Measuring doctors' insight on their performance at the General Medical Council's Tests of Competence pilot exams: a cross-sectional study

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2013-004131
Article Type:	Research
Date Submitted by the Author:	23-Oct-2013
Complete List of Authors:	Mehdizadeh, Leila; University College London, UCL Medical School Sturrock, Alison; University College London, UCL Medical School Myers, Gil; University College London, UCL Medical School Khatib, Yasmin; Queen Mary University of London, Centre for Psychiatry Dacre, Jane; University College London, UCL Medical School
Primary Subject Heading:	Medical education and training
Secondary Subject Heading:	Medical education and training
Keywords:	EDUCATION & TRAINING (see Medical Education & Training), MEDICAL EDUCATION & TRAINING, STATISTICS & RESEARCH METHODS

GMC/ACME research questionnaire

We are considering how accurately doctors are able to judge their own performance, and that of other doctors, in formal testing. Please answer these questions as honestly as possible. We are expecting you to use your judgement to estimate the scores.

These answers will be considered independently from today's validation data and will not impact on your feedback whatsoever. All information will be anonymised before being analysed.

Candidate number

Date of exam

Speciality

A. Knowledge test:

1. Please estimate your own ranking in the written test today, compared to **all the doctors** who are eligible to sit this written test (i.e. fitting the GMC criteria of FY2 to Consultant level, and have worked in the specialty for at least four months within the last year)
*(**0** = lowest rank and **100** = highest rank)*

2. Please estimate your own ranking in the written test today, compared to the other doctors who sat the written test **today**
*(**0** = lowest rank and **100** = highest rank)*

3. There were 120 questions in today's exam, each is worth one mark. There is no negative marking. Please estimate your total score in today's written test.
*(total **120 possible marks**)*

B. OSCE:

4. Please estimate your own ranking in the OSCE today, compared to **all the doctors** who are eligible to sit this OSCE (i.e. fitting the GMC criteria of FY2 to Consultant level, and have worked in the specialty for at least four months within the last year)
*(**0** = lowest rank and **100** = highest rank)*

5. Please estimate your own ranking in the OSCE today, compared to the other doctors who sat the OSCE **today**
*(**0** = lowest rank and **100** = highest rank)*

6. There were 12 stations in today's OSCE each with a maximum of 40 marks. Please estimate your total score in today's OSCE exam.
*(total **480 possible marks**)*

Measuring doctors' insight on their performance at the General Medical Council's Tests of Competence pilot exams: a cross-sectional study

Leila Mehdizadeh[1], Alison Sturrock[2], Gil Myers[3], Yasmin Khatib[4], Jane Dacre[5]

[1] Research Associate in Medical Education, Division of UCL Medical School, University College London, Royal Free Hospital, room GF/664, Hampstead, NW3 2PF.

[2] Senior Lecturer in Medical Education, Division of UCL Medical School, University College London, Royal Free Hospital, room GF/664, Hampstead, NW3 2PF.

[3] Honorary Clinical Teaching Fellow, Division of UCL Medical School, University College London, Royal Free Hospital, room GF/664, Hampstead, NW3 2PF.

[4] Lecturer in Mental Health, Barts and The London School of Medicine and Dentistry, Centre for Psychiatry, Queen Mary University of London, Old Anatomy Building, Charter House Square, London EC1M 6BQ.

[5] Professor of Medical Education, Division of UCL Medical School, University College London, Medical School Building, 74 Huntley Street, London WC1E 6AU.

Corresponding author: Leila Mehdizadeh, l.mehdizadeh@ucl.ac.uk, Tel 020 31089218, UCL Medical School, Royal Free Hospital, room GF/664, Hampstead, NW3 2PF.

Abstract

Objective: To investigate how accurately doctors estimated their performance on the General Medical Council's Tests of Competence pilot exams.

Design: A cross sectional survey design using a questionnaire method.

Setting: University College London Medical School.

Participants: 524 medical doctors working in a range of clinical specialties between foundation year two (FY2) and consultant level.

Main outcome measures: Estimated and actual total scores on a knowledge test and Observed Structured Clinical Examination (OSCE).

Results: The pattern of results for OSCE performance differed to results for knowledge test performance. The majority of volunteers significantly underestimated their OSCE performance. Whereas estimated knowledge test performance differed between high and low performers. Those who did particularly well significantly underestimated their knowledge test performance ($t(196) = -7.70$, $p<0.01$) and those who did less well significantly overestimated ($t(172) = 6.09$, $p<0.01$). There were also significant differences between estimated and/or actual performance by gender, ethnicity and region of Primary Medical Qualification.

Conclusions: Volunteers were more accurate in predicating their knowledge test performance than their OSCE performance. The association between estimated and actual knowledge test performance support the established differences between high and low performers described in the behavioural sciences literature. This was not the case for the OSCE. The implications of the results to the revalidation process are discussed.

Article summary

Article focus

- Revalidation has a strong component of self evaluation. Doctors must provide evidence at their annual appraisal that they continue to meet the standards set out in Good Medical Practice.
- Evidence from the behavioural sciences has demonstrated the deficits in the ability to accurately self-assess one's competencies.
- The purpose of this study was to investigate how accurately doctors estimate their performance on the GMC's Tests of Competence pilot exams.

Key messages

- The literature shows that on tests that discriminate between excellent and poor performance, high performers tend to underestimate their performance whilst low performers overestimate.
- In this study, high performers significantly underestimated their knowledge test scores whilst lower performers significantly over estimated. Whereas most volunteers significantly underestimated their OSCE performance.
- Need to consider whether doctors can and should be trained in accurate self-assessment to reduce the insight gap, particularly as revalidation gets underway.

Strengths and limitations of this study

- This is one of the first studies to look at how well a group of doctors think they perform on a set of exams that have potentially significant real world consequences.
- The large sample means that it has greater power than previous similar studies.
- The results have important implications to the revalidation process of doctors working in the UK.
- The majority of doctors performed well on both exams, therefore patterns between high and lower performers should be interpreted with caution.
- Results are not necessarily applicable to the wider medical community in the UK.

Introduction

The revalidation process of all UK doctors who hold a licence to practise is now underway. Introduced by the General Medical Council (GMC), doctors must provide evidence at their annual appraisal that they continue to meet the standards set out in Good Medical Practice. [1-2] Whilst medical education in the UK aims to produce doctors who are reflective practitioners, [2-3] evidence from the behavioural sciences has demonstrated the deficits in the ability to assess one's own competencies. [4-10] Previous studies showed that psychology students who performed lowest on intellectual and social tasks displayed the least insight by over estimating their own performance. In contrast, the highest performers underestimated their performance. [7]

This pattern has been replicated in the medical context. General practitioners were unable to accurately assess their knowledge of 20 typical clinical conditions. [11] Family medicine residents who performed best on a breaking bad news scenario were more likely to make accurate self estimates than the lowest performers. [12] A systematic review of studies on the accuracy of doctors' self-assessment compared with objective measures of competence showed that doctors who performed the least well also self-assessed the least well. [13] In clinical practice, the dangers of poor performance coupled with unawareness of one's deficiencies are concerning; this doctor may pose a risk to patient safety and may not engage in appropriate professional development activities. With the introduction of revalidation, the number of doctors referred to the GMC for Fitness to Practise investigation may change. Revalidation has a strong component of self evaluation and enforced reflection will offer doctors more opportunities to consider their performance and ways to remedy any weak areas.

In previous studies [11-13] the performance tests were designed specifically for research purposes, to discriminate excellent from poor performance and the results had no real world consequences. We wondered whether the same pattern would emerge between high and low performing doctors on tests of a different nature. Since 1996, the GMC and University College London Medical School have been working in collaboration

to design and deliver the Tests of Competence in FtP investigations.[14] As part of the validation process for these tests, doctors volunteer to take 'pilot exams' that include a knowledge test and an OSCE. These tests differ from those used in previous studies as they have not been designed for the purpose of the present study; rather they are part of a rigorous evaluation process for the GMC's FtP procedures. The written paper consists of a single best answer knowledge test which is machine marked. The OSCE is assessed by trained assessors using a generic domain based mark scheme of 'acceptable', 'cause for concern' and 'unacceptable'. These assessments are designed to test minimum competence; therefore doctors have no opportunity to demonstrate excellence. The volunteer doctors' performance has potential consequences as they are informed that they will be referred to the GMC if their performance fails to meet the minimum standard. The purpose of this study was to investigate how accurately doctors estimate their performance on the GMC's Tests of Competence pilot exams. We were also interested in whether differences in accuracy existed between high and low performers, genders, ethnic backgrounds and Primary Medical Qualification (PMQ) regions.

Methods

Ethics: We received written confirmation from University College London's Research Ethics Committee in October 2008 that the study was exempt from ethical approval. However volunteers explicitly consented to their data being used anonymously for research purposes.

Design: This was a cross sectional study using a questionnaire method.

Sample: Doctors who volunteered to take a GMC pilot exam between August 2011 and July 2012 were invited to participate. The data was collected from 30 pilot events, a total of 524 doctors were included. Volunteers were recruited through advertisement in medical journals, specialty specific newsletters and word of mouth. The sample included

doctors that worked in paediatrics, child psychiatry, anaesthetics, old age psychiatry, forensic psychiatry, general medicine, emergency medicine, orthopaedics, general practice, obstetrics and gynaecology, surgery, cardiology, radiology and care of the elderly. Volunteers ranged from FY2 to consultant level.

Materials: Data was collected from three sources 1) knowledge test, 2) OSCE and 3) self completed questionnaire. The knowledge tests consisted of 120 specialty specific items in a single best answer format with a maximum score of 120. The OSCE included 12 specialty specific stations. Each station was scored by a trained assessor who was usually a consultant in the relevant specialty or a clinical skills nurse. The maximum score for the OSCE was 480. A study specific questionnaire was designed to obtain volunteers' estimated knowledge test and OSCE scores (appendix 1). The questionnaire asked where participants thought they ranked in comparison to other doctors who completed the exams on the same day, in comparison to all doctors who were eligible to sit the GMC pilot exams as well as an estimation of their total knowledge test and OSCE scores.

Outcome measures: We compared volunteers' estimated and actual total scores on the knowledge test and OSCE.

Procedure: Over the course of one day, volunteers had two hours to complete the knowledge test and 96 minutes for the OSCE (eight minutes per station). When all exams had finished volunteers were asked to complete the questionnaire about their performance.

Analyses: Actual and estimated exam scores were compared using SPSS for windows version 19. We split volunteers' estimated and actual scores into tertiles (top, middle, bottom) to see whether differences existed between high and low performers. Results were analysed using descriptive statistics, correlations, t-tests and ANOVAs.

Results

The study was completed with 524 doctors, who volunteered to take the GMC's Tests of Competence pilot exams. Table 1 summarises their demographic characteristics. There were notable differences in the demographics of the volunteers compared with all doctors on the 2011 List of Registered Medical Practitioners (LRMP). In this study, men were under represented and women were over represented. There was also a higher proportion of Asian/Asian British doctors amongst our volunteers and overseas trained doctors were under represented amongst volunteers.

Table 1: Demographic characteristics of sample compared to demographics of doctors on 2011 LRMP

Variable	**Levels**	**Total in this study (N=524)**	**Total on 2011 LRMP(N=245,903)**
Gender	Male	231 (44.1%)	141,369 (57%)
	Female	293 (55.9%)	104,534 (43%)
Ethnicity	White	266 (50.8%)	118,822 (48%)
	Black/Black British	17 (3.2%)	6,812 (2.8%)
	Asian/Asian British	147 (28.1%)	46,664 (4.3%)
	Mixed	25 (4.8%)	3,643 (1.5%)
	Other ethnic groups	20 (3.8%)	9,002 (3.7%)
	Not stated (includes prefer not to say)	49 (9.4%)	60,960 (25%)
Primary Medical Qualification region	UK	408 (77.9%)	155,264 (63%)
	EU country	22 (4.2%)	24,031 (10%)
	Non-EU country	94 (17.9%)	66,608 (27%)

General patterns

Overall, volunteers were more accurate in predicating their total knowledge test score than their OSCE score. There was a moderately strong positive relationship between the difference in estimated and actual knowledge test scores with actual scores; r=0.43, p<0.01. Figure 1 shows a roughly equal distribution of volunteers that over and underestimated their knowledge test scores. Those who over estimated (negative numbers on the y axis) tended to score lower than those who under estimated their knowledge test scores. There was no significant difference between the estimated and actual scores on the knowledge test; t (521) = 1.33, p=0.19.

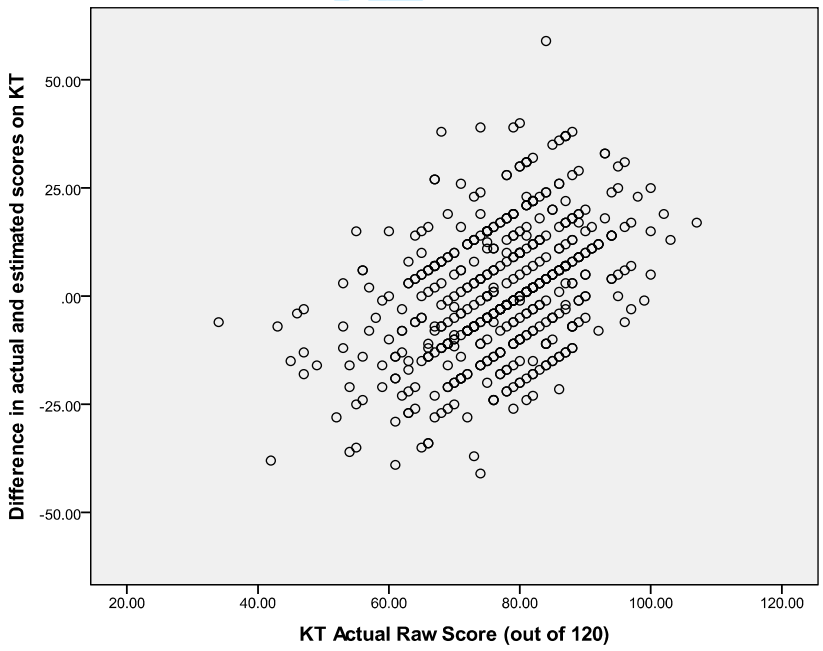


Figure 1: Scatterplot of difference in actual and estimated knowledge test scores against actual knowledge test scores

There was also an association between difference in actual and estimated scores for the OSCE; r=0.33, p<0.01. Figure 2 shows that only a few people overestimated their OSCE scores. The majority under estimated their performance. The few who overestimated their OSCE scores performed less well than those who under estimated

their OSCE performance. There was a significant difference between the estimated and actual total OSCE scores; t (520 = -37.76, p<0.01).

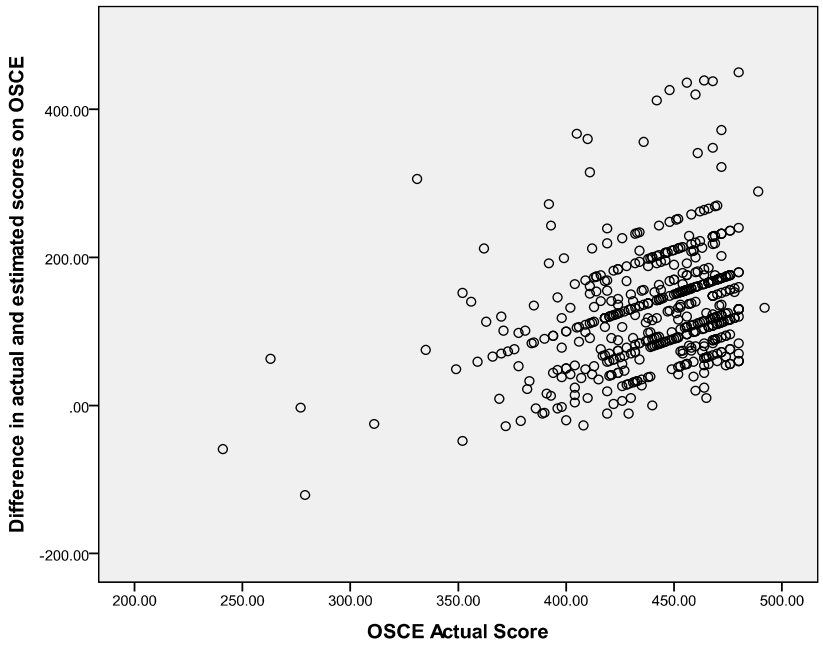


Figure 2: Scatterplot of difference in actual and estimated OSCE scores against actual OSCE scores.

Difference between high and lower performers on their estimated and actual exam scores

There were significant differences between high and low performers on their estimated knowledge test performance. The highest performers significantly underestimated their knowledge test scores by an average of 8 marks (t (196) = -7.70, p<0.01) and the lower performers significantly overestimated by an average of 7 marks (t (172) = 6.09, p<0.01).

Both high and lower performers significantly underestimated their OSCE performance; t (180) = -26.28, p<0.01 and t (172) = -16.20, p<0.01 respectively. Those in the top percentile underestimated their OSCE performance to the greatest extent.

Gender differences on estimated and actual exam scores

Men predicted a higher knowledge test score than women, with mean estimates of 79 (SD 15) and 74 (SD 14) respectively. Levene's test confirmed there was homogeneity of variance and the t-test revealed that this gender difference in mean estimates was significant; t(522) = 4,27, p=<0.01. Women performed slightly better on the knowledge test than men but their means scores were not significantly different; 77 (SD 11) and 76 (SD 10) respectively. Men also predicted a higher overall OSCE performance than women with mean estimates of 329 (SD 59) and 300 (SD 81) respectively. This was a significant difference; t(516) = 4.65, p<0.01. Women outperformed men on OSCE performance and the difference was significant; t(482) = -2.82, p<0.01. A three way ANOVA showed that there was a significant effect of gender on estimated knowledge test (F (1,508) = 4.62, p=0.03) and OSCE performance (F(1,505)=11.74, p<0.01). This means that men, irrespective of their ethnicity or PMQ region, estimated higher than women on both exams.

Ethnic differences on estimated and actual exam scores

There were no significant differences in estimated exam scores between volunteers of different ethnic backgrounds. However there was a tendency for Asian/Asian British volunteers to estimate highest on their knowledge test performance (M=78, SD=15). The highest estimated OSCE performance came from white volunteers (M=317, SD=73) and 'other' ethnic groups (M=316, SD=59).

A one way ANOVA confirmed that there were significant differences by ethnic background on actual knowledge test performance (F=(5,516) = 4.46, p<0.01) and OSCE performance (F=(5,518)=11.48,p<0.01). Fisher's least significant difference test (LSD) was used to explore where these differences occurred. Volunteers who did not specify their ethnicity performed highest on the knowledge test (M=78, SD=11). Asian/Asian British volunteers scored lowest on the knowledge test, particularly in comparison to white volunteers (p<0.01) and volunteers who did not state their ethnicity (p=0.02). Table 2 shows that white volunteers outperformed all other ethnic groups on OSCE performance.

Table 2: Differences in OSCE performance between white and other ethnic groups

Ethnicity		OSCE performance	Significance
White (M=451,SD=29)	Not stated	M=440, SD=23	p=0.04
	Other	M=439, SD=27	p=0.12*
	Mixed	M=433, SD=39	p=0.08*
	Black/Black British	M=428, SD=33	p=0.03
	Asian/Asian British	M=427, SD=37	p<0.01

*Non significant difference

Differences by primary medical qualification region on estimated and actual exam scores

The majority of volunteers gained their PMQ from the UK (78%). Of the non UK trained volunteers, 18% were from a non EU county and 4% were from an EU country. There were significant differences in estimated knowledge test performance between volunteers of different PMQ regions. Non UK trained volunteers estimated significantly higher than UK trained volunteers ($F(2,521)=6.06, p<0.01$). However there were no actual differences in knowledge test performance between volunteers of different PMQ regions ($F(2,519)=2.28$, $p=0.10$). A reverse pattern was true of OSCE performance. Estimated OSCE performance did not differ by PMQ region, although EU trained volunteers tended to make the highest estimates. Actual OSCE performance did significantly differ, with UK trained volunteers outperforming non UK trained volunteers ($F(5,521)=37.96$, $p<0.01$).

Discussion

Principal findings:

In general volunteers performed well on the knowledge test and OSCE (usually 70% and above) so the number of actual low performers was small. They were more accurate in predicting their knowledge test performance than OSCE performance. Differences in predictions between high and low performers were found on the knowledge test but not on the OSCE. In keeping with the results found for psychology students [4-10] high performers significantly underestimated their knowledge test performance whilst lower performers significantly overestimated. Most volunteers significantly underestimated their OSCE performance irrespective of how well they actually did. Differences between estimated and actual performance were apparent between men and women. On both exams, women's estimated performance on both exams was lower than men's but they actually did better than men on both exams, particularly the OSCE. Estimated performances on both exams did not significantly differ between ethnicities, but there was a tendency for Asian/Asian British participants to estimate slightly higher than other groups. Actual performance for both exams did significantly differ by ethnic group. Volunteers of white and unspecified ethnicity performed highest on the knowledge test, while white volunteers outperformed all others on OSCE performance. UK trained volunteers outperformed overseas trained volunteers on OSCE performance but there were no differences by PMQ region on knowledge test performance.

Findings in relation to literature:

This study shows that doctors were moderately accurate in predicting their knowledge test performance but less accurate in predicting their OSCE performance. These results provide support for previously reported patterns between high and low performers as there was a tendency for doctors who performed particularly well on the knowledge test to significantly underestimate their score whilst lower performers significantly

overestimated their score. [7, 13, 15-17] However OSCE performance did not follow this pattern. This discrepancy may be explained by the difference in nature between the knowledge test and OSCE. The knowledge test was in single best answer format and the volunteers either answered the item correctly or incorrectly. The OSCE used a domain based mark scheme which was designed to test for minimum competence. Although volunteers were familiar with the OSCE format, they may not have been familiar with the concept of assessing for minimum competence as their previous experiences of OSCE feedback is usually based on an assessment of excellent performance. It is possible that volunteers assumed that the threshold for good performance was higher than it actually was and this may have made it more difficult for volunteers to predict their OSCE score compared to knowledge test score. The gender and ethnic differences found in this study support that of previous findings to an extent. Women tend to underestimate and men tend to overestimate their performance in medicine. [16, 18-20] Our findings supported this pattern on knowledge test performance but not on OSCE performance. White medical students consistently outperform non white medical students in the UK, [21] the US, [22-26] and in other English speaking countries. [27-28] In this study white volunteers performed substantially higher than non-whites on the OSCE but not on the knowledge test. Further, there was an interaction between ethnicity and PMQ region. OSCE performance was higher in white UK trained doctors than white non-UK trained as well as non-white UK trained doctors.

Implications of findings:

This study lends support to a growing body of evidence that suggests medical students and doctors have limited ability to accurately assess their own performance. [12-13, 15, 17, 29]. In response to such findings there have been suggestions to consider whether formal training in accurate self assessment can be addressed in medical education. (13) One potential benefit of revalidation is that it will enforce doctors to reflect on their clinical knowledge and practice as well as identify areas for improvements. [1] This process may prove to be a good opportunity for doctors to become more self aware.

Overall, most volunteers did not appear to have an inflated view of their exam performance. It is reasonable to assume that doctors who are not overly confident are likely to exercise more caution in their clinical practice than those who are over confident. However, roughly half of the sample did overestimate their knowledge test performance. This is potentially a problem as overconfidence in doctors is associated with poor clinical judgement and decision making. [30-31] Furthermore, research has shown that overconfidence in medicine is more likely to be a male rather than female characteristic. [16, 18] Women are more likely to perceive themselves and be perceived by others as less confident in clinical knowledge and skills. [16, 18]. This pattern was found in our study, despite women out performing men on both the knowledge test and OSCE. In practice, lack of confidence may disadvantage female doctors with patient interaction and career progression. [18] A common pattern of ethnic differences was also found with white volunteers outperforming non white volunteers on the OSCE. The reasons for this performance gap are unclear but cross-cultural differences in communication styles may explain some of the variation in performance on OSCE type exams; assessors may also have been influenced by ethnic stereotypes. [32] This performance gap is in line with the recent controversy around higher rates of failure amongst international medical graduates taking the clinical skills assessment portion of the MRCGP exam [33]. Further research is necessary to understand why these ethnic differences persist in medicine and what can be done to reduce this discrepancy. [21, 32, 34]

Strengths and weaknesses of the study:

This study furthers our understanding of doctors' insight into their own performance on a set of familiar and clinically relevant exams. It is one of the first studies to look at how well a group of doctors think they perform on a set of voluntary exams that have potentially significant consequences. The large sample means that it has greater power than previous similar studies. The tests included were all part of a validation process for the GMC's Fitness to Practice Procedures. The study has practical relevance, as demonstrated by volunteers seeking feedback on their performance and commenting that they used the process as future exam preparation. The main limitations of the study

are that senior doctors were not well represented. The sub-groups between ethnic background and PMQ region were not equal and white UK trained doctors were over represented. This means that the results are not necessarily generalisable to all doctors and that the results may have differed had there been equal numbers of doctors by different ethnic background and PMQ regions. The context in which the exams were taken may have impacted on self estimated scores. Volunteers may have felt it was more acceptable to present themselves as modest, self critical doctors when taking GMC affiliated exams. Therefore the study's findings could be biased by social desirability. Finally, the majority of doctors performed well on both exams, even those who scored in the bottom percentiles. Therefore patterns between high and lower performers should be interpreted with caution. Those who performed less well than the majority may have struggled because of the type of exam material. However, it is likely that the few who performed lower in comparison to a cohort of volunteers who sat the same exam material, were actual low performers.

Future directions and conclusions:

Volunteers were more accurate in predicting their knowledge test performance than OSCE performance. Differences in estimates were found between high and low performers on an objective test of knowledge but not on a practical test of minimum clinical competence (OSCE). This may suggest that doctors have better insight of their knowledge compared to their practical and communication skills. Alternatively, the unfamiliarity of the OSCE as a test of minimum competence may have made it a more difficult exam by which to judge their performance. Estimated and actual performance differed by gender and ethnicity but less so by where a doctor had gained their primary medical qualification. Doctors who are overconfident in their performance are potentially a greater risk to patient safety than those who underestimate. A future study may wish to clarify whether insight does differ between doctors who are identified from the outset as poor performers versus those who do particularly well. Also the reasons for why medical students and doctors assign themselves a particular score has yet to be explored.

Acknowledgments We are grateful to Dr Henry Potts for his statistical support and help with interpreting results.

Funding This work was supported by the General Medical Council.

Competing interests None.

Ethics approval Not required, but volunteers consented to their anonymised data being used for research purposes

Contributors LM analysed the results and wrote the manuscript. AS contributed intellectually to the study and helped revise the manuscript drafts. GM designed the research questionnaire and helped recruit participants. YK recruited participants and entered the data. JD oversaw the study and provided intellectual contribution at all phases. All authors provided feedback on manuscript drafts and approved its final version.

Provenance and peer review Not commissioned, externally peer reviewed.

Data sharing statement Due to our confidentiality agreement with the General Medical Council, data is unavailable for sharing.

References

1. General Medical Council. *Ready for Revalidation: supporting information for appraisal and revalidation*, 2012.
2. General Medical Council. *Good Medical Practice,* 2012.
3. Hays, R., Jolly, B., Caldon, L., McCrorie, P., McAvoy, P., McManus, IC and Rehans, JJ. *Is insight important? Measuring capacity to change performance.* Medical Education, 2002. **36**: p. 965–971.
4. Cross, P., *Not can but will college teaching be improved?* New Directions for Higher Education, 1977. **17**: p. 1-15.
5. Dunning, D., Heath, C and Suls, J. *Flawed self-assessment: implications for health, education, and the workplace.* Psychological Science in the Public Interest, 2004. **5**(3): p. 69-106.
6. Ehrlinger, J., Johnson, K., Banner, M., Dunning, D and Kruger, J. *Why the unskilled are unaware: further explorations of (absent) self-insight among the incompetent.* Organizational Behavior and Human Decision Processes, 2008. **105**: p. 98-121.
7. Kruger, J. and Dunning, D. *Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments.* Journal of Personality and Social Psychology 1999. **77**(6): p. 1121-1134.
8. Marottoli, R. and Richardson, E. *Confidence in, and selfrating of, driving ability among older drivers.* Accident Analysis and Prevention, 1998. **30**: p. 331–336.
9. Rutter, D., Quine, L and Albery, I. *Perceptions of risk in motorcyclists: unrealistic optimism, relative realism and predictions of behaviour.* British Journal of Psychology, 1998. **89**: p. 681–696.
10. Zenger, T. *Why do employers only reward extreme performance? Examining the relationships among performance, pay, and turnover.* Administrative Science Quarterly, 1992. **37**: p. 198–219.

11. Tracey, J., Arroll, B., Barham, P and Richmond, D. *The validity of general practitioners' self assessment of knowledge: cross sectional study.* British Medical Journal, 1997. **315**: p. 1426.
12. Hodges, B., Regehr, G and Martin, D. *Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it* Academic Medicine, 2001. **76**(10): p. S87-89.
13. Davis, D., Mazmanian, PE., Fordis, M., Van Harrison., Thorpe KE and Perrier, L. *Accuracy of physician self-assessment compared with observed measures of competence: a systematic review.* The Journal of the American Medical Association, 2006. **296**: p. 1094-1102.
14. General Medical Council. *The GMC's Fitness to Practise procedures*. 2012.
15. Barnsley, L., Lyon, PM., Ralston, SJ., Hibbert, EJ., Cunningham, I., Gordon, FC and Field MJ. *Clinical skills in junior medical officers: a comparison of self-reported confidence and observed competence.* Medical Education, 2004. **38**(4): p. 358-367.
16. Blanch-Hartigan, D. *Medical students' self-assessment of performance: results from three meta-analyses.* Patient Education and Counseling, 2011. **84**(1): p. 3-9.
17. Tousignant, M. and Desmarchais, J. *Accuracy of student self-assessment ability compared to their own performance in a problem-based learning medical program: a correlation study.* Advances in Health Sciences Education, 2002. **7**: p. 19-27.
18. Blanch, D., Hall, JA., Roter, DL and Frankel, RM. *Medical student gender and issues of confidence.* Patient Education and Counseling, 2008. **72**(3): p. 374-381.
19. Rees, C. And Shepherd, M. *Students' and assessors' attitudes towards students' self-assessment of their personal and professional behaviours.* Medical Education 2005. **39**: p. 30-39.
20. Minter, R., Gruppen, LD., Napolitano, KS and Gauger, PG. *Gender differences in the self-assessment of surgical residents.* American Journal of Surgery, 2005. **189**: p. 647–650.

21. Woolf, K.,Potts, H and McManus, IC. *Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis.* British Medical Journal, 2011. **342**: p.901

22. Xu, G., Veloski, J., Hojat, M., Gonnella, J and Bacharach, B. *Longitudinal comparison of the academic performances of Asian-American and white medical students.* Academic Medicine, 1993. **68**(1): p. 82-86.

23. Dawson, B., Iwamoto, CK., Ross, LP., Nungester, RJ., Swanson, DB and Volle, RL. *Performance on the national board of medical examiners .1. examination by men and women of different race and ethnicity.* Journal of the American Medical Association, 1994. **272**: p. 674-679.

24. Koenig, J., Sireci, S and Wiley, A. *Evaluating the predictive validity of MCAT scores across diverse applicant groups.* Academic Medicine, 1998. **73**: p. 1095-1106.

25. Veloski, JJ., Callahan, CA., Xu, G., Hojat, M and Nash, DB. *Prediction of students' performances on licensing examinations using age, race, sex, undergraduate GPAs, and MCAT scores.* Academic Medicine, 2000. **75**: p. S28-S30.

26. Colliver, J., Swartz, M and Robbs, R. *The effect of examinee and patient ethnicity in clinical-skills assessment with standardized patients.* Advances in Health Sciences Education, 2001. **6**: p. 5-13.

27. Kay-Lambkin, F., Pearson, S and Rolfe, I. *The influence of admissions variables on first year medical school performance: a study from Newcastle University, Australia.* Medical Education, 2002. **36**: p. 154-159.

28. Liddell, M. and Koritsas, S. *Effect of medical students' ethnicity on their attitudes towards consultation skills and final year examination performance.* Medical Education, 2004. **38**: p. 187-198.

29. Eva, K., Cunnington, J., Reiter, HI., Keane, DR and Norman, GR. *How can I know what I don't know? Poor self assessment in a well-defined domain.* Advances in Health Sciences Education, 2004. **9**: p. 211–224.

30. Croskerry, P. *Achieving Quality in Clinical Decision Making: Cognitive strategies and detection of bias.* Academic Emergency Medicine, 2002. **9**: p. 1184-1204.

31. Croskerry, P. *The importance of cognitive errors in diagnosis and strategies to minimize them.* Academic Medicine, 2003. **78**(8): p. 775-780.

32. Woolf, K., Cave, J., Greenhalgh, T and Dacre, J. *Ethnic stereotypes and the underachievement of UK medical students from ethnic minorities: qualitative study.* British Medical Journal, 2008. **337** p.1220.

33. Kaffash, J. *RCGP faces legal threat over international GP trainee failure rates*, Pulse. 13 November 2012. Available at http://www.pulsetoday.co.uk/practice-business/practice-topics/education/rcgp-faces-legal-threat-over-international-gp-trainee-failure-rates/20000852.article (Accessed 4 February 2013)

34. Woolf, K., McManus, IC., Potts, HWW and Dacre, J. *The mediators of minority ethnic underperformance in final medical school examinations.* British Journal of Educational Psychology, 2011. **83**(1) p.135-159

How well do doctors think they perform on the General Medical Council's Tests of Competence pilot exams? A cross-sectional study

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2013-004131.R1
Article Type:	Research
Date Submitted by the Author:	17-Dec-2013
Complete List of Authors:	Mehdizadeh, Leila; University College London, UCL Medical School Sturrock, Alison; University College London, UCL Medical School Myers, Gil; University College London, UCL Medical School Khatib, Yasmin; Queen Mary University of London, Centre for Psychiatry Dacre, Jane; University College London, UCL Medical School
Primary Subject Heading:	Medical education and training
Secondary Subject Heading:	Medical education and training
Keywords:	EDUCATION & TRAINING (see Medical Education & Training), MEDICAL EDUCATION & TRAINING, STATISTICS & RESEARCH METHODS

How well do doctors think they perform on the General Medical Council's Tests of Competence pilot exams? A cross-sectional study

Leila Mehdizadeh[1], Alison Sturrock[2], Gil Myers[3], Yasmin Khatib[4], Jane Dacre[5]

[1] Research Associate in Medical Education, Division of UCL Medical School, University College London, Royal Free Hospital, room GF/664, Hampstead, NW3 2PF.

[2] Senior Lecturer in Medical Education, Division of UCL Medical School, University College London, Royal Free Hospital, room GF/664, Hampstead, NW3 2PF.

[3] Honorary Clinical Teaching Fellow, Division of UCL Medical School, University College London, Royal Free Hospital, room GF/664, Hampstead, NW3 2PF.

[4] Lecturer in Mental Health, Barts and the London School of Medicine and Dentistry, Centre for Psychiatry, Queen Mary University of London, Old Anatomy Building, Charter House Square, London EC1M 6BQ.

[5] Professor of Medical Education, Division of UCL Medical School, University College London, Medical School Building, 74 Huntley Street, London WC1E 6AU.

Corresponding author: Leila Mehdizadeh, l.mehdizadeh@ucl.ac.uk, Tel 020 31089218, UCL Medical School, Royal Free Hospital, room GF/664, Hampstead, NW3 2PF.

Abstract

Objective: To investigate how accurately doctors estimated their performance on the General Medical Council's Tests of Competence pilot exams.

Design: A cross sectional survey design using a questionnaire method.

Setting: University College London Medical School.

Participants: 524 medical doctors working in a range of clinical specialties between foundation year two and consultant level.

Main outcome measures: Estimated and actual total scores on a knowledge test and Observed Structured Clinical Examination (OSCE).

Results: The pattern of results for OSCE performance differed to results for knowledge test performance. The majority of doctors significantly underestimated their OSCE performance. Whereas estimated knowledge test performance differed between high and low performers. Those who did particularly well significantly underestimated their knowledge test performance ($t(196) = -7.70$, $p<0.01$) and those who did less well significantly overestimated ($t(172) = 6.09$, $p<0.01$). There were also significant differences between estimated and/or actual performance by gender, ethnicity and region of Primary Medical Qualification.

Conclusions: Doctors were more accurate in predicating their knowledge test performance than their OSCE performance. The association between estimated and actual knowledge test performance support the established differences between high and low performers described in the behavioural sciences literature. This was not the case for the OSCE. The implications of the results to the revalidation process are discussed.

Article summary

Article focus

- Revalidation has a strong component of self-evaluation. Doctors must provide evidence at their annual appraisal that they continue to meet the standards set out in Good Medical Practice.
- Evidence from the behavioural sciences has demonstrated the deficits in the ability to accurately self-assess one's competencies.
- The purpose of this study was to investigate how accurately doctors estimate their performance on the General Medical Councils Tests of Competence pilot exams.

Key messages

- The literature shows that on tests that discriminate between excellent and poor performance, high performers tend to underestimate their performance whilst low performers overestimate.
- In this study, high performers significantly underestimated their knowledge test scores whilst lower performers significantly overestimated. In contrast, most doctors significantly underestimated their OSCE performance.
- We need to consider how doctors can be trained to develop accurate self-perception, particularly as revalidation gets underway.

Strengths and limitations of this study

- This is one of the first studies to look at how well a group of doctors think they perform on a set of exams that have potentially significant real world consequences.
- The large sample means that it has greater power than previous similar studies.
- The results have important implications to the revalidation process of doctors working in the UK.
- The majority of doctors performed well on both exams, therefore patterns between high and lower performers should be interpreted with caution.
- Results are not necessarily applicable to the wider medical community in the UK.

Introduction

The revalidation process of all UK doctors who hold a licence to practise is now underway. Introduced by the General Medical Council (GMC), doctors must provide evidence at their annual appraisal that they continue to meet the standards set out in Good Medical Practice.[1,2] This requires the doctor to reflect on their own knowledge and practise to be able to demonstrate their strengths and areas that need further development. Whilst medical education in the United Kingdom (UK) aims to produce doctors who are reflective practitioners, [2,3] evidence from the behavioural sciences has demonstrated the deficits in the ability to assess one's own competencies.[4-10] Previous studies showed that psychology students who performed lowest on intellectual and social tasks displayed the least insight by over estimating their own performance. In contrast, the highest performers underestimated their performance.[7] This pattern has been replicated in the medical context.

General practitioners were unable to accurately assess their knowledge of 20 typical clinical conditions.[11] Family medicine residents who performed best on a breaking bad news scenario were more likely to make accurate self-estimates than the lowest performers.[12] A systematic review of studies on the accuracy of doctors' self-assessment compared with objective measures of competence showed that doctors who performed the least well also self-assessed the least well. These studies indicate that people in general as well as medical doctors specifically, have a limited ability to accurately assess their own competence.[13] In clinical practice, the idea that a poorly performing doctor could also lack awareness of their problems is concerning. Such a doctor may pose a risk to patient safety and may not engage in appropriate professional development activities. With the introduction of revalidation, the number of doctors referred to the GMC for Fitness to Practise (FtP) investigation may change. Revalidation has a strong component of self-evaluation and enforced reflection will offer doctors more opportunities to consider their performance and ways to remedy any weak areas.

There is a debate in the self-assessment literature concerning use of terminology and outcome measures. One perspective that has gained more recent attention is the necessity to distinguish self-assessment from self-monitoring as different approaches to the investigation of people's perception of their personal competence. [14-17] The self-assessment approach investigates how well individuals can judge their personal competence against an objective measure of competence.[16] Alternatively, the self-monitoring approach is interested in the extent that people show awareness of the limits in their competence during a given situation, and this can be measured according to an individual's behaviour e.g. taking longer to think about a question of which they are unsure.[16] It is important to clarify that this study looks at doctors' self-assessment ability after completing a set of exams, not how well they can monitor their performance during the exams. Therefore this study does not include any outcomes of doctors' behaviour.

We measured the accuracy of doctors' self-predicted performance on the GMC's Tests of Competence (ToC) pilot exams. Tests of competence are used by the GMC to assess poorly performing doctors under FtP investigation. Before implementation, test content is piloted on volunteer doctors who have no known FtP concerns. Doctors volunteer to take a knowledge test and an Observed Structured Clinical Examination (OSCE) in their relevant specialty. The written paper consists of a single best answer knowledge test which is machine marked. The OSCE is marked by trained assessors using a generic domain based mark scheme of 'acceptable', 'cause for concern' and 'unacceptable'. In the self-assessment literature cited previously, [6-13] the tests that participants were asked to predict their scores on differed in three key ways from the tests used in the present study. They had been designed specifically for research purposes, to discriminate excellent from poor performance and the results had no real world consequences. This was not the case in the present study. Participants were assessed on tests that were established for the purpose of FtP investigations, they are tailored to assess the expected minimum level of competence of a practising doctor, and there were potentially real world consequences if an individual's performance failed to meet the minimum standard.

The purpose of this study was to investigate how well doctors think they perform on the GMC's ToC pilot exams. In particular we wondered whether the established

differences in the literature between high and low performers would emerge; do those who perform well underestimate their performance and do those who perform less well overestimate? We were also interested in whether differences in self-estimates existed by gender, ethnic background and Primary Medical Qualification (PMQ) region.

Methods

Ethics: We received written confirmation from University College London's Research Ethics Committee in October 2008 that the study was exempt from ethical approval. However study participants explicitly consented to their data being used anonymously for research purposes.

Design: This was a cross sectional study using a questionnaire method.

Sample: Doctors who volunteered to take a GMC ToC pilot exam between June 2011 and July 2012 were invited to participate in this study. Volunteers for the pilot exams were recruited through advertisement in medical journals, specialty specific newsletters and word of mouth. The study sample included doctors that worked in paediatrics, child psychiatry, anaesthetics, old age psychiatry, forensic psychiatry, general medicine, emergency medicine, orthopaedics, general practice, obstetrics and gynaecology, surgery, cardiology, radiology and care of the elderly. They ranged from foundation year two to consultant level.

Materials: A study specific questionnaire was designed to obtain participants' estimated knowledge test and OSCE scores (Appendix 1). The questionnaire asked where participants thought they ranked in comparison to other doctors who completed the exams on the same day, in comparison to all doctors who were eligible to sit the GMC test of competence pilot exams as well as an estimation of their total knowledge test and OSCE scores.

Outcome measures: We compared participants' self-estimated and actual total scores on the knowledge test and OSCE. The knowledge tests consisted of 120 specialty specific items in a single best answer (SBA) format with a maximum score

of 120. The OSCE included 12 specialty specific stations. Each station was scored by a trained assessor who was usually a consultant in the relevant specialty or a clinical skills nurse. The maximum score for the OSCE was 480.

Procedure: Doctors who volunteered to take a GMC pilot exam between June 2011 and July 2012 were invited to participate in this study. Once volunteer doctors had completed the knowledge test and OSCE, the study questionnaire was distributed by GM who was a facilitator at the piloting events. Doctors were briefed about the purpose of the study and how their data would be used. They were assured that the completion of the questionnaire was voluntary and would only take 5-10 minutes.

Analyses: Actual and estimated exam scores were compared using SPSS for windows version 19. We split estimated and actual scores into tertiles (top, middle, bottom) to see whether differences existed between high and low performers. Results were analysed using descriptive statistics, correlations, t-tests and ANOVAs.

Results

Between June 2011-July 2012, 689 doctors volunteered to take a pilot ToC and 524 participated in present the study (76% rate of participation). During this period, most were junior and middle grade doctors who qualified in the UK. As compared with all doctors on the 2011 List of Registered Medical Practitioners (LRMP) [18] men were under represented and women were over represented (Table 1). There was also a higher proportion of Asian/Asian British doctors in this study and overseas trained doctors were under represented (Table 1).

Table 1: Demographic characteristics of sample compared to demographics of doctors on 2011 LRMP

Variable	Levels	Total in this study (N=524)	Total on LRMP 2011 (N=245,903)
Gender	Male	231 (44.1%)	141,369 (57%)
	Female	293 (55.9%)	104,534 (43%)
Ethnicity	White	266 (50.8%)	118,822 (48%)
	Black/Black British	17 (3.2%)	6,812 (2.8%)
	Asian/Asian British	147 (28.1%)	46,664 (4.3%)
	Mixed	25 (4.8%)	3,643 (1.5%)
	Other ethnic groups	20 (3.8%)	9,002 (3.7%)
	Not stated (includes prefer not to say)	49 (9.4%)	60,960 (25%)
Primary Medical Qualification region	UK	408 (77.9%)	155,264 (63%)
	EU country	22 (4.2%)	24,031 (10%)
	Non-EU country	94 (17.9%)	66,608 (27%)

General patterns

Overall, participants were more accurate in predicting their total knowledge test score than their OSCE score. There was a moderately strong positive relationship between the difference in estimated and actual knowledge test scores with actual scores; $r=0.43$, $p<0.01$. There was a roughly equal distribution of participants that over and underestimated their knowledge test scores (Figure 1). Those who overestimated (negative numbers on the y axis) tended to score lower than those who underestimated their knowledge test scores. There was no significant difference between the estimated and actual scores on the knowledge test; $t(521) = 1.33$, $p=0.19$.

There was also an association between difference in actual and estimated scores for the OSCE; r=0.33, p<0.01. The vast majority of doctors underestimated their OSCE performance (Figure 2). The few who overestimated their OSCE scores performed less well than those who underestimated their OSCE performance (Figure 2). There was a significant difference between the estimated and actual total OSCE scores; t (520 = -37.76, p<0.01).

Differences between high and lower performers on their estimated and actual exam scores

There were significant differences between high and low performers on their estimated knowledge test performance. The highest performers significantly underestimated their knowledge test scores by an average of 8 marks (t (196) = -7.70, p<0.01) and the lower performers significantly overestimated by an average of 7 marks (t (172) = 6.09, p<0.01).

Both high and lower performers significantly underestimated their OSCE performance; t (180) = -26.28, p<0.01 and t (172) = -16.20, p<0.01 respectively. Those in the top percentile underestimated their OSCE performance to the greatest extent.

Gender differences on estimated and actual exam scores

Men predicted a higher knowledge test score than women, with mean estimates of 79 (SD 15) and 74 (SD 14) respectively. Levene's test confirmed there was homogeneity of variance and the t-test revealed that this gender difference in mean estimates was significant; t(522) = 4,27, p=<0.01. Women performed slightly better on the knowledge test than men but their means scores were not significantly different; 77 (SD 11) and 76 (SD 10) respectively. Men also predicted a higher overall OSCE performance than women with mean estimates of 329 (SD 59) and 300 (SD 81) respectively. This was a significant difference; t(516) = 4.65, p<0.01. Women outperformed men on OSCE performance and the difference was significant; t(482) = -2.82, p<0.01. A three way ANOVA showed that there was a significant effect of gender on estimated knowledge test (F (1,508) = 4.62, p=0.03) and OSCE

performance (F(1,505)=11.74, p<0.01). This means that men, irrespective of their ethnicity or PMQ region, estimated higher than women on both exams.

Ethnic differences on estimated and actual exam scores

There were no significant differences in estimated exam scores between doctors of different ethnic backgrounds. However there was a tendency for Asian/Asian British doctors to estimate highest on their knowledge test performance (M=78, SD=15). The highest estimated OSCE performance came from white doctors (M=317, SD=73) and 'other' ethnic groups (M=316, SD=59).

A one way ANOVA confirmed that there were significant differences by ethnic background on actual knowledge test performance (F=(5,516) = 4.46, p<0.01) and OSCE performance (F=(5,518)=11.48,p<0.01). Fisher's least significant difference test was used to explore where these differences occurred. Doctors who did not specify their ethnicity performed highest on the knowledge test (mean=78, SD=11). Asian/Asian British doctors scored lowest on the knowledge test, particularly in comparison to white doctors (p<0.01) and those who did not state their ethnicity (p=0.02). Table 2 shows that white doctors outperformed all other ethnic groups on OSCE performance.

Table 2: Differences in OSCE performance between white and other ethnic groups

Ethnicity		OSCE performance	Significance
White (M=451,SD=29)	Not stated	M=440, SD=23	p=0.04
	Other	M=439, SD=27	p=0.12*
	Mixed	M=433, SD=39	p=0.08*
	Black/Black	M=428, SD=33	p=0.03

	British		
	Asian/Asian British	M=427, SD=37	p<0.01

*Non-significant difference

Differences by primary medical qualification region on estimated and actual exam scores

The majority of doctors gained their PMQ from the UK (78%). Of the non- UK trained doctors, 18% were from a non-EU county and 4% were from an EU country. There were significant differences in estimated knowledge test performance between doctors of different PMQ regions. Non-UK trained doctors estimated significantly higher than UK trained doctors ($F(2,521)=6.06, p<0.01$). However there were no actual differences in knowledge test performance between doctors of different PMQ regions ($F(2,519)=2.28$, $p=0.10$). A reverse pattern was true of OSCE performance. Estimated OSCE performance did not differ by PMQ region, although EU trained doctors tended to make the highest estimates. Actual OSCE performance did significantly differ, with UK trained doctors outperforming non-UK trained doctors ($F(5,521)=37.96$, $p<0.01$).

Discussion

Principal findings:

In general study participants performed well on the knowledge test and OSCE (usually 70% and above) so the number of actual low performers was small. They were more accurate in predicting their knowledge test performance than OSCE

performance. Differences in predictions between high and low performers were found on the knowledge test but not on the OSCE. In keeping with previous literature [6-13] high performers significantly underestimated their knowledge test performance whilst lower performers significantly overestimated. Most doctors significantly underestimated their OSCE performance irrespective of how well they actually did. Differences between estimated and actual performance were apparent between men and women. On both exams, women's estimated performance was lower than men's but they actually did better than men, particularly on OSCE performance. Estimated performance on both exams did not significantly differ between ethnicities, but there was a tendency for Asian/Asian British participants to estimate slightly higher than other groups. Actual performance for both exams did significantly differ by ethnic group. Doctors of white and unspecified ethnicity performed highest on the knowledge test, while white doctors outperformed all others on OSCE performance. UK trained doctors outperformed overseas trained doctors on OSCE performance but there were no differences by PMQ region on knowledge test performance.

Findings in relation to literature:

Our results are in line with the literature that demonstrates the limited ability people have, including doctors, to accurately self-assess their performance.[5-13,19] Further, this study provides support for previously reported patterns between high and low performers.[7,13,20-22] There was a tendency for doctors who performed particularly well on the knowledge test to significantly underestimate their score whilst lower performers significantly overestimated their score. However OSCE results did not support this pattern. Perhaps it was easier for doctors to predict their own performance on a machine marked test of knowledge rather than a practical skills test that is marked by an assessor. Further, it is likely that they were unfamiliar with tests that are designed to assess minimum competence and may have assumed the threshold for good performance to be higher than it actually was. From a previous study that we recently conducted, we know that many in this cohort of doctors volunteered to sit a ToC in preparation for their forthcoming postgraduate exams. Perhaps a lack of confidence explains the underestimation of OSCE scores that most doctors showed. Alternatively some authors share the view that people will

always be unable to accurately estimate their performance and that this is a methodologically flawed approach to measuring peoples' self-perception.[14-17] Whilst actual performance is poorly correlated with self-ratings (score prediction), there is evidence to suggest it better correlates with behavioural measures. Several studies have shown that when behavioural measures are used, people demonstrate better awareness of their own performance. Psychology students showed awareness of the limits of their knowledge by spending longer on questions they were unsure about and avoiding answering questions they knew they would get incorrect. [15] A study with medical students who took a qualifying exam reported similar findings. [17] Candidates' self-monitoring was measured according to time taken to respond to each question, the number of questions flagged for further consideration and the likelihood of changing their initial answer. This study found that high performers demonstrated better self-monitoring than poorer performers on the exam.[17] Following this evidence, there are recommendations to pursue this line of research approach instead of asking people to estimate their own exam scores. [14-17]

The gender and ethnic differences found in this study support that of previous findings to an extent. Women tend to underestimate and men tend to overestimate their performance in medicine. [21,23-25] Our findings supported this pattern on knowledge test performance but not on OSCE performance. White medical students consistently outperform non-white medical students in the UK, [26] the US, [27-31] and in other English speaking countries.[32,33] In this study white doctors performed substantially higher than non-whites on the OSCE but not on the knowledge test. Further, there was an interaction between ethnicity and PMQ region. OSCE performance was higher in white UK trained doctors than white non-UK trained as well as non-white UK trained doctors.

Implications of findings:

Overall, most doctors did not appear to have an inflated view of their exam performance. It is reasonable to assume that doctors who are not overly confident are likely to exercise more caution in their clinical practice than those who are over confident. However, roughly half of the sample did overestimate their knowledge test

performance. This is potentially a problem as overconfidence in doctors is associated with poor clinical judgement and decision making.[34, 35]. Furthermore, research has shown that overconfidence in medicine is more likely to be a male rather than female characteristic.[21,23] Women are more likely to perceive themselves and be perceived by others as less confident in clinical knowledge and skills.[21,23]. This pattern was found in our study, despite women out performing men on both the knowledge test and OSCE. In practice, lack of confidence may disadvantage female doctors with patient interaction and career progression.[23] A common pattern of ethnic differences was also found with white doctors outperforming non-white doctors on the OSCE. The reasons for this performance gap are unclear but cross-cultural differences in communication styles may explain some of the variation in performance on OSCE type exams; assessors may also have been influenced by ethnic stereotypes.[36] This performance gap is in line with the recent controversy around higher rates of failure amongst international medical graduates taking the clinical skills assessment portion of the MRCGP exam. [37] Further research is necessary to understand why these ethnic differences persist in medicine and what can be done to reduce this discrepancy.[26, 36, 38]

Medical education could facilitate the development of doctors' accurate self-perception by including formal training on the biases that affect the self-perception of all individuals.[16, 34, 35] Doctors would learn about the inherent heuristics they are likely to use when reflecting on strengths and weaknesses of their performance [34, 35]. Medical educators should also establish how feedback can be delivered in a way that is likely to be internalised to encourage the necessary behavioural changes. One potential benefit of revalidation is that it will enforce doctors to reflect on their clinical knowledge and practice as well as identify areas for improvements.[1] This process may prove to be a good opportunity for doctors to become more self-aware. This remains to be seen in future research once the current round of revalidation ends in 2016. In practice, the revalidation process is interested in doctors' awareness and monitoring of the limits of their knowledge and clinical skills, rather than their ability to accurately predict their assessment scores. Therefore an understanding of the universal biases that affect self-perception, coupled with appropriate behaviour changes in response to feedback is likely to improve doctors' self-perception and capacity to self-monitor.

Strengths and weaknesses of the study:

This study lends further support to the literature that suggests doctors have limited ability to estimate their exam performance, even when the exams are in a familiar format (SBA and OSCE). It is one of the first studies to look at how well a group of doctors think they perform on a set of voluntary exams that have potentially significant consequences. The large sample means that it has greater power than previous similar studies. The tests included were part of a validation process for the GMC's fitness to practice procedures. Therefore the study has practical relevance, as demonstrated by doctors seeking feedback on their performance and commenting that they used the process as future exam preparation. There is concern that perhaps only those who thought they had performed well on the exams would have participated in this study, thus introducing selection bias into the sample. However, most of the doctors who were invited to participate did so (76%), and we know from the results that most doctors did not have an inflated view of their exam performances. A limitation of this study is that no measures of behaviour were included that could have demonstrated the extent to which doctors monitor their performance in a given clinical situation. We recognise the value in this alternative approach for extending the present findings in future research. However, in the case of this study, the doctors primarily volunteered to take a ToC than to be in a research study. For this reason, a questionnaire asking for self-estimated scores on the exams they had just taken was a feasible way for us to obtain data on this topic. The results are not necessarily generalisable to all doctors and may have differed had there been equal numbers of doctors by different ethnic background, PMQ regions and seniority. Finally, the majority of doctors performed well on both exams, even those who scored in the bottom percentiles. Therefore patterns between high and lower performers should be interpreted with caution. Those who performed less well than the majority may have struggled because of the type of exam material. However, it is likely that the few who performed lower in comparison to their peers who sat the same exam material, were actual low performers.

Conclusions and future directions

Doctors were more accurate in predicting their knowledge test performance than OSCE performance. High and low performers self-estimated differently on an objective test of knowledge but almost everyone underestimated their performance on a practical skills test (OSCE). Estimated and actual performance differed by gender and ethnicity but less so by where a doctor had gained their PMQ. A follow up to this study may wish to explore in more depth how doctors come to assign themselves a particular score and their reasoning underpinning this judgement. Such data may lead to further understanding of why high performers tend to underestimate their own performance and lower performers overestimate. Anecdotally we know that doctors undergoing FtP investigation often lack sufficient recognition of their problems. Further study on how poor performers in particular can successfully alter their self-perception as a first step towards remediation, is warranted. It will be interesting to monitor the impact revalidation has on the number of complaints to the GMC and whether this formal exercise in self-reflection affords an opportunity for borderline problematic doctors to rectify their deficiencies.

Figure legends

Figure 1: Scatterplot of difference in actual and estimated knowledge test (KT) scores against actual knowledge test scores

Figure 2: Scatterplot of difference in actual and estimated OSCE scores against actual OSCE scores.

Acknowledgments We are grateful to Dr Henry Potts for his statistical support and help with interpreting results.

Funding This work was supported by the General Medical Council.

Competing interests None.

Ethics approval Not required, but study participants consented to their anonymised data being used for research purposes

Contributors LM analysed the results and wrote the manuscript. AS contributed intellectually to the study and helped revise the manuscript drafts. GM designed the research questionnaire and helped recruit participants. YK recruited participants and entered the data. JD oversaw the study and provided intellectual contribution at all phases. All authors provided feedback on manuscript drafts and approved its final version.

Provenance and peer review Not commissioned, externally peer reviewed.

Data sharing statement Due to our confidentiality agreement with the General Medical Council, data is unavailable for sharing.

References

1. General Medical Council. Ready for Revalidation: supporting information for appraisal and revalidation, 2012.

2. General Medical Council. Good Medical Practice, 2012.

3. Hays, R., Jolly, B., Caldon, et al. Is insight important? Measuring capacity to change performance. *Medical Education*, 2002. 36: p. 965–971.

4. Cross, P., Not can but will college teaching be improved? *New Directions for Higher Education*, 1977. 17: p. 1-15.

5. Dunning, D., Heath, C and Suls, J. Flawed self-assessment: implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 2004. 5(3): p. 69-106.

6. Ehrlinger, J., Johnson, K., Banner, M., et al. Why the unskilled are unaware: further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, 2008. 105: p. 98-121.

7. Kruger, J. and Dunning, D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 1999. 77(6): p. 1121-1134.

8. Marottoli, R. and Richardson, E. Confidence in, and self-rating of, driving ability among older drivers. *Accident Analysis and Prevention*, 1998. 30: p. 331–336.

9. Rutter, D., Quine, L and Albery, I. Perceptions of risk in motorcyclists: unrealistic optimism, relative realism and predictions of behaviour. *British Journal of Psychology,* 1998. 89: p. 681–696.

10. Zenger, T. Why do employers only reward extreme performance? Examining the relationships among performance, pay, and turnover. *Administrative Science Quarterly*, 1992. 37: p. 198–219.

11. Tracey, J., Arroll, B., Barham, P and Richmond, D. The validity of general practitioners' self assessment of knowledge: cross sectional study. *British Medical Journal,* 1997. 315: p. 1426.

12. Hodges, B., Regehr, G and Martin, D. Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it. *Academic Medicine*, 2001. 76(10): p. S87-89.

13. Davis, D., Mazmanian, PE., Fordis, M., et al. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *The Journal of the American Medical Association*, 2006. 296: p. 1094-1102.

14. Ward, M., Gruppen, L and Regehr, G. Measuring self-assessment: current state of the art. *Advances in Health Sciences Education*, 2002. 7: p. 63-80.

15. Eva, K. and Regehr, G. Knowing when to look it up: a new conception of self-assessment ability. *Academic Medicine*, 2007. 82(Supplement 10): p. S81-S84.

16. Eva, K. and Regehr, G. "I'll never play professional football" and other fallacies of self-assessment. *Journal of Continuing Education in the Health Professions*, 2008. 28(1): p. 14-19.

17. McConnell, M., Regehr, G., Wood, T, et al. Self-monitoring and its relationship to medical knowledge. *Advances in Health Sciences Education*, 2012. 17: p. 311-323.

18. General Medical Council. The State of Medical Education and Practice in the UK, 2012.

19. Dunning, D., Johnson, K., Ehrlinger, et al. Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, 2003. 12 (3): p. 83-87.

20. Barnsley, L., Lyon, P., Ralston, S., et al. Clinical skills in junior medical officers: a comparison of self-reported confidence and observed competence. *Medical Education*, 2004. 38(4): p. 358-367.

21. Blanch-Hartigan, D., Medical students' self-assessment of performance: results from three meta-analyses. *Patient Education and Counseling*, 2011. 84(1): p. 3-9.

22. Tousignant, M. and Desmarchais, J. Accuracy of student self-assessment ability compared to their own performance in a problem-based learning medical program: a correlation study. *Advances in Health Sciences* 2002. 7: p. 19-27.

23. Blanch, D., Hall, J., Roter, D., et al. Medical student gender and issues of confidence. *Patient Education and Counseling*, 2008. 72(3): p. 374-381.

24. Minter, R., Gruppen, L., Napolitano, et al. Gender differences in the self-assessment of surgical residents. *The American Journal of Surgery*, 2005. 189: p. 647-650.

25. Rees, C. and Shepherd, M. Students' and assessors' attitudes towards students' self-assessment of their personal and professional behaviours. *Medical Education*, 2005. 39: p. 30-39.

26. Woolf, K., Potts, H et al. Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. *British Medical Journal*, 2011. 342: p. 901

27. Colliver, J., Swartz, M and Robbs, R. The effect of examinee and patient ethnicity in clinical-skills assessment with standardized patients. *Advances in Health Sciences Education*, 2001. 6: p. 5-13.

28. Dawson, B., Iwamoto, C., Ross, L., et al. Performance on the national board of medical examiners .1. examination by men and women of different race and ethnicity. *Journal of the American Medical Association*, 1994. 272: p. 674-679.

29. Koenig, J., Sireci, S and Wiley, A. Evaluating the predictive validity of MCAT scores across diverse applicant groups. *Academic Medicine*, 1998. 73: p. 1095-1106.

30. Veloski, J., Callahan, C., Xu, G., et al. Prediction of students' performances on licensing examinations using age, race, sex, undergraduate GPAs, and MCAT scores. *Academic Medicine*, 2000. 75: p. S28-S30.

31. Xu, G., Veloski, J., Hojat, M., et al. Longitudinal comparison of the academic performances of Asian-American and white medical students. *Academic Medicine*, 1993. 68(1): p. 82-86.

32. Kay-Lambkin, F., Pearson, S and Rolfe, I. The influence of admissions variables on first year medical school performance: a study from Newcastle University, Australia. *Medical Education*, 2002. 36: p. 154-159.

33. Liddell, M. and Koritsas, S. Effect of medical students' ethnicity on their attitudes towards consultation skills and final year examination performance. *Medical Education*, 2004. 38: p. 187-198.

34. Croskerry, P. Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Academic Emergency Medicine*, 2002. 9: p. 1184-1204.

35. Croskerry, P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*, 2003. 78(8): p. 775-780.

36. Woolf, K., Cave, J., Greenhalgh, et al. Ethnic stereotypes and the underachievement of UK medical students from ethnic minorities: qualitative study. *British Medical Journal*, 2008. 337.

37. Kaffash, J. RCGP faces legal threat over international GP trainee failure rates, in Pulse 2012.

38. Woolf, K., McManus, IC., Potts, et al. The mediators of minority ethnic underperformance in final medical school examinations. *British Journal of Educational Psychology*, 2011. 83(1): p. 135-159.

How well do doctors think they perform on the General Medical Council's Tests of Competence pilot exams? A cross-sectional study

Leila Mehdizadeh[1], Alison Sturrock[2], Gil Myers[3], Yasmin Khatib[4], Jane Dacre[5]

[1] Research Associate in Medical Education, Division of UCL Medical School, University College London, Royal Free Hospital, room GF/664, Hampstead, NW3 2PF.

[2] Senior Lecturer in Medical Education, Division of UCL Medical School, University College London, Royal Free Hospital, room GF/664, Hampstead, NW3 2PF.

[3] Honorary Clinical Teaching Fellow, Division of UCL Medical School, University College London, Royal Free Hospital, room GF/664, Hampstead, NW3 2PF.

[4] Lecturer in Mental Health, Barts and the London School of Medicine and Dentistry, Centre for Psychiatry, Queen Mary University of London, Old Anatomy Building, Charter House Square, London EC1M 6BQ.

[5] Professor of Medical Education, Division of UCL Medical School, University College London, Medical School Building, 74 Huntley Street, London WC1E 6AU.

Corresponding author: Leila Mehdizadeh, l.mehdizadeh@ucl.ac.uk, Tel 020 31089218, UCL Medical School, Royal Free Hospital, room GF/664, Hampstead, NW3 2PF.

Abstract

Objective: To investigate how accurately doctors estimated their performance on the General Medical Council's Tests of Competence pilot exams.

Design: A cross sectional survey design using a questionnaire method.

Setting: University College London Medical School.

Participants: 524 medical doctors working in a range of clinical specialties between foundation year two and consultant level.

Main outcome measures: Estimated and actual total scores on a knowledge test and Observed Structured Clinical Examination (OSCE).

Results: The pattern of results for OSCE performance differed to results for knowledge test performance. The majority of doctors significantly underestimated their OSCE performance. Whereas estimated knowledge test performance differed between high and low performers. Those who did particularly well significantly underestimated their knowledge test performance ($t(196) = -7.70$, $p<0.01$) and those who did less well significantly overestimated ($t(172) = 6.09$, $p<0.01$). There were also significant differences between estimated and/or actual performance by gender, ethnicity and region of Primary Medical Qualification.

Conclusions: Doctors were more accurate in predicating their knowledge test performance than their OSCE performance. The association between estimated and actual knowledge test performance support the established differences between high and low performers described in the behavioural sciences literature. This was not the case for the OSCE. The implications of the results to the revalidation process are discussed.

Article summary

Article focus

- Revalidation has a strong component of self-evaluation. Doctors must provide evidence at their annual appraisal that they continue to meet the standards set out in Good Medical Practice.
- Evidence from the behavioural sciences has demonstrated the deficits in the ability to accurately self-assess one's competencies.
- The purpose of this study was to investigate how accurately doctors estimate their performance on the General Medical Councils Tests of Competence pilot exams.

Key messages

- The literature shows that on tests that discriminate between excellent and poor performance, high performers tend to underestimate their performance whilst low performers overestimate.
- In this study, high performers significantly underestimated their knowledge test scores whilst lower performers significantly overestimated. In contrast, most doctors significantly underestimated their OSCE performance.
- We need to consider how doctors can be trained to develop accurate self-perception, particularly as revalidation gets underway.

Strengths and limitations of this study

- This is one of the first studies to look at how well a group of doctors think they perform on a set of exams that have potentially significant real world consequences.
- The large sample means that it has greater power than previous similar studies.
- The results have important implications to the revalidation process of doctors working in the UK.
- The majority of doctors performed well on both exams, therefore patterns between high and lower performers should be interpreted with caution.
- Results are not necessarily applicable to the wider medical community in the UK.

Introduction

The revalidation process of all UK doctors who hold a licence to practise is now underway. Introduced by the General Medical Council (GMC), doctors must provide evidence at their annual appraisal that they continue to meet the standards set out in Good Medical Practice.[1,2] This requires the doctor to reflect on their own knowledge and practise to be able to demonstrate their strengths and areas that need further development. Whilst medical education in the United Kingdom (UK) aims to produce doctors who are reflective practitioners, [2,3] evidence from the behavioural sciences has demonstrated the deficits in the ability to assess one's own competencies.[4-10] Previous studies showed that psychology students who performed lowest on intellectual and social tasks displayed the least insight by over estimating their own performance. In contrast, the highest performers underestimated their performance.[7] This pattern has been replicated in the medical context.

General practitioners were unable to accurately assess their knowledge of 20 typical clinical conditions.[11] Family medicine residents who performed best on a breaking bad news scenario were more likely to make accurate self-estimates than the lowest performers.[12] A systematic review of studies on the accuracy of doctors' self-assessment compared with objective measures of competence showed that doctors who performed the least well also self-assessed the least well. These studies indicate that people in general as well as medical doctors specifically, have a limited ability to accurately assess their own competence.[13] In clinical practice, the idea that a poorly performing doctor could also lack awareness of their problems is concerning. Such a doctor may pose a risk to patient safety and may not engage in appropriate professional development activities. With the introduction of revalidation, the number of doctors referred to the GMC for Fitness to Practise (FtP) investigation may change. Revalidation has a strong component of self-evaluation and enforced reflection will offer doctors more opportunities to consider their performance and ways to remedy any weak areas.

There is a debate in the self-assessment literature concerning use of terminology and outcome measures. One perspective that has gained more recent attention is the necessity to distinguish self-assessment from self-monitoring as different approaches to the investigation of people's perception of their personal competence. [14-17] The self-assessment approach investigates how well individuals can judge their personal competence against an objective measure of competence.[16] Alternatively, the self-monitoring approach is interested in the extent that people show awareness of the limits in their competence during a given situation, and this can be measured according to an individual's behaviour e.g. taking longer to think about a question of which they are unsure.[16] It is important to clarify that this study looks at doctors' self-assessment ability after completing a set of exams, not how well they can monitor their performance during the exams. Therefore this study does not include any outcomes of doctors' behaviour.

We measured the accuracy of doctors' self-predicted performance on the GMC's Tests of Competence (ToC) pilot exams. Tests of competence are used by the GMC to assess poorly performing doctors under FtP investigation. Before implementation, test content is piloted on volunteer doctors who have no known FtP concerns. Doctors volunteer to take a knowledge test and an Observed Structured Clinical Examination (OSCE) in their relevant specialty. The written paper consists of a single best answer knowledge test which is machine marked. The OSCE is marked by trained assessors using a generic domain based mark scheme of 'acceptable', 'cause for concern' and 'unacceptable'. In the self-assessment literature cited previously, [6-13] the tests that participants were asked to predict their scores on differed in three key ways from the tests used in the present study. They had been designed specifically for research purposes, to discriminate excellent from poor performance and the results had no real world consequences. This was not the case in the present study. Participants were assessed on tests that were established for the purpose of FtP investigations, they are tailored to assess the expected minimum level of competence of a practising doctor, and there were potentially real world consequences if an individual's performance failed to meet the minimum standard.

The purpose of this study was to investigate how well doctors think they perform on the GMC's ToC pilot exams. In particular we wondered whether the established

differences in the literature between high and low performers would emerge; do those who perform well underestimate their performance and do those who perform less well overestimate? We were also interested in whether differences in self-estimates existed by gender, ethnic background and Primary Medical Qualification (PMQ) region.

Methods

Ethics: We received written confirmation from University College London's Research Ethics Committee in October 2008 that the study was exempt from ethical approval. However study participants explicitly consented to their data being used anonymously for research purposes.

Design: This was a cross sectional study using a questionnaire method.

Sample: Doctors who volunteered to take a GMC ToC pilot exam between June 2011 and July 2012 were invited to participate in this study. Volunteers for the pilot exams were recruited through advertisement in medical journals, specialty specific newsletters and word of mouth. The study sample included doctors that worked in paediatrics, child psychiatry, anaesthetics, old age psychiatry, forensic psychiatry, general medicine, emergency medicine, orthopaedics, general practice, obstetrics and gynaecology, surgery, cardiology, radiology and care of the elderly. They ranged from foundation year two to consultant level.

Materials: A study specific questionnaire was designed to obtain participants' estimated knowledge test and OSCE scores (Appendix 1). The questionnaire asked where participants thought they ranked in comparison to other doctors who completed the exams on the same day, in comparison to all doctors who were eligible to sit the GMC test of competence pilot exams as well as an estimation of their total knowledge test and OSCE scores.

Outcome measures: We compared participants' self-estimated and actual total scores on the knowledge test and OSCE. The knowledge tests consisted of 120 specialty specific items in a single best answer (SBA) format with a maximum score

of 120. The OSCE included 12 specialty specific stations. Each station was scored by a trained assessor who was usually a consultant in the relevant specialty or a clinical skills nurse. The maximum score for the OSCE was 480.

Procedure: Doctors who volunteered to take a GMC pilot exam between June 2011 and July 2012 were invited to participate in this study. Once volunteer doctors had completed the knowledge test and OSCE, the study questionnaire was distributed by GM who was a facilitator at the piloting events. Doctors were briefed about the purpose of the study and how their data would be used. They were assured that the completion of the questionnaire was voluntary and would only take 5-10 minutes.

Analyses: Actual and estimated exam scores were compared using SPSS for windows version 19. We split estimated and actual scores into tertiles (top, middle, bottom) to see whether differences existed between high and low performers. Results were analysed using descriptive statistics, correlations, t-tests and ANOVAs.

Results

Between June 2011-July 2012, 689 doctors volunteered to take a pilot ToC and 524 participated in present the study (76% rate of participation). During this period, most were junior and middle grade doctors who qualified in the UK. As compared with all doctors on the 2011 List of Registered Medical Practitioners (LRMP) [18] men were under represented and women were over represented (Table 1). There was also a higher proportion of Asian/Asian British doctors in this study and overseas trained doctors were under represented (Table 1).

Table 1: Demographic characteristics of sample compared to demographics of doctors on 2011 LRMP

Variable	Levels	Total in this study (N=524)	Total on LRMP 2011 (N=245,903)
Gender	Male	231 (44.1%)	141,369 (57%)
	Female	293 (55.9%)	104,534 (43%)
Ethnicity	White	266 (50.8%)	118,822 (48%)
	Black/Black British	17 (3.2%)	6,812 (2.8%)
	Asian/Asian British	147 (28.1%)	46,664 (4.3%)
	Mixed	25 (4.8%)	3,643 (1.5%)
	Other ethnic groups	20 (3.8%)	9,002 (3.7%)
	Not stated (includes prefer not to say)	49 (9.4%)	60,960 (25%)
Primary Medical Qualification region	UK	408 (77.9%)	155,264 (63%)
	EU country	22 (4.2%)	24,031 (10%)
	Non-EU country	94 (17.9%)	66,608 (27%)

General patterns

Overall, participants were more accurate in predicting their total knowledge test score than their OSCE score. There was a moderately strong positive relationship between the difference in estimated and actual knowledge test scores with actual scores; $r=0.43$, $p<0.01$. There was a roughly equal distribution of participants that over and underestimated their knowledge test scores (Figure 1). Those who overestimated (negative numbers on the y axis) tended to score lower than those who underestimated their knowledge test scores. There was no significant difference between the estimated and actual scores on the knowledge test; $t(521) = 1.33$, $p=0.19$.

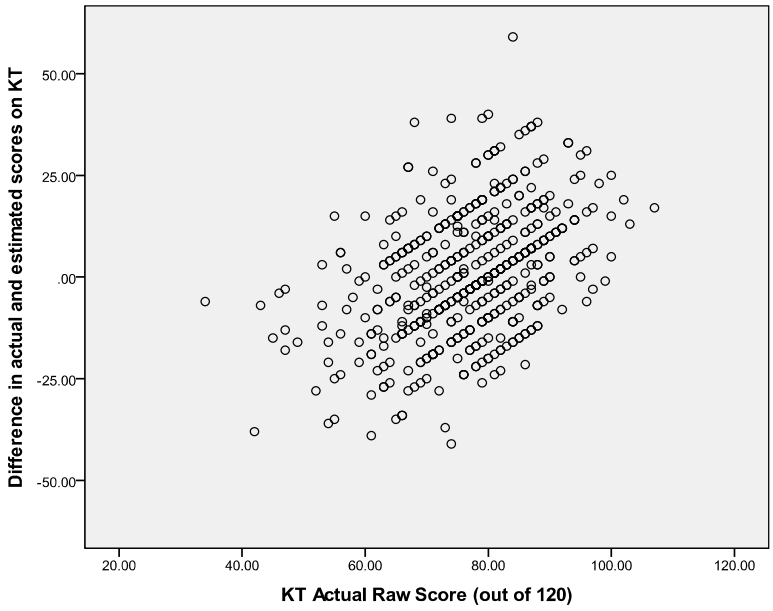


Figure 1: Scatterplot of difference in actual and estimated knowledge test (KT) scores against actual knowledge test scores

There was also an association between difference in actual and estimated scores for the OSCE; r=0.33, p<0.01. The vast majority of doctors underestimated their OSCE performance (Figure 2). The few who overestimated their OSCE scores performed less well than those who underestimated their OSCE performance (Figure 2). There was a significant difference between the estimated and actual total OSCE scores; t (520 = -37.76, p<0.01).

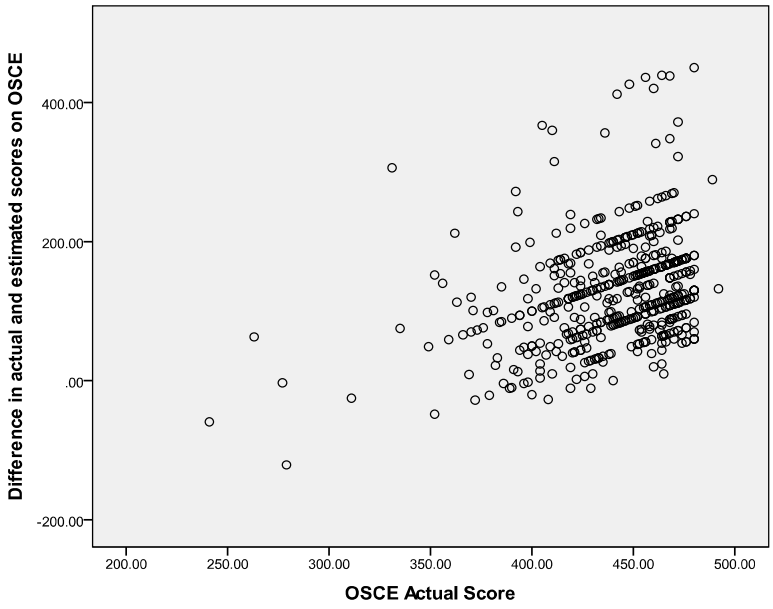


Figure 2: Scatterplot of difference in actual and estimated OSCE scores against actual OSCE scores.

Differences between high and lower performers on their estimated and actual exam scores

There were significant differences between high and low performers on their estimated knowledge test performance. The highest performers significantly underestimated their knowledge test scores by an average of 8 marks (t (196) = -7.70, p<0.01) and the lower performers significantly overestimated by an average of 7 marks (t (172) = 6.09, p<0.01).

Both high and lower performers significantly underestimated their OSCE performance; t (180) = -26.28, p<0.01 and t (172) = -16.20, p<0.01 respectively. Those in the top percentile underestimated their OSCE performance to the greatest extent.

Gender differences on estimated and actual exam scores

Men predicted a higher knowledge test score than women, with mean estimates of 79 (SD 15) and 74 (SD 14) respectively. Levene's test confirmed there was homogeneity of variance and the t-test revealed that this gender difference in mean estimates was significant; t(522) = 4,27, p=<0.01. Women performed slightly better

on the knowledge test than men but their means scores were not significantly different; 77 (SD 11) and 76 (SD 10) respectively. Men also predicted a higher overall OSCE performance than women with mean estimates of 329 (SD 59) and 300 (SD 81) respectively. This was a significant difference; $t(516) = 4.65$, $p<0.01$. Women outperformed men on OSCE performance and the difference was significant; $t(482) = -2.82$, $p<0.01$. A three way ANOVA showed that there was a significant effect of gender on estimated knowledge test ($F(1,508) = 4.62$, $p=0.03$) and OSCE performance ($F(1,505)=11.74$, $p<0.01$). This means that men, irrespective of their ethnicity or PMQ region, estimated higher than women on both exams.

Ethnic differences on estimated and actual exam scores

There were no significant differences in estimated exam scores between doctors of different ethnic backgrounds. However there was a tendency for Asian/Asian British doctors to estimate highest on their knowledge test performance (M=78, SD=15). The highest estimated OSCE performance came from white doctors (M=317, SD=73) and 'other' ethnic groups (M=316, SD=59).

A one way ANOVA confirmed that there were significant differences by ethnic background on actual knowledge test performance ($F=(5,516) = 4.46$, $p<0.01$) and OSCE performance ($F=(5,518)=11.48$, $p<0.01$). Fisher's least significant difference test was used to explore where these differences occurred. Doctors who did not specify their ethnicity performed highest on the knowledge test (mean=78, SD=11). Asian/Asian British doctors scored lowest on the knowledge test, particularly in comparison to white doctors ($p<0.01$) and those who did not state their ethnicity ($p=0.02$). Table 2 shows that white doctors outperformed all other ethnic groups on OSCE performance.

Table 2: Differences in OSCE performance between white and other ethnic groups

Ethnicity		OSCE performance	Significance
White (M=451,SD=29)	Not stated	M=440, SD=23	p=0.04
	Other	M=439, SD=27	p=0.12*
	Mixed	M=433, SD=39	p=0.08*
	Black/Black British	M=428, SD=33	p=0.03
	Asian/Asian British	M=427, SD=37	p<0.01

*Non-significant difference

Differences by primary medical qualification region on estimated and actual exam scores

The majority of doctors gained their PMQ from the UK (78%). Of the non- UK trained doctors, 18% were from a non-EU county and 4% were from an EU country. There were significant differences in estimated knowledge test performance between doctors of different PMQ regions. Non-UK trained doctors estimated significantly higher than UK trained doctors ($F(2,521)=6.06, p<0.01$). However there were no actual differences in knowledge test performance between doctors of different PMQ regions ($F(2,519)=2.28$, $p=0.10$). A reverse pattern was true of OSCE performance. Estimated OSCE performance did not differ by PMQ region, although EU trained doctors tended to make the highest estimates. Actual OSCE performance did significantly differ, with UK trained doctors outperforming non-UK trained doctors ($F(5,521)=37.96$, $p<0.01$).

Discussion

Principal findings:

In general study participants performed well on the knowledge test and OSCE (usually 70% and above) so the number of actual low performers was small. They were more accurate in predicting their knowledge test performance than OSCE performance. Differences in predictions between high and low performers were found on the knowledge test but not on the OSCE. In keeping with previous literature [6-13] high performers significantly underestimated their knowledge test performance whilst lower performers significantly overestimated. Most doctors significantly underestimated their OSCE performance irrespective of how well they actually did. Differences between estimated and actual performance were apparent between men and women. On both exams, women's estimated performance was lower than men's but they actually did better than men, particularly on OSCE performance. Estimated performance on both exams did not significantly differ between ethnicities, but there was a tendency for Asian/Asian British participants to estimate slightly higher than other groups. Actual performance for both exams did significantly differ by ethnic group. Doctors of white and unspecified ethnicity performed highest on the knowledge test, while white doctors outperformed all others on OSCE performance. UK trained doctors outperformed overseas trained doctors on OSCE performance but there were no differences by PMQ region on knowledge test performance.

Findings in relation to literature:

Our results are in line with the literature that demonstrates the limited ability people have, including doctors, to accurately self-assess their performance.[5-13,19] Further, this study provides support for previously reported patterns between high and low performers.[7,13,20-22] There was a tendency for doctors who performed particularly well on the knowledge test to significantly underestimate their score whilst lower performers significantly overestimated their score. However OSCE results did not support this pattern. Perhaps it was easier for doctors to predict their own performance on a machine marked test of knowledge rather than a practical skills test that is marked by an assessor. Further, it is likely that they were unfamiliar

with tests that are designed to assess minimum competence and may have assumed the threshold for good performance to be higher than it actually was. From a previous study that we recently conducted, we know that many in this cohort of doctors volunteered to sit a ToC in preparation for their forthcoming postgraduate exams. Perhaps a lack of confidence explains the underestimation of OSCE scores that most doctors showed. Alternatively some authors share the view that people will always be unable to accurately estimate their performance and that this is a methodologically flawed approach to measuring peoples' self-perception.[14-17] Whilst actual performance is poorly correlated with self-ratings (score prediction), there is evidence to suggest it better correlates with behavioural measures. Several studies have shown that when behavioural measures are used, people demonstrate better awareness of their own performance. Psychology students showed awareness of the limits of their knowledge by spending longer on questions they were unsure about and avoiding answering questions they knew they would get incorrect. [15] A study with medical students who took a qualifying exam reported similar findings. [17] Candidates' self-monitoring was measured according to time taken to respond to each question, the number of questions flagged for further consideration and the likelihood of changing their initial answer. This study found that high performers demonstrated better self-monitoring than poorer performers on the exam.[17] Following this evidence, there are recommendations to pursue this line of research approach instead of asking people to estimate their own exam scores. [14-17]

The gender and ethnic differences found in this study support that of previous findings to an extent. Women tend to underestimate and men tend to overestimate their performance in medicine. [21,23-25] Our findings supported this pattern on knowledge test performance but not on OSCE performance. White medical students consistently outperform non-white medical students in the UK, [26] the US, [27-31] and in other English speaking countries.[32,33] In this study white doctors performed substantially higher than non-whites on the OSCE but not on the knowledge test. Further, there was an interaction between ethnicity and PMQ region. OSCE performance was higher in white UK trained doctors than white non-UK trained as well as non-white UK trained doctors.

Implications of findings:

Overall, most doctors did not appear to have an inflated view of their exam performance. It is reasonable to assume that doctors who are not overly confident are likely to exercise more caution in their clinical practice than those who are over confident. However, roughly half of the sample did overestimate their knowledge test performance. This is potentially a problem as overconfidence in doctors is associated with poor clinical judgement and decision making.[34, 35]. Furthermore, research has shown that overconfidence in medicine is more likely to be a male rather than female characteristic.[21,23] Women are more likely to perceive themselves and be perceived by others as less confident in clinical knowledge and skills.[21,23]. This pattern was found in our study, despite women out performing men on both the knowledge test and OSCE. In practice, lack of confidence may disadvantage female doctors with patient interaction and career progression.[23] A common pattern of ethnic differences was also found with white doctors outperforming non-white doctors on the OSCE. The reasons for this performance gap are unclear but cross-cultural differences in communication styles may explain some of the variation in performance on OSCE type exams; assessors may also have been influenced by ethnic stereotypes.[36] This performance gap is in line with the recent controversy around higher rates of failure amongst international medical graduates taking the clinical skills assessment portion of the MRCGP exam. [37] Further research is necessary to understand why these ethnic differences persist in medicine and what can be done to reduce this discrepancy.[26, 36, 38]

Medical education could facilitate the development of doctors' accurate self-perception by including formal training on the biases that affect the self-perception of all individuals.[16, 34, 35] Doctors would learn about the inherent heuristics they are likely to use when reflecting on strengths and weaknesses of their performance [34, 35]. Medical educators should also establish how feedback can be delivered in a way that is likely to be internalised to encourage the necessary behavioural changes. One potential benefit of revalidation is that it will enforce doctors to reflect on their clinical knowledge and practice as well as identify areas for improvements.[1] This process may prove to be a good opportunity for doctors to become more self-aware. This remains to be seen in future research once the current round of revalidation

ends in 2016. In practice, the revalidation process is interested in doctors' awareness and monitoring of the limits of their knowledge and clinical skills, rather than their ability to accurately predict their assessment scores. Therefore an understanding of the universal biases that affect self-perception, coupled with appropriate behaviour changes in response to feedback is likely to improve doctors' self-perception and capacity to self-monitor.

Strengths and weaknesses of the study:

This study lends further support to the literature that suggests doctors have limited ability to estimate their exam performance, even when the exams are in a familiar format (SBA and OSCE). It is one of the first studies to look at how well a group of doctors think they perform on a set of voluntary exams that have potentially significant consequences. The large sample means that it has greater power than previous similar studies. The tests included were part of a validation process for the GMC's fitness to practice procedures. Therefore the study has practical relevance, as demonstrated by doctors seeking feedback on their performance and commenting that they used the process as future exam preparation. There is concern that perhaps only those who thought they had performed well on the exams would have participated in this study, thus introducing selection bias into the sample. However, most of the doctors who were invited to participate did so (76%), and we know from the results that most doctors did not have an inflated view of their exam performances. A limitation of this study is that no measures of behaviour were included that could have demonstrated the extent to which doctors monitor their performance in a given clinical situation. We recognise the value in this alternative approach for extending the present findings in future research. However, in the case of this study, the doctors primarily volunteered to take a ToC than to be in a research study. For this reason, a questionnaire asking for self-estimated scores on the exams they had just taken was a feasible way for us to obtain data on this topic. The results are not necessarily generalisable to all doctors and may have differed had there been equal numbers of doctors by different ethnic background, PMQ regions and seniority. Finally, the majority of doctors performed well on both exams, even those who scored in the bottom percentiles. Therefore patterns between high and lower performers should be interpreted with caution. Those who performed less well than

the majority may have struggled because of the type of exam material. However, it is likely that the few who performed lower in comparison to their peers who sat the same exam material, were actual low performers.

Conclusions and future directions

Doctors were more accurate in predicting their knowledge test performance than OSCE performance. High and low performers self-estimated differently on an objective test of knowledge but almost everyone underestimated their performance on a practical skills test (OSCE). Estimated and actual performance differed by gender and ethnicity but less so by where a doctor had gained their PMQ. A follow up to this study may wish to explore in more depth how doctors come to assign themselves a particular score and their reasoning underpinning this judgement. Such data may lead to further understanding of why high performers tend to underestimate their own performance and lower performers overestimate. Anecdotally we know that doctors undergoing FtP investigation often lack sufficient recognition of their problems. Further study on how poor performers in particular can successfully alter their self-perception as a first step towards remediation, is warranted. It will be interesting to monitor the impact revalidation has on the number of complaints to the GMC and whether this formal exercise in self-reflection affords an opportunity for borderline problematic doctors to rectify their deficiencies.

Acknowledgments We are grateful to Dr Henry Potts for his statistical support and help with interpreting results.

Funding This work was supported by the General Medical Council.

Competing interests None.

Ethics approval Not required, but study participants consented to their anonymised data being used for research purposes

Contributors LM analysed the results and wrote the manuscript. AS contributed intellectually to the study and helped revise the manuscript drafts. GM designed the research questionnaire and helped recruit participants. YK recruited participants and entered the data. JD oversaw the study and provided intellectual contribution at all phases. All authors provided feedback on manuscript drafts and approved its final version.

Provenance and peer review Not commissioned, externally peer reviewed.

Data sharing statement Due to our confidentiality agreement with the General Medical Council, data is unavailable for sharing.

References

1. General Medical Council. Ready for Revalidation: supporting information for appraisal and revalidation, 2012.

2. General Medical Council. Good Medical Practice, 2012.

3. Hays, R., Jolly, B., Caldon, L., McCrorie, P., McAvoy, P., McManus, IC and Rehans, JJ. Is insight important? Measuring capacity to change performance. *Medical Education*, 2002. 36: p. 965–971.

4. Cross, P., Not can but will college teaching be improved? *New Directions for Higher Education*, 1977. 17: p. 1-15.

5. Dunning, D., Heath, C and Suls, J. Flawed self-assessment: implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 2004. 5(3): p. 69-106.

6. Ehrlinger, J., Johnson, K., Banner, M., Dunning, D and Kruger, J. Why the unskilled are unaware: further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, 2008. 105: p. 98-121.

7. Kruger, J. and Dunning, D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 1999. 77(6): p. 1121-1134.

8. Marottoli, R. and Richardson, E. Confidence in, and self-rating of, driving ability among older drivers. *Accident Analysis and Prevention*, 1998. 30: p. 331–336.

9. Rutter, D., Quine, L and Albery, I. Perceptions of risk in motorcyclists: unrealistic optimism, relative realism and predictions of behaviour. *British Journal of Psychology,* 1998. 89: p. 681–696.

10. Zenger, T. Why do employers only reward extreme performance? Examining the relationships among performance, pay, and turnover. *Administrative Science Quarterly*, 1992. 37: p. 198–219.

11. Tracey, J., Arroll, B., Barham, P and Richmond, D. The validity of general practitioners' self assessment of knowledge: cross sectional study. *British Medical Journal,* 1997. 315: p. 1426.

12. Hodges, B., Regehr, G and Martin, D. Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it. *Academic Medicine*, 2001. 76(10): p. S87-89.

13. Davis, D., Mazmanian, PE., Fordis, M., Van Harrison., Thorpe KE and Perrier, L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *The Journal of the American Medical Association*, 2006. 296: p. 1094-1102.

14. Ward, M., Gruppen, L and Regehr, G. Measuring self-assessment: current state of the art. *Advances in Health Sciences Education*, 2002. 7: p. 63-80.

15. Eva, K. and Regehr, G. Knowing when to look it up: a new conception of self-assessment ability. *Academic Medicine*, 2007. 82(Supplement 10): p. S81-S84.

16. Eva, K. and Regehr, G. "I'll never play professional football" and other fallacies of self-assessment. *Journal of Continuing Education in the Health Professions*, 2008. 28(1): p. 14-19.

17. McConnell, M., Regehr, G., Wood, T and Eva, K. Self-monitoring and its relationship to medical knowledge. *Advances in Health Sciences Education*, 2012. 17: p. 311-323.

18. General Medical Council. The State of Medical Education and Practice in the UK, 2012.

19. Dunning, D., Johnson, K., Ehrlinger, Kruger, J. Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, 2003. 12 (3): p. 83-87.

20. Barnsley, L., Lyon, P., Ralston, S., Hibbert, E., Cunningham, I., Gordon, F., and Field, M. Clinical skills in junior medical officers: a comparison of self-reported confidence and observed competence. *Medical Education*, 2004. 38(4): p. 358-367.

21. Blanch-Hartigan, D., Medical students' self-assessment of performance: results from three meta-analyses. *Patient Education and Counseling*, 2011. 84(1): p. 3-9.

22. Tousignant, M. and Desmarchais, J. Accuracy of student self-assessment ability compared to their own performance in a problem-based learning medical program: a correlation study. *Advances in Health Sciences* 2002. 7: p. 19-27.

23. Blanch, D., Hall, J., Roter, D., Frankel, R. Medical student gender and issues of confidence. *Patient Education and Counseling*, 2008. 72(3): p. 374-381.

24. Minter, R., Gruppen, L., Napolitano, K and Gauger, P. Gender differences in the self-assessment of surgical residents. *The American Journal of Surgery*, 2005. 189: p. 647-650.

25. Rees, C. and Shepherd, M. Students' and assessors' attitudes towards students' self-assessment of their personal and professional behaviours. *Medical Education*, 2005. 39: p. 30-39.

26. Woolf, K., Potts, H and McManus, I. Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. *British Medical Journal*, 2011. 342: p. 901

27. Colliver, J., Swartz, M and Robbs, R. The effect of examinee and patient ethnicity in clinical-skills assessment with standardized patients. *Advances in Health Sciences Education*, 2001. 6: p. 5-13.

28. Dawson, B., Iwamoto, C., Ross, L., Nungester., R., Swanson., D and Volle., R. Performance on the national board of medical examiners .1. examination by men and women of different race and ethnicity. *Journal of the American Medical Association*, 1994. 272: p. 674-679.

29. Koenig, J., Sireci, S and Wiley, A. Evaluating the predictive validity of MCAT scores across diverse applicant groups. *Academic Medicine*, 1998. 73: p. 1095-1106.

30. Veloski, J., Callahan, C., Xu, G., Hojat, M and Nash, D. Prediction of students' performances on licensing examinations using age, race, sex, undergraduate GPAs, and MCAT scores. *Academic Medicine*, 2000. 75: p. S28-S30.

31. Xu, G., Veloski, J., Hojat, M., Gonnella, J and Bacharach, B. Longitudinal comparison of the academic performances of Asian-American and white medical students. *Academic Medicine*, 1993. 68(1): p. 82-86.

32. Kay-Lambkin, F., Pearson, S and Rolfe, I. The influence of admissions variables on first year medical school performance: a study from Newcastle University, Australia. *Medical Education*, 2002. 36: p. 154-159.

33. Liddell, M. and Koritsas, S. Effect of medical students' ethnicity on their attitudes towards consultation skills and final year examination performance. *Medical Education*, 2004. 38: p. 187-198.

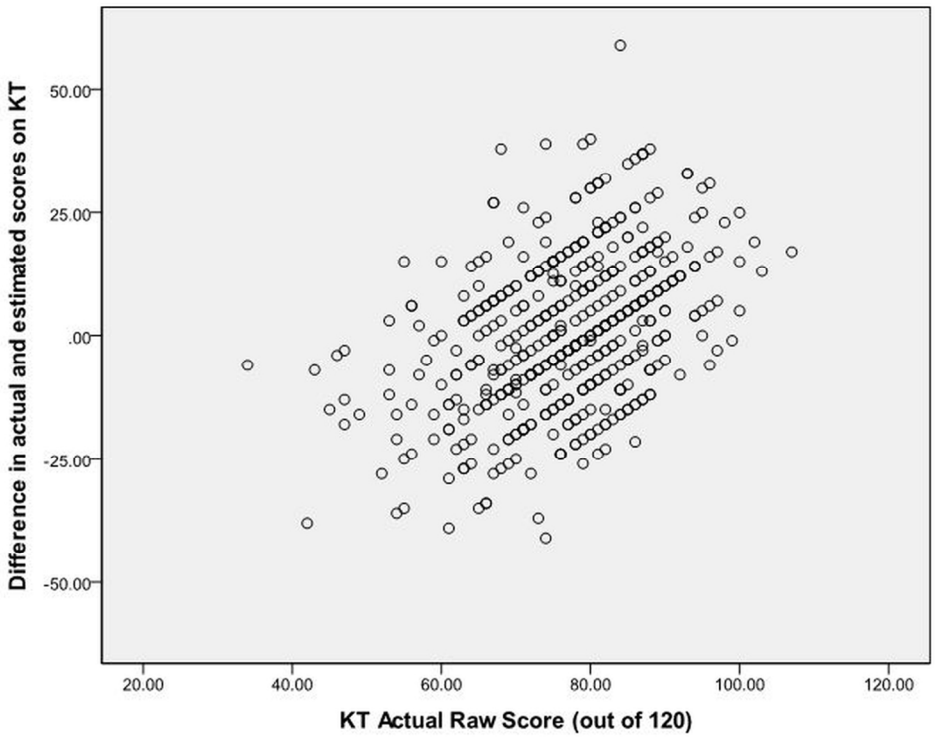
34. Croskerry, P. Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Academic Emergency Medicine*, 2002. 9: p. 1184-1204.

35. Croskerry, P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*, 2003. 78(8): p. 775-780.

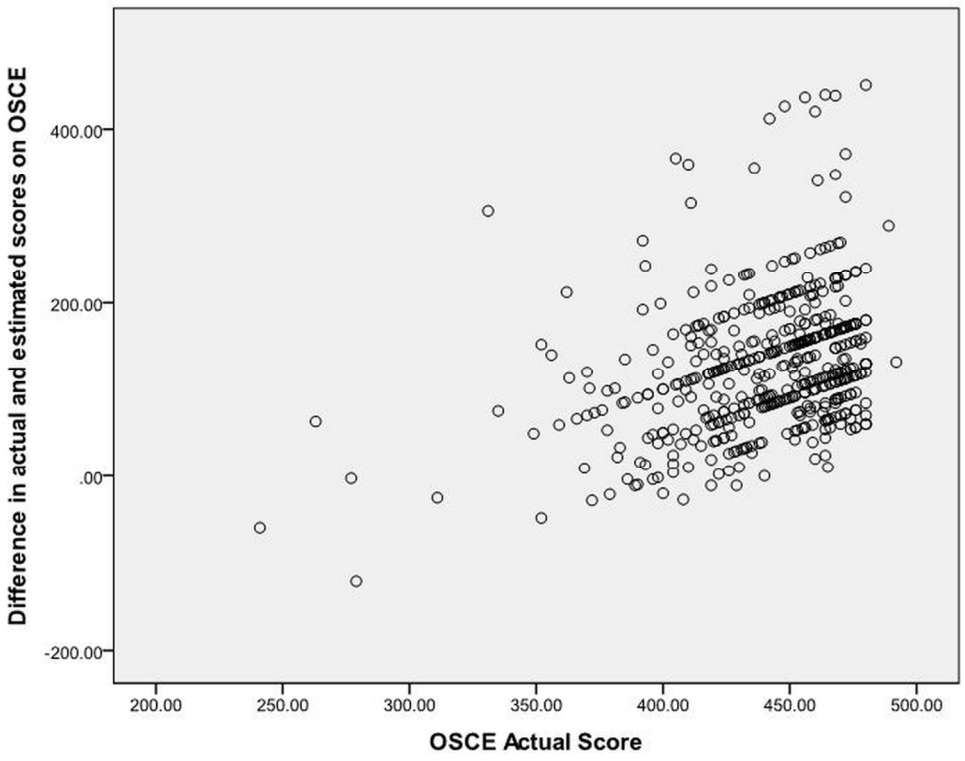
36. Woolf, K., Cave, J., Greenhalgh, T and Dacre, J. Ethnic stereotypes and the underachievement of UK medical students from ethnic minorities: qualitative study. *British Medical Journal*, 2008. 337.

37. Kaffash, J. RCGP faces legal threat over international GP trainee failure rates, in Pulse 2012.

38. Woolf, K., McManus, IC., Potts, H and Dacre, J. The mediators of minority ethnic underperformance in final medical school examinations. *British Journal of Educational Psychology*, 2011. 83(1): p. 135-159.



119x90mm (300 x 300 DPI)



114x90mm (300 x 300 DPI)

GMC/ACME research questionnaire

We are considering how accurately doctors are able to judge their own performance, and that of other doctors, in formal testing. Please answer these questions as honestly as possible. We are expecting you to use your judgement to estimate the scores.

These answers will be considered independently from today's validation data and will not impact on your feedback whatsoever. All information will be anonymised before being analysed.

Candidate number

Date of exam

Speciality

A. Knowledge test:

1. Please estimate your own ranking in the written test today, compared to **all the doctors** who are eligible to sit this written test (i.e. fitting the GMC criteria of FY2 to Consultant level, and have worked in the specialty for at least four months within the last year)
*(**0** = lowest rank and **100** = highest rank)*

2. Please estimate your own ranking in the written test today, compared to the other doctors who sat the written test **today**
*(**0** = lowest rank and **100** = highest rank)*

3. There were 120 questions in today's exam, each is worth one mark. There is no negative marking. Please estimate your total score in today's written test.
*(total **120 possible marks**)*

B. OSCE:

4. Please estimate your own ranking in the OSCE today, compared to **all the doctors** who are eligible to sit this OSCE (i.e. fitting the GMC criteria of FY2 to Consultant level, and have worked in the specialty for at least four months within the last year)
*(**0** = lowest rank and **100** = highest rank)*

5. Please estimate your own ranking in the OSCE today, compared to the other doctors who sat the OSCE **today**
*(**0** = lowest rank and **100** = highest rank)*

6. There were 12 stations in today's OSCE each with a maximum of 40 marks. Please estimate your total score in today's OSCE exam.
*(total **480 possible marks**)*