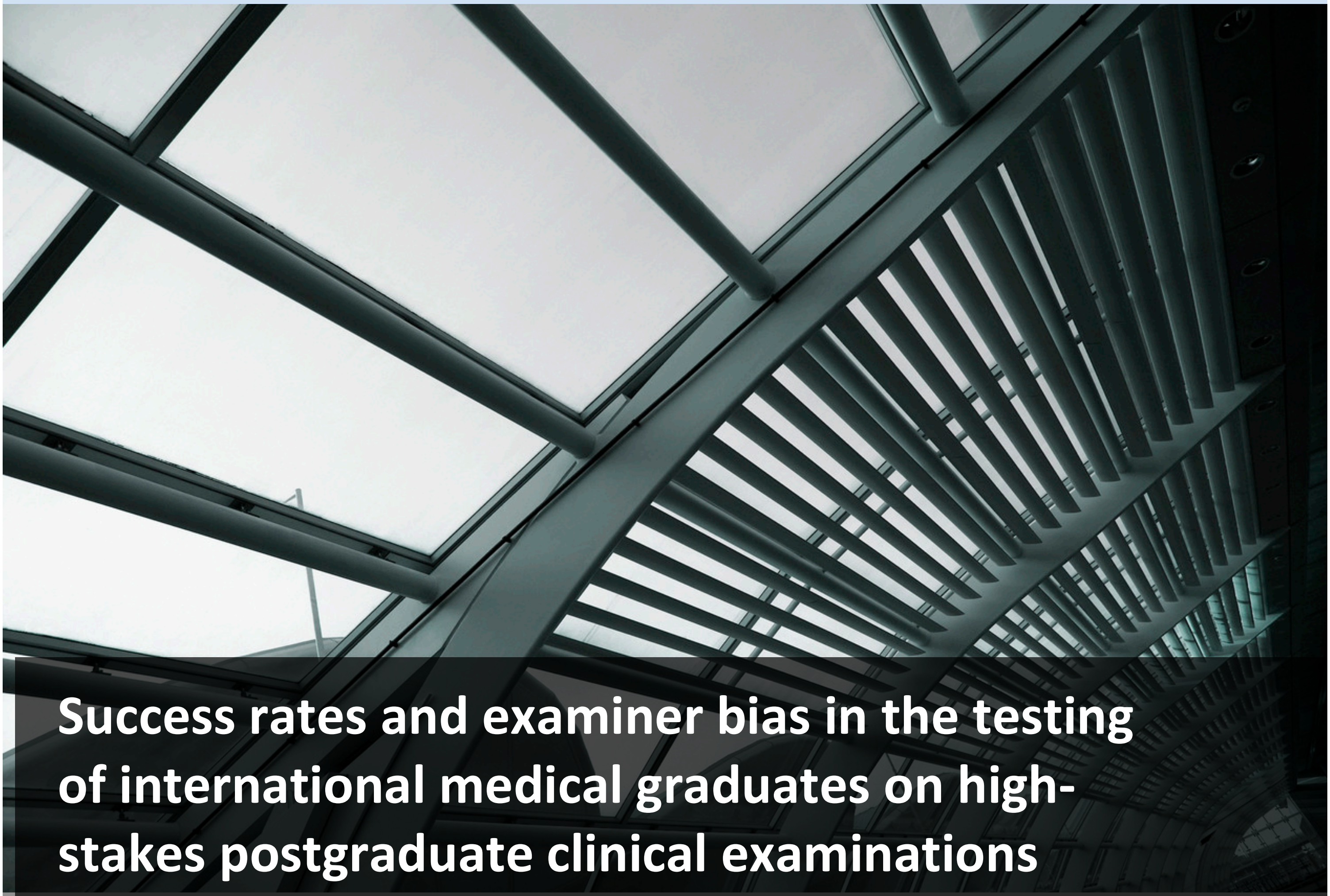


Success rates and examiner bias in the testing of international medical graduates on high-stakes postgraduate clinical examinations

Professor Liz Farmer

Report prepared for the Australasian College for Emergency Medicine

16 August 2017

Table of contents

List of Tables and Figures **4**

Table of abbreviations and acronyms **5**

Definitions of IMGs in the Australian context **5**

Introduction to the OSCE **6**

The nature of expert judgment in clinical examinations **7**

Focus of the review **9**

Data sources for the review **9**

Research ethics **10**

Success rates and examiner bias in the testing of international medical graduates on high-stakes clinical examinations **10**

International studies – North America **10**

International studies – United Kingdom **13**

National studies - Australia and New Zealand **27**

Key findings **32**

List of Tables and Figures

Figure 1. Analysis of examiner biases for the 26 PACES examinations. Reproduced from McManus et al.

Figure 2. Mean station scores of candidates by their three principal demographics (sex, ethnicity, source of primary medical degree) by parallel examiner demographics. Reproduced from Denney et al.

Figure 3. Pass rates in the Royal Australian College of General Practitioners Fellowship exam from 1999 to 2004 according to route of entry and IMG status. Reproduced from Jasper et al.

Table 1: Number of physicians and percentage of board certified in the 10 practice disciplines with the largest number of diplomats. Reproduced from Norcini et al.

Table 2. Pass rates of local and IMG candidates in the Australia and New Zealand College of Anaesthetists final examination (extracted from Higgins et al).

Table of abbreviations and acronyms

Abbreviation	Full term
ACEM	Australasian College for Emergency Medicine
ANZCA	Australia and New Zealand College of Anaesthetists
BAPIO	British Association of Physicians of Indian Origin
BME	Black and minority ethnic
CFPC	College of Family Physicians of Canada
CHS	comparable health care systems
CICM	College of Intensive Care Medicine
CSA	clinical skills assessment
CT scan	computerised tomography
ECG	electrocardiogram
GMC	General Medical Council
IMG	international medical graduate
ITER	in training evaluation reports
MRCP (UK)	Membership of the Royal College of Physicians in the United Kingdom
MRI	magnetic resonance imaging
non-USIMG	non-United States citizens trained internationally
nPACES	new PACES
OSCE	objective structured clinical examination
PACES	practical assessment of clinical examination skills
RCA	UK Royal College of Anaesthetists
RCPSC	Royal College of Physicians and Surgeons of Canada
RP	role-play patient
SP	simulated or standardised patient
UKG	United Kingdom Graduate
USIMG	United States citizens trained internationally
USMG	United States medical graduates

Definitions of IMGs in the Australian context

An international medical graduate (IMG) is a doctor trained outside Australia who is a licensed practitioner in a different country. This term is usually used to depict Australian citizens who trained in overseas institutions as well as non-Australian citizens who trained overseas. Some studies report these groups differentially. Where this occurs, this is noted in the review.

Introduction to the OSCE

Harden and Gleeson first described the Objective Structured Clinical Examination (OSCE) in 1975[1]. It has been widely adopted by educational bodies for assessing clinical skills in medicine and the health professions. It is in broad use in high stakes examinations for undergraduate entry to health professional practice and also for conferring specialist qualifications.

An OSCE consists of a number of "stations" that candidates undertake sequentially under examination conditions. Stations vary in length and complexity.

At each station the candidate is presented with a relevant clinical task usually involving:

- A patient who may be a real patient or a simulated patient/role player (SP/RP)
- a model (simulation) or
- another health professional.

Stations may include a wide variety of accessory materials including:

- Images eg Xrays, Ultrasound, CT, magnetic resonance imaging (MRI), photographs of clinical signs, ECG
- test results
- video recordings
- specialised equipment for demonstration.

Advantages of this format include:

- Clinical skills can be directly observed in realistic situations
- a broad range of skills can be assessed in a relatively short period of time
- assessment tasks are predetermined and 'standardised' so that all candidates receive the same challenges
- the reliability of the examination is moderately high
- the ability to use multiple examiners aims to reduce the effects of examiner (or rater)[2] bias.

One or more examiners are present in the station and observe and mark the candidate. Stations are 'scored' or 'marked' using a written pro-forma, on paper or on electronic tablet. Usually two scores are obtained, the 'station score' and a final 'global rating'.

[1] Harden RM, Gleeson FA. "Assessment of clinical competence using an objective structured clinical

[2] Examiners and raters are terms that are often used interchangeably in the literature

The station score usually comprises either checklists (dichotomous yes/no answer) or rating scales for example " history taking skills", "physical examination skills", which are rated on a scale usually involving five or more points. Some scoring methods employ both approaches.

At the conclusion of the station one or more examiners allocate a global score on a predetermined scale, indicating the proficiency of the candidate on the tasks required at the level required. If two examiners are present, they will either allocate their marks independently, or will achieve a consensus mark for the station.

The nature of expert judgment in clinical examinations

The final score in a clinical examination such as an OSCE, however created, remains an expert judgement in the mind of the examiner(s). Ideally, examiners should carry out their scoring in ways that are completely consistent with the construct of the station and its measurement goals[3]. This ensures that the station scoring is valid ie it tests what it is intended to test and nothing else.

Bias of clinical examiners against some types of candidate, based on characteristics such as gender, ethnicity, previously seen candidate performance[4] or first impressions[5], would represent a threat to the validity of an examination, since such biases are 'construct-irrelevant' characteristics.

It is widely known and universally expected however, that experts will disagree when marking OSCE stations, that is there will often be a difference between the examiners' views of the candidate ability on any given set of tasks and on the overall judgment.

One such and probably the best-known variability is the leniency - stringency tendency effect also referred to as 'Doves versus Hawks' or 'easy markers' versus 'hard markers' [67]. Candidates are frequently reminded that the non-verbal behaviour of examiners may disguise their leniency – stringency tendency! This is known as, for stringent examiners, the "smiling death" phenomenon[8].

[3] Bejar II. Rater Cognition: Implications for Validity. Educational Measurement: Issues and Practice. 2012; 31: 2–9.

[4] Yeates P, et al. 'You're certainly relatively competent': Assessor bias due to recent experiences. Medical Education 2013 47(9): 910-922.

[5] Wood T. Exploring the role of first impressions in rater-based assessments Advances in Health Sciences Education August 2014, Volume 19, Issue 3, pp 409–427.

[6] 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.

[7] 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.

[8] An Aid to the MRCP PACES: Volume 1: Stations 1 and 3. Robert E. J. Ryder, M. Afzal Mir, E. Anne Freeman. John Wiley & Sons.

Considerable effort has been placed into reducing sources of examiner variability in clinical examinations in the past 30 years, including examiner training/calibration with limited success[9] [10].

Gingerich et al concluded:

> It is with good intentions that steps have been taken to make rater-based assessments more consistent through increasingly structured dimensional assessment tools. Changes to rating scales, assessment procedures, and rater training have been based on solid reasoning and rigorous study. It is important to have psychometrically sound assessments that are defensible, useful, and meaningful. But the outcomes from this dedicated work have not entirely met expectations.[11]

Checklists, as part of the scoring approach, have been used widely to increase objectivity, however they have been criticised for rewarding thoroughness over expertise and have not been shown to be superior to rating scales as a method of measuring the candidate's ability[12] [13]. In addition, increasing the mental workload of examiners by using more detailed checklists may be counter-productive.[14]

More recently, researchers have become interested in unravelling the issues surrounding examiner judgment by looking at the ways that examiners think during their assessment process (rater cognition)[15]:

> "Rater cognition has become an important area of inquiry in the medical education assessment literature generally, and in the OSCE literature specifically, because of concerns about potential compromises of validity." [16]

[9] 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.

[10] Eckes T. Introduction to many-facet rasch measurement: analysing and evaluating rater-mediated assessments. 2011, Frankfurt am Main: Internationaler Verlag der Wissenschaften.

[11] Gingerich A, Eva KW. Rater-based assessments as social judgments: Rethinking the etiology of rater errors. 2011. Academic Medicine*,* 86, S1–S7.

[12] Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. Academic Medicine. 1999; 74(10):1129-1134.

[13] RegehrG, Reznick R K, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Academic Medicine. 1998 73(9), 993–997.

[14] Tavares W, Eva KW. Exploring the impact of mental workload on rater-based assessments. Adv Health Sci Educ Theory Pract 2013;18 (2):291–303.

[15] Berendonk C, Stalmeijer RE, Schuwirth LWT. Expertise in performance assessment: assessors' perspectives. Advances in Health Science Education*,* 2012 18, 559–571.

[16] Chahine S, Holmes B, Kowalewski Z. In the minds of OSCE examiners: uncovering hidden assumptions**.** Adv Health Sci Educ Theory Pract. 2016 Aug;21(3):609-25.

There is now an increasing literature surrounding perspectives on examiner cognition that is beyond the scope of this review [17]. Delandshere and Petrosky sum up thus:

> 'Judges' values, experiences, and interests are what makes them capable of interpreting complex performances, but it will never be possible to eliminate those attributes that make them different, even with extensive training and "calibration". [18]

Therefore, variations in assessor judgments may very well represent variations in the way performance can be understood, experienced and interpreted[19]. This phenomenon is not yet fully understood or explained.

The present review is therefore situated in an emerging environment of research into examiner or rater cognition and its effects on judgment.

Focus of the review

This desktop review studies concerning the outcomes of high-stakes postgraduate specialty examinations using the OSCE format. It focuses specifically on two main questions:

- What is known about the success rates of international medical graduates (IMGs) compared to non-IMG candidates?
- Is there any literature investigating the presence of specific examiner bias for or against international medical graduates (IMGs) compared to non-IMG candidates when forming a judgement on the ability of an IMG candidate?

Data sources for the review

Data sources included a professional librarian search using MeSH terms and text word searches of online databases Education Research Complete + ERIC, Scopus, WOS and PsycInfo for maximal retrieval (Appendix 1 shows the full search terms); academic searching in PubMed, high citation medical journals such as the BMJ, Google and Google Scholar; backwards and forwards citations from sourced literature; specific searches of high citation medical education journals, websites, online legal documents, commissioned reviews, personal opinions and/or letters to the Editor, personal knowledge and medical education conference abstracts.

[17] Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014a). Seeing the "black box" differently: Assessor cognition from three research perspectives. Medical Education 2014: 48: 1055–1068.

[18] Delandshere G, Petrosky AR. Capturing teachers' knowledge: performance assessment a) and post-structuralist epistemology, b) from a post-structuralist perspective, c) and post-structuralism, d) none of the above. Educ Res1994;**23** (5):11–8.

[19] Govaerts MJB, Schuwirth L, Van der Vleuten CP, Muijtjens AMM. Workplace-based assessment: Effects of rater expertise. Adv Health Sci Educ Theory Pract. 2011;16:151–165.

All relevant publication dates were included. Greatest effort was placed into finding studies from the UK and countries with comparable medical education systems (USA, Canada, New Zealand, Australia). All papers were published in the English language. Studies using any reputable methodology singly or in combination were considered. Postgraduate high-stakes specialty examinations were given the highest priority. Specifically, examinations in emergency medicine were sought, however only one paper provided any data relating to emergency medicine differential attainment.

Disciplines outside medicine were excluded from the review.

Research ethics

The research for this review was desktop-based only and ethical permission was not required.

Success rates and examiner bias in the testing of international medical graduates on high-stakes clinical examinations

International studies – North America

MacLellan et al investigated the success rates of IMGs in Quebec on the family medicine certification (specialist) clinical examinations from 2001 to 2008. Success rates were significantly lower than those of Canadian trained graduates. Between 2001 and 2008, the average success rate for IMGs was 56%. Quebec medical school graduates who completed the same family medicine residency programs had an average success rate of 93.5% on the same certification examinations during this period.[20] The authors pointed out that the IMGs in this study have already passed several screening competency examinations and have successfully completed a 2-year accredited family medicine residency program. The authors make a number of hypotheses concerning IMGs' poorer performance. They suggest it could be related to how and when IMGs learn to translate their knowledge and integrate it with clinical decision-making or to the diversity of IMGs as a group and the variability of their undergraduate training experiences.

When candidates' performance on the different components of the OSCE was analysed, the authors reported that IMGs performed poorly in multiple aspects of clinical skills including history taking, investigation, differential diagnosis, and treatment. When standardized scores were used to compare IMGs' and Canadian graduates' performances, the IMGs' *z* scores were approximately 1.5 to 2 standard deviations lower than Canadian graduate scores on the same skills. The IMGs' mean aggregate global scores were 15% to 20% lower than CMGs' scores.

[20] MacLellan A, Brailovsky C, Rainsberry P, Bowmer I, Desrochers M. Examination outcomes for international medical graduates pursuing or completing family medicine residency training in Quebec Canadian Family Physician Sep 2010, 56 (9) 912-918.

Schabort et al have pointed out the substantially higher rates of failure on college certification examinations for IMGs when compared with graduates from North American medical schools. In Canada, for instance, retrospective review of the Royal College of Physicians and Surgeons of Canada (RCPSC) and College of Family Physicians of Canada (CFPC) examination outcomes has identified that IMGs have substantial difficulty achieving certification. For example, in 2007, only 66% of IMG residents were successful on the CFPC examination, compared with 90% of their Canadian medical graduate counterparts. The overall success rate for IMGs on the CFPC Certification examination rose to 74% in 2008 but dropped to 64% and 51% in 2009 and 2010, respectively[21].

The problem is equally troubling among IMG residents in other specialties: between 2005 and 2010, only 75% of IMG residents passed the Royal College of Physicians and Surgeons of Canada (RCPSC) examination, while 96% of Canadian medical graduate residents demonstrated the required performance for certification.[22]

In their study, Schabort et al found that first language, birthplace, country of medical training, and previous professional experience were among the most important factors for predicting IMG success. The effects described for first language and birthplace were counter to what might be expected. IMGs who spoke English as a first language and IMG Canadians who studied abroad performed worse on elements of the certification examinations than their counterparts did, although the effects were small.

The effects of previous professional experience were discipline specific, with it being a negative predictor for family medicine IMG success and a positive predictor for Royal College certification success. They concluded that more research is required to improve understanding of the low 'IMG certification success phenomenon', help residency programs identify at-risk residents, and underpin the development of specific educational and remedial interventions to assist IMGs to be more successful.

Norcini et al have conducted a major study of Specialty board certification (equivalent to the Fellowship examinations of the specialist colleges in Australia) among United States citizen graduates and non-United States citizen graduates of international medical schools[23].

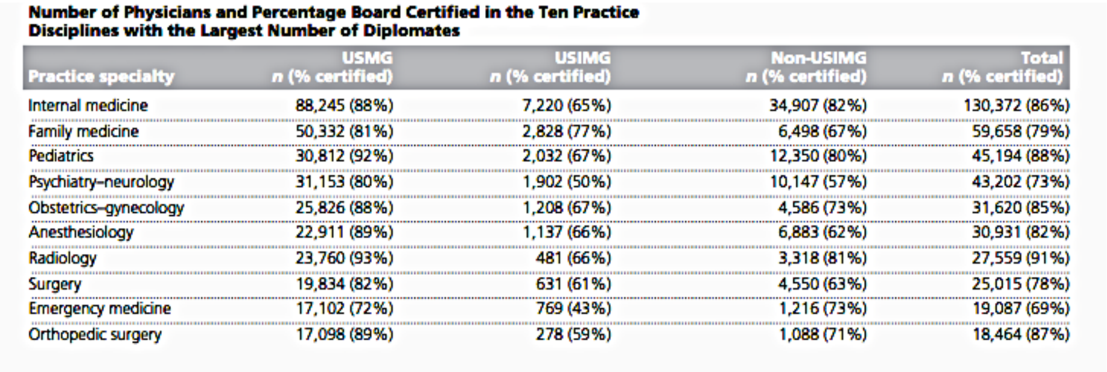
[21] Schabort I, Mercuri M, Grierson LEM. Predicting international medical graduate success on college certification examinations: Responding to the Thomson and Cohl judicial report on IMG selection. Canadian Family Physician. 2014;60(10):e478-e484.

[22] Boulet JR, Swanson D, Cooper R, Norcinii J, McKinley D. A comparison of the characteristics and examination performances of US and non-US citizen international medical graduates who sought Educational Commission for Foreign Medical Graduates certification. Acad Med 2006;81(10 Suppl):S116-9.

[23] Norcini JJ et al. Specialty Board Certification Among U.S. Citizen and non-U.S. Citizen Graduates of International Medical Schools. 2005 Acad Med 80 (10 Suppl), S42-S45.

Table 1 in their paper shows the number of positions and the percentage that have the board certification examinations in the 10 practice disciplines with the largest numbers of diplomats.

Table 1: Number of physicians and percentage of board certified in the 10 practice disciplines with the largest number of diplomats. Reproduced from Norcini et al.



Number of Physicians and Percentage Board Certified in the Ten Practice Disciplines with the Largest Number of Diplomates

Practice specialty	USMG *n* (% certified)	USIMG *n* (% certified)	Non-USIMG *n* (% certified)	Total *n* (% certified)
Internal medicine	88,245 (88%)	7,220 (65%)	34,907 (82%)	130,372 (86%)
Family medicine	50,332 (81%)	2,828 (77%)	6,498 (67%)	59,658 (79%)
Pediatrics	30,812 (92%)	2,032 (67%)	12,350 (80%)	45,194 (88%)
Psychiatry–neurology	31,153 (80%)	1,902 (50%)	10,147 (57%)	43,202 (73%)
Obstetrics–gynecology	25,826 (88%)	1,208 (67%)	4,586 (73%)	31,620 (85%)
Anesthesiology	22,911 (89%)	1,137 (66%)	6,883 (62%)	30,931 (82%)
Radiology	23,760 (93%)	481 (66%)	3,318 (81%)	27,559 (91%)
Surgery	19,834 (82%)	631 (61%)	4,550 (63%)	25,015 (78%)
Emergency medicine	17,102 (72%)	769 (43%)	1,216 (73%)	19,087 (69%)
Orthopedic surgery	17,098 (89%)	278 (59%)	1,088 (71%)	18,464 (87%)

Of most interest are the figures for emergency medicine. Of 19,087 doctors qualified as emergency medicine physicians, the percentage certified overall was the lowest at 69%. Within those certified, over 17,000 doctors were United States medical graduates (USMG) and 72% of these doctors were certified. 769 doctors were United States citizens trained internationally (USIMG) of whom 43% were certified, and 1216 were non-United States citizens trained internationally (non-USIMG), of whom 73% were certified. This shows that the certification rates of USMG and non-USIMG in emergency medicine were similar although the numbers were small in the latter sample. The study does not refer however to the number of times that various groups undertook the examination before being successful, investigate any further candidate demographics or attribute any factors concerning lack of certification. In other specialties, some showed lower Board certification rates especially for USIMGs.

The authors concluded that:

> "It is particularly noteworthy that among recent graduates, non-USIMGs have certification rates that are comparable to USMGs. The exact reasons for this are unclear and among areas meriting further study are the possibility of higher standards for entry and selection to residency, higher motivation among non-USIMGs to achieve certification, or a decline in the ability of the USMGs."

McKendry et al have investigated three specialties in the Royal College of Physicians Surgeons of Canada (RCPSC) certifying examinations (written and oral). These three specialties had a significantly lower written exam pass rate for candidates training in small compared to large programs (neurology, neurosurgery and community medicine). By amalgamating results from 10 specialties, they showed that candidates from small programs (three or fewer residents had lower pass rates (11%) on written examinations compared to candidates in larger programs (10 or more residents).

However, there was no effect of training program size on the pass rate for the oral component of the examinations. The authors highlight the importance of a critical mass of trainees in postgraduate training programs[24].

Andrew reported a study comparing IMGs with Canadian medical school graduates in a general practice training program (known as a family medicine residency program in Canada)[25]. The in training evaluation reports (ITER) and the certification (specialty) examination results for two cohorts of IMG and Canadian trained graduates were examined between the years 2006 and 2008 in British Columbia only. Very small numbers limited the study, however in terms of in training evaluations by supervisors, the figures were similar. Canadian trained residents had 99% of in training evaluations designated as meeting or exceeding expectations compared with 97.6% in the IMG group. When it came to the examination however, only 58% (7 of 12) of the IMG candidates passed the examination compared with 95% (59 of 62) of the Canadian family practice residents. The authors concluded that further research was required to elucidate this differential.

International studies – United Kingdom

In Britain, a difference between candidates' success in undergraduate and postgraduate high-stakes examinations by ethnic background and gender has been reported over the past two decades.[26] [27] [28] [29] [30] [31] [3233].

A 2002 systematic review of the factors influencing medical school success found evidence of underperformance in minority ethnic candidates. This is only 1 of 2

[24] McKendry RJR, Dale P. Does the number of trainees in a postgraduate training program influence the pass rates on certifying examinations? Clin & Investigative Med. 1995. 18:1, 73-79.

[25] Andrew RF. How do IMGs compare with Canadian medical school graduates in a family practice residency program? Canadian Fam Phys. 2010 September; 56 (9): e318-322.

[26] Tyrer SP, Leung W-C, Smalls J, Katona C. The relationship between medical school of training, age, gender and success in the MRCPsych examinations. Psychiatr Bull 2002;26:257-63.

[27] Hurst NG, McManus IC, Mollon J, Dacre JE, Vale JA. Performance in the MRCP(UK) Examination 2003–4: Analysis of pass rates of UK graduates in the Clinical Examination in relation to self-reported ethnicity and gender. BMC Medicine 2007, 5:8.

[28] Wakeford R. International medical graduates' relative under-performance in the MRCGP AKT and CSA examinations. Educ Prim Care 2012;23:148-52.

[29] Wakeford R, Farooqi A, Rashid A, Southgate L. Does the MRCGP examination discriminate against Asian doctors? BMJ 1992;305:92-4.

[30] Esmail A. Ethnicity and academic performance in the UK BMJ 2011; 342 :d709

[31] Woolf, K., et al. Exploring the underperformance of male and minority ethnic medical students in first year clinical examinations. Advances in Health Sciences Education 2008 13(5): 607-616.

[32] Schleicher I et al. Examiner Effect on the Objective Structured Clinical Exam - A Study at Five Medical Schools. 2017 BMC Med Educ 17 (1), 71.

[33] Woolf K, Potts HWW, McManus IC. The relationship between ethnicity and academic performance in UK-trained doctors and medical students: a systematic review and meta-analysis. Brit Med J. 2011;342.

systematic reviews found in this literature review, but was limited in its applicability to the review questions, and so is not considered further.[34]

Although conducted in final examinations at exit from medical school, the study by Wass et al is the only one found in this literature review, concerned with the effect of ethnicity on student performance in OSCE stations, that investigated for any form of discrimination in video recordings of the stations, and concluded that no examples of overt discrimination were found in 309 recordings. [35]

The more recent and increasingly more sophisticated and rigorous postgraduate literature is now reviewed in this paper as pertinent to the research questions.

In 2006, Bessant et al reported a study of factors that may predict success of candidates taking a revision course in preparation for the MRCP (UK) PACES (practical assessment of clinical examination skills) examination[36].

The authors administered a questionnaire survey of candidates attending a PACES revision course prior to sitting the examination. Results were correlated with subsequent pass lists published by the College of Physicians.

Candidates attending courses in 2002 were surveyed and 523 candidates completed questionnaires, evenly balanced between UK and overseas graduates.

Overall, 483 candidates took the examination immediately after the course, and 219 just less than half (45.3%) passed. Results showed that UK graduates were much more likely to pass (67.0%) than overseas graduates (26.2%) (p = 0.003, odds ratio [OR] 5.72). For UK graduates, pass rates were also higher for white candidates (73%) than for ethnic minorities (56%) (p = 0.012, OR 2.15) and for those who passed at the first attempt in the MRCP (UK) part 2 written paper (p = 0.003, OR 2.90).

For IMGs, those who had been qualified for less than eight years were significantly more likely to pass (p = 0.001, OR 2.78). More overseas (45.7%) than UK (30.8%) graduates were confident that they would pass, but unfortunately confidence did not predict success.

In summary, among candidates taking a revision course, UK graduates were more likely to pass the PACES examination than non-UK graduates. Ethnic minority UK graduates seem to have a significantly poorer success rate, although the authors concluded that this finding required confirmation.

[34] Ferguson E, James D, Madeley L. Factors associated with success in medical school: systematic review of the literature. BMJ 2002;324:952-7.

[35] Wass, V, et al. (2003). Effect of ethnicity on performance in a final objective structured clinical examination: qualitative and quantitative study. British Medical Journal 326(7393): 800-803.

[36] Bessant R, Bessant D, Chesser A, Coakley G. Analysis of predictors of success in the MRCP (UK) PACES examination in candidates attending a revision course. Postgrad Med J 2006;82:145-9.

Watmough and Bowhay summarised the performance of graduates by country of primary medical qualification in part one (multiple choice knowledge test) of the UK Royal College of Anaesthetists (RCA) examination from 1999 to 2008. The analysis showed that candidates from the UK, Australia, New Zealand, South Africa and Zimbabwe performed better than those from Egypt, Iraq, Ireland or Pakistan. The authors concluded that some graduates may require additional support prior to undertaking these examinations[37].

Woolf and colleagues have conducted a major systematic review and meta-analysis investigating whether the ethnicity of UK trained doctors and medical students is related to their academic performance[38]. This follows a PhD study conducted by Woolf concerning undergraduate medical students.[39]

Studies that were selected included quantitative reports that measured the performance of medical students or UK trained doctors from different ethnic groups in undergraduate or postgraduate assessments of any kind including written tests. Exclusions were non-UK assessments, only non-UK trained candidates, only self-reported assessment data, only dropouts or another non-academic variable, obvious sampling bias, or insufficient details of ethnicity or outcomes.

In all, 23 meta-analyses of effect sizes could be calculated from 22 reports (n=23 742) and these mainly indicated medium effect sizes. Candidates of "non-white" ethnicity underperformed compared with white candidates (Cohen's d=−0.42, 95% confidence interval −0.50 to −0.34; P<0.001). Effects in the same direction and of similar magnitude were found in meta-analyses of all types of examinations; undergraduate assessments only, postgraduate assessments only, machine marked written assessments only, practical clinical assessments only, assessments with pass/fail outcomes only, assessments with continuous outcomes only, and in a meta-analysis of white *v* Asian candidates only. Heterogeneity was present in all meta-analyses.

The authors conclude that ethnic differences in academic performance are widespread across different medical schools, different types of examination, and in undergraduates and postgraduates. The separate analysis of machine marked written assessments and practical assessments permitted possible effects of examiner bias and verbal communication skills on ethnic differences in attainment in the clinical exams to be investigated. The authors stated:

[37] Watmough S, Bowhay A. An evaluation of the impact of country of primary medical qualification performance in the UK Royal College of Anaesthetists' examinations. *Med Teach* 2011; 33: 938- 40.

[38] Woolf K, Potts HWW, McManus IC. The relationship between ethnicity and academic performance in UK-trained doctors and medical students: a systematic review and meta-analysis. Brit Med J. 2011;342.

[39] Woolf K. The academic underperformance of medical students from ethnic minorities. [PhD thesis]. University of London, 2009.

> " That an ethnic attainment gap was found in both machine marked and face to face assessments suggests that those factors are unlikely to be primarily responsible, although effects might still be present." (p 9).

Given the size of this review it shows that differential attainment has persisted for many years and cannot be dismissed as ‘atypical or local’ problems. The authors proposed that this as a widespread issue, that probably affects all of UK medical and higher education. The authors advocated for further research to ensure a fair and just method of training and of assessing current and future doctors.

Dewhurst et al[40] examined the effects of ethnicity and gender on pass rates in UK medical graduates sitting the Membership of the Royal Colleges of Physicians in the United Kingdom [MRCP (UK)] Examination in 2003–4.

Pass rates for each part of the examination were analysed for differences between graduate groupings based on self-declared ethnicity (84 to 90% declared) and gender (100%) declarations.

In all three parts of the examination, written and clinical, white candidates performed better than other ethnic groups (P < 0.001). In the MRCP(UK) Part 1 and Part 2 Written Examinations, there was no significant difference in pass rate between male and female graduates, nor was there any interaction between gender and ethnicity. Performance of non-white male trainees was particularly poor across all sections of the examination.

In the Part 2 Clinical Examination (Practical Assessment of Clinical Examination Skills, PACES), which is an OSCE-style station based examination, women again performed significantly better overall than men (P < 0.001). Non-white men performed more poorly than expected, relative to white men or non-white women. This cannot be explained readily in terms of generally poorer communicative ability, as their relative performance on the history taking station was equivalent to that in clinical skills stations. As all candidates in this study graduated in the UK, the command and comprehension of English should not be a factor.

Analysis of individual station marks showed significant interaction between candidate and examiner ethnicity for performance on communication skills (P = 0.011), but not on clinical skills (P = 0.176). Analysis of overall average marks showed no interaction between candidate gender and the number of assessments made by female examiners (P = 0.151). Potential examiner prejudice or bias was significant only in the cases where there were two non-white examiners examining a non-white candidate.

The relative underperformance on the communication skills and ethics station may represent, however, a specific problem of cross-cultural interpretation or

[40] Dewhurst NG, McManus IC, Mollon J, Dacre JE, Vale JA. Performance in the MRCP(UK) Examination 2003–4: Analysis of pass rates of UK graduates in the Clinical Examination in relation to self-reported ethnicity and gender. BMC Medicine 2007, 5:8.

understanding. Clinical examinations generate much interest in examiner fairness. In PACES, individual examiner bias is minimised by using objective rather than subjective criteria ("anchor statements") offering candidates of both sexes equal opportunity to demonstrate competence. Examiners are advised to follow the same line of questioning for each candidate-surrogate interaction minimising any potential for bias in individual encounters.

A review of MRCP(UK) examiner performance has shown non-white examiners to have a higher stringency score (harder markers) but analysis of the joint effect of examiner ethnicity and candidate ethnicity showed a significant interaction. More detailed analysis showed that the effect is primarily occurring in the "talking stations" i.e. communication skills stations, and there is no evidence of interaction on clinical skills stations.

Any simplistic explanation in terms of examiner prejudice can be excluded by this finding, as a systematic bias would also be expected to be evident in clinical skills stations as well. The effect was statistically significant in the communication stations, but only in cases where two non-white examiners meet a non-white candidate. Here the data suggested that the non-white candidate was given a higher score than when compared with candidates seeing white/white or white/ non-white examiner pairs. This might reflect different cultural interpretations of judgements being made, particularly when communication skills and ethics are being assessed. Thus, the authors suggested that when two non-white examiners encounter a non-white candidate, the style of discourse may be more consistent, resulting in an opportunity for inadvertent positive bias.

The authors posited that cause of these differences is most likely to be multifactorial, but cannot be readily explained in terms of previous educational experience or differential performance on particular parts of the examination. Potential examiner prejudice, significant only in the cases where there were two non-white examiners and the candidate was non-white, might indicate different cultural interpretations of the judgements being made, primarily in communication and ethics stations in favour of non-white candidates.

Esmail and Roberts have investigated the difference in failure rates in the postgraduate examinations of the Royal College of General Practitioners (MRCGP) UK by ethnic or national background, and attempted to identify any factors that might be associated with these differences in the clinical skills (OSCE) component of the examinations[41].

In this large study of 5095 candidates, data were examined concerning candidates sitting the written knowledge test and the clinical skills assessment components of the MRCGP examination between November 2010 and November 2012. A further

[41] Esmail A, Roberts C. Academic performance of ethnic minority candidates and discrimination in the MRCGP examinations between 2010 and 2012: analysis of data. The BMJ. 2013;347.

analysis was carried out on 1175 candidates not trained in the United Kingdom, who sat an English language capability test (IELTS) and the Professional and Linguistic Assessment Board (PLAB[42]) examination (as required for full medical registration), controlling for scores on these examinations and relating

The authors reported that, after controlling for age, gender, and level of performance in the written knowledge test, significant differences persisted between white UK graduates and other candidate groups. Black and minority ethnic graduates (BME) even though they were trained in the UK were significantly more likely to fail the clinical skills assessment at their first attempt than their white UK colleagues (odds ratio 3.536 (95% confidence interval 2.701 to 4.629), P<0.001; failure rate 17% v 4.5%). Black and minority ethnic candidates who trained abroad were also more likely to fail the clinical skills assessment than white UK candidates (14.741 (11.397 to 19.065), P<0.001; 65% v 4.5%). For candidates not trained in the UK, black or minority ethnic candidates were more likely to fail than white candidates, but this difference was no longer significant after controlling for scores in the applied knowledge test, IELTS, and PLAB examinations (adjusted odds ratio 1.580 (95% confidence interval 0.878 to 2.845), P=0.127).

The authors postulated that "subjective bias due to racial discrimination" in the clinical skills assessment may be one cause of failure for UK trained candidates and international medical graduates. They suggested that changes to the clinical skills assessment could improve the perception of the examination as being biased against black and minority ethnic candidates. They propose that:

> "…it cannot be ascertained if the standardised patients (played by actors) behaved differently in front of candidates from non-white ethnic groups. Nor can we confidently exclude bias from the examiners in the way that they assessed non-white candidates. (*Mitigating factors include that*) there is mandatory training of RCGP examiners in equality and diversity issues, and there is training and monitoring of the actors to ensure consistency in the presentation of the cases. There is also a well developed programme of continuing training and feedback to examiners of their performance."

They suggest examining the diversity of examiners and trained simulated patients, the type of cases included in the examination and the feedback given to candidates as areas for possible improvement. They recommended that the RCGP should investigate how both standardised patients and examiners of black and minority ethnic origin would score candidate physicians who are racially and ethnically concordant and compare that to how non-concordant standardised patients and examiners score candidates of black and minority ethnic origin.

They also postulated that examination success could be affected by extraneous factors, for example, in training experience and in other cultural factors between

[42] This is an OSCE style examination similar to the AMC clinical examination required for general registration.

candidates trained in the UK and abroad. They finally recommended that consideration should be given to strengthening postgraduate training for international medical graduates.

An organization representing some ethnic minority doctors (BAPIO: the British Association of Physicians of Indian Origin) brought a claim to court that the Royal College of General practitioners in the UK (RCGP) was unlawfully discriminating against Black and Minority Ethnic (BME) doctors in their specialty exit clinical OSCE, known as the clinical skills assessment (CSA) both directly and indirectly. A judicial review was conducted, and the claim was unsuccessful, the judge concluding that there was:

> "no basis for contending that the small number who fail [the CSA] ultimately do so for any reason apart from their own shortcomings as prospective general practitioners" [43] .

Mr JUSTICE Mitting also pointed out that on the evidence presented:

> "In summary, the extensive research undertaken so far has identified the problem of differential outcomes which are only partly explicable by known factors and produced tentative suggestions for making alterations: within the competence of the Royal College, the encouragement of and cooperation with the Deaneries to educate candidates in the requirements of the Clinical Skills Assessment and an effort to secure a more representative profile of examiner." (para 24).

The General Medical Council (GMC) commissioned a review by Esmail and Roberts in 2013, evidence from which formed part of the information presented in the judicial review[44]. The clinical skills assessment (CSA) results of the MRCGP examination were studied as part of the review. The results showed significant differences in failure rate between different groups in the examination.

Even after controlling for age, gender and performance at the prior knowledge test, significant differences persisted between white UK graduates and BME UK graduates. BME UK graduates were nearly four times more likely to fail the CSA examination at their first attempt than their white UK colleagues (OR = 3.536, c.i [confidence interval] 2.701-4.629, p= <0.001). BME IMG candidates were nearly

[43] The Queen on the application of Bapio Action Ltd [Claimant] v Royal College of General Practitioners [First Defendant] and General Medical Council [Second Defendant], in the High Court of Justice, Queen's Bench Division, The Administrative Court. 10th April 2014. EWHC 1416 (Admin) 2014, information available at http://lexisweb.co.uk/cases/2014/april/r-on-the-application-of-bapio-action-limited-v-royal-college-of-general-practitioners-and-another. Accessed August 2017.

[44] Esmail and Roberts. At: http://www.gmcuk.org/MRCGP_Final_Report__18th_September_2013.pdf_53516840.pdf accessed July 2017.

fifteen times more than likely to fail this exam than their white UK colleagues (OR= 14.741, c.i. 11.397-19.065, p=<0.001).

In explanation, the authors suggested that it was the preparedness of UK graduates that may have accounted for the differences. Significant differences in the raw pass rates by training location were noted, suggesting that the place of training was relevant to success. This discrepancy has also been shown in US Family Medicine postgraduate examinations[45]. Patterson et al have also pointed to an anomaly in the UK settings:

> "In practice, the combination of selection and training placement systems often operate against the interests of the weaker recruits (that is, those candidates performing least well at selection are assigned to the least popular training placements, thereby encouraging a cycle of educational deprivation)[46]." (p713).

Esmail and Roberts also noted that the diversity of examiner background was not fully representative of general practitioners in the UK despite efforts to improve this but that examiners received cultural diversity training. Finally they discussed the issue of limited feedback given to candidates who fail. They recommended that improved mechanisms for providing formative feedback to failed candidates should be developed. However they also found in their extensive analyses that differences between white and BME UK graduates disappear on the second attempt on the CSA, and also reduce for non-UK trained IMGs. This may reflect on feedback provided and better preparation, also self-directed learning from the examination experience despite limited feedback[47].

Following the judicial review, Wakeford et al published a major British study concerning the difference in performance in specialty examinations for two different colleges (physicians and general practitioners). The focus of the study was the difference between white candidates and black minority ethnic (BME) candidates. The main difference in this study compared to others is that both knowledge tests and clinical tests were examined, and these also were compared for two entirely independent testing organisations.

MRCGP and MRCP(UK) are the main entry qualifications for UK doctors entering the specialties of general practice or hospital [general internal] medicine. The authors were able to compare the performance of MRCP(UK) candidates who had subsequently taken the MRCGP examinations. Both examinations consist of machine marked knowledge tests and clinical skills assessments. Information on performance

[45] Falcone JL, Middleton DB. Performance on the American Board of Family Medicine Certification examination by country of medical training. *J Am Board Fam Med* 2013; 26: 78-81.

[46] Patterson F, Denney ML, Wakeford R, Good D. Fair and equal assessment in postgraduate training? A future research agenda. Br J Gen Pract 2011; 6 (593):712-713.

[47] Haider I, et al. Perceptions of final professional MBBS students and their examiners about objective structured clinical examination (OSCE): A combined examiner and examinee survey. Journal of Medical Sciences (Peshawar) 2016 24(4): 206-211.

on all these tests were included in the study of 2284 candidates who had taken one or more parts of both assessments, MRCP(UK) typically being taken 3.7 years before MRCGP.

The authors identified analysed performance on written knowledge-based multiple-choice tests (MCQs) in the MRCP(UK) Parts 1 and 2 and the similar MRCGP written Applied Knowledge Test (AKT)) and clinical examinations (MRCGP Clinical Skills Assessment (CSA) and MRCP(UK) Practical Assessment of Clinical Skills (PACES).

The authors found that correlations of attainment between MRCGP and MRCP(UK) were high, disattenuated correlations for MRCGP AKT with MRCP (UK) Parts 1 and 2 being 0.748 and 0.698, and for CSA and PACES being 0.636.

Overall, BME candidates performed significantly lower on all five assessments (P < .001). The authors concluded that the high correlations between MRCGP and MRCP(UK) written and clinical tests support the validity of each, suggesting they assess knowledge cognate to both assessments.

The detailed analyses by candidate ethnicity show that although white candidates out-performed BME candidates, the differences were largely mirrored across the two different sets of examinations. Whilst the reason for the differential performance is unclear, the authors concluded that:

> " …the similarity of the effects in independent knowledge and clinical examinations in different specialty colleges suggests the differences are unlikely to result from specific features of either assessment and most likely represent true differences in ability." (p 1).

Their study shows that there is a negative ethnicity effect at each stage of both independent examinations; BME candidates performing less well even after taking performance at previous stages into account. The authors considered that these effects are therefore unlikely to be due to particular features of any one assessment, component of an assessment or style of assessment. The finding that similar effects are found on written tests suggest that these effects cannot be explained simply by bias on the part of clinical examiners.

McManus et al[48] have assessed gender and ethnic bias in over 2000 examiners who had taken part in the clinical skills examinations known as PACES and nPACES (new PACES) examinations of the MRCP(UK).

In these examinations there are two examiners at each station who mark candidates independently. This has provided an important opportunity for the study of examiner variation, in this case by gender and ethnicity of candidate. As both

[48] McManus I, Elder A, Dacre J. Investigating possible ethnicity and sex bias in clinical examiners: an analysis of data from the MRCP(UK) PACES and nPACES examinations BMC Medical Education 2013, 13:103

examiners of the exactly the same interaction, any difference in the interpretation of the performance of the candidate is entirely due to the examiners' views.

Differences between examiners may result from bias or unreliability on the part of the examiners. By comparing each examiner against a 'basket' of all of their co-examiners, it is possible to identify examiners whose behaviour is anomalous. The method assessed the well-known effect known as stringency - leniency bias (referred to in this analysis as hawkishness-doveishness), gender bias, ethnic bias and, as a control condition to assess the statistical method, 'even-number bias' (i.e. treating candidates with odd and even exam numbers differently). Significance levels were Bonferroni corrected because of the large number of examiners being considered. This is a proper statistical correction that is used to avoid over-inflation of the results.

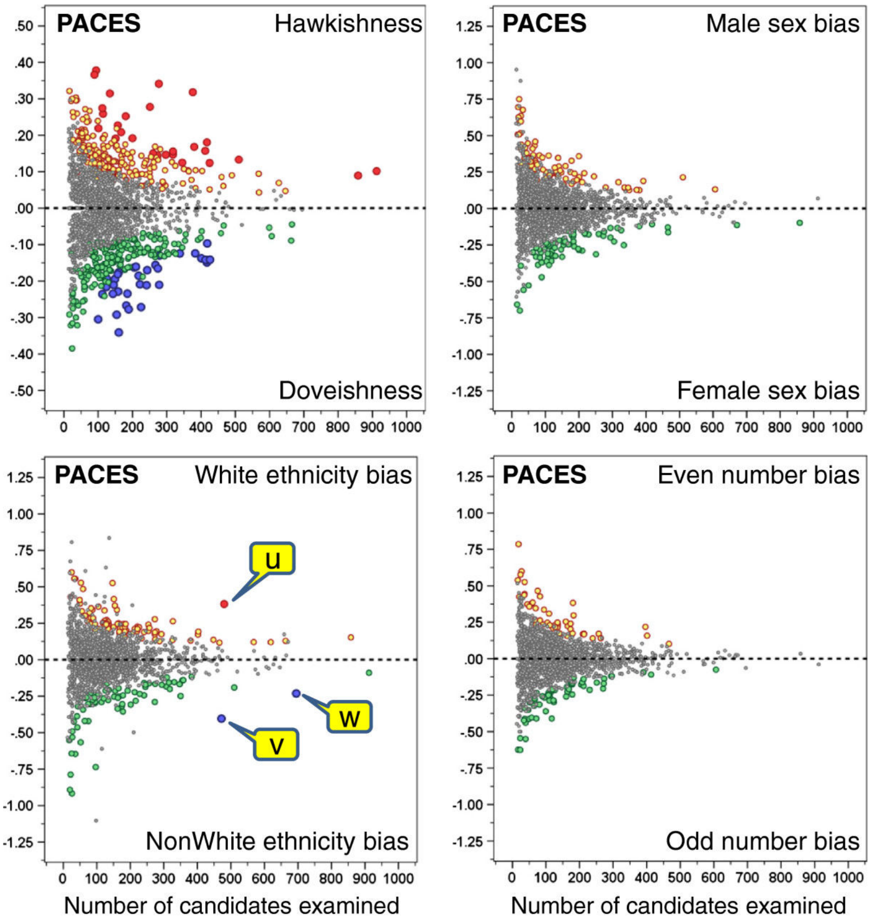
This substantial study used a very large data set, being 26 of PACES and six examinations of nPACES. Data were examined statistically to assess the extent of stringency (hawkishness), as well as gender bias and ethnicity bias in individual examiners.

As in previous studies[49] some examiners were more lenient or more stringent relative to their peers. Importantly for this review, in over 2000 examiners no examiners showed significant gender bias, and only a single examiner showed evidence consistent with ethnic bias.

The following figure is included as it clearly demonstrates the results.

Figure 1. Analysis of examiner biases for the 26 PACES examinations. Reproduced from McManus et al.

[49] 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 modelling. BMC Medical Education. 2006, 6: 42-10.1186/1472-6920-6-42.



The individual graphs show for PACES examinations 1–26 the indices for hawkishness (top left), sex bias (top right), ethnic bias (lower left), and even-number bias (lower right). Each point represents an individual examiner, plotted against the number of candidates examined, and with the significance indicated (grey, NS; orange and green p < .05 uncorrected; red and blue, p < .05 Bonferroni corrected).

The ethnic bias data in the above Figure show that for the 26 PACES examinations there were three examiners, labelled *u*, *v* and *w*, who were significant on the Bonferroni corrected criterion with p < .05. One examiner (*u*) is a white examiner in favour of white candidates and two examiners (*v* and *w*) are non-white examiners in favour of non-white candidates; *w* would not however reach significance on a stricter (p < .001) criterion. When this analysis was repeated for the new clinical examination (nPACES) only one of these examiners (*v*) reached the Bonferroni-corrected p < .05 significance level in both PACES and nPACES.

This examiner was non-white and appeared to be systematically awarding relatively higher marks to non-white candidates. Therefore, in the new version of the clinical examination, the marking of only one examiner (non-white) was consistent with bias towards non-white candidates.

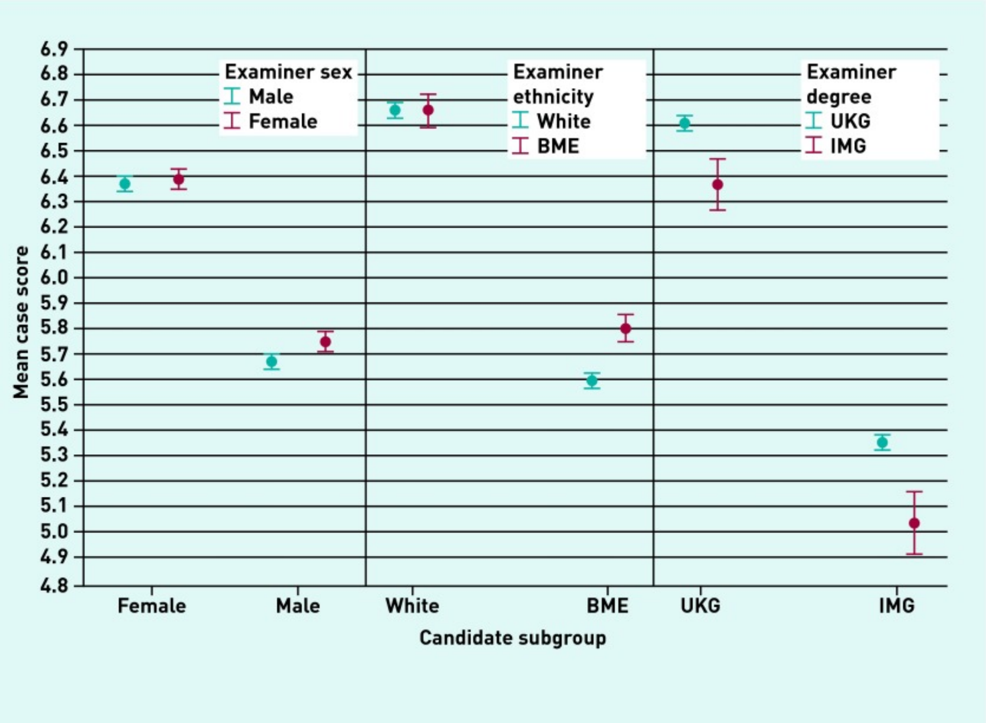
In examinations where there are two independent examiners at a station, the method employed in this study can assess the extent of bias against candidates with particular characteristics. Importantly, the method would be far less sensitive in examinations with only a single examiner per station as examiner variance would be confounded with candidate performance variance. Equally, the method would not be sensitive in examinations where examiners are required to achieve a consensus.

Denney et al published a study of examiner like for like bias by gender, ethnicity and degree source. [50] Data on 4000 candidates (52 000 cases) sitting the MRCGP clinical skills assessment (CSA) in 2011–2012 were obtained. This examination is an OSCE comprising 13 stations with a single examiner. The basis of the investigation was, 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.

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.

The Figure reproduced below shows the relationship of mean station scores of candidates by their three principal demographics (sex, ethnicity, source of primary medical degree) by parallel examiner demographics.

Figure 2. Mean station scores of candidates by their three principal demographics (sex, ethnicity, source of primary medical degree) by parallel examiner demographics. Reproduced from Denney et al.



[50] Denney ML, Freeman A, Wakeford R. MRCGP clinical skills assessment: are the examiners biased, favouring their own by sex, ethnicity and degree source? Br J Gen Pract. 2013;63:e718–25.

The mean effect size of the various differences shown is as follows, in order of size. (effect on station score (maximum marks per station is 9 marks), and effect size (%):

- IMGs receive a higher mark from UKG examiners (green) than from IMG examiners (mean difference = 0.32 marks, 3.6%) *P*<0.001.

- UKGs receive a higher mark from UKG examiners (green) than from IMG examiners (mean difference = 0.24 marks, 2.7%) *P*<0.001.

- BME candidates receive a higher mark from a BME examiner (red) 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 authors also noted the importance of the variables being confounded. For example, ethnicity may be confounded with source of medical degree, hence a multivariate analysis was also performed. 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.

Of all possible interactions, only one was significant at *P*<0.05 (Bonferroni corrected): examiner ethnicity by examiner gender. Male examiners gave similar grades, whether the examiner was 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%).

In this study, interactions between candidates and examiner demographics 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 or crude effects” found in this study, regarding any individual candidate situation, 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.

The authors postulated that the crude effects demonstrate 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 however 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, the authors propose that 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.

This study is limited by the need to group all candidates and examiners from different countries in a single "BME" group for the purposes of analysis. However it has been shown that candidate BME subgroups do not comprise a performance-homogenous whole, in the MRCGP clinical skills examination (p 29)[51].

Denny and Wakeford have also examined the influence of the role player (simulated patient) in the MRCGP clinical skills assessment (CSA). [52] This study investigated the contribution of role-players to and a possible systematic unfairness in the assessment. Using multiple linear regression, data from all 52,702 case scores from the MRCGP CSA for the academic year 2012-2013 were analysed. Candidate data were dichotomised by sex, by ethnicity and by source of primary medical qualification (PMQ); role-players were dichotomised by gender and binary ethnicity; and the transaction of candidate/roleplayer encounters were classified as 'same' or 'different' in terms of the two parties' gender and of their ethnicity. Neither examiner nor role-player characteristics were found to predict any statistically significant portion of case score variance. The most significant ($p < .001$) predictors were source of PMQ (UK or elsewhere: 11% of case score variance), candidates' ethnicity (1%), and candidates' gender (0.6%). This study did not demonstrate any substantial degree of support for the proposition that role-player subgroups systematically influence candidate subgroups' scores.

Richens et al have studied the UK and Ireland intercollegiate specialty board fellowship examinations, developed and conducted in the UK and Ireland by the intercollegiate boards of the Royal College of Surgeons UK and Ireland. [53] They explored effects of gender, ethnic origin, first language, and training status on scores in these examinations across the computer-marked written section and in the face-to-face oral and clinical section. 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).

[51] Wakeford R. MRCGP statistics 2011–12: Annual Report on the AKT and CSA Assessments. 2012. At 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 July 2017).

[52] Denney, M. Wakeford R. "Do role-players affect the outcome of a high-stakes postgraduate OSCE, in terms of candidate sex or ethnicity? Results from an analysis of the 52,702 anonymised case scores from one year of the MRCGP clinical skills assessment." Education for Primary Care 2016 27(1): 39-43.

[53] Richens, D., et al. (2016). "Racial and Gender Influences on Pass Rates for the UK and Ireland Specialty Board Examinations." Journal of Surgical Education **73**(1): 143-150.

The study found that 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).

The authors concluded that 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 two sections of the examination, suggesting that the differential attainment by these factors cannot be attributed to examiner bias in a face-to-face examination.

National studies - Australia and New Zealand

Jasper et al have reported on the different pass rates of the Royal Australian College of General Practitioners (RACGP) Fellowship examination from 1999 to 2004, for various candidate routes to the examination including local graduates and IMGs who have undertaken prescribed general practice postgraduate training programs across Australia[54]. The following table is taken from their paper.

While the figures and any tests of significance were not reported, the table clearly demonstrates that there was a substantial difference in the pass rates between local graduates and IMGs despite having undertaken the same postgraduate training programs. Similarly, the pass rates of IMGs who had not undertaken the training program (practice eligible route) were lower than Australian graduates who had not undertaken the training program.

Figure 3. Pass rates in the Royal Australian College of General Practitioners Fellowship exam from 1999 to 2004 according to route of entry and IMG status. Reproduced from Jasper et al.

[54] Jasper A, Hinchy J, Atkinson K, Rawlin M. The RACGP Examination--changes from 1999-2004. Aust Fam Physician. 2005 Nov;34(11):967-9.

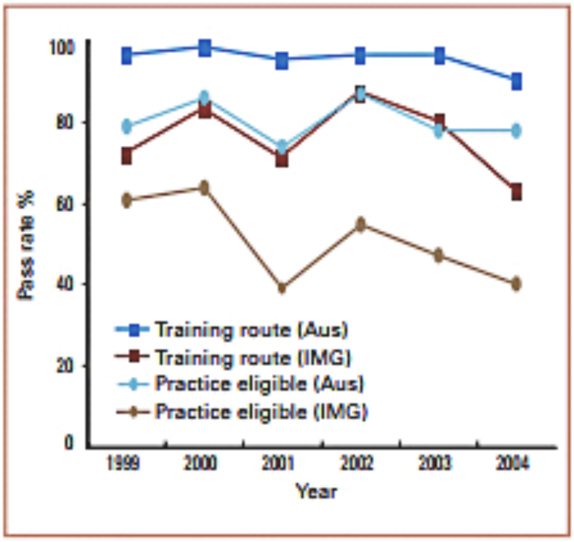


Figure 3. Pass rate for major candidate groups, 1999–2004

Karnik et al have recently published an Australian study relevant to this review that attempted to distinguish international medical graduates by broad categories of country of training [55]. They analysed the performance and predictors of success in the final fellowship examination of the College of Intensive Care Medicine (CICM), and specifically compared the outcomes for international medical graduates attempting the CICM fellowship exam with those of local trainees, defined as those from Australia, New Zealand and Hong Kong (ANZ-HK). They also compared the performance of IMGs from countries with comparable health care systems (CHS) with those from other countries (non-CHS).

Data from six fellowship exam presentations collected prospectively between 2009 and 2011. Pass rates in the final fellowship exam were the index of study. The final examination consists of three parts namely, a written test, a clinical test and a viva. Candidates must exceed a minimum mark on the written test before being invited to undertake the clinical and viva tests. Candidates who had completed an ANZ primary exam were defined as those who had passed CICM, Australian and New Zealand College of Anaesthetists [ANZCA], Australasian College for Emergency Medicine [ACEM], Royal Australian College of Physicians [RACP] or Royal Australasian College of Surgeons [RACS] examinations.

In all, 233 candidates presented to the exam 334 times, and most (73%) were IMGs. ANZ-HK trainees performed significantly better at the exam (79% v 46%, P<0.0001). IMG trainees from CHS performed significantly better than trainees from non-CHS (60% v 40%, P<0.01).

Any candidate completing an ANZ primary exam performed significantly better than non-ANZ primary candidates (74% v 41%, P<0.0001). IMG candidates successful at a

[55] Karnik A, Venkatesh B, Angelico D. Analysis of performance and predictors of success in the final fellowship examination of the College of Intensive Care Medicine [online]. Critical Care and Resuscitation, Vol. 17, No. 1, Mar 2015: 47-50.

postgraduate exam from a CHS country performed significantly better than candidates from a non-CHS country (56% v 34%, P=0.005).

The success rate of IMGs improved to 64% after obtaining an ANZ primary. Candidates taking the exam while working in an intensive care unit had a higher pass rate of 57% compared with 48% of candidates working in non-ICU posts (P=0.23). This was not statistically significant.

The authors concluded that a significant proportion of candidates appearing for the CICM fellowship examination are IMGs. There were differential pass rates of IMGs according to whether their country of graduation was from a comparable health system or not.

In addition, pass rates for trainees who graduated from the ANZ- HK systems had a higher success rate in the fellowship examination. IMGs from a CHS country, or those who completed an ANZ primary also had a much higher success rate compared with other IMGs. There was little difference in pass rates by gender.

Higgins et al conducted an Australian study aimed at understanding the workforce education issues surrounding IMG anaesthetists in rural and remote areas in Australia [56]. As part of this study, they report the differential pass rates of the Australia and New Zealand College of Anaesthetists (ANZCA) IMG candidates compared to local candidates. They also report the results of a small survey of rural and remote IMG anaesthetist candidates.

In terms of differential pass rates, there was a marked difference between IMG candidates and local candidates. The following table is amalgamated from tables two and three in their paper (p 248). Comparative data were provided from May 2006, after changes permitted UK and Irish trained IMG specialists to be exempt from the examination. The average pass rate of IMG doctors, based on these data, was 37.3%.

Table 2. Pass rates of local and IMG candidates in the Australia and New Zealand College of Anaesthetists final examination (extracted from Higgins et al).

Date of examination	Number of local candidates	Pass rate of local candidates	Number of IMG candidates	Pass rate of IMG candidates
May 2007	164	87	21	33
September 2006	84	80	30	50
May 2006	128	90	34	29

[56] Higgins NS, Taraporewalla K, Steyn M, et al. Workforce education issues for international medical graduate specialists in anaesthesia. Aust Health Rev 2010; 34: 246-51.

They also provided information on the pass rate of all IMG candidates for seven examinations from May 2004 to May 2007. The best pass rate was 53% and the lowest pass rate was 29%. The average pass rate was 42.29%.

The authors attributed lower examination performance to a lack of effective study, lack of study time and geographic isolation from other candidates and supervisors. They conclude:

> "…the college Final Examination reports show that the IMG specialist pass rate is considerably lower than local candidates and this is of particular concern.
>
> "… there is a need for further research in the area of distance education-support for IMG specialists and the continuing education needs for this workforce. IMG specialists come from many different backgrounds and training programs in comparison to local trainees. There is also a high variability in anaesthetic training and number of years. In addition, these specialists have had a different exam focus in the past that was more in keeping with the anaesthetic education provided in their own country."

McGrath has also argued that there is no national approach supporting the integration of IMGs into Australian practice.[57]

[57] McGrath BP. Integration of overseas-trained doctors into the Australian medical workforce. Med J Aust 2004; 181: 640-2.

Key findings

Studies have revealed a persisting difference in attainment between local and international medical graduates in respect of high-stakes examinations ranging from medical school to postgraduate specialty examinations. These differences have been reported consistently and repeatedly over several decades in multiple countries and in multiple disciplines with similar patterns including in Australia.

Investigators have reported that these differences are present for both written (such as machine marked multiple-choice questions) and clinical examinations involving expert judgement of examiners.

In postgraduate specialty examinations reported in the world literature, graduates who are citizens of that country and trained in that country perform better than graduates who are trained in other countries. The difference persists even among graduates who are citizens of the country but belong to different ethnic groups, for example British studies have shown that white UK graduates perform better than South Asian or BME UK graduates.

The literature demonstrates that findings of differential attainment cannot be dismissed as atypical or local to any one country or specialty examination.

As scoring is an independent expert judgement, it is theoretically open to conscious or subconscious bias, for example, gender bias, language bias, or other bias such as perceived country of original training or ethnic minority status. Examiner bias is a potential risk in any examination.

Of relevance to this review, very few studies have investigated the potential for examiner bias specifically arising from perceived country of training or ethnicity in high-stakes examinations in relation to the lower success rates of international medical graduates on these examinations when compared to local graduates. Most studies were quantitated, with a focus on testing for bias using actual examination data in large-scale datasets.

A large British comparative study involving over 2000 paired independent examiners in a postgraduate clinical skills test conducted by McManus et al found that no examiners showed significant gender bias, and only a single examiner showed continued evidence consistent with ethnic bias. This examiner was non-white and appeared to be systematically awarding relatively higher marks to non-white candidates.

In the postgraduate study conducted by Dewhurst et al, a significant interaction between candidate and examiner ethnicity was found for performance on communication skills stations in a clinical examination (P = 0.011), but not on clinical skills stations in the same examinations (P = 0.176). Analysis of overall average marks showed no interaction between candidate gender and the number of assessments made by female examiners (P = 0.151). Potential examiner prejudice or bias was significant only in the cases where there were two non-white examiners examining a non-white candidate. Any simplistic explanation in terms of examiner prejudice can be excluded, as a systematic bias would also be expected to be evident in clinical skills stations as well. The effect was statistically significant in the communication stations, in cases where two non-white examiners meet a non-white candidate. Here the data suggested that the non-white candidate was given a higher score than when compared with candidates seeing white/white or white/ non-white examiner pairs.

In the UK, a large study of a postgraduate clinical skills examination conducted by Denney et al attempted to determine whether examiners were biased, 'favouring their own' by gender, ethnicity and degree source. There were four main findings on univariate analyses shown in order of importance (effect on station score (maximum marks per station is 9 marks), and effect 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.

Causes of differential attainment remain unclear and while some effects of interactions by examiner and candidate background have been determined, they appear relatively small in comparison with the size of the differences between groups of local and international graduates around the world. It appears that these persistent differences in outcomes are at present only partly explicable by the factors presented in this review.

Authors have postulated a wide range of factors for further consideration and exploration other than bias in assessment tools as the major determinant of group differences.[58] While beyond the scope of this review, it appears that the problem of differential pass rates is likely multifactorial. Fruitful areas of research appear to consist of examining less tangible educational and social factors and challenges in transition as contributing to examination performance.

This has been recognised as a critical issue in medical education at all levels of training and in higher education generally. It is essential to continue to acknowledge

[58] Patterson F, Denney ML, Wakeford R, Good D. Fair and equal assessment in postgraduate training? A future research agenda. Br J Gen Pract 2011; 61(593):712-713.

and explore this challenging problem, and continue with further research into its underlying causes.

In terms of the findings presented in respect of Specialty College examinations, it seems prudent to engage a diverse group of examiner backgrounds and apply these various examiner groups (eg male/female, Caucasian/non-Caucasian, local graduates/IMGs, older clinicians/younger clinicians) as fairly as possible across OSCE administrations, so that no candidate experiences, for example, a preponderance of examiner backgrounds.

Similarly Colleges should ensure that all examiners and simulated patients have cultural diversity training and robust and regular calibration training. Examiners should receive regular feedback on their performance.

Finally the standard required for a pass in every station should be clearly articulated and understood by all examiners.

Appendix 1: Search Terms for University of Wollongong Library Search

SCOPUS (TITLE-ABS-KEY ((assessor OR examiner* OR observer) W/5 (bias* OR influen* OR discriminat* OR variation OR variance)) AND TITLE-ABS-KEY (exam* OR osce*) AND TITLE-ABS-KEY (racis* OR racial* OR ethnic* OR culture OR international OR foreign) AND TITLE-ABS-KEY ("Medical graduate" OR "medical trainee" OR candidate OR examinee* OR "medical student*")) - 24 results	**WOS** TOPIC: ("observer variation" OR (examiner* AND (bias* OR influen* OR discriminat*))) AND TOPIC: ("Clinical exam*" OR osce*) AND TOPIC: (racis* OR racial* OR ethnic* OR culture OR international OR foreign) AND TOPIC: (“Medical graduate” OR “medical trainee” OR candidate OR examinee* OR “medical student*”) – 7 results
EDR COMPLETE ((assessor OR examiner* OR observer) AND (bias* OR influen* OR discriminat* OR variation OR variance)) AND exam* AND AB (racis* OR racial* OR ethnic* OR culture OR international OR foreign) AND (“Medical graduate” OR “medical trainee” OR candidate OR examinee* OR “medical student*”) - 21	**PsycInfo** ("observer variation" OR (assessor OR examiner*) AND (bias* OR influen* OR discriminat*)) AND ("Clinical exam*" OR osce*) AND (racis* OR racial* OR ethnic* OR culture OR international OR foreign) AND (“Medical graduate” OR “medical trainee” OR candidate OR examinee* OR “medical student*”) - 4 results

Response to EAG Specific Queries 22 August 2017.

1. Examiner diversity by age, gender and country of qualification in the 2016.2 examination

Following the desktop review provided earlier this month, after discussion with Ms Emma Turner, information was requested from ACEM regarding examiner diversity by age, gender and country of qualification in the 2016.2 OSCE.

A spreadsheet was provided containing de-identified information on examiner age, gender and country of qualification for all 100 examiners who participated in the 2016.2 OSCE.

These data were investigated using the same grouping as provided for candidates in the previous ACEM report concerning differential pass rates, namely:

- Group A: Australia, Canada, Ireland, New Zealand, UK and USA, which are commonly considered to be countries with comparable standards.
- Group B: all other countries.

Group A contained 93 examiners (93%). 76 of these examiners were Australian by country of qualification and seven were New Zealand, giving a total of 83%.

Group B contained seven examiners (7%) as follows:

- India (1) Female
- Sri Lanka (1) Male
- South Africa (2) Male
- Iraq (1) Male
- Israel (1) Male
- Denmark (1) Male

This indicates a disparity in examiner diversity by country of qualification between Group A examiners and Group B examiners in this examination.

The mean age of examiners was 51.5 years, range 37 years to 72 years.

There were 31 female examiners (31%) and 69 male examiners (69%) providing an approximately 1:2 ratio of female to male examiners. Only one of the Group B examiners was female. This indicates a disparity in examiner diversity by gender.

The most common examiner profile was therefore male and Australian trained.

2. The borderline regression method

The following question has been put:

> *Can you please address in your review whether ACEM's current practice of not having specific marking criteria for a given station (for example no checklists to arrive at a mark for that station) and the award of a global score as to whether a candidate is at standard is:*
>
> *a) standard practise across other postgraduate high-stakes exit exams;*
>
> *b) commensurate with use of the borderline regression method?*

I have received from the college four different examples of stations used in the 2016.2 examination and a copy of the score determination document provided by the college statistician.

The marking criteria for the stations involve two steps, a ‘station score’ and a separate ‘global rating’. This is standard practice for a borderline regression method, where the station scores are regressed against the global ratings.

Station score

There are no checklist or dichotomous scores in the ACEM scoring method, however a station score **is** produced using a rating scale approach. This is commensurate with best practice in the literature.

It is common or standard practice to use checklists only or rating scales only or a combination of checklists and rating scales to create a station score in high-stakes examinations of this type. This has been addressed briefly in my literature review.

In the examples given, the station score is computed from a series of three “in-station” seven point rating scales where the minimum mark is one and the maximum mark seven. Weighting is also employed.

At each point, there is a written description of the meaning of that point, for example the lowest point on the scale equates to "very poor level of competence displayed" and the highest point of the scale equates to "very high level of competence displayed". The minimum competence level is at the fourth point, which is labelled "minimum level of competence displayed".

The rating scales are selected on the basis of relevance to the station, and the criteria to be considered in each scale are determined for each station individually and are written in detail below the rating scales for examiner reference.

Global score

After the station has been completed examiners are asked to consider the station performance in its entirety and provide a global score, this time on a five-point scale as follows:

“In terms of a safely practising junior emergency physician, select the ONE option that best reflects the candidate’s performance in this station”:

1 Well Below Standard

2 Below Standard

3 Just at Standard

4 Above Standard

5 Well Above Standard

A score of three is taken as the point of intersection with the regression line in the ACEM borderline regression method, i.e. **just at** standard of junior emergency physician.

In terms of the use of borderline regression of this type, it is standard practice, and the global rating scale used in the ACEM examination is one of a number of variations of scales used for global ratings for the purpose of the borderline regression method.

The method employed to obtain an overall pass mark is also a standard method. The method used takes the pass mark for each station obtained from the borderline regression method and averages it to obtain an overall pass mark for the examination. This is adjusted by adding one standard error (SEM), which is also common practice.

The final pass mark therefore allows candidate to compensate for performance between stations. This is known as a compensatory method. In brief, this means that a candidate may perform below the pass mark in one or more stations and above the pass mark in other stations, allowing them, once all stations are taken into account, to still achieve an overall score that is above the overall pass mark for the examination and therefore achieve a pass. As such, the pass mark is created without reference to the absolute number of stations passed by each candidate.

This standard setting method heavily relies on the examiners using their overall judgement of the candidate’s performance to award a global score rather than counting marks or referring simply to the station scores.

The borderline method is simple to execute in the time pressure of an OSCE. It is critical nonetheless that all examiners understand clearly the definition of a “borderline” candidate or in the case of the ACEM examination the meaning of the third point "just at standard", and are able to make a confident expert decision about whether a candidate falls below, at or above this category.

In order to make this judgement as robust as possible, and to ensure that examiners clearly understand the cognitive differences between a station score and a global rating, examiner training, including practice in using the marking method and calibration discussions, is essential.