posts in the UK or Ireland to enter the examination.

Standard 17 of the 2010 GMC standards for curricula and assessments[1] relates to equality and diversity and states that Colleges must have equal opportunities and antidiscrimination policies in place in relation to trainees and trainers, together with an indication of how these will be implemented and monitored. Under the Equality Act 2010[2] the Colleges must, in carrying out their functions, have due regard to the need to eliminate unlawful discrimination, harassment and victimization, advance equality of opportunity and foster good relations between people who share a protected characteristic and those who do not. In response

Correspondence: Inquiries to David Richens, MB, BS, FRCS, Joint Committee on Intercollegiate Examinations, 2 Hill Place, Edinburgh, EH8 9DS, UK; E-mail: m.gray@jcie.org.uk, drichens@mac.com

to the Act and to the GMC standard, the Joint Surgical Royal Colleges published an Equality and Diversity Policy in July 2013.[3]

Concerns have existed for some time around the known differential attainment by protected characteristics in medical postgraduate examinations.[4] In 2006, a British Medical Association survey of Royal College examinations showed a disparity between examiner, workforce, and candidate profiles.[5] In 2013, Esmail and Roberts authored a report, commissioned by the GMC, concerning the Royal College of General Practitioner examinations.[6] This demonstrated differential attainment in the examination according to ethnicity and the authors speculated that this may be owing to a lack of familiarity with UK general practice and examination structure/process. Additionally the different ethnic cohorts scrutinized were not thought to be academically equivalent. Contemporaneously, an article in the *British Medical Journal* based on the same data concluded that "subjective bias owing to racial discrimination in the clinical skills assessment may be a cause of failure for UK trained candidates and international medical graduates."[7]

Following this publication the British Association of Physicians of Indian Origin obtained a judicial review of the methodology of the Royal College of General Practitioners examination. Justice John Mitting rejected the claim that aspects of the clinical skills assessments should be declared unlawful and ruled that the College was neither racially discriminatory nor in breach of its public sector equality duty. However, he did rule that there was a disparity in results between different groups and the Royal College of General Practitioners must take action to address this.[8]

The recent widely reported concerns about bias in Royal College examinations highlighted the need to review existing assessment practices in all medical specialty examinations.[9,10]

In this study, we asked:

1. Whether differences in pass rates could be demonstrated between candidates by gender, ethnic origin, first language, or training status.
2. Whether candidates performed similarly when marked by a computer and by examiners.

Gender and ethnic origin were selected for study as recognizable characteristics that could therefore be subject to examiner bias in clinical and oral examinations. First language and training status were selected for study based on earlier reports exploring differences in performance by these characteristics.[11-14]

METHODS

During the time frame of this study (2009-2013) the 9 surgical specialties were cardiothoracic surgery, general surgery, neurosurgery, oral and maxillofacial surgery, otolaryngology, pediatric surgery, plastic surgery, trauma and orthopedic surgery, and urology.

All the surgical specialty examinations consist of 2 sections. Section 1 is a multiple choice examination sat online and consisting of 2 papers marked by computer—a single best answer paper of 110 questions and an extended matching item paper of 135 questions. Scores on the 2 papers are combined to give an overall score. Candidates have to pass Section 1 before being allowed to proceed to the Section-2 examination, which consists of a combination of patient-based clinical examinations and scenario-based structured orals stations double marked by trained examiners. The examination standard is that of the level of a day-1 consultant working in the UK and Ireland.

Candidates

The analysis included 9987 candidate attempts (4952 in Section 1 and 5035 in Section 2) across 177 sittings (84 sittings of Section 1 and 93 sittings of Section 2). Candidates who sat the examination multiple times appeared multiple times in the data set. The 93 sittings of Section 2 included 3567 examiners (660 unique examiners).

Data Collection

As part of the application process, candidates voluntarily declare their gender, ethnic origin, first language, training status, primary medical qualification, and possession of the Membership of Royal College Surgeons qualification as part of the application process. For each category, candidates had the option "prefer not to say". The candidates' declared ethnic origins were grouped into 5 categories according to the 2011 UK government ethnicity categories: Asian, black, mixed, white, and other. For the purpose of this study First language was categorized as either English or Other. Gender was categorized as either men or women. Training status had 3 levels: candidates with National Training numbers in recognized Training Posts (termed "in training towards a Certificate of Completion of Training [CCT]"); candidates in recognized training posts but without a national training number (termed "in training not towards a CCT"); and candidates without a training number who are not in recognized training posts (termed "not in training"). Core candidates were defined as those in training towards CCT who were on their first attempt.

Additional information stored for each candidate entry includes their score in the examination, their attempt number, and the pass mark.

Data Analysis

To conduct this analysis, it was necessary to make the data comparable across different sittings and across different

specialties, where there may be a different number of maximum points in the examination, and a different pass mark. Candidates' scores were transformed by first subtracting the pass mark from the candidates' total scores and then *z*-transforming these difference scores after splitting the data by sitting. This process, which is commonly used to standardize scores, captures where each score sits within the distribution of scores. Hereafter in the article, candidates' "*z*-transformed difference scores" are referred to as "scores". We explored candidates' scores in a 2 (section) × 3 (gender) × 3 (first language) × 3 (training) × 6 (ethnic origin) analysis of variance.

RESULTS

Candidate and Examiner Demographics

Across all characteristics, the number of candidates and examiners answering "prefer not to say" has decreased over the years, demonstrating a greater willingness to provide data.

Of the 9987 candidates attempts, 5727 (57%) were classed as "in training working towards a CCT", 4022 (40%) were classed as "not in training", and 238 (2%) were classed as "in training not working towards a CCT". Looking only at first attempt candidates (7083 in total), the numbers were, respectively, 4878 (69%), 2075 (29%), and 130 (2%). The higher proportion of candidates classed as "not in training" in the full data set is an indicator of the greater likelihood of these candidates being unsuccessful on their first attempt at the examination.

The proportion of candidates (and examiners) describing themselves within each broad ethnic origin category across this 5-year period is as follows:

1. Asian: 36% (14%)
2. Black: 3% (0.4%)
3. Mixed: 2% (0.6%)
4. White: 27% (77%)
5. Other: 11% (2%)
6. Prefer not to say: 22% (6%).

This breakdown by ethnic origin shows a greater proportion of Asian candidates than candidates from any other ethnic group, and a greater proportion of white examiners than examiners from any other ethnic group.

Figure 1 shows that there has been a steady 50% increase in the proportion of examiners describing their ethnic origin as Asian from 12% in 2009 to 18% in 2013, whereas the proportion of candidates describing themselves as belonging to different ethnic groups has remained relatively stable over this period.

With regards to first language, the proportions for candidates (and examiners) are as follows:

1. English: 45% (85%)
2. Other: 35% (11%)
3. Prefer not to say: 20% (4%).

Although most examiners have English as their first language, the number of candidates for whom this is the case falls short of the majority. The evolution of this characteristic over time, captured in Figure 2, shows a

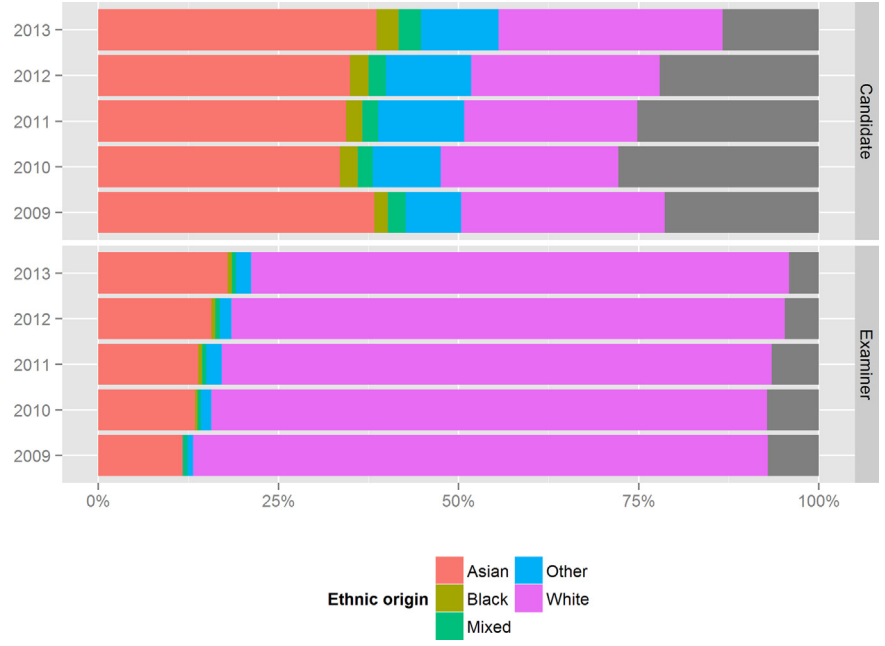

FIGURE 1. Evolution of ethnic origin data over time in candidates and examiners.

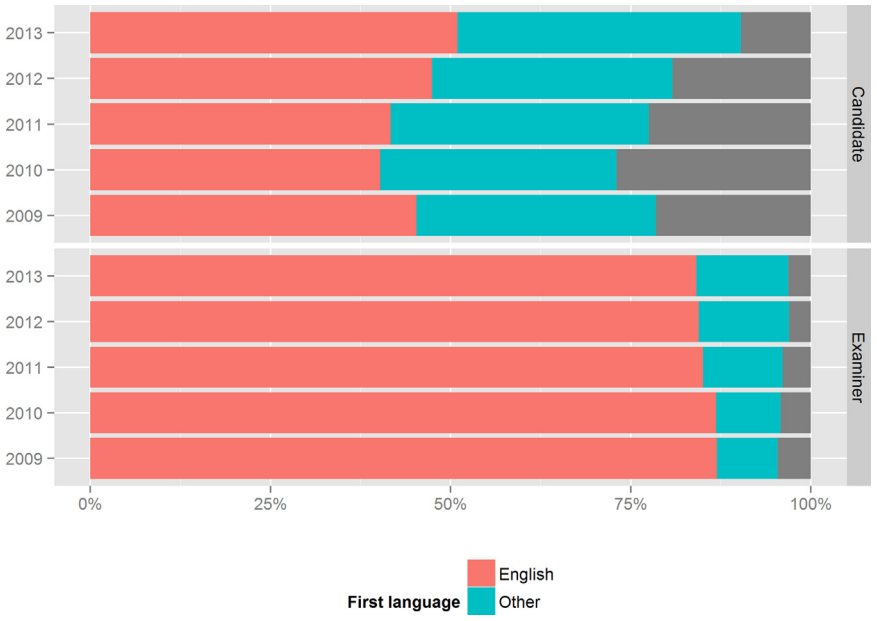

FIGURE 2. Evolution of first language over time in candidates and examiners.

63% increase from 8% in 2009 to 13% in 2013 examiners who do not have English as their first language.

With regards to gender, we find a greater proportion of men in the examiners than in the candidates:

1. Men: 87% (94%)
2. Women: 12% (4%)
3. Prefer not to say: 2% (2%).

The evolution over time shows little change in the proportion of male and female candidates and examiner over this period.

Candidate' Scores by Demographic Characteristic, Examination Component and Training Status

Candidates' scores were analyzed in a 2 (section) × 3 (gender) × 3 (first language) × 3 (training) × 6 (ethnic origin) analysis of variance. This analysis revealed that the variable which accounted for the largest part of the variance in candidates' scores was Training (F[2, 9818] = 27.66, $p < 0.001$, where F denotes the ratio of variances). Post hoc pairwise comparisons were conducted using a Bonferroni adjustment (the correction made to p values when several dependent or independent tests are being performed simultaneously on a single data set). This showed that candidates in training towards a CCT achieved significantly higher scores than those in training not towards a CCT ($p < 0.001$), who in turn achieved a significantly higher score than candidates not in training ($p = 0.003$). No significant interaction was found between training and section ($F < 1$, $p = 0.95$), meaning that the effect of Training was similar across Section 1 and Section 2, as illustrated in Figure 3.

Because Training accounted for so much of the variance in this model and because of the difficulty of interpreting effects of demographic characteristics compounded by number of attempts and training status, subsequent analyses were carried out using only the "core candidates" (in training towards a CCT and on their first attempt).

Within the core group of candidates ($N = 4878$), no significant main effect for gender was found ($F < 1$, $p > 0.05$). Significant main effects were found for first language (F[2, 4809] = 5.29, $p = 0.005$) and Ethnic Origin (F[5, 4809] = 2.34, $p = 0.04$). Post hoc Bonferroni analyses included a category for "prefer not to say", but only comparisons between declared characteristics are reported for ease of interpretation.

Post hoc analyses on the First Language data showed that candidates whose first language is English achieved significantly higher scores than candidates with another first language ($p < 0.001$). Post hoc analyses on the ethnic origin data showed that candidates who described their ethnic origin as "white" achieved significantly higher scores than candidates in all other ethnic groups did (all comparisons $p < 0.01$). No other pairwise comparisons were statistically significant.

We found no significant interaction effects of section × ethnic origin ($F < 1$, $p > 0.05$) or section × first language ($F < 1$, $p > 0.05$). This is reflected in the trends shown in Figures 4 and 5; similar differences between groups are seen in Section 1 and Section 2.

Finally, the analysis revealed a significant section × gender interaction effect in "core candidate" (F[2, 4809] = 5.70, $p = 0.003$), meaning that the difference between

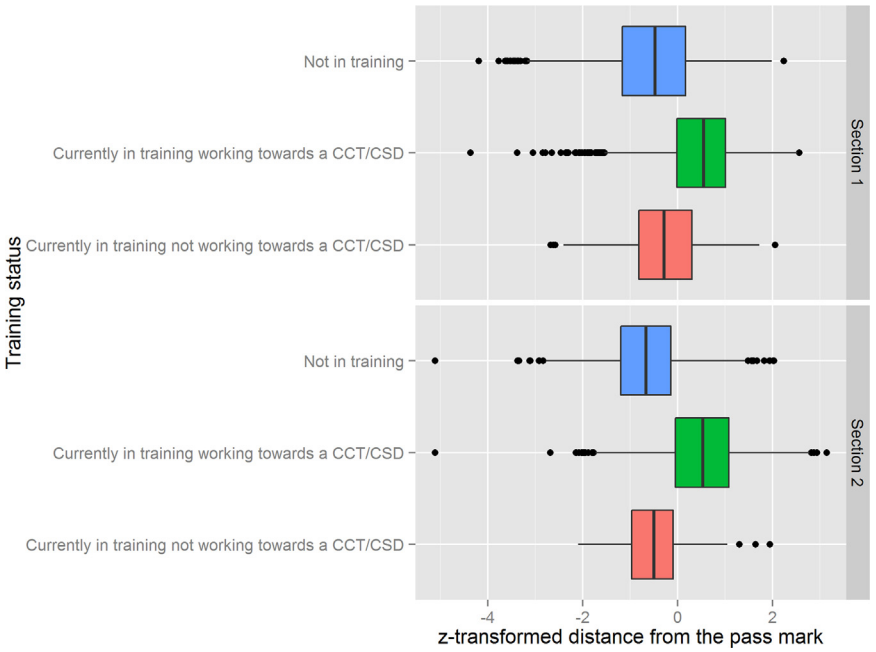

FIGURE 3. Candidate scores by training status across Section 1 and Section 2. CSD, certificate of specialist doctor.

candidates by gender was different in the 2 sections, as illustrated in Figure 6. Simple effects analyses carried out on the data split by section revealed a significant main effect for gender in both sections (Section 1: $F[2, 2374] = 7.99$, $p < 0.001$); Section 2: $F[2, 2498] = 8.90$, $p < 0.001$).

Bonferroni pairwise comparisons on the Section-1 data revealed that the male candidates achieved significantly higher scores than the female candidates ($p < 0.001$). The same analysis on the Section-2 data showed that the female candidates performed slightly better than the male candidates did but that the difference was not significant. The only significant pairwise comparisons in the Section-2 data were between the candidates who did not declare their gender and those who did. Removing the "prefer not to say"

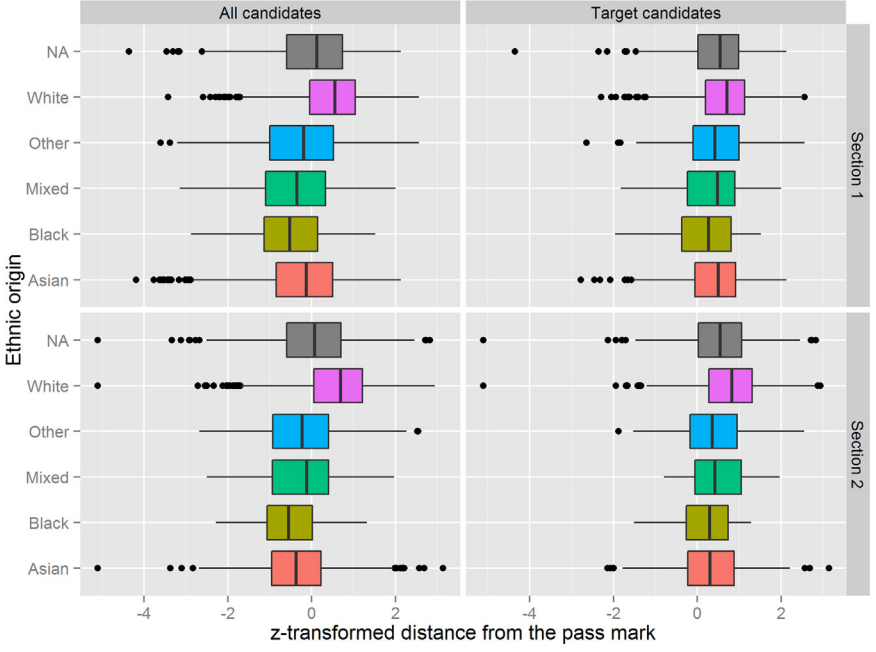

FIGURE 4. Scores by ethnic origin in all candidates and in "core candidates" in Section 1 and Section 2. NA, not applicable.

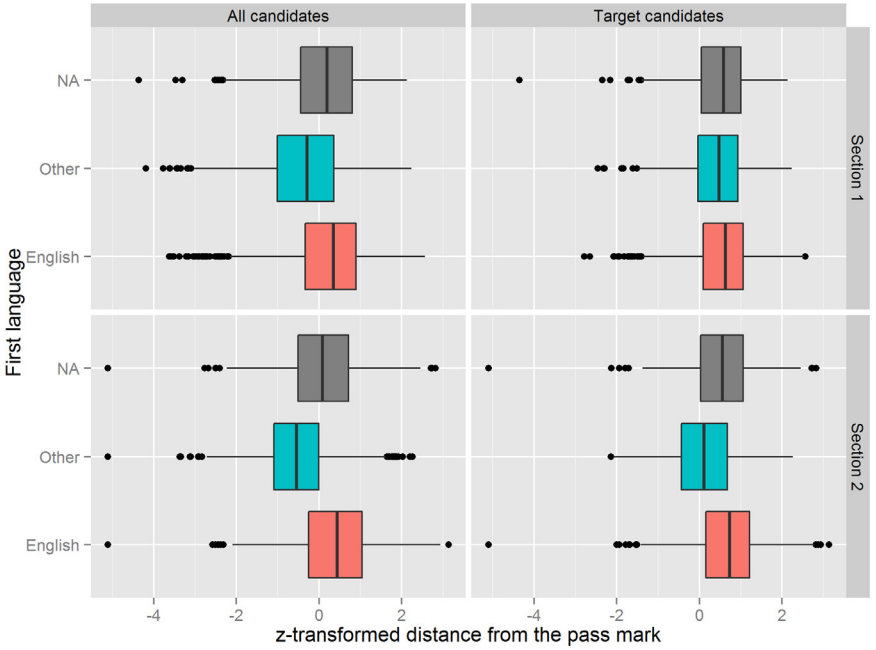

FIGURE 5. Scores by first language in all candidates and in "core candidates" in Section 1 and Section 2. NA, not applicable.

candidates from the data set and repeating the simple effects analysis revealed again a significant main effect for gender in both sections. The effect was greater in Section 1, where the male candidates achieved a significantly higher score than the female candidates ($t[2347] = 4.00$, $p < 0.001$), than in Section 2 where the female candidates achieved a significantly higher score than the male candidates ($t[2475] = 2.20$, $p = 0.03$).

DISCUSSION

Assessment at this level should ensure that only those who have demonstrated the required level of knowledge, competence and application thereof are allowed to pass, both for the reassurance of patients and for the accountability of the profession. The outcomes have far-reaching effects on the career opportunities of those who do not pass. The

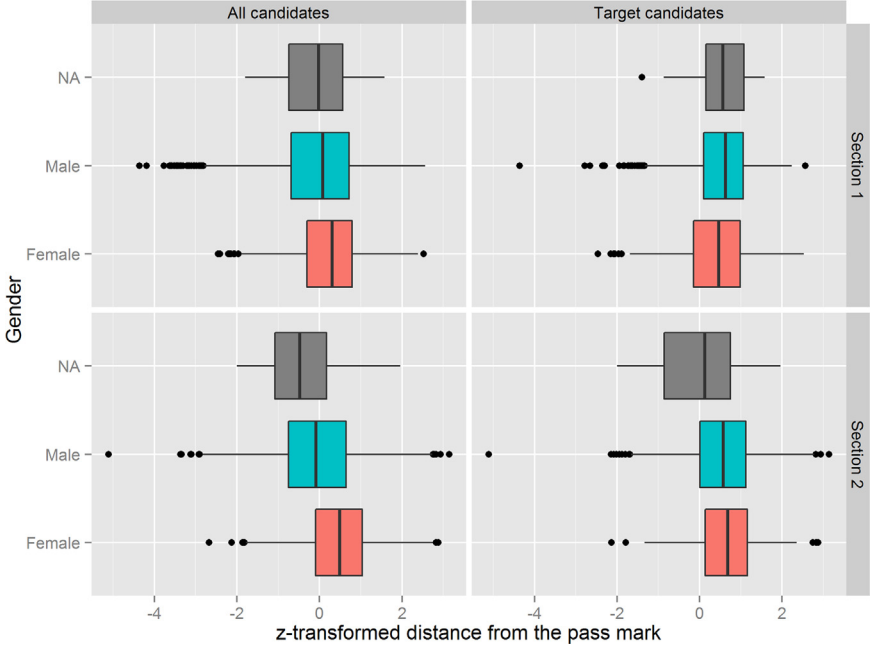

FIGURE 6. Scores by gender in all candidates and in "core candidates" in Section 1 and Section 2. NA, not applicable.

examinations must therefore allow for fair and robust decision making based solely on the candidates' abilities.

The JCIE complies with UK law (Equality Act) and is committed to meeting, and if possible exceeding, GMC standards on equality and diversity. The JCIE Equality and Diversity Policy required an initial scoping exercise, the results of which formed the basis of this first reported study of performance by demographic characteristics in the UK and Ireland surgical specialty Board examinations.

The exploration of demographic characteristics showed that there was a gradual increase in examiners describing themselves as having a nonwhite ethnic origin, with a 50% increase in the proportion of examiners from Asian ethnic origins, and a 63% increase in the proportion of examiners who do not have English as their first language during this period. These trends are encouraging, and reflect the desire to make the ethnic mix of examiners more representative of the various stakeholders in these examinations (the UK and Irish patient population, the medical workforce and the candidates). These changes have taken place as part of the natural evolution of examiner panels without any positive action in examiner recruitment or selection over this time frame.

The exploration of scores showed that the variable, which had the largest effect was training status. As these examinations were specifically developed to assess surgical training in the UK and Ireland against the Intercollegiate Surgical Curriculum, with examination standards set at the level of a day-1 consultant working in these countries, this result is not surprising. National trainees have undergone a minimum training of 6 years with support from a Deanery mentor, structured around a defined national curriculum with work-based assessments and formal annual review of competence progression. At the completion of their training, the training program director is required to certify that the candidate is at the level of a day-1 consultant before they are eligible to sit the examination. The effect training had on candidates' outcomes is part of the body of evidence supporting the validity of these examinations, whereby examinees with advanced training in the topic outperform those with less training.

When looking only at core candidates (in training towards a CCT and on their first attempt), we found smaller but significant main effects for ethnic origin and for first language. However, these effects were statistically indistinguishable between Section 1 (computer marked) and Section 2 (face to face). This provides reassurance that the differential attainment across these candidate groups could not be attributed to examiner bias in a face-to-face examination. The gender differences noted in Section 1 and Section 2 mirror a common observation of women performing better in constructed response formats and men performing better in selected response formats,[15-17] though more research is needed on this topic within postgraduate medical assessments.

These findings echo the findings from other investigations into differential attainment and reinforce the call for more research underpinning factors in the wider context of postgraduate medical training.

AUTHOR CONTRIBUTIONS

D.R., T.G., H.T., and P.T. conceived of the study. D.R., T.G. and H.T. provided the review of the literature. J.J. provided guidance on the gathering and reporting of equality and diversity data. C.F. provided statistical expertise in study design and conducted the statistical analysis. D. R., T.G., and C.F. provided the discussion of findings. All authors contributed to the refinement of the manuscript, approved the final manuscript and are accountable for all aspects of the work in ensuring that questions relating to the accuracy or integrity of any parts of the work are appropriately investigated and resolved.

REFERENCES

1. *General Medical Council*. Standards for curricula and assessment systems. 2010.
2. Equality Act 2010, United Kingdom.
3. Joint Surgical Colleges Equality and diversity Policy. Available at: ⟨jcie.org.uk⟩; 2013.
4. Wakeford R, Farooqi A, Rashid A, Southgate L. Does the MRCGP examination discriminate against Asian doctors? *Br Med J*. 1992;305(6845):931-935.
5. *British Medical Association*. Examining equality: a survey of royal college examinations. 2006.
6. Esmail A, Roberts C. Independent review of the Membership of the Royal College of General Practitioners (MRCGP) and the Clinical Skills Assesment (CSA) examination. University Manchester, 2003.
7. Esmail A., Roberts C. Academic performance of ethnic minority candidates and discrimination in the MRCGP examinations between 2010 and 2012: analysis of data. Br Med J. 2013;347: f5662.
8. The Queen On The Application Of BAPIO Action Ltd v Royal College General Practitioners and the General Medical Council [2014] EWHC 1416 (Admin).
9. White C. Regulator starts review of MRCGP pass rate amid claims of bias. *BMJ Careers*. 2013;19. Available at: ⟨http://careers.bmj.com/careers/advice/view-article.html?id=20011323⟩.

10. Talwar D. Foreign-trained doctors "face GP exam discrimination." *BBC News Health*. Available at: ⟨http://www.bbc.co.uk/news/health-23245607⟩; 2013.

11. Woolf K, Potts HWW, McManus IC. Impact of student ethnicity and primary childhood language on communication skill assessment in a clinical performance examination. *Br Med J*. 2011;342:d901.

12. Slick GL. Performance on the American Osteopathic Board of Internal medicine Certifying Examination 1986-2002 of various demographic groups and the impact of AOA re-entry resolutions on allopathic-trained candidates taking the examination. *J Am Osteopath Assoc*. 2004;104(7):294-300.

13. Fernandez A, Wang F, Braveman M, Finkas LK, Hauer KE. Impact of student ethnicity and primary childhood language on communication skills assessment in a clinical performance examination. *J Gen Intern Med*. 2007;22(8):1155-1160.

14. Spike NA, Hays RB. Analysis by training status of performance in the certification examination for Australian family doctors. *Med Educ*. 1999;33(8):612-615.

15. Mazzeo J, Schmitt AP, Bleistein CA. Sex-related performance differences on constructed-response and multiple-choice sections of advanced placement examninations. *College Board Report No. 92-97*, EST RR No. 93-95, 1993.

16. Kelly S, Dennick R. Evidence of gender bias in true-false-abstain medical examinations. *BMC Med Educ*. 2009;9:32.

17. Everaert P, Arthur N. Constructed-response versus multiple choice: the impact on performance in combination with gender. *Working Papers of Faculty of Economics and Business Administration, Ghent University*. Belgium: Ghent University, Faculty of Economics and Business Administration, 2012.