Mei Ling Denney, Adrian Freeman and Richard Wakeford

# MRCGP CSA:

## are the examiners biased, favouring their own by sex, ethnicity, and degree source?


## Abstract

**Background**
Concern exists regarding differential performance of candidates in postgraduate clinical assessments by ethnicity, sex, and country of primary qualification. Could examiner bias be responsible?

**Aim**
To explore whether candidate demographics affect examiners' judgements, by investigating candidates' case performances by candidates' and examiners' demographics.

**Design and setting**
Data on 4000 candidates (52 000 cases) sitting the MRCGP clinical skills assessment in 2011–2012.

**Method**
Univariate analyses were undertaken of subgroup performance (male/female, white/black and minority ethnic (BME), UK/non-UK graduates) by parallel examiner demographics. Due to confounding of variables, these were complemented by multivariate ANOVA and multiple regression analyses.

**Results**
Univariate analysis showed some differences between outcomes between the same-group and other-group examiners: these were contradictory regarding examiners 'favouring their own', for example, males received higher marks from female examiners than from males: maximum effect size was 3.6%. A six-way ANOVA confirmed all three candidate and examiner variables as having significant effects individually, identifying one significant interaction (examiner sex by examiner ethnicity). Stepwise regression showed candidate variables predicting 12% of score variance, parallel examiner demographics adding little (approximately 0.2% of variance). One 'transactional' variable proved significant, explaining 0.06% of score variance.

**Conclusion**
Examiners show no general tendency to 'favour their own kind'. With confounding between variables, as far as the impact on candidates' case scores, substantial effects relate to candidate and not examiner characteristics. Candidate–examiner interaction effects were inconsistent in their direction and slight in their calculated impact.



**Keywords**
educational measurement; foreign medical graduates; general practice; racism; sexism.


## INTRODUCTION

Clinical medical examinations are subject to a variety of potential threats to their reliability: while candidates' scores vary according to their ability, leading to 'true variance' in their scores, 'error variance' can result from a variety of sources, including variable case difficulty, variable differences in behaviour between patients, real or simulated, and from differential marking behaviour among and within examiners. This research addresses the concern that examiners may grade candidates differentially so as to discriminate unfairly by sex, ethnicity, or whether they were domestic or foreign medical graduates, by their own demographics.

Performance of subgroups of candidates often varies in terms of these variables. Presently, underperformance by international medical graduates (IMGs) in the high-stakes specialty qualifying examinations of the Royal College of General Practitioners (RCGP) is leading to claims of bias and to litigation.[1] In clinical examinations where examiners judge performance of candidates 'live' and thus can identify candidates' sex and ethnicity and possibly infer where their initial degree was obtained, the potential for unfair treatment could arise from systematic bias of parallel subgroups of examiners, who could favour their own kind by sex, ethnicity, or source of degree, especially if examining alone. To prevent such discriminatory behaviour, the medical profession has long advocated that examinations attempt to mimic the demographic characteristics of their candidature by those of their examiners.[2]

Unexplained differential performance in medical examinations among otherwise comparable subgroups of candidates has been described by McManus *et al* as a phenomenon that 'provides a serious challenge to the discipline of medical education'.[3] It is poorly researched: there are only two substantial studies of the potential bias of examiners by sex, ethnic group, or source of primary medical degree, and these concern the examinations of the MRCP(UK) examination where two examiners work together in pairs.[4,5]

### Differential candidate performance

Dewhurst *et al*[4] found sex and ethnicity differences among UK graduates (UKGs) in the intercollegiate MRCP Part 2 Clinical Examination (PACES) in 2003–2004 (male candidates failing at 1.5 times the rate of female candidates, and black and minority ethnic (BME) candidates failing at 1.7 times the rate of the white candidates); the latest (2012) published statistics[6] show sex differences (UKG male candidates failing at 1.3 times the rate of UKG female candidates), ethnicity differences (BME UKG candidates failing at 1.3 times the rate of white UKG candidates), and differences with regard to source of primary medical degree (IMGs failing at 3.1 times the rate of UKGs).

The clinical examination of the Royal College of Paediatrics and Child Health,


**ML Denney**, FRCGP, clinical lead for research, MRCGP, Royal College of General Practitioners, London. **A Freeman**, FRCGP, senior lecturer in medical education, University of Exeter Medical School, Exeter. **R Wakeford**, MA, CPsychol, Life Fellow, Hughes Hall, University of Cambridge, Cambridge.

**Address for correspondence**
Richard Wakeford, Hughes Hall, University of Cambridge, Cambridge CB1 2EW.
**E-mail:** rew5@cam.ac.uk



**Submitted:** 22 June 2013; **Editor's response:** 23 July 2013; **final acceptance:** 14 August 2013.





This is the full-length article (published online 9 Oct 2013) of an abridged version published in print. Cite this article as: **Br J Gen Pract 2013; DOI: 10.3399/bjgp13X674396**

## How this fits in

There is considerable variation in performance between demographically-defined candidate subgroups (for example, male/female) in postgraduate UK medical examinations. In objective structured clinical examinations (OSCEs) there is potential for unfair bias associated with parallel examiner demographics (such as male/female), but very little is known about such effects. This study analysed a year's worth of cases from the MRCGP clinical skills assessment in terms of candidates' and examiners' sex, ethnicity, and whether they qualified in the UK or abroad. It identified some differences but no general effect of 'examiners favouring their own kind'; a significant general finding. The differences suggest that it would be prudent to review OSCE circuits routinely to ensure examiner heterogeneity, but they offer no support for the presumed desirability of seeking to match the demographic characteristics of the examiner group to those of the candidates.

(the MRCPCH) reports differences (2010–2011) with regard to sex ('pass-rate is between 15% and 25% higher for females as compared to males') and with regard to source of degree, with IMGs failing at approximately 1.9 times the rate of UKGs.[7]

The clinical examination of the Royal College of Psychiatrists (the MRCPsych CASC) reports (2008–2010) sex differences (UKG male candidates failing at 1.5 times the rate of UKG female candidates), ethnicity differences (BME UKG candidates failing at 2.4 times the rate of white UKG candidates), and differences with regard to source of primary medical degree (IMGs failing at 4.7 times the rate of UKGs on first attempt).[8]

In the MRCGP in first attempts at the clinical skills assessment (CSA) the following differences have been reported: sex differences (UKG male candidates failing at 2.1 times the rate of UKG female candidates); ethnicity differences (BME UKG candidates failing at 3.2 times the rate of white UKG candidates); and differences with regard to source of primary medical degree (IMGs failing at 6.6 times the rate of UKGs).[9]

A major review and meta-analysis by Woolf *et al* of performance by ethnic group in UK examinations showed consistent underperformance by BME medical students and postgraduates,[10] across UK specialties. Differential performance by candidate subgroups is not limited to UK examinations and has been reported in the US[11] and Canada.[12]

Differential performance is not limited to clinical examinations. In the MRCGP's computer-delivered test of applied clinical knowledge (AKT) in 2011–2012 among first-attempt UKGs the male failure rate was 1.7 times the female failure rate, the BME failure rate was 2.9 times the white failure rate, and first-attempt IMGs failed at 4.0 times the UKG rate.[9]

### Examiners' and their marking behaviour

It has long been recognised that individual examiners may vary, often systematically, in the marks that they award to candidates[13] and considerable work has been undertaken to identify (and possibly train or expel) outliers or to correct their marks.[14]

When McManus and colleagues[15] adjusted marks for examiner stringency in the MRCP(UK) PACES they found that about 4% of candidates' outcomes would have changed. Harasym and colleagues[16] used a similar approach in undergraduate family medicine objective structured clinical examinations (OSCEs) and found that removing hawkish (stringent) and doveish (lenient) influences potentially changed the outcome for 11% of candidates. In a recent study,[5] McManus and colleagues investigated the marks awarded by about 2000 examiners working in pairs in the MRCP(UK) PACES (and nPACES) and found that approximately 2% of examiners were statistically significant hawks and 2% significant doves; none showed significant sex bias; and only one showed ethnicity bias across both PACES and nPACES (a BME examiner who appeared to be giving relatively high marks to BME than white candidates).

In a study of the MRCP(UK) Dewhurst *et al*[4] reported no association between candidate and examiner sex and a small but highly significant interaction of candidate and examiner ethnicity on stations assessing communication skills and ethics. McManus and colleagues detected 'possible sex bias in no examiners and possible ethnic bias in only one' in the MRCP(UK).[5]

### The MRCGP clinical skills assessment

The CSA is single marked. It comprises a circuit of 13 stations, blueprinted against the curriculum and changed daily, with each case being depicted by role players who are accompanied by the examiner who marks the candidate's performance. Normally, three circuits run in parallel in the morning, and the same set of circuits/stations runs in the afternoon with appropriate 'quarantining' of candidates over lunchtime. Systematic prior training

of examiners and role players takes place, concluding in the hour before the morning examination, with the three sets of role players and examiners coordinating their patient portrayal and marking scheme to ensure parallel deliveries and assessments. The examiners grade the candidates on their own, without a co-examiner present.

In common with other UK medical Royal Colleges, the RCGP seeks to recruit its examiners from as diverse a base as possible, in part towards representativeness of their examinations' candidature. However, all must succeed in passing a recruitment process which involves a successful retake of the AKT multiple choice knowledge test and a day's observed individual and group performance working up and marking videoed cases. The College then notes information on examiner (and candidate) demographics to facilitate the later exploration of issues surrounding 'fairness': the examining body being mindful of its duty in law 'to promote fairness' across defined population groups, including by sex and ethnicity (Equality Act 2010). Self-reported ethnicity and sex are recorded, and whether the doctor graduated in the UK or elsewhere. Typically the candidate data are 99% complete for ethnicity and 100% complete for sex and source of medical degree; and complete data exist on all three demographics regarding the examiners.

## METHOD

There were 4029 candidate attempts at the CSA in 2011–2012; with 13 cases comprising an attempt, this represented a total of 52 377 assessed candidate cases. However, ethnicity data were absent on 29 candidate attempts (<1%), and these candidates' cases were deleted from the analysis. The remaining candidates' data were complete and the database comprised 52 000 candidate cases. Due to the relatively small numbers of black, Chinese, and other minority ethnic candidates, BME candidates were conflated into a single group, the system used by Woolf *et al*.[10] Thus candidates and examiners were all classified as male or female, UKG or IMG, and white or BME.

A demographic summary of the candidature, the 251 examiners who participated in the examinations during 2011–2012, and the 'transactional ethnicity' of the case pairs are contained within Table 1. Some relatively small cell sizes result from the currently small numbers of IMG examiners, although no cell contains less than 981 occurrences: the small numbers are due to the increase in foreign graduates being relatively recent, building up only in the first decade of the 21st century,[17] and thus causing numerical disparity between older IMGs (potential examiners) and younger IMGs (candidates). Here, 'IMG' is used synonymously with non-UK, and includes those candidates or examiners from the EEA/EU.

Examiners mark each case that they see on three domains — data-gathering skills, interpersonal skills, and clinical management skills — which are each graded 'clear pass' (3 marks), 'pass' (2 marks), 'fail' (1 mark) or 'clear fail' (0 marks), providing a 'case score' of between 0 and 9, and a maximum possible examination total of 117. For standard-setting purposes only, examiners award an additional 'borderline' grade, but this is not used in the case score or the examination total.

The overview case score statistics were calculated first by subgroups of candidates and examiners. 'Error bar' graphs were then drawn which show the mean subgroup candidate case scores (with 95% confidence interval) by sex (male/female), ethnic group (white/BME) and source of medical degree (UKG/IMG), cross-tabulated by the same examiner subgroups. Next, in view of the lack of previous work (and thus information) on the general topic of examiner–candidate encounters with variable demographics on both sides, a six-way univariate analysis of variance (ANOVA) was performed on the three demographic characteristics of candidates and examiners to assess the importance of the 64 possible interactions. Finally, two stepwise linear regression analyses of candidates' case scores were undertaken to estimate the importance and extent of error variance introduced as a result of these candidate and examiner demographics, and of the possible influence of the three 'transactional' variables, indicating whether the transactional encounter was homogenous (for example, male examiner, male candidate) or heterogeneous (for example, BME examiner, white candidate) with respect to sex, ethnicity, and source of medical degree.

## RESULTS

### Raw candidate and examiner group differences on case scores

The comparisons of raw candidate and examiner group case score data show significant differences (by ANOVA) between all groups compared, relatively small as between examiner subgroups (mean differences 1–4%) but relatively large as between candidate subgroups (mean

**Table 1. Sample numbers, percentages and mean case scores of candidate, examiner subgroups, encounters in which examiner–candidate transaction demographics were the same, and where they were different**

| Demographic variable | | | Candidates in demographic subgroup | Examiners in demographic subgroup | Cases with *same* transactional demographic | Cases with *different* transactional demographic |
|---|---|---|---|---|---|---|
| **Ethnicity** | White | *n* | 1586 | 215 | 17 433 | 3185 |
| | | Total, % | 39.4 | 85.7 | 33.5 | 6.1 |
| | | Mean | 6.67 | 6.03 | 6.67 | 6.66 |
| | BME | *n* | 2414 | 36 | 4721 | 26 661 |
| | | Total, % | 60.6 | 14.3 | 9.1 | 51.3 |
| | | Mean | 5.64 | 6.15 | 5.81 | 5.61 |
| **Sex** | Female | *n* | 2044 | 96 | 9921 | 16 651 |
| | | Total, % | 51.1 | 38.2 | 19.1 | 32.0 |
| | | Mean | 6.38 | 6.08 | 6.39 | 6.37 |
| | Male | *n* | 1956 | 155 | 16 179 | 9249 |
| | | Total, % | 48.9 | 61.8 | 31.1 | 17.8 |
| | | Mean | 5.70 | 6.03 | 5.67 | 5.75 |
| **Degree source** | UKG | *n* | 2224 | 241 | 27 472 | 1440 |
| | | Total, % | 55.6 | 96.0 | 52.8 | 2.8 |
| | | Mean | 6.61 | 6.06 | 6.62 | 6.38 |
| | IMG | *n* | 1776 | 10 | 981 | 22 107 |
| | | Total, % | 44.4 | 4.0 | 1.9 | 42.5 |
| | | Mean | 5.35 | 5.84 | 5.04 | 5.36 |
| **Total** | | *n* | 4000 | 251 | 52 000 | |
| | | Total, % | 100.0 | 100.0 | 100.0 | |

*BME = black and minority ethnic. IMG = international medical graduate. UKG = UK graduate.*

differences 12–24%). Table 1 includes these in the first two columns of data. Four comparisons were significant at $P<0.001$, with the data on candidates demonstrating the directional differences also seen in the MRCGP Annual Reports.[9]

Cross-tabulation of mean case scores of candidates by their three principal demographics (sex, ethnicity, source of primary medical degree) are shown visually in the error bar plots below (Figure 1) by parallel examiner demographics: Table 1 shows the actual data (mean scores) in the final two columns.

The mean effect size of the various differences evident in the chart are as follows, in order of size:

- IMGs receive a higher mark from UKG examiners than from IMG examiners (mean difference = 0.32 marks, 3.6%) $P<0.001$.
- UKGs receive a higher mark from UKG examiners than from IMG examiners (mean difference = 0.24 marks, 2.7%) $P<0.001$.
- BME candidates receive a higher mark from a BME examiner than from a white examiner (mean difference = 0.20 marks, 2.2%) $P<0.001$.
- Male candidates receive a higher mark from a female examiner than from a male examiner (mean difference = 0.08 marks, 0.9%) $P<0.001$.

The remaining two differences (female candidates receiving a higher mark from a female examiner than from a male examiner; white candidates receiving a higher mark from a white examiner than from a BME examiner) were not significant.

### Six-way univariate analysis of variance

The six-way univariate ANOVA showed that as main effects the three demographic characteristics all had predictive ability, although to various levels of statistical significance: in order, candidate degree source ($F = 232.9$, degrees of freedom [df] = 1, $P<0.0001$), examiner degree source ($F = 60.9$, df = 1, $P<0.0001$), candidate sex ($F = 30.2$, df = 1, $P<0.0001$), examiner

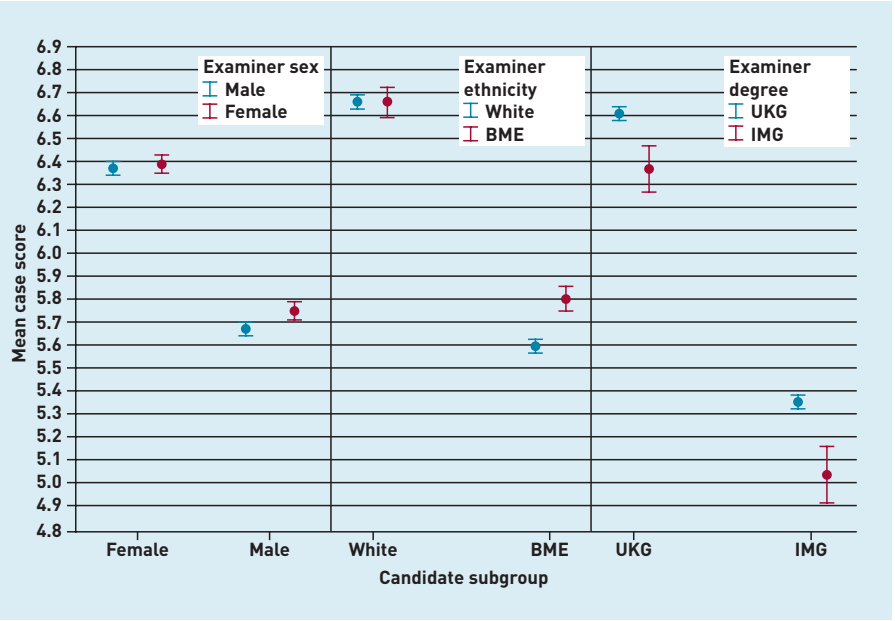


***Figure 1. Mean case scores (0–9 with 95% CI) of candidate subgroups by parallel examiner subgroups. BME = black and minority ethnic. IMG = international medical graduate. UKG = UK graduate.***

ethnicity (F = 30.0, df = 1, *P*<0.0001), examiner sex (F = 5.3, df = 1, *P*<0.02), and candidate ethnicity (F = 4.0, df = 1, *P*<0.05). Of all possible interactions, only one transpired to be significant at *P*<0.05 (Bonferroni corrected): examiner ethnicity by examiner sex. Male examiners gave similar grades, whether white or BME, whereas BME female examiners gave higher grades than white female examiners (effect size approximately 0.8 marks out of 9 or 8.9%).

### Stepwise linear regression analyses

The results of the stepwise multiple regression analysis using the six candidate and examiner demographics as predictors of the case score are shown in Table 2. This analysis shows that the principal sources of systematic variance were candidate demographics, which accounted for about 12% of total case score variance. Examiner demographics added very little (approximately 0.2% of the total variance).

Finally, to establish the possible systematic influence of the pairing of candidates and examiners, a transactional classification variable was calculated for each of the three aspects, sex, ethnicity, and degree source. These indicated whether the transactional encounter was homogenous (such as male examiner, male candidate) or heterogeneous (such as BME examiner, white candidate). In this second stepwise regression analysis, these three derived variables replaced the three 'examiner' variables listed in Table 2.

In the event, only one of these three 'transactional' variables proved significant in this second analysis: whether ethnicity was the same, candidate to examiner, or different. This showed that a slightly lower mark was awarded to candidates of the opposite ethnicity than to those of the same; but the impact of this was minimal, accounting for 0.06% of total variance in case score, with effect size, computed as semi-partial R = 0.025. 'Transactional' source of degree and sex were not significant predictors.

## Table 2. Stepwise multiple regression of predictors of case score (0–9)

| Model | Variable entered | Multiple R | $R^2$ | $R^2$ change | B | β | 95% CI | F change | *P*-value | Variance explained |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Candidate: UK/non-UK graduate | 0.321 | 0.103 | 0.103 | −0.955 | −0.244 | −0.993 to −0.916 | 5965.41 | <0.001 | 10.3% |
| 2 | Candidate: male/female | 0.338 | 0.114 | 0.011 | 0.378 | 0.097 | 0.346 to 0.410 | 669.72 | <0.001 | 1.1% |
| 3 | Candidate: white/BME | 0.348 | 0.121 | 0.007 | −0.396 | −0.100 | −0.435 to −0.357 | 395.29 | <0.001 | 0.7% |
| 4 | Examiner: UK/non-UK graduate | 0.349 | 0.122 | 0.001 | −0.440 | −0.048 | −0.522 to −0.359 | 53.84 | <0.001 | 0.1% |
| 5 | Examiner: white/BME | 0.351 | 0.123 | 0.001 | 0.236 | 0.043 | 0.187 to 0.284 | 87.78 | <0.001 | 0.1% |
| 6 | Examiner: male/female | 0.351 | 0.123 | 0.000 | 0.041 | 0.010 | 0.008 to 0.074 | 6.09 | <0.020 | <0.1% |

## DISCUSSION

### Summary

In this study, candidate–examiner interaction effects were inconsistent in their direction in terms of examiners 'favouring their own' (the statistically significant effects were that BME examiners favoured BME candidates, female examiners favoured male candidates, and IMG examiners gave lower marks to both UKG and IMG candidates) and also slight in their calculated impact. The effect size of the potential significant raw effects found in this study, regarding any individual candidate case, could result in, for example, male candidates receiving a 0.9% enhancement of their case score under a female examiner and any candidate receiving, irrespective of their source of degree, a 2.4% enhancement of their case score under a UKG examiner as opposed to an IMG examiner (Table 1).

These are crude potential effects and the variables concerned are confounded (ethnicity with source of medical degree, for example), hence the need for multivariate analysis. However, what the crude effects demonstrate is the need to apply the various examiner groups (male/female, white/BME, UKGs/IMGs) as fairly as possible across the days and circuits of candidates, so that no candidate experiences, for example, all female or all IMG examiners.

This study provides no support for equating examiner representation to that of candidates from the point of view of delivering a fair assessment to all groups of candidates. Nevertheless incorporating a variety of subgroups of examiners in the examiner panel has benefits for collegiality and examination development, and incorporating approaches to practice which may themselves vary between these subgroups.

As far as the impact on candidates' case scores is concerned, the only substantial sources of variance in this examination relate to their own, and not to examiners', demographic characteristics; 12% and 0.2% of case score variance, respectively. When the confounding of these variables was accounted for, the effects were slight in their impact, the only significant 'transactional' effect explaining 0.06% of case score variance, or, for instance, about one-third of that accounted for by the sequence in which a case is taken.

### Strengths and limitations

One of the main strengths of this study was its large dataset which included all candidates and examiners involved in a year's worth of a high-stakes national examination and the comprehensive information collected on candidates' and examiners' demographics. This study also addressed an issue on which the existing evidence base is very small and, therefore, makes a significant contribution to this small knowledge base. Thus, it should feed into contemporary real-world discussions and decisions.

The study's principal weaknesses include:

1. Non-equivalent participation by examiners. The 251 examiners assessed an average of 207 cases in the study, but the range was from 22 to 610. Very much more complex analyses would have been necessary to account for this 'clustering'.
2. The need to conflate BMEs into single groups (candidates and examiners). The number of examiners in certain BME subgroups made this necessary.
3. The fact that, determined largely by differential application rates to read medicine at UK universities, representation by BMEs is unrepresentative of population norms.
4. That it is concerned, necessarily, with an examination that is single-marked. Some high stakes examinations have two examiners present.
5. That the unit of analysis is the 'case' and not the examination. This study examined the performance of candidates (and examiners) on individual cases, not on their overall outcomes.
6. That it does not examine the issue of possible bias introduced by simulated patients or role players.

An additional issue may be the application of Bonferroni corrections (for repeated testing) to the significance levels in the six-way ANOVA: are important 'true' effects being concealed? No evidence was found for this.

### Comparison with existing literature

Differential performance among student subgroups is by no means unknown in higher education generally in the UK, although little researched. A 2008 review[18] by Richardson for the Higher Education Quality Assurance Agency concluded, as regards sex, that 'since 1990 women have been much more likely to obtain good degrees than men' and as regards ethnicity that 'white students are both more likely to obtain good degrees and more likely to obtain first-class honours than are students from other ethnic groups'. This


### Funding

None.

### Ethical approval

Ethical approval for service evaluation and audit work was not required.

### Provenance

Freely submitted; externally peer reviewed.

### Competing interests

All three authors are involved in the design, delivery, and quality assurance of the MRCGP Clinical Skills Assessment.

### Acknowledgements

We thank the Assessment Development Committee of the RCGP for their encouragement to undertake and publish this work, and Professor Chris McManus for statistical advice.


### Discuss this article

Contribute and read comments about this article on the Discussion Forum: **http://www.rcgp.org.uk/bjgp-discuss**

concurs with Woolf *et al*'s review and meta-analysis findings from medicine.[10]

This comment emphasises the fact that much work on differential performance by ethnicity, including the present study, conflates all BME groups. That this is unsatisfactory can be seen in the differential performance by candidate subgroups in both the examination components of the MRCGP[9]: the BME subgroups do not comprise a performance-homogenous whole. Conflation was necessary here, not only to permit comparison with the major recent work,[10] but particularly because of small numbers of examiners in most BME groups other than South Asian.

It is important to acknowledge that differential performance in examinations may occur as a result of selection procedures which recruit subgroups of differential ability, and from training programmes of varying effectiveness, as well as from possible unfairness in the examinations themselves. Indeed, recent research[17] has suggested that there are marked differences between UKGs' performance and that of IMGs in the selection procedures for general practice that would themselves predict very different results in similar assessments later on in training.

**Implications for practice and research**

As far as the delivery of OSCEs is concerned, it is important to note that there may be systematic bias introduced by different subgroups of examiners, although in unpredictable directions (for example, women examiners marking male candidates higher than do male examiners). The practice implication of this is that circuits of OSCE examiners should routinely be subject to an 'equality audit' to ensure that none are unusually imbalanced with regard to sex, ethnicity, or source of medical qualification, with examiners being swapped across circuits if necessary to improve the balance.

A second implication for practice arises from discussions with colleagues from other similar examinations, namely a recommendation that a standard comprehensive dataset of information be recorded for each and every case at an OSCE delivery. The dataset should include:

- For candidates: GMC number (which permits recording of sex and source of primary medical qualification), ethnic group, and age.
- For examiners: (the same as for candidates).
- For patients/role players: a unique permanent reference number, ethnic group, age, sex.
- For candidate cases: the sequence in the OSCE in which the candidate encountered the case.

Further studies should address the statistical issues raised by differential examiner participation and investigate candidate examination outcomes (such as the impact on candidates' CSA scores as opposed to individual case scores). Investigating differences within the BME candidate group should certainly form a topic for future research, as should further qualitative research towards understanding the detailed genesis of performance differences. Further work is also needed to compare the potential biases within and costs of single versus dual marking of such OSCE examinations: where does the desirable balance lie?

The overarching conclusion of this analysis is that systematic bias by examiner subgroup does not explain any substantial differential candidate subgroup performance. Although there are some significant differences in examiner subgroup marking behaviour, poorly performing candidate subgroups do not arise from the marking behaviour of specific examiner subgroups. Examiners show no general tendency to 'favour their own kind'.

## REFERENCES


1. Bostock N. Fairness of CSA tests faces judicial review scrutiny. *GP* 2013; **4 Mar:** http://www.gponline.com/News/article/1173247 (accessed 1 Oct 2013).
2. British Medical Association. *Examining equality: a survey of royal college examinations*. London: BMA, 2006.
3. McManus IC, Woolf K, Dacre J. The educational background and qualifications of UK medical students from ethnic minorities. *BMC Med Educ* 2008; **8:** 21. doi: 10.1186/1472-6920-8-21.
4. Dewhurst NG, McManus C, Mollon J, *et al*. Performance in the MRCP(UK) examination 2003-4: analysis of pass rates of UK graduates in relation to self-declared ethnicity and gender. *BMC Med* 2007; **5:** 8.
5. McManus IC, Elder AT, Dacre J. Investigating possible ethnicity and sex bias in clinical examiners: an analysis of data from the MRCP(UK) PACES and nPACES examinations. *BMC Med Educ* 2013; 1**3:** 103. doi: 10.1186/1472-6920-13-103.
6. Federation of Royal Colleges of Physicians of the UK. *MRCP(UK) and Specialty Certificate Examinations: pass rates by gender and ethnicity – 2012*. http://www.mrcpuk.org/SiteCollectionDocuments/MRCP(UK)%20Pass%20Rates%20by%20Gender%20Ethnicity%202012.pdf (accessed 1 Oct 2013).
7. General Medical Council. *Annual Specialty Report for 2010/11: Royal College of Paediatrics and Child Health*. http://www.gmc-uk.org/RCPCH_ASR.pdf_48272553.pdf (accessed 1 Oct 2013).
8. Royal College of Psychiatrists. *MRCPsych Examinations Cumulative Results 2008-2010*. http://www.rcpsych.ac.uk/pdf/MRCPsych%20Cumulative%20Results%20Report%20-%20August%202011.pdf (accessed 1 Oct 2013).
9. Wakeford R. *MRCGP statistics 2011–12: Annual Report on the AKT and CSA Assessments*. 2012. http://www.rcgp.org.uk/gp-training-and-exams/mrcgp-exam-overview/mrcgp-annual-reports/~/media/Files/GP-training-and-exams/Annual%20reports/MRCGP%20Statistics%202011-12%20final%20121212.ashx (accessed 4 Oct 2013).
10. Woolf K, Potts HWW, McManus IC. Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. BMJ 2011; **342:** :d901.
11. American Board of Family Medicine. Examination pass rates. 2013 pass rates. https://www.theabfm.org/about/2013passrates.pdf (accessed 4 Oct 2013).
12. MacLellan A-M, Brailovsky C, Rainsberry P, *et al*. Examination outcomes for international medical graduates pursuing or completing family medicine residency training in Quebec. *Can Fam Physician* 2010; **56(9):** 912–918.
13. Burt C. The analysis of examination marks. In: Hartog P, Rhodes EC (eds). *The marks of examiners*. London: MacMillan, 1936; 309–310.
14. Roberts C, Rothnie I, Zoanetti N, Crossley J. Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? *Med Educ* 2010; **44(7):** 690–698. doi: 10.1111/j.1365-2923.2010.03689.x.
15. McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modeling. *BMC Med Educ* 2006; **6:** 42.
16. Harasym P, Woloschuk W, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Adv Health Sci Educ Theory Pract* 2008; **13(5):** 617–632. Epub 2007 Jul 3.
17. Wakeford R. International medical graduates' relative under-performance in the MRCGP AKT and CSA examinations. *Educ Prim Care* 2012; **23(3):** 148–152.
18. Richardson JTE. The attainment of ethnic minority students in UK higher education. *Studies in Higher Education* 2008; **33(1):** 33–48.